Maryland School Assessment (MSA)
Science

Grades 5 and 8

Technical Report
2012 Operational Test

September 2012

ALWAYS LEARNING PEARSON

Table of Contents

List of Tables

List of Figures

Table of Appendices

Test Overview and Design

Introduction

The Maryland School Assessment (MSA) tests are measures of students' knowledge relative to the Maryland State Curriculum at grades 5 and 8. The MSA Science test was 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 2011-2012 test administration cycle.

Purpose

The purpose of this MSA Technical Report is to provide objective information regarding technical aspects of the 2012 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.

From test development to final reporting, each of these facets of the MSA Science test contributes to the validity of the inferences made about the test results. This technical manual addresses these topics for the 2011-2012 testing year.

Test Overview

In 2002, the Maryland State Department of Education adopted the testing program known as the Maryland School Assessment (MSA). The first two subjects to be established under this new

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

Purpose and Use

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

Test Content, Specifications and Design

The MSA Science test was designed to align to the Maryland State Curriculum (MSC) that specifies curricular indicators and objectives that contributed directly to measuring content standards. 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, thereby narrowing the focus for teachers further. Finally, objectives provide teachers with very clear information about what specific learning should occur. The MSC is widely disseminated to Maryland educational stakeholders, including teachers, central office staff, students, parents and other stakeholders.

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

The basic test specifications were established by MSDE and provided to Pearson to guide the test development and administration. Since the inception of the Science test, there have been six test administrations—a census field test in 2007 and four operational tests (2008 through 2012). All administrations were conducted under the same testing conditions. Accordingly, the field test was designed to match the requirements of the operational administration test blueprint, i.e., a student taking the census field test and the operational test would respond to the same number and type of items. However, because of embedding of field test items on the operational form, there were fewer scored items on the operational form, even with the same number of overall items. Beginning with the 2008 operational test, two base forms (i.e., two forms of scored operational items) were used. Each form had a total of 77 items on the grade 5 form and 75 items on the grade 8 form. Grade 5 tests had 66 operational (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 2012 operational MSA Science included two types of items: selected response (SR) and brief constructed response (BCR). SR items require students to select a correct answer from several alternatives. For the 2012 MSA Science tests, students selected an answer from four options. Each SR item was scored dichotomously (i.e., 0 or 1). BCR items require students to provide a short answer using words, numbers, and/or symbols. All BCR items are scored using a generic rubric and scores range from 0-3 based on concordant scores from two independent raters. In cases where the scores differ by one point, the higher score is used. In cases where the rater scores differ by two or more points, a third expert rater's independent score is used as a resolution.

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

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

MSA Science Test Blueprints

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

Table 1. Grade 5 MSA Science Standards Assessed

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

Table 2. Grade 8 MSA Science Standards Assessed

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

MSA Science 2012 Operational Test Construction

The 2012 operational tests were created according to the test blueprints (see Table 1 and 2) and reflective of the Maryland State Curriculum for Science in the form of measureable Indicators and Objectives. As such, each of the two operational forms yielding student scores has the same test composition as that of the 2008 tests in terms of content, total number of items/score points, and item types. Additionally, each operational form was created with five unique sets of embedded field test items (see MSA Science 2012 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 2012 MSA Science operational test forms was an iterative process primarily involving Pearson content experts, MSDE, and Pearson psychometricians. Initial test forms were created to meet the respective blueprints, reflect the MSC measureable Indicators and Objectives, and align with statistical characteristics of the 2008 operational tests. Only items deemed eligible after being administered live (field tested) and reviewed by content experts based on statistical indicators (see Data Review of the Field Test Items) were used. Additional content-related characteristics that were part of the creation of the operational test forms had to do with ensuring there was no cuing from one item to the next. That is, items were scrutinized to make sure nothing in any one question or passage would provide information relevant to answering any other item correctly.

Classical item statistics were used in conjunction with item response theory (IRT) statistics to help target the overall test forms. The guiding principles were choosing items with reasonably strong point biserial correlations (ideally >.30) and matching a spread of item difficulties in line with the 2008 forms. Items flagged for any reason based on the data review criteria (also including differential item functioning, as described later) were identified as such, and staff members were discouraged from using them. Item level statistical targets based on overall test, by standard, and by item type were also used for guidance. IRT test characteristic curves (TCCs), test information functions (TIFs), and conditional standard error 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 2012 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. 2012 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 2012 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 2012 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 2012 field test. Overall, development was structured to spread the items across the six standards specified within the Maryland State Curriculum 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 2011 for the 2012 administration was approximately 180 items for each grade: 155 SR and 25 BCR items.

During 2010 published technical passages to be approved for item development were selected and reviewed by Pearson content staff and MSDE content experts. An item writer training was held in early December 2010. 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 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;
- 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 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 2011, 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 June 2011. 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 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 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 2012, 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 2013 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 2013 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	32262	53.08	38540	63.08
Paper	28515	46.92	22560	36.92
Form				
1	5779	9.51	5809	9.51
2	5747	9.46	5886	9.63
3	5775	9.50	5745	9.40
4	5829	9.59	5812	9.51
5	5721	9.41	5907	9.67
6	8487	13.96	8636	14.13
7	5889	9.69	5776	9.45
8	5832	9.60	5779	9.46
9	5823	9.58	5842	9.56
10	5895	9.70	5908	9.67
Gender				
Female	29832	49.09	29918	48.98
Male	30941	50.91	31169	51.02
Unknown	4	**	13	**
Ethnicity				
Hispanic/ Latino	7214	11.87	6445	10.55
Non-Hispanic/ Latino	53559	88.13	54642	89.45
Race				
American Indian	177	0.33	188	0.34
Asian/Pacific Islander	3663	6.84	3526	6.45
African American	21358	39.88	21994	40.25
Native Hawaiian	64	0.12	78	0.14
White	25833	48.23	26775	49.00
Two or More Races	2464	4.60	2081	3.81
All	**60777**	**100**	**61100**	**100.00**

Note: differences in values reflect missing data
** less than 0.001

Distribution of Students across Forms

As described, MSA Science test forms are composed of a set of operational items and field test items. Ideally, each respective test form will be administered to randomly equivalent groups of students. This helps ensure that any item and test level statistics are more directly comparable. The administration of multiple test forms is commonly referred to as "spiraling." The MSA Science test forms were spiraled at the student level and within mode of administration so that there would be an even distribution of tests across forms. Table 5 presents this distribution of tests across forms by mode of administration at each grade. Within-form overages (i.e. Grade 5 online Form 6) 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	2958	2933	2977	2994	2893	5493	3051	2970	2970	3023
	Paper	2821	2814	2798	2835	2828	2994	2838	2862	2853	2872
	Overall	5779	5747	5775	5829	5721	8487	5889	5832	5823	5895
Grade 8	Online	3679	3709	3563	3597	3704	5811	3587	3560	3632	3698
	Paper	2130	2177	2182	2215	2203	2825	2189	2219	2210	2210
	Overall	5809	5886	5745	5812	5907	8636	5776	5779	5842	5908

Key Check Analysis of Operational Test Data

Using preliminary data collected from the 2012 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 2012 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).

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

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

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, . . . , m_j−1), the GPC model gives the probability of response r as a function of latent variable θ as

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

where

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

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

Calibration of the mixed test format (3PL/GPC model) items was conducted using 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
- had a student score that 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 2012 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 consisted 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 through 2011 field test calibrations or from the 2008 through 2011 operational administrations.

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 track changes in item parameters and to 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 2012 MSA Science tests to the base scales were included in this stability check.

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

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

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

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

Flagged items were further reviewed through examination of the 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 three items removed from grade 5 and four items removed from the grade 8 common item sets. 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 (12 OP items >> 07 base scale)	1.035406	0.265267	1.069871	0.284065

The transformation constants were applied to the 2012 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.

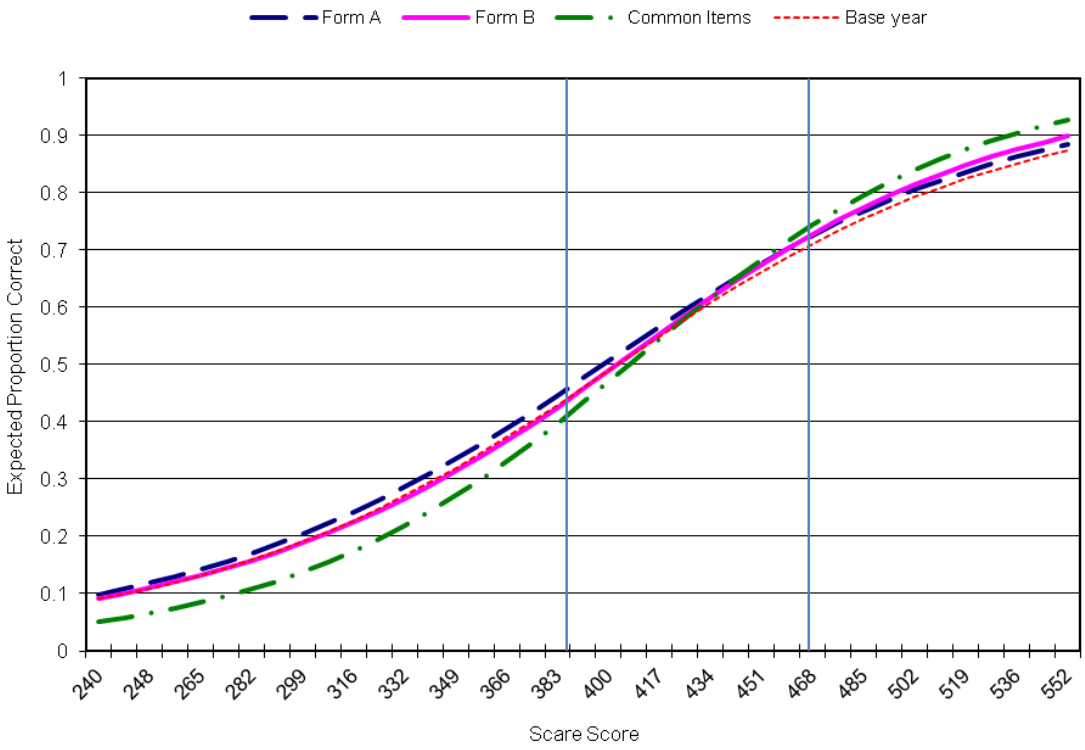
Test Analysis, Operational Scaling and Scoring

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 2012 Form A and B curves and curves reflective of operational items from the 2008 (base year) administration.

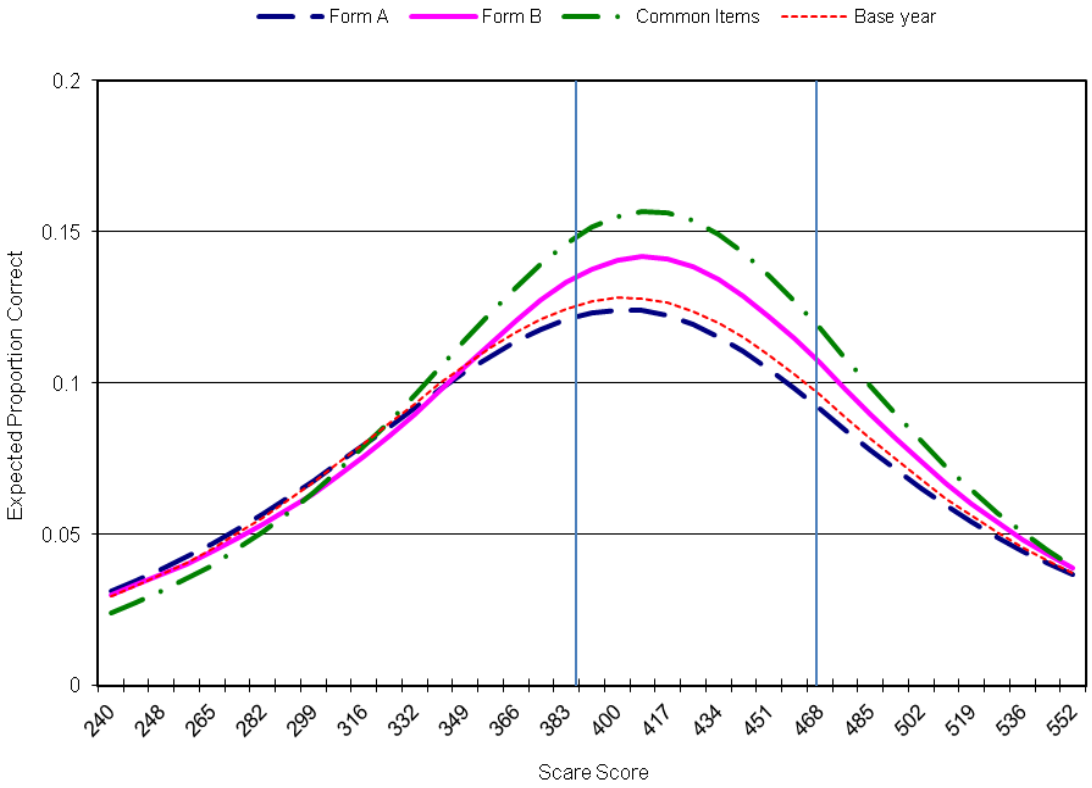
Figure 1 shows the overlaid TCC plots for Form A, Form B, form-to-form linking items and base-year item pool for grade 5. Figure 2 also displays test information curves for Form A, Form B, form-to-form linking items and the base-year. Figure 3 illustrates the conditional standard error of measurements for the four item sets. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reporting scale metric (each performance level is denoted at the top of the plot: Basic, Proficient, and Advanced). It should also be noted that each curve is presented according to the MSA Science scale score metric, which is described in the Defining Scale Ranges section.

Figure 1. Test Characteristic Curves of the Grade 5 Science Test



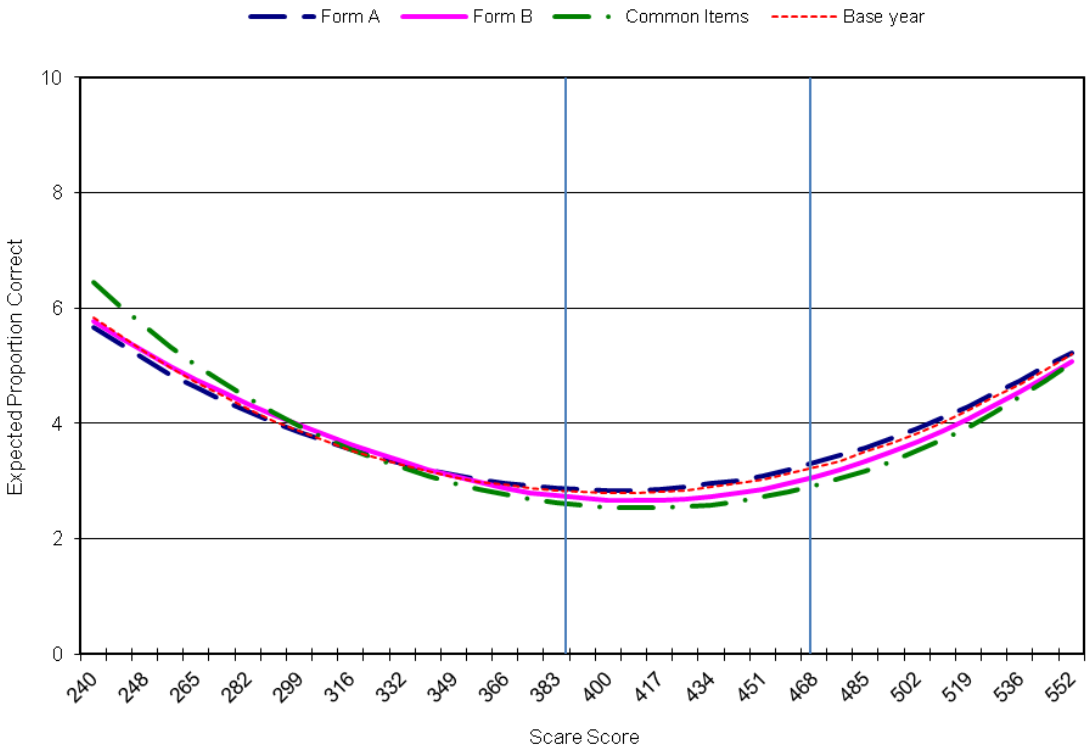
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

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).

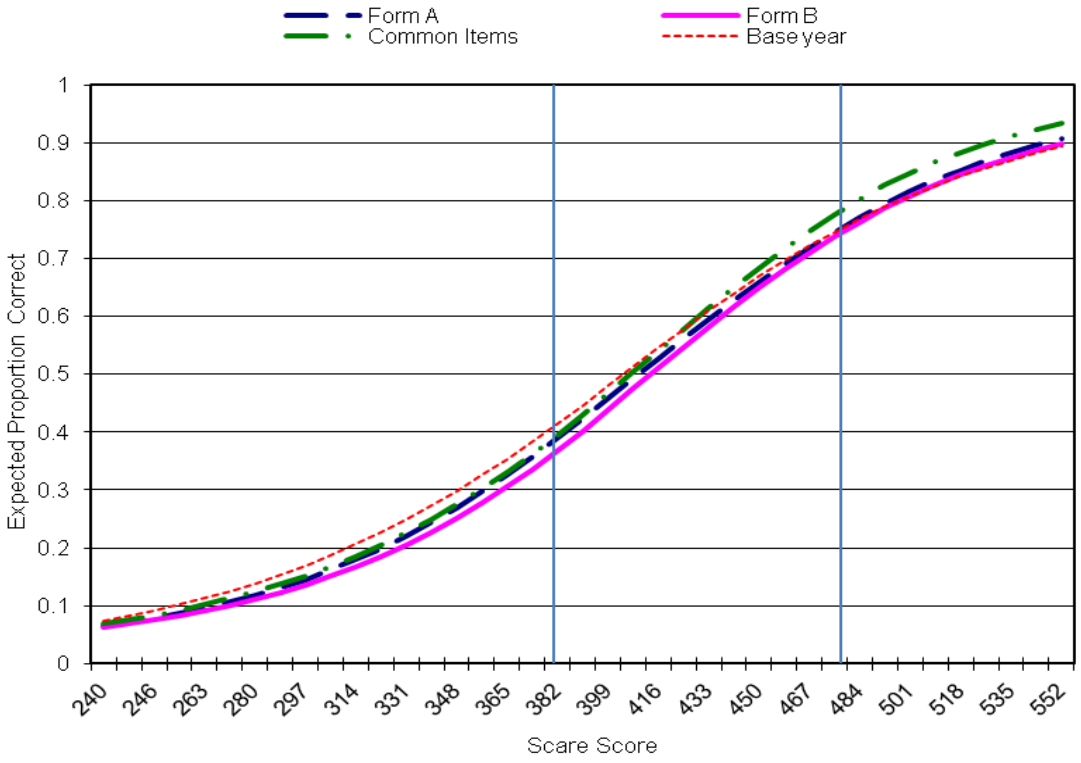
Figure 3. Conditional Standard Error of Measurement for the Grade 5 Science Test



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).

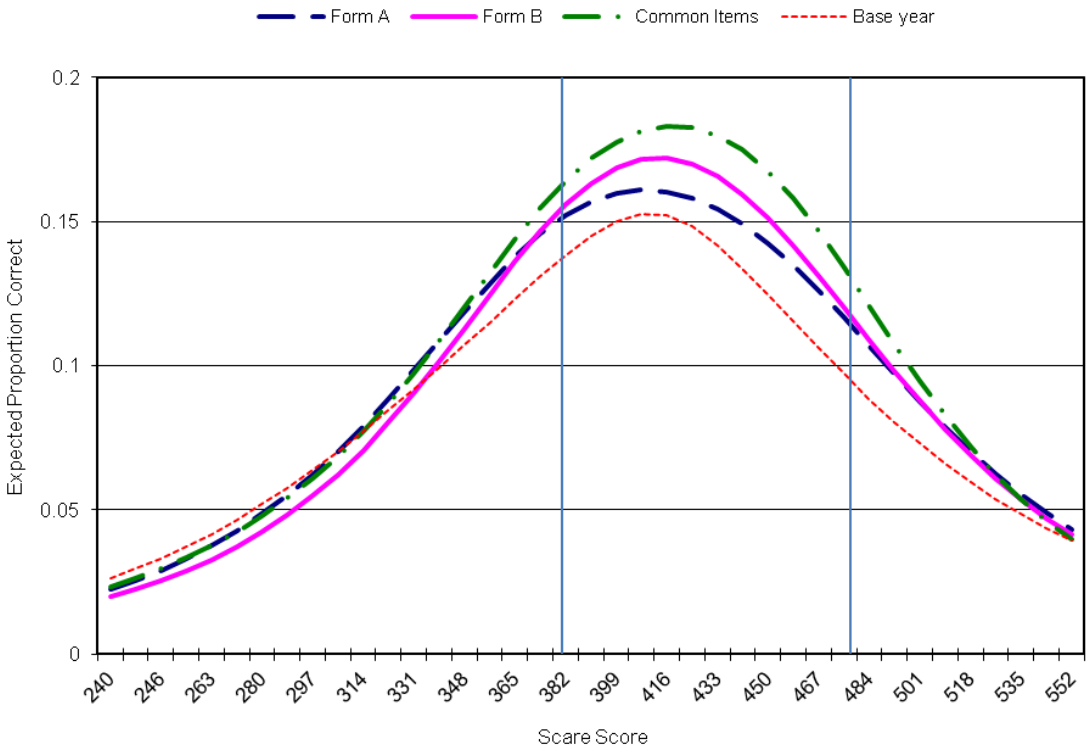
As with grade 5, IRT item parameter estimates were used to generate characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the base forms, form-to-form linking items, and base-year operational test for grade 8. Figure 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 reporting scale metric. Note that each curve is presented relative to the scale score metric described in the Defining Scale Ranges section.

Figure 4. Test Characteristic Curves of the Grade 8 Science Test



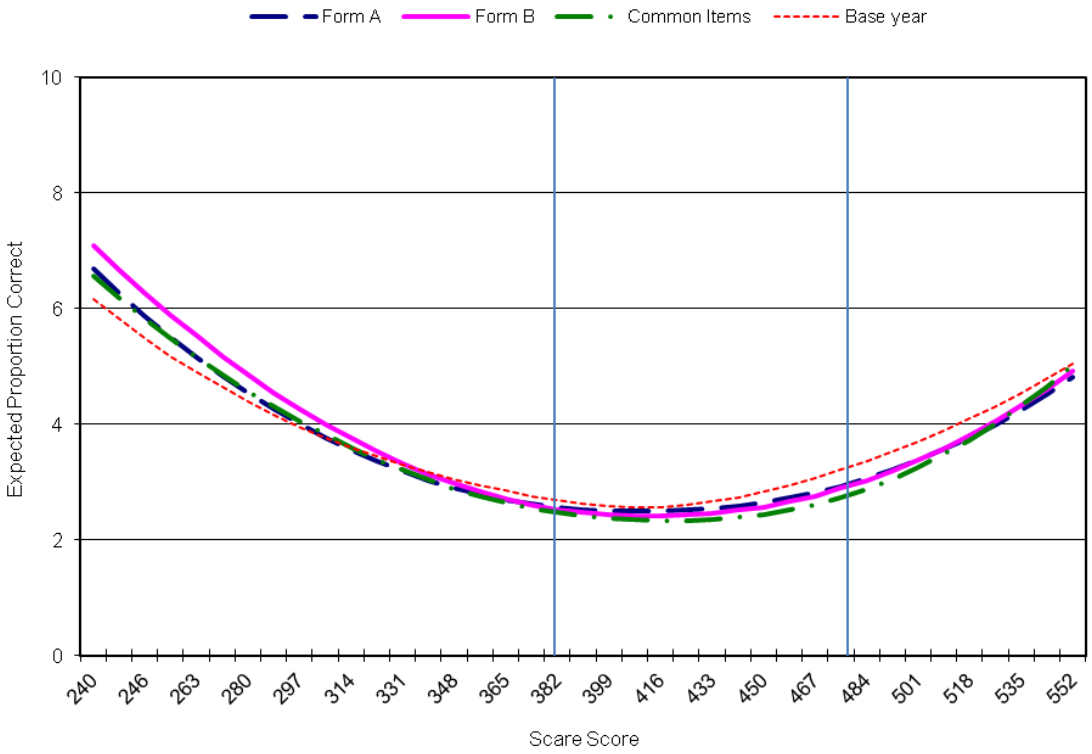
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

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 that can be more meaningfully interpreted by students, teachers and other stakeholders. In order to facilitate the use and interpretation of the results of the 2012 MSA Science operational administration, scale scores were created through the application of scaling constants determined from the base 2007 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 2012 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, the technical manual, and the evaluation report, which are available upon request.

The 2012 operational scores were estimated by the pattern scoring approach. The 2012 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. It should be noted that one SR item in grade 5 was not used for equating or scoring purposes because it had been previously released and overall impact was negligible. Final theta estimates from ISE 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 (i.e., they effectively result in estimates of $\pm \infty$). Because of this it is typical to establish ability values or scale scores that are in line with the respective overall scale. For the MSA Science tests, the LOSS and HOSS values reflect the values associated with these extreme scores. Additionally, there are instances in which certain score patterns close to zero and perfect scores will provide ability estimates where the respective conditional standard errors of measurement (CSEM) are very large. These inflated CSEM estimates are problematic in that they are out of line with estimates from different score patterns but of the same ability. In addition to establishing reasonable scale scores for these points, it is also desirable to provide some reasonable associated standard error to promote appropriate score interpretation.

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

First of all, consideration was given to the magnitude of standard errors relative to the overall scale score range. The current scale ranges from 240 to 650 (410 total points). When standard

errors exceed 40% of a scale range, the utility of a test score interpretation is limited. With this in mind, the initial 2007 MSA Science base 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 denotes the variances for the k individual items; s_{sum}^2 denotes the variance for the sum of all items.

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

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

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

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	0.93	0.93
Gender	Female	0.92	0.92	0.92	0.93
	Male	0.92	0.93	0.93	0.94
Ethnicity	Hispanic/ Latino	0.91	0.91	0.92	0.92
	Non-Hispanic/ Latino	0.92	0.92	0.93	0.93
Race	African American	0.90	0.90	0.90	0.91
	American Indian	0.91	0.92	0.91	0.92
	Asian/Pacific Islander	0.91	0.92	0.92	0.94
	Native Hawaiian	0.97	0.94	0.91	0.93
	White	0.90	0.91	0.91	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 2012 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 2012 MSA Science tests yield scale scores that range between 240 and 650. As a result of calibration, equating, and scaling the scale scores from the two base forms are comparable within the same grade, but not across grade levels. The only inferences that can be appropriately drawn from scale scores are that higher scale scores represent higher performance on the MSA Science test. Thus, performance levels and descriptions can give a specific interpretation other than a simple interpretation because they were developed to bring meaning to the scale scores.

Performance Levels and Descriptions

Performance levels and descriptions provide specific information about students' performance levels and help interpret the 2012 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

Table 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	31	57	11	412	46.7	60777	29	60	11	414	44.1	32262	34	54	11	409	49.5	28515
Gender																		
Male	32	57	11	411	47.9	30941	29	59	11	414	45.2	16407	35	53	12	408	50.5	14534
Female	31	58	11	412	45.5	29832	29	61	10	414	42.9	15855	34	55	11	410	48.2	13977
Ethnicity																		
Hispanic/ Latino	41	54	5	398	42.9	7214	40	55	5	400	42.0	3428	42	54	5	397	43.7	3786
Non-Hispanic/ Latino	30	58	12	413	46.9	53559	28	61	12	415	44.0	28834	33	54	12	411	50.0	24725
Race																		
American Indian	33	60	7	409	41.9	177	28	67	5	411	40.3	97	39	53	9	406	43.8	80
Asian/Pacific Islander	15	63	22	434	46.3	3663	15	65	20	433	44.0	1692	15	61	24	435	48.2	1971
African American	51	46	3	389	42.4	21358	47	50	3	393	41.0	10321	55	43	2	384	43.2	11037
Native Hawaiian	47	50	3	384	62.1	64	32	65	3	407	39.9	34	63	33	3	358	72.7	30
White	16	66	18	430	41.5	25833	17	67	16	428	40.1	15277	15	65	20	433	43.2	10556

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

Table 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 8 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	29	63	7	411	49.4	61100	28	66	6	411	46.7	38540	32	59	9	409	53.6	22560
Gender																		
Male	30	62	9	411	51.8	31169	28	65	8	412	49.3	19705	32	57	10	410	55.7	11464
Female	29	65	6	410	46.6	29918	28	67	5	411	43.9	18835	31	61	8	409	51.0	11083
Ethnicity																		
Hispanic/ Latino	40	57	3	395	48.4	6445	41	57	3	393	48.2	3666	39	57	4	396	48.7	2779
Non-Hispanic/ Latino	28	64	8	412	49.1	54642	26	67	7	413	46.2	34874	31	59	10	411	53.8	19768
Race																		
American Indian	31	65	3	406	44.4	188	34	64	3	404	44.4	140	25	71	4	412	44.5	48
Asian/Pacific Islander	11	69	20	440	47.6	3526	12	72	16	436	46.4	1763	11	65	24	443	48.6	1763
African American	49	50	1	385	44.4	21994	47	52	1	388	43.0	12811	52	47	1	382	46.0	9183
Native Hawaiian	36	62	3	399	47.3	78	35	61	4	403	48.9	49	38	62		391	44.4	29
White	14	75	12	430	42.2	26775	14	76	10	428	40.5	18734	13	71	17	437	45.5	8041

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 2012 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 2012 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 2012 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 (12 FT items >> 12 OP items)	1.033233	0.266260	1.065505	0.290947

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., |Δ|) 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 (|Δ|) is less than 1.5 or the absolute value of the MH delta (|Δ|) 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 k th stratification variable, $m_{Fk} = (1/n_{F+k}) F_k$ is the mean item score for the focal group in the k th stratum, and $m_{Rk} = (1/n_{R+k}) R_k$ is the analogous value for the reference group. 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:

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

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. It should be noted that "Total" in Table 13 reflects total items flagged based on the largest DIF classification level. That is, items flagged at both the B and C would be counted as "C" in Table 13.

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

	DIF Classification Level				
Grade	**B**	**BB**	**C**	**CC**	**Total**
5	12	4	1	2	19
8	5	3	0	4	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 2012 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
- B or 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
- BB or 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 71 items in grade 5 and 70 items in grade 8 were inspected during data review as a result of the item not meeting the statistical flagging criteria. Eleven of the 71 total flagged items were rejected from the grade 5 pool and eight of the 70 flagged items for grade 8 were rejected.

Validity

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

Messick (1989) defined validity as follows:

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

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

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

To investigate the validity evidence of the 2012 MSA-Science tests, content-related evidence, differential item functioning (DIF) analysis on gender and ethnicity, and evidence based on internal structure were collected.

Content-related Evidence

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

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

The Test Overview and Design section details the connection between the MSA Science blueprint and the MSC. The 2012 MSA Science tests were constructed exclusively using items that met not only the statistical criteria described in this report, but also verified as aligning to the MSC by Maryland science content experts. As described, tests were constructed according to the test blueprints and as such, scores provided are reflective of overall Science ability as defined within the state standards.

Differential Item Functioning (DIF)

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

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

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

Inter-Correlations among Standards

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

Table 15 reports the correlations among the six standards based on scale scores. The standard-level (subtest) inter-correlations ranged from 0.53 to 0.85 where most are greater than .60. The standard subscores are moderately highly related to one another and more strongly related to the total test score. This suggests there is some uniqueness to items grouped by standard but that they are collectively measuring a dominant overall construct (science).

Table 15. Correlation among MSA Science content standards

Grade 5 Form A	Mean	SD		Str1	Str2	Str3	Str4	Str5	Str6	Total
	414.80	62.60	**Str1**	1.000						
	422.20	79.25	**Str2**	0.577	1.000					
	413.40	62.05	**Str3**	0.645	0.580	1.000				
	413.41	62.08	**Str4**	0.638	0.568	0.621	1.000			
	429.23	83.79	**Str5**	0.561	0.526	0.530	0.548	1.000		
	414.89	74.69	**Str6**	0.618	0.578	0.607	0.607	0.547	1.000	
	412.39	46.91	**Total**	0.836	0.769	0.819	0.816	0.729	0.806	1.000
Grade 5 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	415.24	61.05	**Str1**	1.000						
	413.56	59.80	**Str2**	0.642	1.000					
	408.57	81.62	**Str3**	0.580	0.592	1.000				
	414.79	73.49	**Str4**	0.613	0.616	0.575	1.000			
	419.68	81.18	**Str5**	0.551	0.563	0.530	0.572	1.000		
	414.17	67.51	**Str6**	0.635	0.638	0.577	0.605	0.568	1.000	
	410.79	46.56	**Total**	0.820	0.830	0.762	0.809	0.750	0.817	1.000
Grade 8 Form A				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	419.46	72.13	**Str1**	1.000						
	418.79	81.37	**Str2**	0.578	1.000					
	414.83	64.17	**Str3**	0.648	0.611	1.000				
	408.40	72.93	**Str4**	0.599	0.567	0.633	1.000			
	409.31	83.91	**Str5**	0.572	0.571	0.598	0.571	1.000		
	415.47	65.88	**Str6**	0.646	0.602	0.663	0.613	0.588	1.000	
	412.61	48.04	**Total**	0.804	0.780	0.842	0.787	0.764	0.838	1.000
Grade 8 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	411.95	68.40	**Str1**	1.000						
	401.14	77.08	**Str2**	0.628	1.000					
	409.51	66.75	**Str3**	0.667	0.656	1.000				
	406.82	78.45	**Str4**	0.585	0.563	0.595	1.000			
	409.41	79.49	**Str5**	0.650	0.642	0.669	0.595	1.000		
	435.99	106.21	**Str6**	0.596	0.591	0.606	0.550	0.599	1.000	
	408.72	50.50	**Total**	0.824	0.817	0.850	0.748	0.825	0.762	1.000

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

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted for the 2012 MSA Science tests to examine the relationship between the subtest scores relative the total test score. 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 that indicate good fit are higher than 0.90 (Tabachnick & Fidell, 2001). The RMSEA is a function of the estimated discrepancy between the population covariance matrix and the model-implied covariance matrix, with a value of less than or equal to .05 indicating close fit and a value between .05 and .08 indicating a "reasonable error of approximation" (Browne & Cudeck,

1993, p. 144). Hu and Bentler (1999) propose an RMSEA ≤ .06 as the guideline for close fit. Table 16 summarizes fit indicators estimated from the confirmatory factor analysis for the 2012 MSA Science tests. The confirmatory factor analysis results provide additional evidence to support the conclusion that scores from the MSA Science tests reflect a single latent trait (Science). For both grades, the lowest AGFI was 0.9873, and the highest RMSEA was 0.0419. 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. 9946	0. 0273
Grade 5 Form B	0. 9985	0. 0138
Grade 8 Form A	0. 9887	0. 0396
Grade 8 Form B	0. 9873	0. 0419

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

Evidence for Scores from Accommodated Testing

Accommodations are offered to students with disabilities that preclude them from being fairly assessed by the tests as they are written (e.g., visually impaired students). In order to examine whether or not these accommodations are effective (i.e., result in valid test scores) the CFA conducted to examine the relationship between standards was repeated using only students testing with accommodations and then again using only students testing without accommodations. The results of this analysis showed comparable levels of model fit based on the two groups (see Table 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 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. 9900	0. 0332	0. 9948	0. 0269
Grade 5 Form B	0. 9993	0. 0000	0. 9982	0. 0149
Grade 8 Form A	0. 9837	0. 0439	0. 9896	0. 0379
Grade 8 Form B	0. 9921	0. 0307	0. 9878	0. 0409

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

References

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

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (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.

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

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

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

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

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

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	d1	d2	d3
50015	OP	0.68	0.27	0.39052	-0.62714	0.13332			
50032	OP	0.91	0.23	0.48568	-2.75027	0.05177			
50033	OP	0.58	0.43	0.85180	0.32414	0.18551			
50041	OP	0.68	0.46	0.84180	-0.22513	0.13487			
50054	OP	0.49	0.33	0.64009	0.88407	0.18059			
50059	OP	0.92	0.36	1.03676	-1.76295	0.05067			
50062	OP	0.38	0.26	0.60214	1.58518	0.17317			
50078	OP	0.44	0.55	0.66942	0.71497	0.00000	2.25240	0.02213	-2.27454
50092	OP	0.42	0.23	0.46482	1.57615	0.18122			
50096	OP	0.70	0.26	0.37468	-0.99514	0.06616			
50109	OP	0.77	0.33	0.65565	-0.50376	0.30716			
50126	OP	0.43	0.22	1.22926	1.56717	0.31436			
50171	OP	0.59	0.37	0.62657	0.20579	0.16090			
50172	OP	0.45	0.38	0.82568	0.93333	0.16950			
50216	OP	0.74	0.45	0.83710	-0.55067	0.13056			
50221	OP	0.65	0.37	0.63088	-0.03068	0.19291			
50228	OP	0.80	0.33	0.70607	-0.51783	0.38631			
50311	OP	0.90	0.37	0.92181	-1.68907	0.04717			
50329	OP	0.78	0.50	1.10006	-0.66826	0.12518			
50345	OP	0.68	0.40	0.81787	-0.00327	0.24576			
50352	OP	0.90	0.34	0.82644	-1.75333	0.06594			
50362	OP	0.65	0.34	0.56042	-0.10303	0.17906			
50399	OP	0.73	0.27	0.39181	-1.35647	0.03077			
50418	OP	0.73	0.49	1.15640	-0.26548	0.21213			
50439	OP	0.65	0.37	0.68731	0.01935	0.21883			
50453	OP	0.73	0.30	0.51515	-0.49754	0.24132			
50475	OP	0.56	0.48	1.27136	0.50100	0.22082			
50477	OP	0.77	0.42	0.76331	-0.87770	0.05995			
50479	OP	0.56	0.52	0.58880	-0.27910	0.00000	2.60616	-0.18212	-2.42404
50546	OP	0.80	0.36	0.73489	-0.76288	0.27638			
50583	OP	0.43	0.38	0.87084	0.99482	0.16174			
50600	OP	0.77	0.43	0.82135	-0.65985	0.16300			
50658	OP	0.65	0.36	0.55900	-0.27784	0.10792			
50677	OP	0.55	0.41	0.77066	0.44613	0.15384			
50678	OP	0.76	0.39	0.69344	-0.78624	0.13666			
50901	OP	0.47	0.41	0.73938	0.67214	0.10750			
50905	OP	0.90	0.24	0.50672	-2.56698	0.06351			
50908	OP	0.42	0.26	0.74875	1.50967	0.24366			
50915	OP	0.60	0.29	0.42503	-0.06245	0.11279			
50916	OP	0.43	0.44	1.02398	0.87383	0.13742			
50926	OP	0.81	0.40	0.86517	-0.71148	0.28372			
50934	OP	0.94	0.27	0.77238	-2.41683	0.03610			
50942	OP	0.49	0.56	0.73541	0.40440	0.00000	2.59038	0.04164	-2.63202
51004	OP	0.82	0.27	0.45761	-1.85185	0.05274			
51006	OP	0.68	0.39	0.64898	-0.33526	0.14618			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
51009	OP	0.68	0.41	0.95705	0.14653	0.30131			
51012	OP	0.57	0.37	0.65442	0.35158	0.16505			
51018	OP	0.87	0.34	0.68971	-1.69013	0.03268			
51019	OP	0.68	0.27	0.37787	-0.87255	0.03695			
51021	OP	0.42	0.41	1.06039	1.00641	0.16540			
51022	OP	0.51	0.36	0.77308	0.80547	0.21551			
51023	OP	0.64	0.44	0.83551	-0.00977	0.15859			
51026	OP	0.53	0.38	0.87141	0.72972	0.23838			
51032	OP	0.57	0.25	0.32633	-0.17242	0.02457			
51034	OP	0.41	0.51	0.71957	0.77583	0.00000	2.98904	-0.55009	-2.43895
55172	OP	0.71	0.30	0.43948	-1.01581	0.03875			
55198	OP	0.54	0.42	1.08803	0.64857	0.24523			
55207	OP	0.91	0.30	0.74439	-1.68997	0.28157			
55234	OP	0.88	0.36	0.82250	-1.55690	0.08806			
55241	OP	0.93	0.32	0.95178	-1.57296	0.34127			
50302_02	OP	0.75	0.44	0.84422	-0.62012	0.14476			
50302_04	OP	0.72	0.43	0.74781	-0.62615	0.07567			
50553_04	OP	0.52	0.34	0.67458	0.74076	0.20684			
50553_06	OP	0.66	0.35	0.55703	-0.30419	0.14312			
50564_01	OP	0.59	0.41	1.00244	0.46885	0.26475			
50564_04	OP	0.72	0.52	1.34921	-0.19081	0.21860			
50564_05	OP	0.61	0.24	0.51312	0.73455	0.33274			
50590_01	OP	0.69	0.39	0.76255	-0.12212	0.23299			
50590_02	OP	0.54	0.50	1.20389	0.47186	0.17591			
50590_04	OP	0.39	0.39	0.98398	1.11216	0.14729			
50620_01	OP	0.41	0.39	1.04653	1.07403	0.17137			
50620_02	OP	0.44	0.31	0.68774	1.18473	0.18997			
50628_01	OP	0.94	0.28	0.78567	-2.31002	0.02972			
50628_03	OP	0.62	0.44	0.68617	-0.20366	0.04899			
50635_03	OP	0.28	0.35	1.20002	1.48454	0.12176			
50635_04	OP	0.35	0.20	0.68392	2.03494	0.22093			
50635_05	OP	0.57	0.35	0.71930	0.54870	0.23793			
50670_01	OP	0.39	0.35	0.83331	1.23214	0.15937			
50670_03	OP	0.68	0.42	1.09684	0.17548	0.33186			
50670_05	OP	0.73	0.39	0.74991	-0.33995	0.23785			
50670_07	OP	0.35	0.52	0.62034	1.20767	0.00000	2.33807	-0.42789	-1.91018
50672_02	OP	0.93	0.32	0.88083	-1.99054	0.07238			
50672_03	OP	0.67	0.49	1.03173	-0.08216	0.16583			
50672_04	OP	0.64	0.36	0.82939	0.40394	0.31353			
50674_02	OP	0.52	0.39	0.73537	0.60583	0.16664			
50674_06	OP	0.39	0.38	0.85200	1.10843	0.13180			
50674_07	OP	0.31	0.44	0.60471	1.65994	0.00000	2.96789	-0.85119	-2.11670
50923_01	OP	0.89	0.38	0.90143	-1.45287	0.15595			
50923_04	OP	0.43	0.31	0.84036	1.30350	0.22262			
50929_01	OP	0.54	0.43	1.23085	0.70744	0.25089			
50929_04	OP	0.81	0.47	1.17423	-0.56731	0.27085			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
50930_01	OP	0.65	0.38	0.55372	-0.38867	0.04507			
50930_02	OP	0.41	0.28	0.57544	1.37460	0.15983			
50934_03	OP	0.71	0.53	1.21556	-0.27355	0.14503			
50934_04	OP	0.59	0.37	0.79881	0.50996	0.26130			
50934_05	OP	0.61	0.36	0.62444	0.10645	0.16443			
50937_03	OP	0.73	0.53	1.32208	-0.29343	0.19355			
50937_05	OP	0.62	0.46	0.87889	0.06244	0.15375			
50941_01	OP	0.47	0.26	0.65906	1.36674	0.26596			
50941_02	OP	0.70	0.52	1.03183	-0.31383	0.10465			
50941_03	OP	0.69	0.45	1.03236	0.01618	0.25838			
51053_02	OP	0.57	0.29	0.49996	0.56869	0.22123			
51053_03	OP	0.54	0.38	1.02735	0.79954	0.27579			
51054_01	OP	0.72	0.30	0.44735	-1.06658	0.03857			
51054_02	OP	0.78	0.38	0.70832	-0.78967	0.19651			
51054_04	OP	0.81	0.32	0.55120	-1.45771	0.07262			
51056_03	OP	0.56	0.29	0.39753	0.12726	0.06587			
51056_04	OP	0.91	0.30	0.76164	-1.59874	0.30198			
55073_01	OP	0.78	0.50	1.24183	-0.48584	0.21474			
55073_05	OP	0.47	0.43	1.14849	0.81280	0.19560			
55073_06	OP	0.62	0.37	0.53674	-0.30015	0.03993			
50481	FT	0.69	0.37	0.60847	-0.40375	0.11414			
50551	FT	0.87	0.38	0.98294	-1.00864	0.30444			
50929	FT	0.30	0.52	0.78091	1.72955	0.00000	2.91253	-0.43660	-2.47593
51008	FT	0.82	0.39	0.77406	-1.24486	0.15163			
51011	FT	0.80	0.44	0.97558	-0.66431	0.22880			
51016	FT	0.67	0.44	0.88577	-0.03284	0.20100			
51025	FT	0.58	0.37	0.51452	-0.07529	0.05956			
51036	FT	0.32	0.43	0.73910	1.96920	0.00000	3.51410	-0.55957	-2.95453
51038	FT	0.39	0.49	0.61029	1.29223	0.00000	2.73707	-0.05337	-2.68370
51040	FT	0.89	0.40	1.31641	-0.91181	0.37184			
51160	FT	0.60	0.39	0.86519	0.48210	0.27088			
51161	FT	0.73	0.20	0.34179	-0.63678	0.28811			
51164	FT	0.75	0.31	0.54156	-0.60863	0.21808			
51167	FT	0.69	0.39	0.73641	-0.14187	0.19929			
51168	FT	0.74	0.34	0.55064	-0.80922	0.10477			
51169	FT	0.66	0.44	0.81719	-0.11815	0.17926			
51170	FT	0.34	0.18	0.87033	1.96870	0.25948			
51171	FT	0.83	0.43	1.07629	-0.69773	0.28690			
51174	FT	0.52	0.35	0.66446	0.76243	0.19153			
51176	FT	0.49	0.21	0.53960	1.58952	0.30300			
51180	FT	0.31	0.58	0.80173	1.42135	0.00000	1.90792	-0.27813	-1.62979
51183	FT	0.71	0.34	0.51168	-0.78310	0.05757			
51184	FT	0.66	0.38	0.72787	0.06483	0.23306			
51187	FT	0.20	0.53	0.69234	2.28708	0.00000	1.55070	0.38237	-1.93307
51188	FT	0.68	0.55	1.28687	-0.06737	0.14670			
51190	FT	0.42	0.33	0.65058	1.19094	0.14374			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
51194	FT	0.47	0.27	0.86398	1.32820	0.28664			
51195	FT	0.62	0.29	0.48749	0.21413	0.19969			
51197	FT	0.81	0.39	0.73716	-1.03133	0.11372			
51199	FT	0.49	0.53	0.66940	0.36884	0.00000	2.45983	-0.13174	-2.32809
50926_01	FT	0.59	0.29	0.68712	0.83951	0.35695			
50926_02	FT	0.81	0.41	0.87581	-0.79465	0.21956			
50926_04	FT	0.61	0.46	0.99624	0.28828	0.20691			
50926_05	FT	0.45	0.38	0.77485	0.89995	0.16398			
51041_02	FT	0.70	0.27	0.41161	-0.70908	0.12915			
51041_03	FT	0.64	0.34	0.50315	-0.34070	0.06701			
51041_04	FT	0.63	0.46	0.83632	0.02446	0.11695			
51041_05	FT	0.69	0.24	0.34308	-1.05558	0.04903			
51051_01	FT	0.87	0.46	1.28497	-1.06161	0.13097			
51051_02	FT	0.55	0.40	0.70636	0.36736	0.11847			
51051_03	FT	0.31	0.25	0.75168	1.74041	0.17309			
51051_04	FT	0.71	0.34	0.50740	-1.10762	0.03767			
51091_01	FT	0.54	0.38	1.43508	0.87500	0.32030			
51091_04	FT	0.50	0.30	1.57536	1.21191	0.35092			
51091_05	FT	0.58	0.40	0.88660	0.52536	0.23032			
51091_06	FT	0.32	0.04	1.61187	3.00384	0.31154			
51091_07	FT	0.33	0.56	0.85009	1.78014	0.00000	2.66277	0.15840	-2.82118
51091_08	FT	0.22	0.55	0.86942	2.28537	0.00000	2.16826	-0.14338	-2.02489
51092_01	FT	0.70	0.35	0.85984	0.27547	0.38625			
51092_02	FT	0.19	0.10	0.98527	2.71592	0.16152			
51092_03	FT	0.29	0.18	0.94304	2.00041	0.20513			
51092_04	FT	0.46	0.37	1.16386	0.89788	0.26146			
51092_05	FT	0.53	0.24	0.77133	1.39321	0.37397			
51092_06	FT	0.36	0.28	0.51251	1.48215	0.10148			
51093_01	FT	0.73	0.39	0.77277	-0.26375	0.22962			
51093_02	FT	0.67	0.19	0.28023	-0.76886	0.11773			
51093_03	FT	0.54	0.32	0.94897	1.03582	0.32871			
51093_04	FT	0.47	0.48	1.24326	0.76859	0.15675			
51093_05	FT	0.75	0.46	0.97504	-0.38309	0.19126			
51093_07	FT	0.26	0.55	0.74792	2.27780	0.00000	2.35165	0.26761	-2.61926
51121_01	FT	0.23	0.15	0.66044	2.76188	0.14914			
51121_03	FT	0.42	0.16	0.23890	1.77882	0.09260			
51121_04	FT	0.10	-.00	2.04256	2.81591	0.09325			
51121_05	FT	0.83	0.31	0.57937	-1.51519	0.10177			
51122_01	FT	0.96	0.31	1.25224	-2.10900	0.11509			
51122_03	FT	0.20	0.08	0.56460	3.81387	0.15813			
51122_04	FT	0.27	0.06	0.29239	5.45705	0.21048			
51122_05	FT	0.59	0.26	0.65837	0.97119	0.35871			
51123_01	FT	0.52	0.37	0.85747	0.79672	0.22065			
51123_02	FT	0.56	0.29	0.63368	0.82290	0.27742			
51123_04	FT	0.45	0.35	0.80321	1.09928	0.19515			
51123_05	FT	0.50	0.14	0.23391	1.72257	0.20284			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
51124_01	FT	0.30	0.29	0.66027	1.56206	0.11947			
51124_02	FT	0.55	0.33	0.48871	0.25079	0.07508			
51124_04	FT	0.60	0.37	0.63192	0.15659	0.14511			
51124_05	FT	0.27	0.13	0.79041	2.46926	0.20922			
51125_01	FT	0.35	0.42	1.20534	1.20461	0.12563			
51125_02	FT	0.28	0.37	1.37299	1.48280	0.11509			
51125_03	FT	0.41	0.46	1.11457	0.95864	0.12337			
51125_05	FT	0.37	0.25	0.35075	1.47441	0.03376			
51126_01	FT	0.09	-.02	1.50910	3.07622	0.08201			
51126_02	FT	0.18	-.12	-0.41707	-3.39140	0.10149			
51126_03	FT	0.53	0.31	0.49085	0.51536	0.12553			
51126_04	FT	0.70	0.40	0.68163	-0.36057	0.13058			
51127_01	FT	0.80	0.31	0.76186	-0.28188	0.45742			
51127_02	FT	0.53	0.23	0.72787	1.40148	0.37133			
51127_03	FT	0.51	0.27	0.53274	0.84583	0.24067			
51127_04	FT	0.66	0.34	0.74990	0.16998	0.34407			
51128_01	FT	0.33	0.13	0.20101	3.39589	0.08503			
51128_02	FT	0.46	0.15	0.21714	1.43372	0.10386			
51128_03	FT	0.48	0.23	0.31521	0.80642	0.07440			
51128_04	FT	0.27	0.25	0.87577	1.95094	0.14265			
51131_01	FT	0.11	-.12	-0.62509	-3.04396	0.06719			
51131_02	FT	0.64	0.30	0.41383	-0.38478	0.07532			
51131_04	FT	0.30	0.27	1.06751	1.75884	0.17290			
51131_05	FT	0.92	0.29	0.72692	-2.07956	0.09268			
51134_01	FT	0.36	0.24	0.41626	1.67943	0.10318			
51134_02	FT	0.36	0.21	0.52704	1.99043	0.19181			
51134_03	FT	0.75	0.13	0.20429	-1.75661	0.23342			
51134_05	FT	0.58	0.35	0.54556	0.20956	0.11355			
51135_01	FT	0.55	0.23	0.30031	0.13300	0.07221			
51135_02	FT	0.62	0.18	0.25614	-0.22257	0.15251			
51135_03	FT	0.84	0.43	0.89747	-1.21087	0.06113			
51135_04	FT	0.65	0.37	0.54578	-0.41058	0.07950			
51138_01	FT	0.62	0.30	0.60871	0.48507	0.28961			
51138_02	FT	0.88	0.35	1.09751	-0.60974	0.51851			
51138_03	FT	0.66	0.35	1.02037	0.58425	0.40180			
51138_05	FT	0.62	0.15	0.19787	-0.76632	0.08445			
55074_01	FT	0.49	0.33	0.64312	0.88725	0.19206			
55074_02	FT	0.27	0.16	0.79707	2.42823	0.19555			
55074_03	FT	0.46	0.37	0.98943	1.00377	0.22915			
55074_04	FT	0.36	0.16	0.27284	2.47987	0.12190			
55074_05	FT	0.50	0.23	0.31553	0.64038	0.07718			
55074_06	FT	0.38	0.28	0.95263	1.55305	0.22417			
55083_01	FT	0.68	0.44	0.92662	-0.02996	0.21936			
55083_02	FT	0.37	0.15	0.65718	2.43419	0.27489			
55083_03	FT	0.65	0.44	0.94344	0.15048	0.22015			
55083_05	FT	0.62	0.48	1.26746	0.29778	0.23852			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
55083_06	FT	0.41	0.64	0.93420	0.87984	0.00000	1.75757	0.06508	-1.82264
55083_07	FT	0.65	0.50	1.06828	0.00806	0.15041			

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	d1	d2	d3
80012	OP	0.83	0.32	0.53840	-1.77905	0.01846			
80013	OP	0.47	0.25	0.48631	1.34325	0.22984			
80024	OP	0.80	0.42	1.21700	-0.28044	0.42928			
80037	OP	0.51	0.34	0.46742	0.40654	0.06687			
80044	OP	0.78	0.51	1.17710	-0.59135	0.15842			
80048	OP	0.34	0.21	1.03858	1.91477	0.23932			
80051	OP	0.90	0.36	1.07706	-0.96477	0.44993			
80055	OP	0.52	0.68	0.65447	0.21295	0.00000	0.35773	0.57241	-0.93015
80064	OP	0.49	0.67	0.80857	0.33205	0.00000	1.57900	0.10352	-1.68252
80072	OP	0.57	0.37	0.79225	0.65788	0.25212			
80090	OP	0.38	0.58	0.78609	1.03712	0.00000	2.30079	-0.27023	-2.03057
80107	OP	0.77	0.55	1.45673	-0.43647	0.18273			
80112	OP	0.59	0.48	0.82156	0.09526	0.11261			
80122	OP	0.39	0.17	1.54410	1.84134	0.31937			
80178	OP	0.53	0.44	0.81166	0.51612	0.15121			
80201	OP	0.69	0.36	0.51885	-0.72901	0.02212			
80206	OP	0.55	0.48	0.86754	0.35676	0.11961			
80209	OP	0.53	0.33	0.66334	0.85402	0.22937			
80214	OP	0.66	0.18	0.23897	-1.28955	0.03567			
80225	OP	0.75	0.38	0.74136	-0.30655	0.28129			
80244	OP	0.61	0.35	1.05320	0.77486	0.38703			
80253	OP	0.41	0.35	0.73113	1.17341	0.15460			
80272	OP	0.68	0.38	0.79898	0.20269	0.30756			
80330	OP	0.77	0.46	0.86180	-0.68969	0.12503			
80344	OP	0.84	0.48	1.33782	-0.78102	0.21741			
80409	OP	0.58	0.30	0.94350	1.03630	0.38113			
80413	OP	0.49	0.32	0.59312	0.95591	0.20678			
80505	OP	0.51	0.51	0.49526	0.41488	0.00000	1.85187	0.48399	-2.33587
80567	OP	0.68	0.50	1.06864	-0.05874	0.21449			
80572	OP	0.65	0.38	0.53033	-0.49815	0.00832			
80585	OP	0.75	0.29	0.40872	-1.41869	0.01612			
80587	OP	0.79	0.41	0.76013	-0.88147	0.16010			
80640	OP	0.59	0.43	0.75491	0.23157	0.17345			
80648	OP	0.68	0.50	1.05794	-0.09149	0.20511			
80662	OP	0.85	0.42	1.03083	-0.81057	0.29948			
80775	OP	0.59	0.54	1.26823	0.27000	0.15928			
80901	OP	0.58	0.42	0.87301	0.46190	0.24330			
80912	OP	0.68	0.36	0.50339	-0.68806	0.01223			
80913	OP	0.79	0.47	1.10110	-0.47677	0.23553			
80918	OP	0.64	0.45	0.74729	-0.09846	0.10864			
80926	OP	0.44	0.37	0.81625	1.06827	0.17381			
80956	OP	0.57	0.34	0.97842	0.95250	0.34491			
81001	OP	0.51	0.25	0.69166	1.38165	0.32557			
81006	OP	0.62	0.40	0.61932	-0.07168	0.07560			
81009	OP	0.84	0.40	0.76122	-1.40713	0.01405			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81010	OP	0.29	0.29	1.21344	1.70982	0.16795			
81015	OP	0.65	0.40	0.71075	-0.00440	0.17391			
81019	OP	0.67	0.40	0.73945	0.01588	0.23253			
81022	OP	0.43	0.31	0.71133	1.31027	0.20067			
81023	OP	0.68	0.44	0.76863	-0.20558	0.12848			
81028	OP	0.44	0.24	0.29840	0.88833	0.02612			
81029	OP	0.66	0.39	0.55552	-0.48187	0.06837			
85201	OP	0.81	0.40	0.76920	-0.87375	0.17629			
80138_01	OP	0.43	0.32	0.97280	1.33842	0.24730			
80138_02	OP	0.53	0.32	0.93872	1.09864	0.31630			
80138_05	OP	0.77	0.49	1.00729	-0.60540	0.10677			
80138_07	OP	0.44	0.59	0.65669	0.57078	0.00000	1.79902	-0.41152	-1.38752
80200_01	OP	0.49	0.40	0.90331	0.84878	0.19413			
80200_05	OP	0.52	0.44	1.12028	0.74836	0.21090			
80467_01	OP	0.84	0.45	0.96230	-1.20459	0.05878			
80467_02	OP	0.57	0.51	0.90157	0.18548	0.10244			
80467_03	OP	0.57	0.42	0.72625	0.30850	0.15618			
80530_01	OP	0.75	0.57	1.45078	-0.45234	0.14858			
80530_03	OP	0.60	0.39	0.74952	0.31831	0.22906			
80530_06	OP	0.30	0.68	0.94027	1.20428	0.00000	1.37585	-0.20090	-1.17495
80534_01	OP	0.51	0.49	1.41266	0.67348	0.20703			
80534_02	OP	0.69	0.53	1.13308	-0.16469	0.17534			
80534_03	OP	0.66	0.50	1.02906	-0.05738	0.18308			
80597_01	OP	0.32	0.28	1.06450	1.70419	0.18749			
80597_02	OP	0.45	0.29	0.78747	1.37223	0.26027			
80597_05	OP	0.45	0.28	0.80357	1.41244	0.26683			
80615_01	OP	0.60	0.59	1.22329	0.08028	0.09338			
80615_02	OP	0.63	0.38	0.88500	0.46885	0.32061			
80615_04	OP	0.43	0.32	0.60452	1.17965	0.15725			
80697_04	OP	0.76	0.42	0.74737	-0.78566	0.13906			
80697_05	OP	0.64	0.37	0.97745	0.53708	0.36689			
80702_01	OP	0.44	0.36	0.90766	1.13991	0.20576			
80702_03	OP	0.66	0.40	0.93876	0.34636	0.32487			
80702_04	OP	0.43	0.40	1.07634	1.10195	0.18619			
80920_01	OP	0.68	0.42	0.71972	-0.23913	0.14681			
80920_02	OP	0.52	0.38	0.71615	0.68790	0.17597			
80928_03	OP	0.51	0.43	0.71253	0.47924	0.11726			
80928_04	OP	0.77	0.35	0.56233	-1.08940	0.11728			
80928_05	OP	0.47	0.27	0.35039	0.68244	0.04241			
80934_02	OP	0.75	0.41	0.87312	-0.24718	0.29726			
80934_05	OP	0.62	0.36	0.70109	0.39123	0.26092			
80935_03	OP	0.60	0.53	1.09569	0.17332	0.11472			
80935_04	OP	0.69	0.52	1.37851	0.06486	0.24865			
81041_01	OP	0.66	0.45	0.80931	-0.11879	0.17593			
81041_02	OP	0.63	0.33	0.45048	-0.31305	0.08292			
81041_05	OP	0.62	0.36	0.75661	0.48852	0.31060			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81044_01	OP	0.53	0.40	1.09538	0.81985	0.27659			
81044_03	OP	0.45	0.41	1.50976	1.04724	0.24062			
81052_02	OP	0.40	0.42	1.16894	1.11593	0.16263			
81052_03	OP	0.62	0.34	0.77278	0.64297	0.32918			
81053_01	OP	0.73	0.43	0.89097	-0.18767	0.25773			
81053_03	OP	0.73	0.33	0.48402	-0.98470	0.04023			
81058_01	OP	0.83	0.47	1.23480	-0.67041	0.25937			
81058_03	OP	0.57	0.36	0.61074	0.45564	0.17636			
81058_04	OP	0.63	0.42	0.70157	0.03152	0.13199			
85056_01	OP	0.37	0.31	1.07423	1.53183	0.20936			
85056_04	OP	0.57	0.31	0.45594	0.22724	0.11453			
85056_06	OP	0.90	0.42	1.26173	-1.26763	0.16508			
85056_08	OP	0.49	0.61	0.75313	0.38372	0.00000	2.01520	0.04140	-2.05660
85076_02	OP	0.77	0.51	1.30285	-0.44168	0.24556			
85076_03	OP	0.62	0.34	0.68707	0.45643	0.30372			
85076_04	OP	0.25	0.39	1.29987	1.52278	0.08906			
85076_08	OP	0.53	0.48	0.53275	0.02167	0.00000	3.20701	0.09968	-3.30670
80773	FT	0.53	0.28	0.34823	-0.04804	0.06339			
80941	FT	0.35	0.62	0.88120	1.30197	0.00000	2.17360	-0.18992	-1.98367
81004	FT	0.19	-.07	-0.17854	-6.85859	0.10101			
81013	FT	0.32	0.06	1.02247	2.77859	0.29160			
81014	FT	0.51	0.44	0.70961	0.48861	0.07392			
81021	FT	0.33	0.20	0.55912	2.30922	0.19362			
81025	FT	0.38	0.27	0.56974	1.66281	0.16452			
81159	FT	0.29	0.38	0.37021	2.16628	0.00000	2.89989	-0.69335	-2.20654
81160	FT	0.59	0.22	0.26496	-0.77496	0.04838			
81161	FT	0.38	0.05	0.07354	7.76212	0.13646			
81162	FT	0.54	0.45	0.85516	0.47090	0.16012			
81165	FT	0.58	0.33	1.06230	0.97530	0.37086			
81166	FT	0.50	0.30	0.90610	1.23412	0.31275			
81167	FT	0.38	0.19	0.85801	1.97696	0.27687			
81168	FT	0.35	0.28	0.83829	1.67216	0.18707			
81169	FT	0.37	0.28	0.78769	1.65046	0.20193			
81171	FT	0.33	0.14	1.05695	2.36055	0.27087			
81172	FT	0.54	0.45	0.95613	0.35711	0.23984			
81176	FT	0.50	0.36	0.62564	0.71095	0.15403			
81178	FT	0.28	0.62	0.78724	1.56888	0.00000	1.41918	0.06138	-1.48056
81179	FT	0.64	0.37	0.59271	-0.09821	0.12048			
81180	FT	0.63	0.53	1.05926	0.03388	0.11937			
81181	FT	0.65	0.43	1.10122	0.37495	0.31358			
81182	FT	0.77	0.39	0.82190	-0.41095	0.27808			
81183	FT	0.42	0.28	0.63284	1.45320	0.19476			
81184	FT	0.13	-.14	-0.44141	-3.37569	0.06448			
81187	FT	0.22	0.61	0.88432	1.94894	0.00000	1.53698	0.00182	-1.53881
81196	FT	0.52	0.31	0.51504	0.72666	0.15190			
81197	FT	0.70	0.55	1.47889	0.00583	0.23029			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81198	FT	0.45	0.36	0.75509	1.02759	0.17909			
80670_02	FT	0.58	0.30	0.62968	0.81960	0.30087			
80670_03	FT	0.27	0.17	0.95838	2.17515	0.20358			
80670_04	FT	0.19	-.01	-0.14949	-15.03837	0.17273			
80670_05	FT	0.30	0.22	0.91838	1.92584	0.21225			
80670_06	FT	0.19	0.14	1.07066	2.55460	0.13827			
80670_07	FT	0.32	0.57	0.79956	1.50936	0.00000	2.99900	-0.34636	-2.65264
80744_01	FT	0.12	-.03	-0.14358	-12.70838	0.08526			
80744_03	FT	0.36	0.01	0.02582	31.41982	0.19655			
80744_04	FT	0.34	0.14	0.61971	2.72518	0.25522			
80744_05	FT	0.50	0.07	0.08373	2.10875	0.10184			
80744_06	FT	0.51	0.30	0.89098	1.25909	0.31062			
80744_07	FT	0.18	0.54	0.77094	6.84952	0.00000	6.26313	4.44776	-10.71090
80745_01	FT	0.67	0.13	0.18766	-1.02978	0.17161			
80745_03	FT	0.77	0.34	0.62263	-0.62537	0.27591			
80745_04	FT	0.18	0.19	0.48216	3.06033	0.06762			
80745_05	FT	0.41	0.14	0.58117	2.55976	0.32122			
80745_07	FT	0.25	0.56	0.65270	1.82914	0.00000	1.66205	0.01214	-1.67419
80745_08	FT	0.43	0.53	0.72909	0.69011	0.00000	2.98400	-0.44657	-2.53741
80747_01	FT	0.30	0.38	0.70440	1.43872	0.04072			
80747_03	FT	0.70	0.33	0.61637	-0.07714	0.28148			
80747_04	FT	0.47	0.48	1.29029	0.80965	0.16712			
80747_05	FT	0.55	0.33	0.87129	0.97487	0.31828			
80747_06	FT	0.44	0.54	0.66961	0.74611	0.00000	2.59329	-0.12433	-2.46895
80747_07	FT	0.31	0.50	0.73904	1.92643	0.00000	3.04857	-0.53989	-2.50869
81046_01	FT	0.78	0.53	1.25276	-0.48316	0.13284			
81046_02	FT	0.57	0.38	0.71690	0.47799	0.19262			
81046_03	FT	0.48	0.25	0.33271	0.73396	0.05225			
81046_04	FT	0.20	0.12	0.75992	2.98407	0.14465			
81120_01	FT	0.58	0.28	0.98392	1.17287	0.40913			
81120_02	FT	0.44	0.43	1.24677	1.05895	0.18740			
81120_03	FT	0.51	0.13	0.17319	1.06959	0.12152			
81120_04	FT	0.66	0.40	0.63163	-0.26602	0.08007			
81121_02	FT	0.58	0.45	0.71104	0.09074	0.05494			
81121_03	FT	0.23	0.20	1.22197	2.12854	0.15534			
81121_04	FT	0.67	0.53	1.01771	-0.19048	0.08639			
81121_05	FT	0.66	0.53	1.02333	-0.13875	0.11414			
81122_01	FT	0.41	0.33	0.49929	1.06954	0.05620			
81122_02	FT	0.70	0.46	0.75399	-0.82899	0.10642			
81122_03	FT	0.54	0.39	0.52501	-0.06142	0.07671			
81122_04	FT	0.93	0.30	0.80024	-2.03422	0.10526			
81123_01	FT	0.65	0.50	0.89880	-0.09347	0.09771			
81123_02	FT	0.63	0.33	0.47577	-0.24376	0.08316			
81123_03	FT	0.35	0.38	1.35146	1.33183	0.16670			
81123_05	FT	0.50	0.48	1.00017	0.59985	0.14326			
81124_01	FT	0.65	0.30	0.42646	-0.37763	0.09550			

UIN	Status	Pvalue	Ptbis	a	b	c	d1	d2	d3
81124_02	FT	0.39	0.34	1.02915	1.36771	0.19760			
81124_03	FT	0.58	0.35	0.52928	0.13847	0.09843			
81124_04	FT	0.40	0.29	1.34405	1.50854	0.26711			
81125_01	FT	0.67	0.32	0.44317	-0.66513	0.04758			
81125_03	FT	0.68	0.45	0.75966	-0.31668	0.08896			
81125_04	FT	0.41	0.41	1.11929	1.08682	0.17653			
81125_05	FT	0.34	0.31	0.52666	1.47805	0.07788			
81127_01	FT	0.74	0.46	0.87999	-0.51928	0.10548			
81127_03	FT	0.81	0.42	0.91756	-0.66253	0.23484			
81127_04	FT	0.69	0.39	0.86837	0.15435	0.30365			
81127_05	FT	0.31	0.04	0.57454	4.50628	0.29413			
81129_01	FT	0.49	0.32	0.62320	0.85147	0.24198			
81129_02	FT	0.29	0.10	0.95555	2.80052	0.25519			
81129_03	FT	0.20	0.12	0.97884	2.61154	0.16342			
81129_04	FT	0.53	0.15	0.19377	0.64883	0.10837			
81131_01	FT	0.35	0.30	0.88795	1.63784	0.17834			
81131_02	FT	0.26	0.28	0.64961	1.99834	0.09271			
81131_03	FT	0.53	0.40	1.18968	0.89436	0.27957			
81131_05	FT	0.16	-.05	-0.30606	-6.07217	0.12726			
81133_01	FT	0.54	0.41	0.74125	0.55266	0.15950			
81133_02	FT	0.57	0.49	1.01853	0.11214	0.19932			
81133_04	FT	0.66	0.30	0.44136	-0.34058	0.12168			
81133_05	FT	0.68	0.47	0.95597	-0.36700	0.23703			
81134_02	FT	0.11	-.11	-0.54108	-3.48021	0.07198			
81134_03	FT	0.39	0.35	0.78857	1.29903	0.14531			
81134_04	FT	0.61	0.46	1.10181	0.40622	0.24227			
81134_05	FT	0.66	0.42	0.68429	-0.12005	0.11248			
81136_01	FT	0.23	0.08	0.39430	4.45640	0.16947			
81136_02	FT	0.64	0.43	0.77179	0.13333	0.16457			
81136_03	FT	0.47	0.26	0.61028	1.44966	0.25276			
81136_04	FT	0.90	0.24	0.50095	-2.52468	0.09531			
81138_01	FT	0.67	0.49	1.03925	0.03385	0.18699			
81138_02	FT	0.85	0.31	0.57795	-1.68889	0.06204			
81138_03	FT	0.71	0.42	0.69951	-0.50273	0.06640			
81138_04	FT	0.59	0.49	1.04925	0.28746	0.15711			
81139_01	FT	0.67	0.43	0.69224	-0.30297	0.07681			
81139_02	FT	0.50	0.43	1.17640	0.85024	0.21744			
81139_03	FT	0.79	0.43	0.80130	-0.90080	0.08162			
81139_04	FT	0.68	0.17	0.23688	-1.23014	0.08187			
85074_01	FT	0.94	0.29	0.88914	-1.94065	0.15774			
85074_02	FT	0.37	0.42	0.95760	1.15039	0.11233			
85074_03	FT	0.35	0.09	1.26250	2.39669	0.30632			
85074_04	FT	0.78	0.49	1.03907	-0.59069	0.11144			
85074_06	FT	0.46	0.21	0.35502	1.43744	0.16643			

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**
50481	-0.66	-0.0631			-0.42	-0.0395			0.45	0.0355		
50551	-0.22	-0.0055			-0.74	-0.0378			0.49	0.0199		
50929	-0.19	-0.0887			0.00	-0.1055	BB	W	-0.23	0.0916		
51008	-1.15	-0.0790	B	W	-0.11	-0.0029			-0.50	-0.0266		
51011	-1.09	-0.0747	B	W	-1.26	-0.0828	B	W	-0.12	-0.0067		
51016	-0.27	-0.0331			-0.17	-0.0118			-0.01	-0.0006		
51025	0.09	-0.0029			-0.25	-0.0317			-0.07	-0.0061		
51036	-0.19	-0.0684			0.00	-0.0287			-0.23	0.0696		
51038	-0.19	-0.0983			0.00	0.0088			-0.23	0.2317	CC	F
51040	-0.80	-0.0346			-0.51	-0.0216			-0.27	-0.0095		
51160	-0.33	-0.0341			0.47	0.0367			-0.64	-0.0563		
51161	-0.54	-0.0448			-0.33	-0.0246			-1.18	-0.0943	B	M
51164	0.18	0.0116			0.12	0.0109			0.42	0.0305		
51167	0.33	0.0318			0.54	0.0422			0.07	0.0056		
51168	-1.36	-0.1022	B	W	-1.04	-0.0792	B	W	-0.05	-0.0038		
51169	0.36	0.0329			0.07	0.0093			0.74	0.0563		
51170	-0.31	-0.0173			0.11	0.0098			-0.49	-0.0468		
51171	-0.59	-0.0286			-0.46	-0.0241			1.03	0.0512	B	F
51174	0.26	0.0109			-0.02	-0.0067			-0.18	-0.0175		
51176	-0.27	-0.0248			-0.33	-0.0334			0.36	0.0363		
51180	-0.19	-0.1073			0.00	-0.0900			-0.23	0.0345		
51183	-0.39	-0.0153			-0.36	-0.0304			0.21	0.0158		
51184	-0.38	-0.0370			0.07	0.0032			-0.23	-0.0191		
51187	-0.19	-0.1702	BB	W	0.00	-0.0227			-0.23	-0.0143		
51188	-0.27	-0.0164			-0.37	-0.0288			0.30	0.0194		
51190	-0.01	-0.0095			-0.44	-0.0396			-0.63	-0.0583		
51194	-0.48	-0.0221			-0.81	-0.0707			0.12	0.0118		
51195	-0.90	-0.0808			-0.94	-0.0936			-0.03	-0.0030		
51197	-1.23	-0.0959	B	W	-0.59	-0.0347			-0.60	-0.0344		
51199	-0.19	-0.1643	BB	W	0.00	-0.1438	BB	W	-0.23	0.2014	CC	F
50926_01	-0.67	-0.0640			-1.18	-0.1171	B	W	-0.09	-0.0088		
50926_02	-0.09	-0.0065			-0.51	-0.0323			0.98	0.0525		
50926_04	-0.31	-0.0220			0.09	0.0143			0.43	0.0350		
50926_05	-0.31	-0.0135			-0.30	-0.0242			0.21	0.0192		
51041_02	-0.68	-0.0485			-0.59	-0.0526			-0.47	-0.0401		
51041_03	-0.14	-0.0140			0.43	0.0358			0.18	0.0162		
51041_04	-1.02	-0.0892	B	W	-0.87	-0.0791			-0.76	-0.0596		
51041_05	0.59	0.0230			0.67	0.0573			0.11	0.0095		
51051_01	-0.58	-0.0332			0.04	0.0018			-0.18	-0.0066		
51051_02	-0.90	-0.0797			-0.61	-0.0560			0.18	0.0163		
51051_03	-0.83	-0.0542			-0.18	-0.0075			-0.21	-0.0170		
51051_04	-0.12	-0.0055			0.22	0.0181			-0.12	-0.0101		
51091_01	-0.52	-0.0468			-0.18	-0.0156			0.28	0.0249		
51091_04	0.07	0.0224			-0.49	-0.0408			0.01	0.0022		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
51091_05	0.10	0.0010			0.56	0.0553			0.14	0.0122		
51091_06	0.10	-0.0010			-0.10	-0.0088			0.06	0.0057		
51091_07	-0.19	-0.0162			0.00	0.0447			-0.23	0.1246	BB	F
51091_08	-0.19	-0.0350			0.00	-0.0548			-0.23	0.0798		
51092_01	-0.94	-0.0789			-1.19	-0.1043	B	W	-0.35	-0.0282		
51092_02	-0.24	-0.0143			-0.01	-0.0026			-0.21	-0.0148		
51092_03	-0.02	0.0030			-0.09	-0.0044			-0.09	-0.0075		
51092_04	-0.28	-0.0247			-0.44	-0.0424			0.14	0.0132		
51092_05	0.34	0.0316			0.25	0.0201			-0.14	-0.0136		
51092_06	-0.40	-0.0441			-0.18	-0.0241			0.20	0.0179		
51093_01	-0.72	-0.0636			-0.42	-0.0340			-0.93	-0.0671		
51093_02	0.13	0.0109			0.07	0.0093			0.06	0.0052		
51093_03	-0.50	-0.0433			-0.53	-0.0486			-1.01	-0.0965	B	M
51093_04	-0.62	-0.0480			0.01	0.0017			-0.24	-0.0193		
51093_05	-0.13	-0.0084			0.08	0.0070			0.22	0.0147		
51093_07	-0.19	0.0008			0.00	0.0008			-0.23	0.0406		
51121_01	0.38	0.0241			-0.71	-0.0493			0.21	0.0154		
51121_03	-0.12	-0.0185			-0.55	-0.0614			0.09	0.0087		
51121_04	0.97	0.0360			0.52	0.0190			0.08	0.0033		
51121_05	-0.21	-0.0121			-0.53	-0.0354			-0.16	-0.0086		
51122_01	-0.83	-0.0122			-1.66	-0.0336	C	W	-0.31	-0.0045		
51122_03	-0.85	-0.0562			-0.27	-0.0182			-0.07	-0.0046		
51122_04	-0.34	-0.0378			-0.49	-0.0413			-0.21	-0.0177		
51122_05	-0.38	-0.0262			-0.17	-0.0163			0.15	0.0138		
51123_01	-0.18	-0.0034			0.08	0.0077			0.44	0.0409		
51123_02	0.31	0.0377			0.14	0.0086			-0.52	-0.0499		
51123_04	0.13	0.0061			-0.05	-0.0132			0.73	0.0697		
51123_05	-0.13	-0.0368			0.30	0.0290			-0.28	-0.0289		
51124_01	-0.30	-0.0183			-0.48	-0.0397			-0.23	-0.0185		
51124_02	-0.41	-0.0298			-0.35	-0.0331			-0.29	-0.0275		
51124_04	0.14	0.0133			0.05	0.0080			0.09	0.0084		
51124_05	0.27	0.0159			0.36	0.0276			-0.23	-0.0195		
51125_01	-0.26	-0.0166			-0.21	-0.0126			0.06	0.0046		
51125_02	-0.47	-0.0253			-0.64	-0.0376			0.01	0.0008		
51125_03	-0.39	-0.0392			0.31	0.0271			-0.62	-0.0494		
51125_05	-0.02	-0.0094			0.01	0.0031			-0.13	-0.0118		
51126_01	0.11	0.0130			-0.13	-0.0001			-0.73	-0.0248		
51126_02	-0.60	-0.0398			-0.26	-0.0174			0.82	0.0506		
51126_03	-0.54	-0.0510			-0.18	-0.0176			-0.93	-0.0901		
51126_04	-0.19	-0.0174			0.00	0.0004			-0.23	-0.0168		
51127_01	0.10	0.0250			0.35	0.0243			-0.35	-0.0220		
51127_02	-0.27	-0.0327			-0.24	-0.0263			-0.03	-0.0032		
51127_03	-0.19	-0.0189			-0.43	-0.0366			0.19	0.0182		
51127_04	0.13	-0.0013			0.38	0.0339			0.54	0.0463		
51128_01	0.39	0.0159			0.05	0.0002			0.31	0.0282		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
51128_02	-0.20	-0.0208			0.34	0.0391			0.46	0.0485		
51128_03	-0.32	-0.0261			-0.34	-0.0365			-0.77	-0.0776		
51128_04	-0.60	-0.0329			0.03	0.0046			-0.63	-0.0488		
51131_01	-0.41	-0.0145			-0.65	-0.0263			-0.70	-0.0279		
51131_02	-0.13	-0.0175			0.78	0.0660			-0.05	-0.0034		
51131_04	-0.69	-0.0326			-0.28	-0.0231			-0.86	-0.0711		
51131_05	-1.21	-0.0440	B	W	-0.62	-0.0176			0.59	0.0167		
51134_01	0.14	0.0036			-0.41	-0.0390			0.42	0.0400		
51134_02	-0.03	-0.0012			-0.31	-0.0238			0.19	0.0171		
51134_03	-0.71	-0.0545			-0.49	-0.0403			-0.54	-0.0432		
51134_05	0.13	0.0069			-0.05	-0.0048			0.05	0.0049		
51135_01	-0.06	-0.0153			-0.66	-0.0687			-0.44	-0.0452		
51135_02	0.93	0.0784			0.56	0.0539			-0.12	-0.0116		
51135_03	-1.01	-0.0525	B	W	-0.78	-0.0432			-0.22	-0.0101		
51135_04	-0.01	-0.0056			-0.02	-0.0004			0.65	0.0549		
51138_01	-0.67	-0.0654			-0.49	-0.0484			-0.40	-0.0363		
51138_02	-0.12	-0.0018			-0.20	-0.0114			-0.12	-0.0047		
51138_03	-0.60	-0.0485			-0.35	-0.0294			-0.35	-0.0294		
51138_05	-0.25	-0.0382			-0.27	-0.0302			-0.10	-0.0097		
55074_01	0.45	0.0374			-0.36	-0.0388			-0.04	-0.0038		
55074_02	-0.15	-0.0171			-0.46	-0.0394			-0.26	-0.0208		
55074_03	-0.52	-0.0349			-0.13	-0.0116			-0.51	-0.0459		
55074_04	-0.35	-0.0390			-0.40	-0.0363			0.07	0.0073		
55074_05	-0.07	-0.0061			-0.17	-0.0131			0.28	0.0286		
55074_06	-0.85	-0.0584			-0.26	-0.0180			0.30	0.0276		
55083_01	-0.93	-0.0780			-0.46	-0.0363			-0.22	-0.0170		
55083_02	-0.28	-0.0304			-0.64	-0.0599			-0.29	-0.0277		
55083_03	-0.95	-0.0766			-0.65	-0.0546			0.60	0.0476		
55083_05	-0.39	-0.0273			0.24	0.0224			0.23	0.0176		
55083_06	-0.19	-0.1182			0.00	-0.0753			-0.23	0.1606	BB	F
55083_07	-0.51	-0.0554			-0.08	-0.0087			0.33	0.0253		

IN=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**
80773	-0.15	-0.0219			-0.42	-0.0462			-0.34	-0.0345		
80941	0.22	-0.2214	CC	W	0.13	-0.1913	CC	W	-0.07	0.1515	BB	F
81004	-0.36	-0.0198			-0.37	-0.0260			-0.32	-0.0212		
81013	0.68	0.0556			-0.01	0.0035			0.14	0.0137		
81014	-0.27	-0.0142			-0.13	-0.0111			-0.09	-0.0076		
81021	-0.01	0.0035			0.04	0.0057			0.01	0.0031		
81025	-0.05	0.0048			-0.01	0.0057			-0.20	-0.0187		
81159	0.22	0.0152			0.13	-0.0716			-0.07	0.1147		
81160	-0.09	0.0070			-0.20	-0.0215			0.28	0.0262		
81161	0.31	0.0235			0.21	0.0248			0.26	0.0252		
81162	0.15	0.0061			0.18	0.0150			0.56	0.0479		
81165	-0.88	-0.0727			-0.04	-0.0039			-0.13	-0.0116		
81166	-0.58	-0.0420			-0.41	-0.0410			-0.89	-0.0872		
81167	-0.42	-0.0194			-0.17	-0.0162			-0.20	-0.0196		
81168	-0.45	-0.0382			-0.02	-0.0099			-0.74	-0.0649		
81169	-0.28	-0.0102			-0.16	-0.0106			0.07	0.0069		
81171	-0.38	-0.0175			0.16	0.0100			-0.60	-0.0558		
81172	0.02	-0.0048			-0.31	-0.0377			-0.05	-0.0032		
81176	-0.50	-0.0434			-0.08	-0.0071			-0.68	-0.0633		
81178	0.22	-0.1196			0.13	-0.1047			-0.07	0.1005		
81179	-0.00	0.0059			-0.09	-0.0081			0.32	0.0286		
81180	-1.20	-0.0971	B	W	-1.30	-0.1075	B	W	-0.60	-0.0431		
81181	-0.60	-0.0273			0.11	0.0078			-0.26	-0.0204		
81182	-0.10	-0.0054			-0.25	-0.0205			-0.18	-0.0113		
81183	0.06	0.0014			0.29	0.0249			0.37	0.0352		
81184	-0.39	-0.0158			0.16	0.0084			-0.37	-0.0182		
81187	0.22	-0.0209			0.13	-0.0405			-0.07	0.2259	CC	F
81196	0.66	0.0749			0.95	0.0932			0.17	0.0164		
81197	-0.03	-0.0015			-0.61	-0.0465			0.09	0.0063		
81198	-0.31	-0.0105			-0.58	-0.0516			-0.54	-0.0491		
80670_02	-0.33	-0.0352			-0.86	-0.0874			-0.35	-0.0328		
80670_03	0.12	0.0066			-0.23	-0.0136			0.04	0.0045		
80670_04	-0.27	-0.0212			-0.49	-0.0368			-0.28	-0.0184		
80670_05	-0.17	-0.0077			-0.00	0.0030			0.22	0.0193		
80670_06	-0.25	-0.0067			0.48	0.0267			-0.56	-0.0347		
80670_07	0.22	-0.0713			0.13	-0.0854			-0.07	0.1149	BB	F
80744_01	0.36	0.0235			-0.16	-0.0021			-0.16	-0.0073		
80744_03	-0.24	-0.0200			-0.00	0.0000			0.73	0.0738		
80744_04	0.32	0.0206			-0.02	-0.0058			0.00	0.0008		
80744_05	0.45	0.0463			-0.46	-0.0447			0.23	0.0246		
80744_06	-0.18	0.0023			0.19	0.0232			0.11	0.0089		
80744_07	0.22	0.0064			0.13	-0.0373			-0.07	0.1414	BB	F
80745_01	-0.22	-0.0108			-0.12	-0.0115			0.24	0.0233		
80745_03	-0.62	-0.0479			-0.77	-0.0648			-0.84	-0.0573		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
80745_04	-0.30	-0.0243			-0.55	-0.0321			0.42	0.0258		
80745_05	-0.06	-0.0104			-0.27	-0.0331			-0.14	-0.0132		
80745_07	0.22	-0.1359	BB	W	0.13	-0.1890	BB	W	-0.07	0.0290		
80745_08	0.22	-0.0889			0.13	-0.0058			-0.07	0.1848	CC	F
80747_01	-0.47	-0.0264			0.07	0.0072			0.08	0.0062		
80747_03	0.42	0.0421			0.00	-0.0011			0.39	0.0314		
80747_04	-0.55	-0.0496			-0.73	-0.0598			0.17	0.0137		
80747_05	-0.39	-0.0410			0.42	0.0381			-0.81	-0.0764		
80747_06	0.22	-0.0649			0.13	-0.0654			-0.07	0.3385	CC	F
80747_07	0.22	-0.1379	CC	W	0.13	-0.0952	BB	W	-0.07	0.0062		
81046_01	-0.12	-0.0150			0.97	0.0550			0.34	0.0182		
81046_02	0.22	0.0207			0.13	0.0105			-0.07	-0.0065		
81046_03	-0.17	-0.0042			-0.17	-0.0142			-0.38	-0.0379		
81046_04	0.13	0.0075			-0.40	-0.0267			-0.53	-0.0338		
81120_01	-0.22	-0.0315			-0.49	-0.0587			-0.84	-0.0810		
81120_02	-0.32	-0.0096			0.08	0.0064			-1.31	-0.1115	B	M
81120_03	0.57	0.0658			-0.05	-0.0014			0.31	0.0330		
81120_04	0.23	0.0164			0.10	0.0094			-0.24	-0.0187		
81121_02	-0.25	-0.0215			0.04	-0.0011			0.25	0.0223		
81121_03	0.11	0.0088			0.40	0.0308			-0.78	-0.0534		
81121_04	-0.05	-0.0170			-0.20	-0.0154			1.25	0.0880	B	F
81121_05	0.61	0.0518			-0.05	-0.0049			0.79	0.0556		
81122_01	0.18	0.0026			-0.59	-0.0517			0.11	0.0117		
81122_02	-0.63	-0.0501			-0.64	-0.0566			-0.76	-0.0531		
81122_03	0.21	0.0169			-0.61	-0.0623			0.05	0.0054		
81122_04	-0.48	-0.0196			-0.95	-0.0322			0.18	0.0043		
81123_01	-0.54	-0.0462			-0.65	-0.0524			-0.42	-0.0325		
81123_02	0.26	0.0247			-0.02	-0.0040			-0.23	-0.0225		
81123_03	-0.44	-0.0322			-0.06	-0.0005			-0.55	-0.0434		
81123_05	-0.28	-0.0310			-0.15	-0.0145			0.37	0.0315		
81124_01	-0.05	-0.0124			-0.30	-0.0281			0.04	0.0041		
81124_02	-0.32	-0.0261			0.14	0.0097			-0.23	-0.0206		
81124_03	-0.35	-0.0390			0.27	0.0230			0.22	0.0187		
81124_04	-0.06	0.0058			0.01	0.0019			-0.09	-0.0102		
81125_01	-0.32	-0.0258			-0.15	-0.0174			0.80	0.0691		
81125_03	-0.41	-0.0171			-0.01	-0.0014			-0.47	-0.0350		
81125_04	-0.76	-0.0405			0.65	0.0546			-0.12	-0.0104		
81125_05	-0.22	-0.0111			-1.15	-0.0872	B	W	0.25	0.0212		
81127_01	-0.29	-0.0215			0.09	0.0043			-0.20	-0.0131		
81127_03	0.37	0.0243			0.78	0.0505			0.02	0.0015		
81127_04	-0.07	0.0071			0.08	0.0068			-0.02	-0.0013		
81127_05	0.21	0.0268			-0.35	-0.0282			-0.32	-0.0292		
81129_01	-0.53	-0.0437			-0.55	-0.0561			-0.28	-0.0268		
81129_02	-0.02	0.0052			-0.36	-0.0298			-0.26	-0.0205		
81129_03	-0.45	-0.0198			-0.40	-0.0242			-0.48	-0.0323		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
81129_04	-0.29	-0.0359			-0.06	-0.0091			0.49	0.0506		
81131_01	-0.77	-0.0501			-0.46	-0.0382			-0.79	-0.0682		
81131_02	-0.16	-0.0021			-0.14	-0.0101			-0.23	-0.0173		
81131_03	0.14	0.0126			0.06	0.0067			-0.53	-0.0483		
81131_05	-0.40	-0.0223			0.04	0.0006			-0.17	-0.0092		
81133_01	-0.80	-0.0767			0.14	0.0038			0.05	0.0043		
81133_02	-0.41	-0.0370			-0.99	-0.0925			-0.64	-0.0517		
81133_04	-0.50	-0.0346			-0.63	-0.0617			0.74	0.0644		
81133_05	-0.52	-0.0402			-1.00	-0.0875			0.17	0.0121		
81134_02	0.49	0.0206			0.55	0.0207			-0.30	-0.0118		
81134_03	-0.59	-0.0637			-0.70	-0.0629			-0.15	-0.0120		
81134_04	-0.65	-0.0525			-0.64	-0.0578			-0.34	-0.0276		
81134_05	-0.40	-0.0434			-0.09	-0.0122			0.01	0.0006		
81136_01	-0.24	-0.0052			0.25	0.0165			0.37	0.0273		
81136_02	0.07	0.0022			-0.53	-0.0490			0.14	0.0124		
81136_03	-0.36	-0.0233			-0.01	-0.0025			0.06	0.0061		
81136_04	0.05	0.0137			0.15	0.0041			0.79	0.0272		
81138_01	-0.34	-0.0232			0.08	0.0048			-0.37	-0.0270		
81138_02	-0.98	-0.0594			-0.51	-0.0287			0.16	0.0073		
81138_03	-0.28	-0.0209			-0.25	-0.0231			0.06	0.0057		
81138_04	-0.24	-0.0202			0.06	0.0029			-0.38	-0.0309		
81139_01	-0.40	-0.0362			-0.58	-0.0509			0.36	0.0286		
81139_02	-0.52	-0.0405			-0.21	-0.0163			-0.72	-0.0628		
81139_03	-1.02	-0.0647	B	W	-0.82	-0.0541			0.62	0.0353		
81139_04	0.22	0.0269			0.94	0.0784			0.30	0.0282		
85074_01	-0.37	-0.0132			-0.26	-0.0057			-0.83	-0.0182		
85074_02	-0.29	-0.0254			0.16	0.0172			-0.30	-0.0252		
85074_03	0.61	0.0381			0.36	0.0314			0.12	0.0121		
85074_04	0.27	0.0095			-0.07	-0.0046			-0.47	-0.0255		
85074_06	-0.08	-0.0053			-0.30	-0.0337			0.23	0.0230		

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)