Maryland School Assessment (MSA)
Science

Grades 5 and 8

Technical Report
2010 Operational Test

October 2010

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 2009-2010 test administration cycle.

Purpose
The purpose of this MSA Technical Report is to provide objective information regarding technical aspects of the 2010 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 covers all of these topics for the 2009-2010 testing year.

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

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

Purpose and Use

By assessing student achievement against the Science academic standards, the MSA Science test serves two important purposes. First, the MSA Science test provides an accountability tool to measures performance levels of students, schools, and districts against the Science academic standards. Second, it provides 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 four test administrations—a census field test in 2007 and three operational tests (2008 through 2010). 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 2010 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 2010 MSA Science tests, students selected an answer from four options. Each SR item was scored dichotomously (i.e., 0 or 1). BCR items require students to provide a short answer using words, numbers, and/or symbols. All BCR items are scored using a generic rubric and scores range from 0-3 based on concordant scores from two independent raters. In cases where the scores differ by one point, the higher score is used. In cases where the rater scores differ by two or more points, a third expert rater's independent score is used as a resolution.

MSA Science 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 2010 Operational Test Construction

The 2010 operational tests were created according to the test blueprints (see Table 1 and 2) and reflective of the Voluntary State Curriculum (VSC) in the form of measureable Indicators and Objectives. As such, each of the two operational forms yielding student scores has the same test composition as that of 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 2010 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 2010 MSA Science operational test forms was an iterative process primarily involving Pearson content experts, MSDE, and Pearson psychometricians. Initial test forms were created to meet the respective blueprints, reflect the VSC measureable Indicators and Objectives, and align with statistical characteristics of the 2008 operational tests. Only items deemed eligible after being administered live (field tested) and reviewed by content experts based on statistical indicators (see Data Review of the Field Test Items) were used. Additional content-related characteristics that were part of the creation of the operational test forms had to do with ensuring there was no cuing from one item to the next. That is, items were scrutinized to make sure nothing in any one question or passage would provide information relevant to answering any other item correctly.

Classical item statistics were used in conjunction with item response theory (IRT) statistics to help target the overall test forms. The guiding principles were choosing items with reasonably strong point biserial correlations (>.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 2010 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. 2010 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 2010 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 2010 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 2010 field test. Overall, development was structured to spread the items across the six standards specified within the Maryland (Voluntary) State Curriculum (VSC/MSC) and across the topics, indicators, objectives and assessment limits within each standard. Targets were developed at both grades 5 and 8; item development began once the development targets were finalized. The target number of items developed in 2009 for the 2010 administration was approximately 180 items for each grade: 155 SR and 25 BCR items.

During 2008 published technical passages to be approved for item development were selected and reviewed by Pearson content staff, MSDE content experts, and three separate Maryland content and bias committees. An item writer training was held in early December 2009. Current or former non-Maryland Science educators were recruited to write items and lab stimuli on behalf of the program. During the training, writers were introduced to a number of topics by both MSDE and Pearson staff. Topics for training included:

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

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

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

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

Additionally, Pearson content specialists reviewed all items within each grade for the full range of item difficulty and consideration of a range of cognitive complexity. Cognitive complexity refers how items are solved. For example, complexity may range from items where students only need to rely on memory to answer a question versus having to evaluate and synthesize something to respond correctly. After this review, items went through an iterative development process between content specialist and copy editors, universal design specialists, and research librarians. In addition, all art and graphical supports for the items were produced. Finally, all BCR items were reviewed by Pearson Performance Scoring Center staff for scorability. Once Pearson completed the internal development, items were released to MSDE for review via Pearson's Item Tracker system. In May of 2009, Pearson and MSDE content experts met to review and discuss each new item and collaborate on revisions. Once revisions were made and reviewed again through the internal Pearson development team, the items were prepared for another series of content and bias reviews in Maryland.

Review panels of Maryland residents were convened in July 2009. Three different panels were convened to review items for each grade. Content review was conducted at each grade by Maryland educators within the appropriate grade range to further confirm content accuracy and grade-level appropriate vocabulary and language and to identify and discuss potential improvements to the item stem or distractors. A separate bias/sensitivity panel at each grade was convened to examine the items for any possible socio-economic, geographical, cultural or gender biases. Finally, another committee of educators reviewed item text and graphics with particular focus on possible issues for blind or visually impaired students. Before reviewing materials, MSDE and Pearson provided an overview to the panelists on the purpose of each panel, the VSC/MSC, and the criteria by which they were asked to evaluate the items. Since the evaluation criteria were different, the content panelists and bias/sensitivity panelists were trained separately.

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

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

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

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

In addition to the panels reviewing the items to be field tested in spring 2010, 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 2011 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 2011 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	31796	53.09	38698	62.69
Paper	28095	46.91	23031	37.31
Form				
1	5731	9.57	5896	9.55
2	5744	9.59	5963	9.66
3	6486	10.83	5870	9.51
4	5656	9.44	7219	11.69
5	5751	9.60	5867	9.50
6	5692	9.50	7128	11.55
7	5747	9.60	6037	9.78
8	7594	12.68	5863	9.50
9	5665	9.46	5966	9.66
10	5825	9.73	5920	9.59
Gender				
Female	29071	48.54	30291	49.56
Male	30809	51.44	31419	50.41
Unknown	11	0.02	19	0.03
Ethnicity				
Native American	225	0.38	198	0.32
Asian	3693	6.17	3749	6.07
African American	22759	38.00	22991	37.25
White	27512	45.94	28996	46.97
Hispanic	5689	9.50	5775	9.36
Unknown	13	0.02	20	0.03
All	**59891**	**100.00**	**61729**	**100.00**

* Differences in values reflect missing data

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. online Form 8) 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	2977	3017	3596	2882	2957	2883	2917	4777	2828	2962
	Paper	2754	2727	2890	2774	2794	2809	2830	2817	2837	2863
	Overall	5731	5744	6486	5656	5751	5692	5747	7594	5665	5825
Grade 8	Online	3676	3731	3650	4978	3627	4318	3772	3599	3712	3635
	Paper	2220	2232	2220	2241	2240	2810	2265	2264	2254	2285
	Overall	5896	5963	5870	7219	5867	7128	6037	5863	5966	5920

Key Check Analysis of Operational Test Data

Using preliminary data collected from the 2010 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 2010 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:

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

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

Polytomous Item Response Theory Model

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

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

where

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

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

Calibration of the mixed test format (3PL/GPC model) items was conducted using 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 2010 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, 2008, or 2009 field test calibrations or from the 2008 and 2009 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 2010 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 four items removed from the grade 5 common item set and six items removed from the grade 8 common item set. The final transformation constants for each grade following this procedure are listed in Table 6.

Table 6. Operational Transformation Constants

	Grade 5		**Grade 8**	
	Slope	**Intercept**	**Slope**	**Intercept**
Operational (10 OP items -> 07 base scale)	1.007342	0.181907	1.075326	0.204608

The transformation constants were applied to the 2010 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.

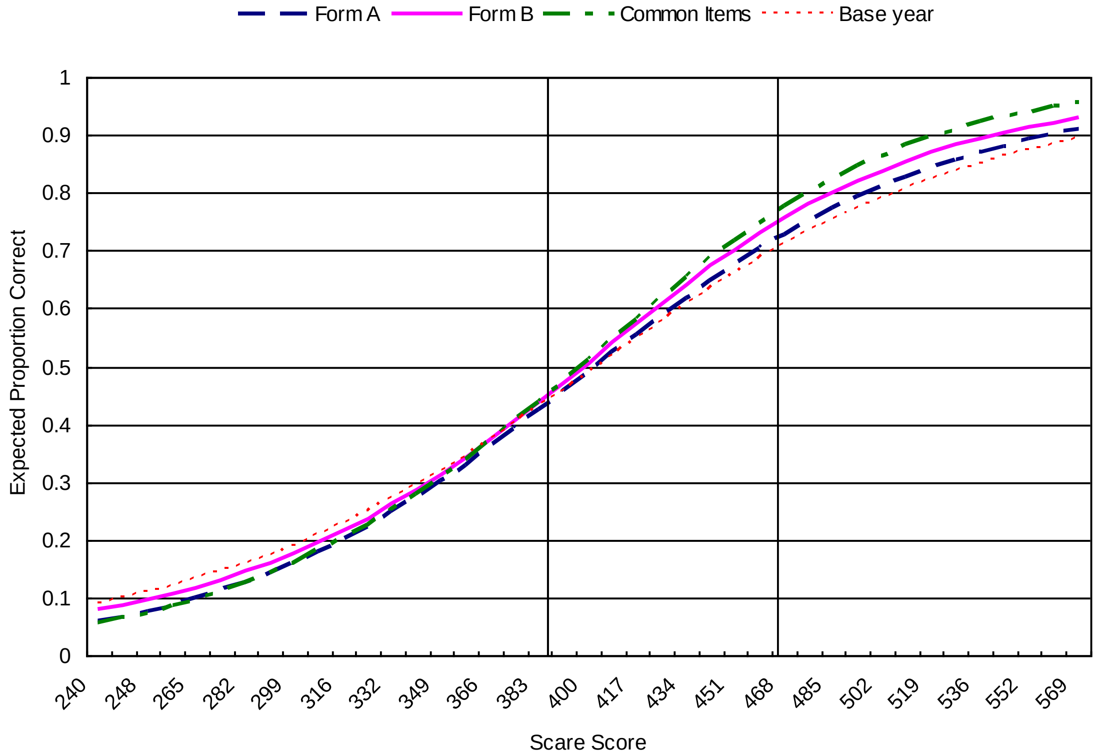
Test Analysis, Operational Scaling and Scoring

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 2010 Form A and B curves and curves reflective of operational items from the 2008 (base year) administration.

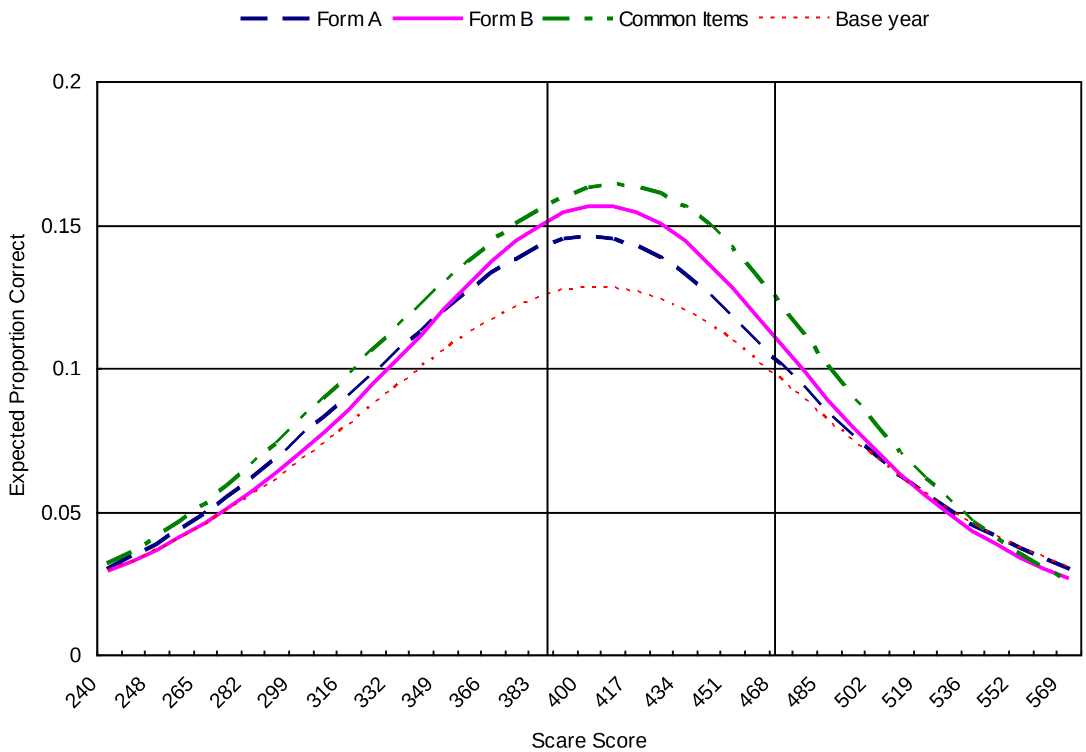
Figure 1 shows the overlaid TCC plots for Form A, Form B, form-to-form linking items and base-year item pool for grade 5. These plots illustrate that the operational form A and B scales are very closely aligned to the base scale (and to each other). Figure 2 also displays test information curves for 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 reportable scale metric (each performance level is denoted at the top of the plot: Basic, Proficient, and Advanced). It should also be noted that each curve is presented according to the MSA Science scale score metric, which is described in the Defining Scale Ranges section.

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



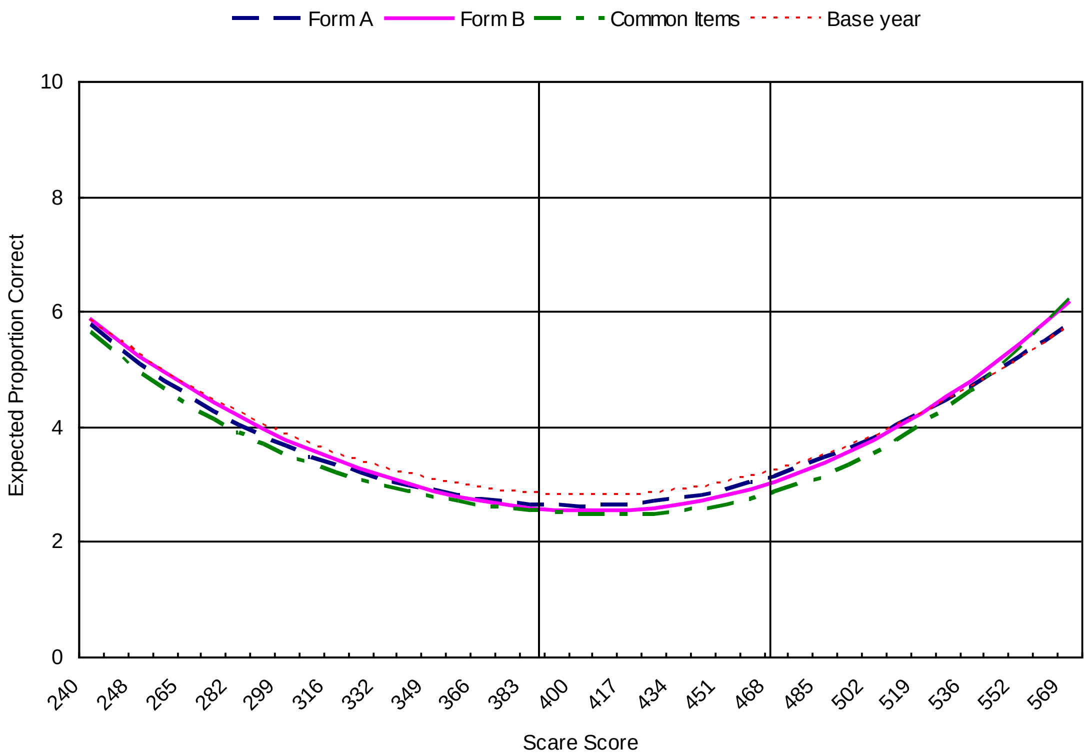
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

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



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

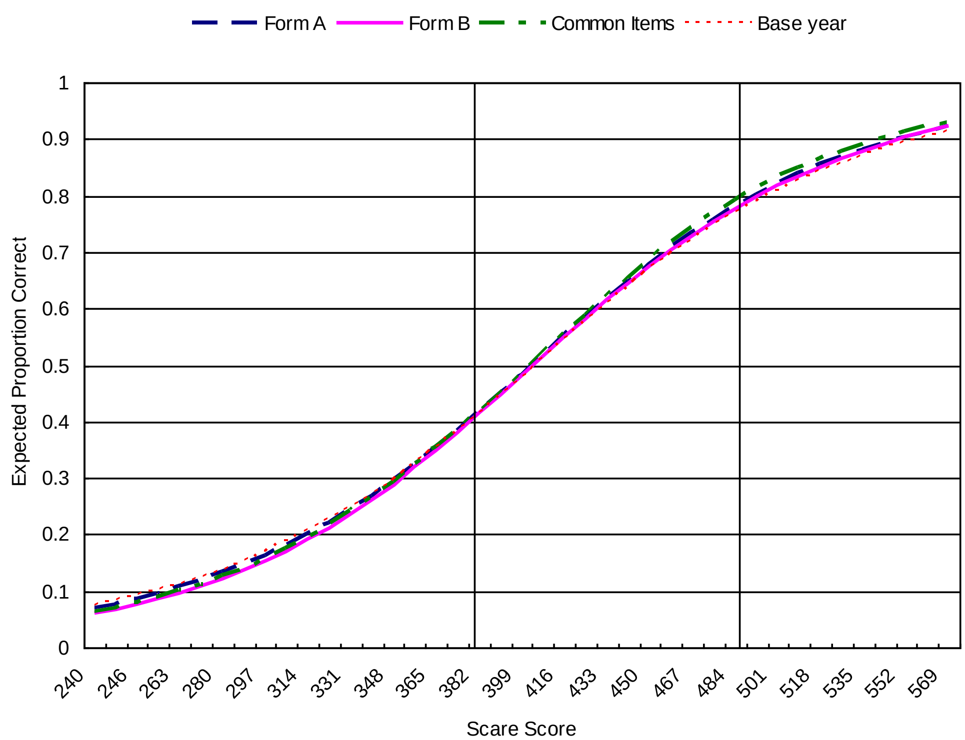
Figure 3. Conditional Standard Error of Measurement for the Grade 5 Science Test



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

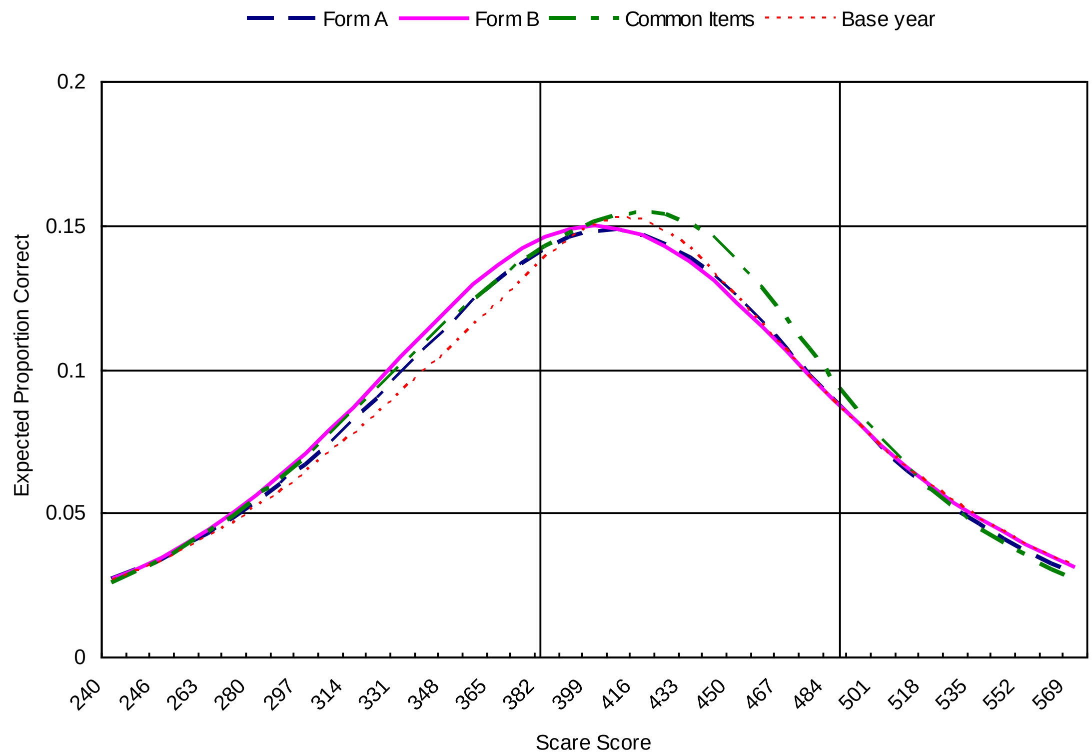
Similar to grade 5, IRT item parameter estimates were used to generate characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the base forms, form-to-form linking items, and base-year operational test for grade 8. Figure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

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



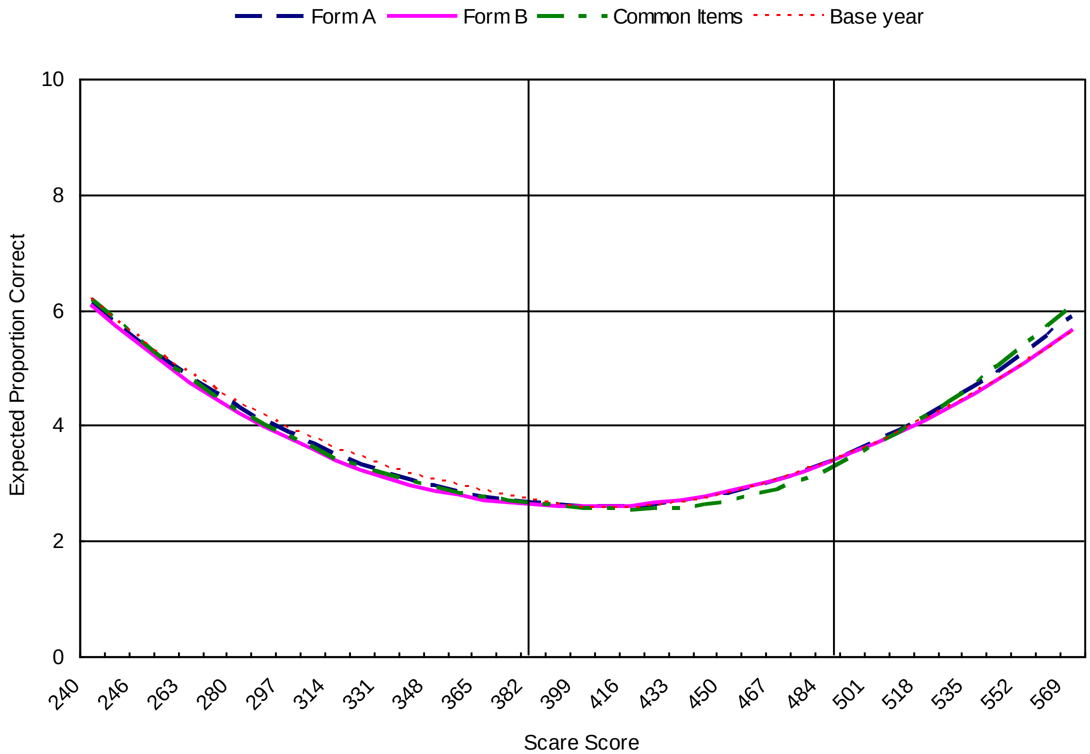
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

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



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

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



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

Defining Scale Ranges

The theta scale is not often used for reporting because of interpretation issues arising from a scale with values typically ranging from -4.0 to +4.0. Therefore, following the calibration and equating phases, the resulting theta values are transformed to a reporting scale 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 2010 MSA Science operational administration, scale scores were created through the application of scaling constants determined from the base 2007 test administration. Scale scores were computed using the following simple linear transformation equation:

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

where, M1 is a multiplicative term, M2 is an additive term, and θ is an IRT based measure of student ability. These scaling constants (M1 and M2) were developed to meet MSDE requirements that the mean and standard deviation (sd) be established in the base year at mean scale score = 400 and sd = 40, while maintaining the lowest obtainable scale score (LOSS) at 240 and the highest obtainable scale score (HOSS) at 650. The LOSS and HOSS set the minimum and maximum values that are possible on the MSA Science test. These scaling constants as well as the LOSS and HOSS for each grade appear in Table 7.

Table 7. Target LOSS, HOSS, and Scaling Constants for Grades 5 and 8.

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

ISE Pattern Scoring

Pearson used an internally developed software program called IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2007) to conduct pattern scoring for the spring 2010 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 2010 operational scores were estimated by the pattern scoring approach. The 2010 operational item parameters were first equated to the base theta scale established in 2007. The equated item parameters were then used to estimate student ability (theta) using Pearson's ISE program. The theta estimates were transformed onto the MSA Science operational scale using the scaling constants described above.

Conditional Standard Errors for LOSS and HOSS

Within ISE, student ability (theta) is determined via maximum likelihood estimation (MLE). One characteristic of MLE is that for students with scores of zero or perfect scores, abilities are not estimable (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	0.93	0.94	0.94
Gender	Male	0.93	0.93	0.94	0.94
	Female	0.92	0.92	0.93	0.93
Ethnicity	Native American	0.90	0.91	0.92	0.92
	Asian	0.92	0.93	0.93	0.93
	Black	0.90	0.91	0.92	0.92
	White	0.91	0.92	0.93	0.92
	Hispanic	0.91	0.91	0.93	0.92

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

Student Performance

Score Interpretation

To help provide appropriate interpretation of the 2010 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 2010 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 2010 MSA Science scale scores. They describe what students at a particular level generally know and are able to do and can be applicable to all students within a grade level.

Performance standards for the MSA Science tests were established in 2007. Details of the standard-setting process and outcomes are provided in MSA Science standard-setting technical report (Pearson, 2007). The Maryland State Board of Education reviewed the performance standards recommended by the standard-setting committee and made a modification in the recommendation. The performance standards approved by the State Board are listed in Table 9. Students whose scale scores are lower than the Proficient cut score are classified as "Basic." The highest performance group whose scale score is equal or higher than Advanced cut score belongs to the "Advanced" group. The middle group is called "Proficient."

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

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

Tables 10 reports percentages of grade 5 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 5 as well as by administration mode.

Table 10. Grade 5 Performance Level Percentages and Summary Statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	34	56	9	408	45.3	59891	33	58	9	409	43.3	31796	35	54	10	408	47.5	28095
Gender																		
Male	34	55	11	409	47.0	30809	33	57	10	410	44.9	16204	36	53	11	408	49.2	14605
Female	34	58	8	408	43.5	29071	33	60	7	408	41.7	15592	35	56	9	408	45.5	13479
Ethnicity																		
Native American	33	62	5	407	40.9	225	33	63	4	405	40.7	125	34	60	6	410	41.2	100
Asian	17	63	19	429	45.2	3693	18	65	17	427	43.8	1736	17	62	21	432	46.3	1957
Black	52	45	3	388	40.7	22759	51	47	2	389	39.3	11187	53	44	3	387	41.9	11572
White	18	66	15	426	41.0	27512	20	67	13	423	40.1	16273	16	66	18	430	41.8	11239
Hispanic	50	48	3	391	41.9	5689	48	49	2	392	40.4	2475	50	46	3	390	43.1	3214
Note: Performance Levels, B=Basic, P=Proficient, A=Advanced																		

Tables 11 reports percentages of grade 8 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 5 as well as by administration mode.

Table 11. Grade 8 Performance Level Percentages and Summary Statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	32	61	7	408	47.9	61729	31	63	6	409	45.6	38698	34	57	8	408	51.5	23031
Gender																		
Male	33	59	8	409	50.1	31419	32	61	7	409	47.8	19521	35	56	9	408	53.6	11898
Female	32	62	6	408	45.5	30291	31	64	5	408	43.2	19177	33	59	7	408	49.1	11114
Ethnicity																		
Native American	27	67	6	410	43.7	198	29	66	6	409	42.5	140	24	71	5	412	47.0	58
Asian	13	69	18	437	46.7	3749	15	71	15	432	45.0	1944	11	67	22	442	47.8	1805
Black	52	47	1	384	41.7	22991	49	49	1	387	40.0	14269	56	43	1	380	43.9	8722
White	16	73	11	428	42.9	28996	16	74	10	426	41.4	19052	16	70	14	430	45.5	9944
Hispanic	47	51	2	391	43.5	5775	48	50	2	390	43.0	3293	46	51	3	392	44.1	2482
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 2010 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 2010 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 2010 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 (10 FT items -> 10 OP items)	1.002247	0.180579	1.046715	0.198166

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 flags by category and not total items flagged.

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

	DIF Classification Level				
Grade	**B**	**BB**	**C**	**CC**	**Total**
5	6	4	1	2	13
8	13	6	2	2	23

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 2010 MSA Science field test data review based on the following criteria:

SR items will be flagged if:

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

BCR items will be flagged if:

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

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

Results of Data Review

A total of 56 items in grade 5 and 66 items in grade 8 were inspected during data review as a result of the item not meeting the statistical flagging criteria. Ten of the 56 total flagged item were rejected from the grade 5 pool and seven of the 66 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 2010 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 Voluntary State Curriculum (VSC). 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 VSC. The 2010 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 VSC 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 VSC 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 0.54 to 0.66 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
	411.39	59.46	**Str1**	1.00000						
	407.97	59.27	**Str2**	0.63403	1.00000					
	410.15	59.82	**Str3**	0.61645	0.65124	1.00000				
	411.82	75.81	**Str4**	0.59414	0.61546	0.59191	1.00000			
	426.26	89.62	**Str5**	0.55969	0.57483	0.53721	0.53496	1.00000		
	418.13	82.95	**Str6**	0.56862	0.61260	0.58315	0.56371	0.53880	1.00000	
	407.98	44.82	**Total**	0.81091	0.85119	0.82228	0.78068	0.73109	0.77460	1.00000
Grade 5 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	409.59	60.84	**Str1**	1.00000						
	413.62	82.20	**Str2**	0.61137	1.00000					
	415.42	65.96	**Str3**	0.65664	0.60883	1.00000				
	414.90	80.51	**Str4**	0.57807	0.55241	0.57513	1.00000			
	412.19	68.97	**Str5**	0.60818	0.57785	0.60127	0.56063	1.00000		
	418.49	76.89	**Str6**	0.60311	0.58634	0.60311	0.53951	0.56419	1.00000	
	408.62	45.83	**Total**	0.83200	0.79273	0.82821	0.74430	0.79414	0.78103	1.00000
Grade 8 Form A				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	411.53	67.84	**Str1**	1.00000						
	410.51	75.53	**Str2**	0.59630	1.00000					
	406.88	63.25	**Str3**	0.66222	0.62134	1.00000				
	411.16	70.22	**Str4**	0.65005	0.61328	0.63911	1.00000			
	411.91	76.17	**Str5**	0.63411	0.61934	0.63979	0.63285	1.00000		
	423.81	92.14	**Str6**	0.60218	0.59311	0.60894	0.58570	0.59904	1.00000	
	407.22	48.07	**Total**	0.83492	0.79636	0.83386	0.81786	0.81021	0.77729	1.00000
Grade 8 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	416.50	72.93	**Str1**	1.00000						
	411.35	70.49	**Str2**	0.60173	1.00000					
	412.73	64.26	**Str3**	0.64646	0.63159	1.00000				
	416.55	76.93	**Str4**	0.60499	0.59358	0.61751	1.00000			
	407.07	69.23	**Str5**	0.62031	0.60519	0.63287	0.60033	1.00000		
	432.12	99.50	**Str6**	0.60445	0.59890	0.63064	0.58765	0.59230	1.00000	
	409.71	47.67	**Total**	0.81321	0.81415	0.83497	0.78985	0.80803	0.78724	1.00000

*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 2010 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 2009 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 0.989, and the highest RMSEA was 0.039. 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	0.994	0.028
Grade 5 Form B	0.994	0.028
Grade 8 Form A	0.989	0.039
Grade 8 Form B	0.995	0.026

*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	.993	.026	.994	.028
Grade 5 Form B	.992	.030	.994	.028
Grade 8 Form A	.994	.023	.989	.039
Grade 8 Form B	.993	.024	.995	.026

*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*
50003	OP	0.78	0.34	0.63322	-0.87064	0.22360			
50016	OP	0.74	0.50	1.25138	-0.35615	0.22067			
50028	OP	0.65	0.37	0.71253	0.03443	0.23535			
50050	OP	0.69	0.27	0.62909	0.31997	0.41046			
50052	OP	0.79	0.31	0.57052	-1.07828	0.19790			
50054	OP	0.50	0.34	0.67062	0.70301	0.17889			
50058	OP	0.76	0.40	0.72639	-0.95211	0.05997			
50066	OP	0.72	0.50	1.03583	-0.49535	0.11413			
50083	OP	0.64	0.49	1.10725	-0.01361	0.18281			
50090	OP	0.60	0.32	0.64650	0.35831	0.26319			
50092	OP	0.41	0.22	0.60040	1.62770	0.24143			
50100	OP	0.25	0.50	0.56026	1.59557	0.00000	1.32252	-0.17118	-1.15133
50107	OP	0.60	0.36	0.60979	0.04676	0.14206			
50108	OP	0.81	0.24	0.37930	-2.12871	0.04107			
50109	OP	0.77	0.35	0.58420	-1.21761	0.03304			
50121	OP	0.60	0.36	0.66546	0.22318	0.20525			
50127	OP	0.47	0.31	0.80021	1.06100	0.24321			
50172	OP	0.51	0.41	0.78501	0.50026	0.14912			
50183	OP	0.90	0.33	0.77354	-1.87399	0.03429			
50216	OP	0.71	0.43	0.79517	-0.51061	0.13818			
50219	OP	0.86	0.39	0.84907	-1.51415	0.02257			
50227	OP	0.72	0.42	0.72056	-0.73075	0.03428			
50228	OP	0.78	0.34	0.76924	-0.43310	0.37934			
50232	OP	0.30	0.26	0.78762	1.73979	0.15447			
50238	OP	0.60	0.24	1.05000	1.10622	0.46417			
50276	OP	0.48	0.37	0.76389	0.74806	0.17243			
50288	OP	0.89	0.36	0.89563	-1.51120	0.15763			
50290	OP	0.63	0.45	0.90841	0.02558	0.18312			
50319	OP	0.63	0.38	0.97000	0.36798	0.31786			
50320	OP	0.44	0.30	0.46388	0.86888	0.09411			
50335	OP	0.73	0.51	1.49947	-0.19535	0.27330			
50348	OP	0.51	0.35	0.74064	0.70764	0.21268			
50350	OP	0.50	0.35	0.65403	0.64759	0.16156			
50352	OP	0.86	0.39	0.94120	-1.17562	0.20421			
50364	OP	0.89	0.33	1.16835	-0.73457	0.56895			
50365	OP	0.57	0.33	0.58188	0.30351	0.18168			
50416	OP	0.52	0.30	0.61386	0.80680	0.23157			
50433	OP	0.47	0.20	0.63137	1.60557	0.32007			
50439	OP	0.63	0.39	0.72933	-0.01478	0.18081			
50447	OP	0.37	0.24	1.21308	1.56031	0.25920			
50468	OP	0.43	0.37	1.19511	1.01689	0.21714			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
50477	OP	0.81	0.40	0.76247	-1.25441	0.02810			
50554	OP	0.59	0.51	1.14008	0.14276	0.15770			
50562	OP	0.50	0.37	1.36213	0.87998	0.28753			
50566	OP	0.68	0.40	0.92210	0.03728	0.30257			
50574	OP	0.52	0.33	0.48843	0.21420	0.04425			
50575	OP	0.88	0.33	0.74233	-1.76751	0.03584			
50583	OP	0.43	0.38	0.84946	0.88584	0.15749			
50600	OP	0.80	0.39	0.81621	-0.85325	0.23108			
50693	OP	0.74	0.46	0.96763	-0.47586	0.18731			
55142	OP	0.75	0.46	0.87293	-0.72839	0.06901			
55148	OP	0.55	0.43	0.79037	0.25195	0.12477			
55167	OP	0.67	0.38	0.61057	-0.55356	0.04846			
55205	OP	0.73	0.31	0.55044	-0.59176	0.23752			
55210	OP	0.43	0.57	0.60702	0.74535	0.00000	1.55843	0.48367	-2.04210
55230	OP	0.60	0.47	1.19214	0.28668	0.23892			
55234	OP	0.85	0.40	0.93088	-1.17722	0.16091			
50001_01	OP	0.56	0.47	0.98751	0.25017	0.15473			
50001_02	OP	0.59	0.51	1.29537	0.24815	0.19942			
50001_06	OP	0.58	0.41	0.85393	0.35692	0.21901			
50049_02	OP	0.56	0.35	0.52564	0.05916	0.06598			
50049_03	OP	0.63	0.46	0.74484	-0.28830	0.03527			
50049_05	OP	0.60	0.39	0.77432	0.24150	0.21294			
50056_01	OP	0.64	0.38	0.59705	-0.37243	0.07185			
50056_02	OP	0.71	0.45	0.94335	-0.33680	0.18574			
50056_03	OP	0.72	0.46	0.97166	-0.34848	0.19896			
50130_04	OP	0.37	0.30	0.79866	1.32730	0.17235			
50130_05	OP	0.59	0.40	0.88270	0.34376	0.23739			
50130_08	OP	0.49	0.56	0.71169	0.36552	0.00000	2.26591	0.19991	-2.46582
50240_02	OP	0.69	0.41	1.04889	0.13356	0.34774			
50240_04	OP	0.62	0.38	0.81574	0.29463	0.27542			
50240_08	OP	0.39	0.53	0.53025	1.07043	0.00000	1.77179	0.56652	-2.33831
50302_01	OP	0.44	0.21	0.56093	1.59180	0.26167			
50302_04	OP	0.70	0.43	0.75502	-0.54528	0.08915			
50502_01	OP	0.67	0.40	0.82949	-0.09532	0.23356			
50502_02	OP	0.75	0.45	0.83214	-0.83738	0.02367			
50502_04	OP	0.94	0.34	1.16140	-1.85964	0.04086			
50553_01	OP	0.81	0.41	0.97337	-0.73866	0.28030			
50553_05	OP	0.50	0.35	1.08775	0.87682	0.28226			
50590_01	OP	0.71	0.35	0.63565	-0.42159	0.21864			
50590_02	OP	0.48	0.44	1.17305	0.65215	0.18200			
50590_04	OP	0.32	0.34	1.05814	1.35044	0.14740			
50591_01	OP	0.56	0.28	0.45365	0.26645	0.14996			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
50591_03	OP	0.67	0.49	1.23030	-0.05196	0.22811			
50616_02	OP	0.71	0.35	0.61125	-0.54549	0.16525			
50616_03	OP	0.94	0.30	0.91220	-2.10246	0.04336			
50616_05	OP	0.48	0.47	1.32413	0.63574	0.18378			
50618_03	OP	0.58	0.29	0.47007	0.25686	0.17683			
50618_04	OP	0.69	0.39	0.63313	-0.64530	0.05849			
50618_05	OP	0.71	0.51	1.00481	-0.51825	0.06658			
50629_01	OP	0.68	0.46	0.93251	-0.22597	0.18431			
50629_03	OP	0.78	0.36	0.65288	-1.03617	0.13827			
50629_05	OP	0.45	0.57	0.57364	0.65611	0.00000	1.18303	0.85629	-2.03933
50632_01	OP	0.58	0.45	0.92952	0.23215	0.17097			
50632_03	OP	0.65	0.52	1.15260	-0.09082	0.15168			
50632_04	OP	0.74	0.51	1.26582	-0.41189	0.19078			
55004_02	OP	0.67	0.40	0.80048	-0.10656	0.22890			
55004_04	OP	0.67	0.39	0.89279	0.10723	0.30065			
55006_01	OP	0.56	0.39	0.70988	0.22916	0.14589			
55006_02	OP	0.72	0.44	0.83475	-0.52857	0.13429			
55006_03	OP	0.75	0.41	0.90544	-0.41712	0.27968			
55007_01	OP	0.95	0.27	0.90176	-2.32482	0.05046			
55007_04	OP	0.29	0.15	0.57796	2.55682	0.19317			
55013_01	OP	0.88	0.28	0.58576	-2.09624	0.02721			
55013_03	OP	0.80	0.43	0.85083	-1.04174	0.07191			
55013_05	OP	0.89	0.25	0.55630	-2.36693	0.03698			
55060_02	OP	0.43	0.31	1.04698	1.23070	0.25882			
55060_06	OP	0.44	0.26	0.95567	1.38974	0.29351			
55060_08	OP	0.42	0.57	0.68838	0.67529	0.00000	2.06308	-0.10791	-1.95517
55080_01	OP	0.52	0.38	0.65012	0.35743	0.11107			
55080_02	OP	0.83	0.43	0.94136	-1.14923	0.07698			
55080_03	OP	0.54	0.44	1.24029	0.56829	0.24647			
50900	FT	0.52	0.37	0.75612	0.53094	0.20052			
50901	FT	0.44	0.43	0.78571	0.59706	0.09326			
50902	FT	0.69	0.39	0.66018	-0.52061	0.10317			
50903	FT	0.25	0.19	1.52153	1.79271	0.17871			
50904	FT	0.58	0.35	0.52845	0.04281	0.09233			
50905	FT	0.86	0.21	0.41061	-2.37724	0.15517			
50906	FT	0.66	0.53	1.28691	-0.05701	0.16668			
50908	FT	0.45	0.28	0.73647	1.14669	0.24049			
50909	FT	0.76	0.43	0.77933	-0.80004	0.05986			
50911	FT	0.36	0.26	1.01049	1.50038	0.23933			
50913	FT	0.61	0.21	0.32492	-0.03490	0.16140			
50915	FT	0.61	0.28	0.40824	-0.24312	0.08634			
50916	FT	0.42	0.42	1.18666	0.88156	0.16787			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
50919	FT	0.86	0.38	0.92661	-1.01454	0.29351			
50920	FT	0.77	0.41	0.83370	-0.70897	0.23289			
50922	FT	0.39	0.55	0.79183	0.99616	0.00000	2.40739	-0.30098	-2.10640
50925	FT	0.54	0.48	0.94377	0.37403	0.13587			
50926	FT	0.73	0.38	0.68568	-0.53742	0.16700			
50927	FT	0.65	0.37	0.64337	-0.21373	0.14051			
50928	FT	0.25	0.22	0.70847	2.13590	0.12870			
50931	FT	0.29	0.43	0.72004	2.50328	0.00000	3.76190	-0.44960	-3.31229
50932	FT	0.64	0.25	0.34752	-0.59042	0.06517			
50933	FT	0.50	0.33	0.85697	0.91172	0.25382			
50934	FT	0.94	0.28	0.84575	-2.38398	0.07284			
50936	FT	0.36	0.23	0.41101	1.62338	0.09847			
50937	FT	0.76	0.40	0.85154	-0.42498	0.26006			
50941	FT	0.37	0.58	0.73440	1.07074	0.00000	1.97807	-0.15398	-1.82409
50942	FT	0.50	0.55	0.74422	0.18803	0.00000	2.39178	-0.05197	-2.33981
50943	FT	0.76	0.16	0.25600	-2.09949	0.12951			
50129_01	FT	0.63	0.26	0.41848	-0.21258	0.14314			
50129_03	FT	0.89	0.34	0.88536	-1.25139	0.31618			
50129_04	FT	0.77	0.32	0.52114	-1.32016	0.05428			
50129_05	FT	0.73	0.36	0.60170	-0.89433	0.11893			
50129_08	FT	0.31	0.60	0.74453	1.33063	0.00000	1.59781	0.17330	-1.77111
50669_02	FT	0.65	0.30	0.43738	-0.44914	0.07977			
50669_03	FT	0.34	0.31	1.07322	1.35295	0.19077			
50669_04	FT	0.48	0.38	0.80759	0.73957	0.18465			
50669_05	FT	0.40	0.20	0.41983	1.81551	0.17571			
50669_06	FT	0.30	0.47	0.60827	1.68040	0.00000	2.77483	-0.65441	-2.12041
50670_01	FT	0.41	0.31	0.70388	1.23844	0.17584			
50670_02	FT	0.81	0.37	0.98670	-0.44679	0.40076			
50670_03	FT	0.80	0.36	0.87200	-0.51365	0.36630			
50670_04	FT	0.57	0.32	0.46466	-0.08038	0.04834			
50670_05	FT	0.71	0.39	0.68299	-0.52699	0.11565			
50670_07	FT	0.39	0.51	0.63661	0.89792	0.00000	2.27511	-0.33196	-1.94315
50675_01	FT	0.74	0.43	0.91883	-0.46131	0.22113			
50675_02	FT	0.86	0.36	0.97448	-0.78777	0.42152			
50675_03	FT	0.50	0.15	0.19966	0.79668	0.08673			
50675_04	FT	0.89	0.22	0.47567	-2.56709	0.10500			
50675_05	FT	0.43	0.22	0.37732	1.25293	0.12308			
50675_06	FT	0.33	0.56	0.66892	1.46387	0.00000	1.75754	0.26144	-2.01898
50920_02	FT	0.53	0.38	0.84373	0.56432	0.20182			
50920_03	FT	0.84	0.45	1.25983	-0.78539	0.27473			
50920_04	FT	0.41	0.26	0.46084	1.29491	0.12030			
50920_05	FT	0.51	0.41	0.86851	0.60057	0.17294			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
50921_01	FT	0.27	0.11	0.70884	2.80896	0.20875			
50921_02	FT	0.33	0.38	0.82621	1.18463	0.08737			
50921_03	FT	0.38	0.18	1.02092	1.80785	0.28792			
50921_04	FT	0.72	0.33	0.54935	-0.70934	0.12150			
50923_01	FT	0.87	0.41	1.05134	-1.37601	0.13603			
50923_02	FT	0.51	0.18	0.44745	1.49385	0.29845			
50923_04	FT	0.39	0.28	0.97452	1.32839	0.23519			
50923_05	FT	0.51	0.21	0.60916	1.36336	0.32604			
50924_01	FT	0.77	0.38	0.70948	-0.92761	0.09183			
50924_02	FT	0.90	0.30	0.76681	-1.67578	0.20412			
50924_03	FT	0.65	0.40	0.78082	-0.02244	0.20965			
50924_04	FT	0.76	0.32	0.55451	-0.99641	0.14140			
50925_01	FT	0.61	0.46	1.21992	0.21798	0.26235			
50925_02	FT	0.58	0.38	0.71630	0.17320	0.18075			
50925_03	FT	0.62	0.30	1.04666	0.75812	0.41694			
50925_04	FT	0.55	0.30	0.44531	0.16606	0.07842			
50927_01	FT	0.33	0.20	0.36633	2.04369	0.11461			
50927_03	FT	0.32	0.25	0.65195	1.80770	0.14806			
50927_04	FT	0.56	0.32	0.46153	-0.05441	0.08265			
50927_05	FT	0.53	0.34	0.52303	0.28854	0.06950			
50928_01	FT	0.81	0.33	0.59088	-1.42246	0.10346			
50928_03	FT	0.45	0.33	0.58932	0.84705	0.12700			
50928_04	FT	0.56	0.16	0.23363	0.20097	0.12397			
50928_05	FT	0.50	0.17	0.57540	1.70608	0.36007			
50929_01	FT	0.49	0.41	1.28146	0.73879	0.24670			
50929_03	FT	0.54	0.25	0.42912	0.57318	0.17957			
50929_04	FT	0.81	0.44	1.16043	-0.50068	0.33614			
50929_05	FT	0.65	0.37	0.60435	-0.33310	0.10774			
50930_01	FT	0.70	0.36	0.57976	-0.68388	0.07446			
50930_02	FT	0.42	0.27	0.55543	1.22706	0.16696			
50930_04	FT	0.25	0.19	0.66210	2.31592	0.14134			
50930_05	FT	0.57	0.29	0.41283	-0.01056	0.07417			
50932_01	FT	0.40	0.09	0.46112	3.12285	0.31967			
50932_02	FT	0.74	0.40	0.76992	-0.52453	0.17436			
50932_04	FT	0.58	0.29	0.39407	-0.20487	0.08624			
50932_05	FT	0.63	0.33	0.46788	-0.56962	0.04964			
50933_01	FT	0.38	0.12	0.18782	2.71622	0.09844			
50933_02	FT	0.53	0.44	0.82402	0.36388	0.11734			
50933_03	FT	0.36	0.23	0.57365	1.64162	0.17548			
50933_04	FT	0.70	0.37	0.72484	-0.25179	0.24532			
50934_02	FT	0.32	0.31	0.87127	1.41523	0.14248			
50934_03	FT	0.72	0.52	1.15110	-0.35792	0.12672			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
50934_04	FT	0.61	0.35	0.88773	0.47505	0.31175			
50934_05	FT	0.61	0.37	0.67403	0.13776	0.18570			
50937_01	FT	0.47	0.39	1.06866	0.89390	0.21721			
50937_02	FT	0.50	0.33	0.84611	0.93386	0.25285			
50937_03	FT	0.71	0.54	1.49205	-0.32692	0.23467			
50937_05	FT	0.50	0.49	0.94242	0.27793	0.10939			
50938_01	FT	0.56	0.33	0.55524	0.33915	0.14241			
50938_02	FT	0.39	0.11	0.18002	2.76893	0.10808			
50938_03	FT	0.61	0.40	0.70331	0.02142	0.13382			
50938_04	FT	0.64	0.21	0.28896	-0.77163	0.07370			
50940_01	FT	0.47	0.42	1.37243	0.75702	0.23325			
50940_02	FT	0.65	0.32	0.47414	-0.66015	0.05319			
50940_03	FT	0.56	0.37	0.78208	0.46876	0.21404			
50940_04	FT	0.79	0.39	0.77764	-0.81319	0.17080			
50940_05	FT	0.28	0.52	0.64527	1.60513	0.00000	1.70950	-0.08993	-1.61957
55073_01	FT	0.79	0.50	1.32120	-0.60304	0.21648			
55073_02	FT	0.35	0.18	0.87794	1.90166	0.25504			
55073_05	FT	0.45	0.42	1.31794	0.81883	0.20598			
55073_06	FT	0.64	0.39	0.65598	-0.21734	0.10924			
55073_07	FT	0.39	0.57	0.85525	1.04686	0.00000	2.35906	-0.13441	-2.22465

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*
80008	OP	0.34	0.28	1.53814	1.52085	0.22509			
80013	OP	0.52	0.27	0.37171	0.33822	0.07595			
80023	OP	0.57	0.42	0.85105	0.39818	0.21156			
80032	OP	0.36	0.28	0.66547	1.59537	0.17545			
80036	OP	0.62	0.44	1.17651	0.40453	0.30226			
80041	OP	0.41	0.63	0.62568	0.56099	0.00000	1.21294	-0.37928	-0.83366
80050	OP	0.86	0.39	0.86127	-1.51869	0.03285			
80061	OP	0.66	0.45	0.77553	-0.19765	0.13878			
80064	OP	0.48	0.67	0.82470	0.38344	0.00000	1.58057	0.14756	-1.72813
80066	OP	0.65	0.56	1.23345	-0.12107	0.13373			
80071	OP	0.53	0.43	0.80860	0.45565	0.17008			
80072	OP	0.50	0.35	0.87520	0.94454	0.26416			
80073	OP	0.56	0.44	0.78620	0.32816	0.14384			
80081	OP	0.71	0.39	0.59977	-0.85179	0.03742			
80083	OP	0.36	0.30	0.61606	1.45876	0.13480			
80102	OP	0.65	0.52	1.14792	0.00261	0.20220			
80108	OP	0.55	0.56	0.74415	-0.06025	0.00000	2.58520	0.10992	-2.69512
80112	OP	0.66	0.42	0.74187	-0.11188	0.19092			
80121	OP	0.78	0.31	0.54054	-0.97622	0.22771			
80131	OP	0.57	0.43	1.19128	0.59009	0.28615			
80178	OP	0.53	0.44	0.92462	0.53722	0.18774			
80196	OP	0.67	0.43	0.91647	0.05983	0.27578			
80201	OP	0.74	0.36	0.55073	-1.07669	0.00930			
80222	OP	0.81	0.39	0.70635	-1.32134	0.03250			
80225	OP	0.72	0.35	0.61268	-0.41828	0.22796			
80227	OP	0.65	0.43	0.74486	-0.15454	0.15139			
80229	OP	0.57	0.44	0.71939	0.15173	0.11641			
80253	OP	0.45	0.42	0.80731	0.78376	0.12630			
80254	OP	0.72	0.34	0.51450	-0.91324	0.06960			
80291	OP	0.60	0.46	1.00068	0.30525	0.21868			
80299	OP	0.57	0.48	0.97874	0.28185	0.18021			
80310	OP	0.48	0.34	0.55454	0.68267	0.12037			
80344	OP	0.78	0.54	1.23465	-0.68134	0.10626			
80347	OP	0.51	0.27	0.54565	1.06827	0.25288			
80414	OP	0.76	0.55	1.33546	-0.53147	0.14016			
80447	OP	0.82	0.46	0.92658	-1.11475	0.02918			
80469	OP	0.31	0.31	0.69750	1.60397	0.11636			
80501	OP	0.68	0.46	0.96654	-0.04414	0.22856			
80503	OP	0.83	0.48	1.21558	-0.93841	0.14534			
80558	OP	0.84	0.44	0.97963	-1.26159	0.02860			
80567	OP	0.65	0.50	1.01309	-0.05162	0.18088			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
80568	OP	0.65	0.50	0.96536	-0.13812	0.13587			
80619	OP	0.84	0.45	0.99782	-1.20072	0.06087			
80620	OP	0.53	0.41	1.07584	0.72389	0.25379			
80625	OP	0.55	0.54	0.95298	0.13605	0.05602			
80632	OP	0.78	0.46	0.89680	-0.88890	0.08133			
80648	OP	0.67	0.49	0.97151	-0.13210	0.16959			
85141	OP	0.60	0.50	0.80474	-0.13150	0.03732			
85232	OP	0.66	0.50	0.99774	-0.14755	0.16912			
80002_01	OP	0.74	0.37	0.72775	-0.34992	0.29307			
80002_04	OP	0.48	0.46	0.92451	0.61168	0.13970			
80002_06	OP	0.46	0.51	1.02233	0.59554	0.10557			
80056_01	OP	0.63	0.48	0.83117	-0.16752	0.10280			
80056_05	OP	0.62	0.34	0.47303	-0.37486	0.05172			
80098_01	OP	0.69	0.41	0.82109	-0.05273	0.26376			
80098_03	OP	0.70	0.47	1.06272	-0.10161	0.25198			
80098_05	OP	0.45	0.31	0.76022	1.20228	0.23251			
80236_01	OP	0.59	0.44	0.69125	-0.07879	0.07008			
80236_04	OP	0.45	0.48	1.17466	0.74321	0.15900			
80236_08	OP	0.46	0.62	0.77473	0.39428	0.00000	2.00193	-0.16494	-1.83699
80248_03	OP	0.66	0.48	0.84220	-0.29739	0.10780			
80248_05	OP	0.45	0.31	0.69201	1.13453	0.21390			
80248_06	OP	0.68	0.46	0.80627	-0.33868	0.13487			
80268_02	OP	0.57	0.30	0.38628	-0.21744	0.01855			
80268_03	OP	0.58	0.42	1.22062	0.59288	0.31152			
80268_04	OP	0.66	0.43	0.72363	-0.30045	0.12452			
80408_01	OP	0.86	0.37	0.78670	-1.61521	0.02036			
80408_03	OP	0.69	0.52	0.99678	-0.38688	0.10208			
80408_04	OP	0.70	0.57	1.25108	-0.37386	0.10983			
80467_01	OP	0.89	0.39	0.93030	-1.58045	0.02571			
80467_02	OP	0.59	0.48	0.80505	0.07358	0.09055			
80467_03	OP	0.59	0.42	0.74534	0.20571	0.16708			
80475_02	OP	0.64	0.39	0.56730	-0.47698	0.03254			
80475_04	OP	0.73	0.52	1.03907	-0.55538	0.09744			
80475_06	OP	0.26	0.58	0.58511	1.26399	0.00000	0.36614	-0.00739	-0.35875
80484_01	OP	0.82	0.40	0.74743	-1.33083	0.02007			
80484_02	OP	0.65	0.38	0.54820	-0.46623	0.02982			
80484_03	OP	0.71	0.53	1.03053	-0.43228	0.09649			
80507_02	OP	0.56	0.40	0.90818	0.61804	0.25776			
80507_04	OP	0.69	0.43	0.73870	-0.35261	0.15302			
80507_05	OP	0.74	0.34	0.50136	-1.18271	0.01709			
80529_03	OP	0.70	0.41	0.76538	-0.22181	0.22066			
80529_04	OP	0.70	0.37	0.70517	-0.08402	0.27988			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
80529_05	OP	0.84	0.46	1.16799	-0.95683	0.18782			
80534_02	OP	0.68	0.51	1.01218	-0.23975	0.13773			
80534_03	OP	0.64	0.50	1.01930	0.00675	0.16930			
80534_08	OP	0.22	0.51	0.52959	1.76051	0.00000	0.98288	-0.19407	-0.78881
80535_02	OP	0.67	0.36	0.72531	0.16218	0.30574			
80535_06	OP	0.44	0.31	0.77721	1.25311	0.22890			
80535_08	OP	0.36	0.63	0.70379	0.84195	0.00000	1.28818	-0.49410	-0.79409
80538_01	OP	0.40	0.28	0.70950	1.50714	0.21020			
80538_03	OP	0.86	0.40	0.89129	-1.45272	0.02555			
80538_04	OP	0.61	0.43	0.75958	0.08618	0.14991			
80595_03	OP	0.64	0.43	0.74858	-0.15511	0.14640			
80595_05	OP	0.46	0.33	0.58118	0.88650	0.14106			
80595_06	OP	0.75	0.47	0.87080	-0.75230	0.07638			
80639_01	OP	0.48	0.39	0.74534	0.76800	0.16133			
80639_02	OP	0.67	0.39	0.56662	-0.64174	0.01088			
80639_05	OP	0.60	0.39	0.78433	0.36993	0.24298			
80697_01	OP	0.63	0.41	0.73309	0.05365	0.18474			
80697_02	OP	0.57	0.46	1.12930	0.45755	0.23363			
80697_04	OP	0.77	0.41	0.70023	-1.04158	0.03276			
80698_02	OP	0.74	0.38	0.66740	-0.68416	0.18394			
80698_03	OP	0.45	0.36	0.82892	0.99923	0.20472			
80698_04	OP	0.56	0.34	0.54607	0.34453	0.16761			
85060_01	OP	0.71	0.32	0.58689	-0.25698	0.28848			
85060_02	OP	0.82	0.30	0.50045	-1.78538	0.02353			
85060_07	OP	0.47	0.49	0.63136	0.41575	0.00000	3.33630	-0.16965	-3.16665
80900	FT	0.57	0.32	0.43685	-0.05539	0.05083			
80901	FT	0.58	0.38	0.90354	0.55290	0.28740			
80904	FT	0.64	0.39	0.81382	0.24492	0.26778			
80907	FT	0.48	0.33	1.12410	1.05041	0.31310			
80908	FT	0.56	0.29	0.38252	0.04803	0.05189			
80911	FT	0.40	0.20	1.05756	1.70405	0.31382			
80912	FT	0.69	0.37	0.54063	-0.76439	0.04219			
80913	FT	0.76	0.49	1.12267	-0.59408	0.22586			
80916	FT	0.16	0.14	1.26315	2.40443	0.11089			
80917	FT	0.28	0.25	0.94037	1.86843	0.15694			
80918	FT	0.62	0.44	0.80748	0.08506	0.16238			
80919	FT	0.50	0.64	0.92811	0.46545	0.00000	1.84585	0.34721	-2.19306
80923	FT	0.52	0.17	0.23521	0.63645	0.09823			
80924	FT	0.48	0.37	1.15325	0.92242	0.27918			
80926	FT	0.43	0.32	0.76051	1.19019	0.19557			
80928	FT	0.63	0.36	0.51835	-0.27596	0.07660			
80929	FT	0.44	0.37	0.59246	0.77283	0.10833			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
80930	FT	0.60	0.25	0.32990	-0.32506	0.06445			
80933	FT	0.21	0.47	0.53731	2.02057	0.00000	1.54294	-0.54203	-1.00091
80935	FT	0.21	0.13	1.09604	2.37323	0.15658			
80938	FT	0.69	0.49	0.91278	-0.35859	0.11760			
80939	FT	0.68	0.50	1.02140	-0.20546	0.14693			
80942	FT	0.33	0.53	0.64786	1.49542	0.00000	2.16807	-0.30657	-1.86150
80943	FT	0.20	0.17	0.44944	3.13122	0.08554			
80944	FT	0.38	0.65	0.77306	1.10394	0.00000	0.80908	0.75033	-1.55942
80955	FT	0.88	0.23	0.43530	-2.48893	0.10027			
80956	FT	0.52	0.33	0.87033	0.95187	0.29342			
80960	FT	0.87	0.41	0.94797	-1.47521	0.06537			
80963	FT	0.63	0.36	0.57737	-0.08173	0.13929			
80138_01	FT	0.44	0.34	0.80288	1.09964	0.19123			
80138_02	FT	0.68	0.36	0.93372	0.40249	0.38289			
80138_03	FT	0.59	0.43	0.62119	-0.05432	0.03677			
80138_04	FT	0.57	0.31	0.88112	0.93398	0.35545			
80138_05	FT	0.79	0.46	0.98396	-0.69833	0.16236			
80138_07	FT	0.51	0.60	0.68615	0.29193	0.00000	1.70974	-0.40040	-1.30934
80746_01	FT	0.51	0.21	0.40573	1.19327	0.24621			
80746_02	FT	0.63	0.45	0.73623	-0.21917	0.09571			
80746_03	FT	0.46	0.38	0.75142	0.79563	0.15729			
80746_04	FT	0.30	0.08	1.26701	2.38964	0.26143			
80746_05	FT	0.53	0.23	0.30663	0.19358	0.06321			
80746_06	FT	0.40	0.48	0.68938	1.05729	0.00000	3.09667	-0.36178	-2.73488
80750_01	FT	0.63	0.17	0.22832	-0.46174	0.13029			
80750_02	FT	0.44	0.37	0.81780	0.99235	0.17102			
80750_03	FT	0.34	0.15	1.28497	2.04193	0.27506			
80750_04	FT	0.68	0.34	0.60239	-0.21730	0.21345			
80750_05	FT	0.48	0.28	0.65196	1.28648	0.26053			
80750_07	FT	0.46	0.53	0.56792	0.70820	0.00000	1.85365	0.20269	-2.05633
80751_01	FT	0.53	0.29	0.68625	0.95722	0.32330			
80751_02	FT	0.36	0.22	0.46500	1.94203	0.15861			
80751_03	FT	0.20	0.10	1.72588	2.22924	0.15789			
80751_04	FT	0.11	-.03	-0.56360	-4.87338	0.10125			
80751_05	FT	0.60	0.29	0.41202	-0.11558	0.08918			
80751_07	FT	0.32	0.61	0.75760	1.31933	0.00000	1.70785	0.01705	-1.72490
80921_01	FT	0.83	0.27	0.48810	-1.55993	0.20269			
80921_02	FT	0.63	0.47	0.84411	-0.17537	0.15916			
80921_04	FT	0.65	0.37	0.88982	0.27631	0.36407			
80921_05	FT	0.29	0.15	0.82475	2.41107	0.21980			
80922_01	FT	0.51	0.32	0.53014	0.66808	0.14213			
80922_02	FT	0.23	0.25	1.00903	2.05292	0.12692			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
80922_04	FT	0.57	0.25	0.34807	0.03035	0.07996			
80922_05	FT	0.52	0.23	0.83717	1.49460	0.38319			
80923_01	FT	0.77	0.30	0.45725	-1.32910	0.05871			
80923_02	FT	0.52	0.50	0.94926	0.39474	0.08980			
80923_03	FT	0.08	-.01	1.59237	3.30236	0.07852			
80923_05	FT	0.57	0.52	1.33657	0.39491	0.18966			
80926_01	FT	0.57	0.37	1.13196	0.77640	0.33923			
80926_03	FT	0.38	0.14	1.07804	2.01125	0.31909			
80926_04	FT	0.28	0.30	0.75066	1.64708	0.12277			
80926_05	FT	0.50	0.23	0.58763	1.42372	0.30805			
80927_01	FT	0.75	0.36	0.81020	-0.04665	0.38855			
80927_02	FT	0.26	0.11	0.57003	3.28685	0.19841			
80927_03	FT	0.64	0.05	0.07470	-1.52043	0.19301			
80927_05	FT	0.18	0.07	1.14092	2.78063	0.15590			
80928_02	FT	0.25	0.10	0.16397	5.38317	0.08030			
80928_03	FT	0.58	0.44	0.78756	0.19649	0.12884			
80928_04	FT	0.75	0.40	0.64379	-1.12909	0.06261			
80928_05	FT	0.50	0.26	0.37869	0.59005	0.09246			
80929_01	FT	0.32	0.20	0.51782	2.26667	0.16780			
80929_02	FT	0.66	0.44	0.72666	-0.30804	0.08294			
80929_03	FT	0.60	0.36	0.57681	0.14038	0.14334			
80929_04	FT	0.64	0.23	0.32111	-0.58880	0.09294			
80930_02	FT	0.75	0.45	0.78859	-0.83694	0.04946			
80930_03	FT	0.70	0.47	1.02954	-0.16824	0.23760			
80930_04	FT	0.70	0.24	0.32558	-1.22928	0.06029			
80930_05	FT	0.37	0.15	0.97412	2.09771	0.30467			
80931_01	FT	0.62	0.48	1.06928	0.09256	0.23402			
80931_02	FT	0.83	0.39	1.15355	-0.41656	0.45496			
80931_03	FT	0.63	0.45	0.85620	-0.08064	0.19820			
80931_04	FT	0.60	0.40	0.58343	-0.16375	0.04933			
80932_01	FT	0.63	0.28	0.39359	-0.44753	0.08729			
80932_02	FT	0.34	0.32	0.66322	1.42916	0.12054			
80932_04	FT	0.68	0.29	0.59173	0.20221	0.35659			
80932_05	FT	0.65	0.18	0.24258	-1.09778	0.07317			
80935_01	FT	0.43	0.41	1.10203	1.02610	0.18445			
80935_02	FT	0.69	0.48	1.04047	-0.12514	0.21073			
80935_03	FT	0.57	0.54	1.10691	0.24492	0.11585			
80935_04	FT	0.69	0.51	1.17299	-0.09136	0.20227			
80936_01	FT	0.82	0.36	0.85695	-0.44157	0.42094			
80936_02	FT	0.24	0.21	1.12249	2.05090	0.15786			
80936_03	FT	0.60	0.19	0.27031	0.01021	0.15243			
80936_04	FT	0.63	0.46	1.05413	0.28073	0.24234			

UIN	Status	Pvalue	Ptbis	*a*	*b*	*c*	*d1*	*d2*	*d3*
80939_01	FT	0.31	0.32	0.53756	1.53756	0.05032			
80939_02	FT	0.62	0.39	0.61796	-0.22786	0.15014			
80939_04	FT	0.80	0.41	0.82189	-0.92019	0.16575			
80939_05	FT	0.82	0.50	1.31512	-1.10622	0.12690			
80940_01	FT	0.67	0.37	0.54099	-0.65232	0.07575			
80940_02	FT	0.69	0.46	0.76752	-0.64471	0.07699			
80940_03	FT	0.54	0.28	0.45553	0.61753	0.18586			
80940_04	FT	0.32	0.04	0.82352	3.54604	0.30653			
80940_05	FT	0.21	0.51	0.57041	1.79263	0.00000	0.82422	0.11142	-0.93564
80941_01	FT	0.48	0.51	1.36263	0.63399	0.17108			
80941_02	FT	0.45	0.25	0.92525	1.44509	0.31252			
80941_03	FT	0.63	0.43	0.77247	-0.00345	0.16071			
80941_04	FT	0.38	0.33	0.90732	1.30950	0.18077			
85056_01	FT	0.33	0.27	1.11324	1.59578	0.21693			
85056_02	FT	0.68	0.47	0.92058	-0.28801	0.21137			
85056_04	FT	0.51	0.26	0.35986	0.40951	0.06796			
85056_05	FT	0.75	0.50	0.95171	-0.87668	0.06262			
85056_06	FT	0.86	0.47	1.34201	-1.03184	0.15081			
85056_08	FT	0.47	0.55	0.69454	0.47186	0.00000	2.21250	-0.14767	-2.06483

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**
50900	-0.42	-0.0333			-0.31	-0.0314			0.11	0.0100		
50901	-0.00	-0.0013			-0.63	-0.0529			-0.81	-0.0711		
50902	0.16	0.0093			0.35	0.0271			0.49	0.0386		
50903	-0.03	0.0096			-0.25	-0.0087			-0.61	-0.0460		
50904	-0.17	-0.0276			-0.05	-0.0018			0.58	0.0529		
50905	0.03	-0.0020			0.47	0.0187			-0.50	-0.0240		
50906	-0.02	-0.0039			-0.48	-0.0448			0.25	0.0179		
50908	-0.31	-0.0232			-0.25	-0.0288			-0.45	-0.0425		
50909	-1.10	-0.0747	B	W	-0.63	-0.0441			-1.03	-0.0661	B	M
50911	-0.15	-0.0034			-0.25	-0.0211			-0.13	-0.0115		
50913	0.16	-0.0054			-0.10	-0.0150			0.37	0.0363		
50915	-0.01	-0.0086			0.23	0.0184			-0.50	-0.0476		
50916	-0.76	-0.0564			0.12	0.0079			-0.98	-0.0826		
50919	-0.54	-0.0357			-1.26	-0.0768	B	W	0.99	0.0445		
50920	-0.60	-0.0481			-0.35	-0.0267			0.05	0.0031		
50922	N/A	-0.1166	BB	W	N/A	-0.1021			N/A	0.0897		
50925	-0.04	-0.0029			-0.41	-0.0381			-1.09	-0.0907	B	M
50926	-1.11	-0.0926	B	W	-0.74	-0.0665			-0.24	-0.0180		
50927	-0.24	-0.0269			-0.52	-0.0513			-0.32	-0.0265		
50928	-0.05	0.0034			-0.00	-0.0012			0.25	0.0188		
50931	N/A	-0.0270			N/A	-0.0241			N/A	0.0976	BB	F
50932	-1.00	-0.0834			-0.40	-0.0350			0.11	0.0110		
50933	-0.01	0.0123			0.62	0.0605			-0.52	-0.0501		
50934	-0.01	-0.0033			-0.43	-0.0150			0.89	0.0195		
50936	0.20	0.0175			-0.28	-0.0225			-0.06	-0.0059		
50937	-0.47	-0.0340			-0.97	-0.0795			-0.56	-0.0384		
50941	N/A	-0.0815			N/A	-0.1130			N/A	0.0338		
50942	N/A	-0.1573	BB	W	N/A	-0.1354	BB	W	N/A	0.1843	CC	F
50943	-0.11	-0.0094			0.01	0.0024			-0.82	-0.0616		
50129_01	0.38	0.0326			0.31	0.0282			0.60	0.0545		
50129_03	-0.47	-0.0227			0.12	0.0028			0.11	0.0041		
50129_04	-0.44	-0.0380			0.10	0.0086			-0.05	-0.0035		
50129_05	-0.11	-0.0122			-0.30	-0.0249			0.36	0.0276		
50129_08	N/A	-0.1209			N/A	-0.1858	BB	W	N/A	0.0352		
50669_02	-0.62	-0.0589			-0.74	-0.0694			0.34	0.0307		
50669_03	-0.56	-0.0388			-0.26	-0.0201			-0.60	-0.0505		
50669_04	-0.46	-0.0378			-0.34	-0.0308			0.08	0.0076		
50669_05	0.05	-0.0046			-0.54	-0.0513			0.21	0.0211		
50669_06	N/A	-0.0262			N/A	-0.0298			N/A	0.0634		
50670_01	0.56	0.0547			-0.14	-0.0134			0.15	0.0133		
50670_02	-0.02	-0.0049			-0.76	-0.0526			-0.55	-0.0316		
50670_03	0.02	-0.0067			0.46	0.0309			0.05	0.0024		
50670_04	0.08	0.0023			0.06	0.0055			0.12	0.0117		

	Black/White				Hispanic/White				Male/Female			
UIN	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor
50670_05	-0.21	-0.0197			-0.54	-0.0484			-0.34	-0.0268		
50670_07	N/A	-0.1113			N/A	-0.0554			N/A	0.1063		
50675_01	0.11	0.0075			0.21	0.0153			0.12	0.0083		
50675_02	0.15	0.0075			-0.12	-0.0062			-0.22	-0.0103		
50675_03	-0.17	-0.0135			0.26	0.0244			0.49	0.0512		
50675_04	-0.37	-0.0167			-0.35	-0.0157			0.05	0.0014		
50675_05	-0.12	-0.0162			-0.55	-0.0519			-0.60	-0.0598		
50675_06	N/A	-0.0051			N/A	-0.1319			N/A	0.2183	CC	F
50920_02	-0.35	-0.0344			0.24	0.0274			-0.08	-0.0071		
50920_03	0.33	0.0163			0.49	0.0289			1.50	0.0671	C	F
50920_04	0.21	0.0209			-0.04	-0.0023			0.56	0.0538		
50920_05	-0.57	-0.0416			-0.56	-0.0482			0.31	0.0275		
50921_01	-0.23	-0.0220			-0.18	-0.0151			-0.32	-0.0265		
50921_02	-0.59	-0.0398			-0.31	-0.0252			0.12	0.0090		
50921_03	0.00	0.0079			0.08	0.0023			-0.43	-0.0413		
50921_04	0.25	0.0215			0.49	0.0433			0.52	0.0413		
50923_01	-0.24	-0.0143			-0.16	-0.0099			-0.61	-0.0251		
50923_02	0.06	0.0166			0.24	0.0293			-0.19	-0.0197		
50923_04	0.08	0.0096			-0.36	-0.0245			-0.24	-0.0224		
50923_05	0.04	0.0094			0.38	0.0321			-0.24	-0.0241		
50924_01	-0.49	-0.0450			-0.39	-0.0277			0.16	0.0103		
50924_02	-0.22	-0.0117			0.54	0.0201			0.26	0.0090		
50924_03	0.16	0.0148			-0.15	-0.0179			0.38	0.0316		
50924_04	0.22	0.0141			-0.01	-0.0052			0.26	0.0182		
50925_01	-0.60	-0.0620			-0.90	-0.0816			-0.02	-0.0004		
50925_02	-0.22	-0.0214			-0.42	-0.0418			0.30	0.0263		
50925_03	-0.65	-0.0518			-0.91	-0.0872			0.09	0.0089		
50925_04	0.40	0.0323			0.58	0.0628			0.01	0.0022		
50927_01	-0.45	-0.0324			-0.13	-0.0121			-0.56	-0.0512		
50927_03	-0.97	-0.0592			-0.32	-0.0236			-0.24	-0.0222		
50927_04	-0.28	-0.0220			-0.37	-0.0352			-0.27	-0.0259		
50927_05	-0.04	-0.0128			-0.22	-0.0222			-0.00	0.0005		
50928_01	-0.20	-0.0226			-0.28	-0.0226			-0.33	-0.0193		
50928_03	-0.83	-0.0631			-0.68	-0.0635			-0.39	-0.0386		
50928_04	-0.66	-0.0633			-0.41	-0.0437			-0.24	-0.0251		
50928_05	-0.75	-0.0665			-0.48	-0.0477			-1.12	-0.1142	B	M
50929_01	-0.84	-0.0735			-0.87	-0.0771			-0.22	-0.0197		
50929_03	-0.17	-0.0194			-0.89	-0.0973			-0.20	-0.0198		
50929_04	-0.29	-0.0232			-0.02	-0.0048			0.17	0.0094		
50929_05	-0.80	-0.0893			-0.84	-0.0846			-0.46	-0.0387		
50930_01	-0.72	-0.0751			-0.81	-0.0770			-0.20	-0.0162		
50930_02	-0.67	-0.0521			-0.41	-0.0346			0.29	0.0284		
50930_04	-0.40	-0.0140			-0.13	-0.0050			-0.17	-0.0127		
50930_05	-0.38	-0.0321			-0.29	-0.0265			0.03	0.0043		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
50932_01	0.15	0.0070			-0.28	-0.0326			-0.18	-0.0172		
50932_02	-0.68	-0.0462			-0.81	-0.0652			0.23	0.0166		
50932_04	0.33	0.0397			0.32	0.0298			-0.19	-0.0177		
50932_05	0.02	0.0049			-0.44	-0.0452			0.10	0.0083		
50933_01	0.14	0.0205			0.78	0.0759			-0.36	-0.0350		
50933_02	0.40	0.0449			0.26	0.0221			0.02	0.0027		
50933_03	-0.09	-0.0029			-0.35	-0.0369			-0.51	-0.0481		
50933_04	0.18	0.0077			-0.63	-0.0607			0.05	0.0044		
50934_02	-0.39	-0.0211			0.18	0.0143			-0.68	-0.0570		
50934_03	-0.29	-0.0248			-0.20	-0.0160			-0.02	-0.0009		
50934_04	-0.77	-0.0649			-0.75	-0.0703			-0.87	-0.0788		
50934_05	-0.27	-0.0235			-0.29	-0.0264			0.31	0.0274		
50937_01	-1.18	-0.1064	B	W	-1.15	-0.1055	B	W	-0.73	-0.0671		
50937_02	-0.01	0.0086			-0.30	-0.0290			-0.24	-0.0217		
50937_03	0.01	0.0024			-0.07	-0.0024			0.43	0.0266		
50937_05	-0.52	-0.0321			-0.16	-0.0127			-0.22	-0.0178		
50938_01	0.15	0.0232			0.18	0.0182			0.56	0.0536		
50938_02	0.16	0.0168			0.15	0.0088			0.32	0.0321		
50938_03	-0.42	-0.0469			-0.66	-0.0655			-0.29	-0.0253		
50938_04	0.01	0.0005			0.11	0.0077			-0.43	-0.0410		
50940_01	-0.18	-0.0117			0.16	0.0160			-0.31	-0.0264		
50940_02	-0.83	-0.0859			-0.67	-0.0676			0.06	0.0066		
50940_03	-0.05	0.0006			0.19	0.0189			0.10	0.0108		
50940_04	-0.12	-0.0135			0.87	0.0635			0.49	0.0311		
50940_05	N/A	-0.0198			N/A	-0.0853			N/A	0.1093		
55073_01	-0.95	-0.0653			-0.81	-0.0573			-0.60	-0.0331		
55073_02	-0.21	-0.0085			0.27	0.0312			-0.11	-0.0102		
55073_05	-0.20	-0.0058			0.17	0.0219			-0.28	-0.0233		
55073_06	-0.44	-0.0472			0.04	0.0008			-0.45	-0.0371		
55073_07	N/A	-0.0717			N/A	-0.0464			N/A	0.1051		

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

Table B.2 Grade 8 DIF results

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
80008	-0.01	-0.0006			0.03	0.0006			-0.20	-0.0162		
80013	-0.33	-0.0344			-0.67	-0.0700			-0.57	-0.0552		
80023	-0.41	-0.0447			-0.83	-0.0788			-0.99	-0.0845		
80032	0.49	0.0353			-0.16	-0.0173			-0.12	-0.0110		
80036	-0.45	-0.0367			0.47	0.0399			-0.71	-0.0564		
80041	N/A	-0.1675			N/A	-0.0441			N/A	0.1259		
80050	-0.56	-0.0301			-1.33	-0.0761	B	W	-0.66	-0.0280		
80061	-0.30	-0.0266			-0.43	-0.0389			1.00	0.0765		
80064	N/A	-0.1884	BB	W	N/A	-0.0711			N/A	-0.0096		
80066	-0.09	-0.0038			-0.35	-0.0292			-0.05	-0.0034		
80071	-0.82	-0.0660			-0.36	-0.0336			0.34	0.0296		
80072	-0.02	0.0073			0.14	0.0164			-0.50	-0.0465		
80073	-0.95	-0.0887			-0.62	-0.0592			-0.71	-0.0605		
80081	-0.28	-0.0268			-0.52	-0.0466			-0.01	-0.0004		
80083	-0.19	-0.0167			0.07	0.0067			-0.85	-0.0757		
80102	-0.39	-0.0295			-0.79	-0.0650			-0.22	-0.0160		
80108	N/A	-0.0811			N/A	-0.0792			N/A	0.1654	BB	F
80112	0.10	0.0131			0.06	0.0034			-0.12	-0.0099		
80121	0.26	0.0191			0.05	0.0000			-0.35	-0.0234		
80131	0.36	0.0292			-0.04	-0.0042			-0.41	-0.0347		
80178	-0.53	-0.0468			-0.06	-0.0082			-1.35	-0.1122	B	M
80196	-0.46	-0.0334			-0.35	-0.0289			-0.22	-0.0170		
80201	0.15	0.0084			-0.12	-0.0127			0.34	0.0249		
80222	-0.83	-0.0587			-0.90	-0.0628			-0.90	-0.0496		
80225	-0.18	-0.0208			-0.51	-0.0466			-0.30	-0.0233		
80227	0.68	0.0579			0.03	0.0032			0.12	0.0099		
80229	-0.07	-0.0087			-0.66	-0.0606			-0.49	-0.0411		
80253	-0.54	-0.0399			-0.50	-0.0449			-0.89	-0.0761		
80254	-0.01	0.0026			-0.03	-0.0049			0.08	0.0062		
80291	-0.42	-0.0415			0.46	0.0356			-0.70	-0.0567		
80299	-0.27	-0.0221			-0.66	-0.0572			-0.43	-0.0344		
80310	0.01	0.0032			0.08	0.0070			-0.02	-0.0018		
80344	-1.30	-0.0922	B	W	-0.97	-0.0644			-0.84	-0.0421		
80347	-0.29	-0.0283			-0.15	-0.0152			0.26	0.0257		
80414	-0.92	-0.0654			-0.43	-0.0292			0.08	0.0043		
80447	-0.19	-0.0066			-0.64	-0.0424			0.86	0.0420		
80469	-0.01	0.0014			0.02	0.0013			0.14	0.0120		
80501	-0.37	-0.0281			-0.31	-0.0269			-0.38	-0.0281		
80503	-0.79	-0.0477			-0.94	-0.0580			0.04	0.0014		
80558	-0.77	-0.0528			-0.85	-0.0507			-0.05	-0.0022		
80567	-0.43	-0.0366			-0.57	-0.0495			-0.23	-0.0163		
80568	0.32	0.0185			0.18	0.0114			0.56	0.0405		
80619	0.43	0.0226			0.13	0.0066			0.46	0.0203		

	Black/White				Hispanic/White				Male/Female			
UIN	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor
80900	-0.17	-0.0228			-0.35	-0.0377			0.25	0.0236		
80901	0.03	0.0088			0.35	0.0309			0.01	0.0009		
80904	-0.31	-0.0297			-0.54	-0.0470			-0.06	-0.0052		
80907	-0.19	-0.0039			-0.21	-0.0170			-0.45	-0.0431		
80908	-0.50	-0.0501			0.01	-0.0008			0.46	0.0440		
80911	-0.47	-0.0401			-0.51	-0.0517			-0.84	-0.0826		
80912	-0.15	-0.0163			-0.31	-0.0312			-0.40	-0.0312		
80913	0.37	0.0314			-0.03	0.0001			0.71	0.0428		
80916	0.51	0.0281			0.10	0.0069			-0.21	-0.0106		
80917	-0.73	-0.0392			-0.29	-0.0200			0.29	0.0242		
80918	-0.28	-0.0173			-0.82	-0.0759			-0.05	-0.0037		
80919	N/A	-0.0532			N/A	-0.0456			N/A	0.1411	BB	F
80923	0.68	0.0666			0.51	0.0533			0.82	0.0851		
80924	-0.12	-0.0085			0.32	0.0349			0.52	0.0473		
80926	-0.05	-0.0006			-0.52	-0.0473			-0.24	-0.0222		
80928	-0.45	-0.0381			-0.92	-0.0899			-0.35	-0.0295		
80929	-0.17	-0.0121			-0.88	-0.0753			-0.04	-0.0048		
80930	0.36	0.0290			-0.17	-0.0217			0.40	0.0370		
80933	N/A	0.0006			N/A	0.0888			N/A	-0.0195		
80935	-0.64	-0.0374			-0.60	-0.0415			-0.70	-0.0477		
80938	-0.42	-0.0301			0.30	0.0195			-0.88	-0.0614		
80939	-0.12	-0.0037			-0.12	-0.0125			1.35	0.0960	B	F
80942	N/A	-0.1268	BB	W	N/A	-0.1275	BB	W	N/A	0.1225		
80943	-0.04	-0.0024			-0.68	-0.0417			0.11	0.0063		
80944	N/A	-0.2210	BB	W	N/A	-0.1340			N/A	-0.0704		
80955	-1.11	-0.0500	B	W	-1.56	-0.0791	C	W	0.28	0.0126		
80956	-1.12	-0.0991	B	W	-1.05	-0.1047	B	W	-0.40	-0.0370		
80960	-1.08	-0.0536	B	W	-1.24	-0.0625	B	W	-0.24	-0.0097		
80963	0.13	0.0216			0.03	-0.0007			0.27	0.0220		
80138_01	-0.45	-0.0368			-0.36	-0.0297			0.10	0.0086		
80138_02	-0.66	-0.0689			0.03	-0.0012			-0.42	-0.0341		
80138_03	-0.15	-0.0180			-0.02	-0.0040			-0.59	-0.0518		
80138_04	-0.12	-0.0057			0.09	0.0061			0.11	0.0105		
80138_05	-0.40	-0.0409			-0.74	-0.0562			0.70	0.0397		
80138_07	N/A	-0.0184			N/A	-0.0161			N/A	0.1498		
80746_01	-0.19	-0.0242			-0.35	-0.0428			0.06	0.0069		
80746_02	-0.35	-0.0307			-0.32	-0.0303			0.69	0.0558		
80746_03	-0.65	-0.0466			-0.44	-0.0398			-0.16	-0.0148		
80746_04	-0.49	-0.0115			0.08	0.0137			-0.06	-0.0056		
80746_05	-0.28	-0.0436			-0.92	-0.0990			-0.16	-0.0159		
80746_06	N/A	0.0021			N/A	-0.0132			N/A	0.1699	CC	F
80750_01	-0.47	-0.0366			-0.10	-0.0131			-0.41	-0.0407		
80750_02	0.27	0.0229			0.20	0.0138			-0.57	-0.0509		
80750_03	0.03	0.0022			-0.16	-0.0180			-0.63	-0.0588		

	Black/White				Hispanic/White				Male/Female			
UIN	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**	**Delta**	**SMD**	**Sig**	**Favor**
80750_04	-0.17	-0.0183			-0.26	-0.0269			0.24	0.0186		
80750_05	-0.13	-0.0018			0.17	0.0114			0.15	0.0135		
80750_07	N/A	-0.2008	BB	W	N/A	-0.2361	CC	W	N/A	0.1531	BB	F
80751_01	-0.66	-0.0550			-0.90	-0.0932			-0.44	-0.0435		
80751_02	0.13	-0.0008			0.02	0.0004			-0.29	-0.0257		
80751_03	-0.12	-0.0003			0.21	0.0138			0.08	0.0064		
80751_04	-0.49	-0.0130			0.27	0.0159			0.12	0.0060		
80751_05	0.50	0.0468			-0.02	0.0016			0.52	0.0488		
80751_07	N/A	-0.0750			N/A	-0.1670	BB	W	N/A	0.0809		
80921_01	-0.12	-0.0036			-0.50	-0.0325			0.12	0.0074		
80921_02	0.05	-0.0032			-0.30	-0.0333			-0.00	-0.0006		
80921_04	-0.26	-0.0169			-0.55	-0.0514			0.07	0.0037		
80921_05	0.57	0.0627			0.70	0.0629			0.27	0.0235		
80922_01	-0.04	0.0016			-0.07	-0.0126			0.46	0.0480		
80922_02	-0.12	-0.0058			-0.01	-0.0020			0.28	0.0208		
80922_04	0.35	0.0318			0.00	0.0011			0.57	0.0577		
80922_05	-0.24	-0.0296			-0.24	-0.0199			-0.14	-0.0121		
80923_01	-0.14	-0.0254			-0.10	-0.0061			0.06	0.0034		
80923_02	0.06	0.0013			0.24	0.0183			0.22	0.0180		
80923_03	-0.21	-0.0029			-1.03	-0.0309	B	W	-0.61	-0.0190		
80923_05	0.32	0.0248			0.23	0.0189			1.04	0.0802	B	F
80926_01	-0.84	-0.0710			-0.30	-0.0298			-0.24	-0.0232		
80926_03	0.20	0.0240			0.01	0.0014			0.22	0.0190		
80926_04	0.99	0.0624			0.95	0.0564			-0.02	-0.0007		
80926_05	-0.41	-0.0392			-0.62	-0.0651			-0.27	-0.0253		
80927_01	-0.60	-0.0442			-0.34	-0.0237			-0.24	-0.0168		
80927_02	-0.37	-0.0345			-0.41	-0.0330			0.29	0.0236		
80927_03	0.20	0.0163			0.16	0.0126			-0.34	-0.0323		
80927_05	-0.34	-0.0096			0.53	0.0340			0.14	0.0089		
80928_02	0.05	0.0076			0.41	0.0323			0.21	0.0162		
80928_03	-0.78	-0.0617			-0.21	-0.0208			0.16	0.0124		
80928_04	-0.42	-0.0318			-0.79	-0.0679			0.25	0.0156		
80928_05	-0.18	-0.0254			-0.54	-0.0612			0.39	0.0401		
80929_01	0.08	-0.0005			-0.02	-0.0027			-0.52	-0.0470		
80929_02	-0.75	-0.0523			-1.24	-0.1068	B	W	-0.05	-0.0035		
80929_03	-0.05	0.0053			-0.55	-0.0548			-0.32	-0.0289		
80929_04	0.34	0.0276			0.10	0.0083			-0.27	-0.0256		
80930_02	-0.13	-0.0248			-1.03	-0.0882	B	W	-0.07	-0.0051		
80930_03	-0.12	-0.0192			-0.25	-0.0264			-0.04	-0.0031		
80930_04	-0.32	-0.0392			-0.55	-0.0538			-0.25	-0.0207		
80930_05	-0.12	-0.0099			-0.14	-0.0136			0.69	0.0672		
80931_01	-0.08	-0.0097			0.90	0.0698			-0.33	-0.0282		
80931_02	0.04	0.0015			-0.48	-0.0326			0.00	0.0005		
80931_03	0.50	0.0437			0.29	0.0249			0.50	0.0401		

	Black/White				Hispanic/White				Male/Female			
UIN	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor	Delta	SMD	Sig	Favor
80931_04	-0.75	-0.0713			-0.63	-0.0664			-0.12	-0.0106		
80932_01	0.18	0.0085			-1.22	-0.1286	B	W	-0.14	-0.0118		
80932_02	-0.18	-0.0238			-0.93	-0.0734			-0.12	-0.0089		
80932_04	-0.52	-0.0424			-0.14	-0.0074			-0.82	-0.0702		
80932_05	0.60	0.0396			0.43	0.0354			1.12	0.1047	B	F
80935_01	-0.04	-0.0003			0.38	0.0308			-0.55	-0.0481		
80935_02	0.31	0.0246			0.18	0.0141			0.21	0.0164		
80935_03	-0.57	-0.0440			-1.33	-0.1086	B	W	0.66	0.0507		
80935_04	-0.32	-0.0182			-0.53	-0.0423			0.11	0.0079		
80936_01	0.24	0.0199			0.50	0.0312			-0.58	-0.0329		
80936_02	-0.79	-0.0453			-0.68	-0.0470			-0.93	-0.0672		
80936_03	0.00	-0.0008			0.65	0.0640			0.31	0.0295		
80936_04	-0.48	-0.0477			-0.10	-0.0090			0.58	0.0477		
80939_01	-0.32	-0.0248			0.36	0.0279			-0.08	-0.0085		
80939_02	-0.08	0.0099			0.02	-0.0057			-0.37	-0.0326		
80939_04	-0.13	-0.0108			-0.77	-0.0538			-0.10	-0.0063		
80939_05	-1.07	-0.0641	B	W	-1.60	-0.1020	C	W	0.70	0.0301		
80940_01	-0.38	-0.0387			-0.42	-0.0396			-0.21	-0.0189		
80940_02	0.09	0.0092			-0.24	-0.0206			0.00	-0.0015		
80940_03	-0.29	-0.0270			-0.26	-0.0221			-0.32	-0.0314		
80940_04	-0.08	-0.0013			0.34	0.0319			0.13	0.0126		
80940_05	N/A	-0.0517			N/A	-0.0991			N/A	0.0738		
80941_01	-0.06	-0.0014			-0.31	-0.0231			-0.05	-0.0032		
80941_02	0.02	0.0262			-0.18	-0.0160			-0.34	-0.0333		
80941_03	0.01	-0.0075			-0.06	-0.0146			0.32	0.0266		
80941_04	-0.37	-0.0152			0.35	0.0297			-0.64	-0.0558		
85056_01	-0.13	-0.0061			0.37	0.0272			-0.28	-0.0257		
85056_02	0.05	0.0023			-0.20	-0.0186			0.27	0.0214		
85056_04	-0.24	-0.0049			-0.13	-0.0065			-0.32	-0.0306		
85056_05	0.06	0.0017			-0.44	-0.0368			0.22	0.0129		
85056_06	-0.37	-0.0146			-0.11	-0.0020			1.20	0.0467	B	F
85056_08	N/A	-0.0252			N/A	-0.0349			N/A	0.1515	BB	F

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)