Technical Manual for the Maryland Integrated Science Assessment

Grades 5 and 8

Academic Year
2018–19

October 2019

Prepared by Pearson

Table of Contents

Index of Tables

Table of Figures

Chapter 1: Introduction

Background

The Maryland Integrated Science Assessment (MISA) are science assessments in grades 5, 8, and High School (this manual addresses the grade 5 and 8 tests). These assessments provide educators, parents, and the public with information on student progress towards science literacy. Administered annually in the spring, MISA was established to meet the requirements of the Every Students Succeeds Act (ESSA) of 2015. ESSA requires that states administer to all students annual assessments in science once in each grade span (3-5, 6-8 and HS) that are aligned to state standards. In 2013, the Maryland State Board of Education adopted the Next Generation Science Standards (NGSS) as the new Maryland Science Standards. Pearson was contracted by Maryland State Department of Education (MSDE) to develop, administer, and maintain the MISA tests. This report provides technical details of work accomplished during the 2018-2019 test administration cycle. This administration marks the second operational administration of MISA. A stand-alone field test was administered in 2016-2017.

Purpose and Uses

By assessing student achievement against the NGSS academic standards, the MISA tests serve two important purposes. First, MISA provides an accountability tool that measures overall performance as well as differing levels of defined performance across students, schools, and districts against the NGSS standards. Second, it provides stakeholders with important information about what students have learned, which, if applied constructively, can foster improvement of instructional programs, classroom education, and school performance. Improved student learning is a key goal of any educational assessment program.

This manual can support educators in using test results to inform and improve instruction, and by extension enhance student learning. In addition, this manual can serve as a resource for educators in explaining assessment information to students, parents, teachers, school boards, and the public. The purpose of this MISA Technical Manual is to provide objective information regarding technical aspects of the 2019 MISA operational tests at grades 5 and 8. It 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 MISA tests. Other sources of information regarding MISA include the MISA 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 MISA tests. Specifically, information was selected for inclusion in this report based on ESSA requirements and standards from the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014).

This manual provides information about the MISA assessments regarding:

1. Content of the tests;
2. Test form design;

3. Identification of ineffective items;
4. Reliability of the tests;
5. Statistical characteristics of the test questions;
6. Calibration of test forms;
7. Detection of item bias;
8. Scoring and reporting the results of the tests.

It should be noted that this report primarily addresses technical details with respect to the student level components and scores (based on all students). Components that incorporate the matrixed elements are included in the design and item level statistics sections. However, they are not included in performance breakouts or test level analyses since these are not administered to every student and are only reported out at aggregate levels.

From test development to final reporting, each of these facets of the MISA tests contribute to the validity of the inferences made about the test results. This technical manual addresses these topics for the 2018-2019 testing year.

Chapter 2: Test Development

MISA Content Coverage

The MISA Science tests are built to align with the Maryland Next Generation Science Standards. According to MSDE's website, the NGSS are composed of three dimensions from the National Research Council (NRC) Framework. The NRC Framework describes a vison of what it means to be proficient in science. It rests on a view of science as both a body of knowledge and an evidence-based model and theory building enterprise that continually extends, refines, and revises knowledge. It presents three dimensions that will combined to form each performance expectation.

Dimension 1: Scientific and Engineering Practices

The Scientific and Engineering Practices (SEP) describe behaviors that scientists engage in as they investigate and build models and theories about the natural world and the key set of engineering practices that engineers use as they design and build models and systems.

The NRC uses the term practices instead of a term like "skills" to emphasize that engaging in scientific investigation requires not only skill but also knowledge that is specific to each practice. Part of the NRC's intent is to better explain and extend what is meant by "inquiry" in science and the range of cognitive, social, and physical practices that it requires.

Dimension 2: Crosscutting Concepts

Crosscutting Concepts describe concepts that bridge disciplinary boundaries, having explanatory value throughout much of science and engineering. These crosscutting concepts have application across all domains of science and are a way of linking the different domains of science. These include:

- Patterns;
- Cause and Effect;
- Scale, Proportion and Quantity;
- Systems and System Models;
- Energy and Matter: Flows, Cycles, and Conservation;
- Structure and Function; and
- Stability and Change.

The Framework emphasizes that these concepts need to be made explicit for students because they provide an organizational schema for interrelating knowledge from various science fields into a coherent and scientifically-based view of the world.

Dimension 3: Disciplinary Core Ideas

Disciplinary core ideas have the power to focus K–12 science curriculum, instruction, and assessments on the most important aspects of science. To be considered core, the ideas met at least two of the following criteria and ideally all four:

1. Have **broad importance** across multiple sciences or engineering disciplines or be a key organizing concept of a single discipline;
2. Provide a **key tool** for understanding or investigating more complex ideas and solving problems;
3. Relate to the **interests and life experiences of students** or be connected to societal or personal concerns that require scientific or technological knowledge;
4. Be teachable and learnable over multiple grades at increasing levels of depth and sophistication.

Disciplinary ideas are grouped in four major domains: physical sciences; the life sciences; the earth and space sciences; and engineering, technology and applications of science.

Performance Expectations and Evidence Statements, Types, and Families

The focus of MISA test development is to create sets of items that are related to a stimulus (phenomenon) and are aligned to one or more of the NGSS performance expectations (PEs) and use them to elicit evidence of student achievement with respect to the NGSS standards.

PEs provide descriptions of what students should be able to do by the end of instruction for a given grade level or grade band, and are designed “to gather evidence of students’ ability to apply the practices and their understanding of the crosscutting concepts in the contexts of specific applications in multiple disciplinary areas.” (National Research Council, 2012, p. 218).

In an effort to describe more specifically what proficient student performance of the PEs would look like, evidence statements were developed for every PE in every grade level. These are intended to provide clear, measurable components that, if met, fully satisfy each PE described within the NGSS (NGSS, 2015). Together, performance expectations and evidence statements are used to guide the development of the MISA tests and add to the framework of reporting MISA results to students, teachers, and others. The performance expectations that are assessed on MISA at each grade level are shown in Tables 1 and 2 below.

Table 2.1. Grade 5 MISA Performance Expectations Assessed by Domain

Earth & Space Science	Life Science	Physical Science
3-ESS2-1	3-LS1-1	3-PS2-1
3-ESS2-2	3-LS2-1	3-PS2-2
3-ESS3-1	3-LS3-1	3-PS2-3
4-ESS1-1	3-LS3-2	3-PS2-4
4-ESS2-1	3-LS4-1	4-PS3-1
4-ESS2-2	3-LS4-2	4-PS3-2
4-ESS3-1	3-LS4-3	4-PS3-3
4-ESS3-2	3-LS4-4	4-PS3-4
5-ESS1-1	4-LS1-1	4-PS4-1
5-ESS1-2	4-LS1-2	4-PS4-2
5-ESS2-1	5-LS1-1	4-PS4-3
5-ESS2-2	5-LS2-1	5-PS1-1
5-ESS3-1		5-PS1-2
		5-PS1-3
		5-PS1-4
		5-PS2-1
		5-PS3-1

Table 2.2. Grade 8 MISA Performance Expectations Assessed by Domain

Earth & Space Science	Life Science	Physical Science
MS-ESS1-1	MS-LS1-1	MS-PS1-1
MS-ESS1-2	MS-LS1-3	MS-PS1-2
MS-ESS1-3	MS-LS1-4	MS-PS1-4
MS-ESS1-4	MS-LS1-5	MS-PS1-5
MS-ESS2-2	MS-LS1-6	MS-PS1-6
MS-ESS2-3	MS-LS1-7	MS-PS2-1
MS-ESS2-4	MS-LS2-1	MS-PS2-3
MS-ESS2-5	MS-LS2-2	MS-PS2-4
MS-ESS2-6	MS-LS2-3	MS-PS2-5
MS-ESS3-1	MS-LS3-2	MS-PS3-1
MS-ESS3-2	MS-LS4-1	MS-PS3-2
MS-ESS3-3	MS-LS4-2	MS-PS3-3
MS-ESS3-4	MS-LS4-3	MS-PS3-4
MS-ESS3-5	MS-LS4-4	MS-PS3-5
	MS-LS4-5	MS-PS4-1
	MS-LS4-6	

While the granularity of the evidence statements for PEs was appropriate in focusing MISA item set development, it was decided that for scoring, reporting, and using MISA for instruction, it would be more useful if evidence statements could be aggregated at a higher level.

This aggregation was done by first extracting from the NGSS the headings that were used to group together the evidence statements for each content domain and performance expectation. These headings – now labelled evidence types – were reviewed by Pearson and MSDE, and then aggregated into clusters of headings for similar kinds of evidence. For example, the evidence types of collecting and organizing data, identifying relationships, and interpreting data were grouped together to form the cluster - or evidence family – of data and information.

The evidence families and the evidence types within them are shown in Table 3.

Table 2.3. MISA Evidence Family Categories

Evidence Family	Evidence Types
Data and information	Collecting and organizing data; identifying relationships; interpreting data
Claims and evidence	Identifying, evaluating, and critiquing evidence; supporting claims
Reasoning	Reasoning and synthesis
Phenomena	Addressing phenomena of the natural world; identifying the phenomenon under investigation; articulating the explanation of a phenomenon
Design solutions and constraints	Using scientific knowledge to generate design solutions; describing criteria and constraints; evaluating potential solutions; modifying the design solution;
Model components, relationships, and connections	Components of the model and their connections and relationships
Representations and analysis	Representations; mathematical modeling and computational analysis
Investigations	Identifying the scientific nature the question; identifying the evidence to address the purpose of the investigation; planning the investigation

MISA Test Design and Development

Overview

In order to assess the three dimensions of the performance expectations found in the standards, a set of interrelated items is required. There are no items on the MISA that are not part of an item set. The MISA uses the item set as the building block of the assessment. Specific items may focus on two of the dimensions, but together in a set, all three dimensions are covered and inferences can be made about a student's three-dimensional learning. Each item set is based on a stimulus (i.e., a scientific phenomenon) with six selected response (SR), technology enhanced (TE), and constructed response (CR) items. Students are administered 11 item sets which are presented to them in four testing sessions called units.

Each item set on the MISA has a stimulus that focuses on a specific real world context or phenomenon. The stimulus and items form a storyline and includes multiple components that work together to partially or fully assess a bundle of chosen Performance Expectations (i.e., a group of related Performance Expectations from the NGSS). This requires students to explicitly use their understanding of the three dimensions to make sense of the information provided to them. The students can refer to the content in the stimulus while answering all the items in the item set.

The stimulus may include technical passages to read, a video, charts/diagrams, or a simulation with which the student interacts. The stimulus may include multiple tabs for student interaction. After the student interacts with the stimulus they will be given six items that are supported by the stimulus.

For the operational MISA tests, the item sets in Units 1, 2, and 3 are referred to as being a part of the core forms (or cores), and will be used to produce individual student scores. Each core form consists of six item sets that are unique to that form, and three item sets that are common across two forms that will be used as equating links between the core forms.

Unit 4 in the operational test contains a combination of two item sets. For some students, this is one of three different matrixed item sets and one field test item set, while other students will take two field test sets. The matrixed item sets are used to provide additional content coverage for the reporting of school-level and above scores. The field test item sets are used to pilot new core or matrixed item sets for inclusion on future MISA forms. The following notes provide additional details regarding the composition of the Units on the MISA test:

Design of Units 1, 2, and 3

- Units 1, 2, and 3 are the core forms that are used to produce individual student scores
- Each Unit contains three item sets – one from each Domain (Life Science, Earth Space Science, Physical Science)
- Each item set contains five to six items and each Unit will thus have 17 - 18 items and total 24 raw score points
- Each item set contains one CR item
- Two of the item sets contains a 3-point CR item
- The third item set contains a 2-point CR item
- The remaining five items within an item set consists of 1-point SR or TE items
- Across the three units, each Domain should total 24 raw score points (total score on the core form of 72)

Design of Unit 4

- Unit 4 is employed in two ways, either for adding a matrixed item set for producing school level scores, or for field-testing new core and matrixed item sets
- Some versions of Unit 4 contain one of three different operational matrixed item sets – one for each Domain – and a field test item set

- The matrixed item set contains a simulation and one 4-point CR item and five 1-point SR items
- The field test item set consists of either a 2- or 3-point CR item and four to five 1 or 2-point SR or TE items
- Other versions of Unit 4 consist of two field test item sets, each with either a 2- or 3-point CR item and five 1-point SR or TE items

To ensure that MSDE is in accordance with the federal law that requires states to align their tests to their content standards, the NGSS 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 MISA assessments.

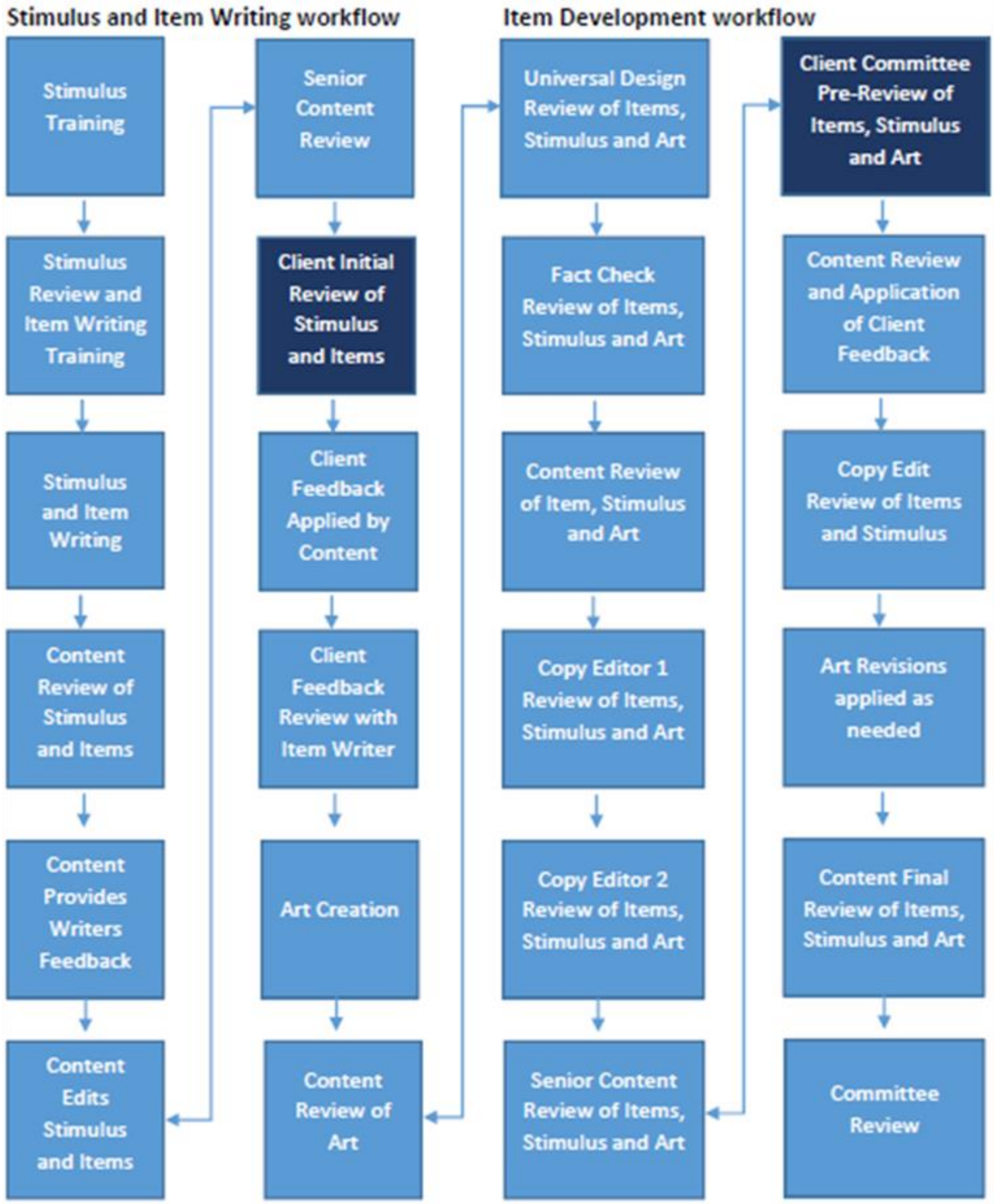
The basic test specifications were established by MSDE with help by Pearson to guide the test development and administration. 2019 marked the second operational administration of MISA. All administrations were conducted under the same testing conditions (see Chapter 4). Accordingly, the field test was 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. To help discourage cheating behavior, two base forms (i.e., two forms of scored operational items) are used for each grade. Each core form has a total of 54 items yielding a total raw score of 72 points. For both grade tests, only core operational items contributed to individual student scores. The two base forms share a set of 18 common items (one item set per domain). These common items provide an internal link used for placing all items onto a common scale via concurrent calibration (described in chapter 7). In 2019, they also provide a means of linking back to the base 2018 scales.

Item and Stimulus Development

Items and stimuli were developed through externally hired professional item writers. These item writers were recruited and trained in late 2017 and early 2018. Training was facilitated by Pearson assessment specialists and included instruction on the MISA test design, NGSS standards, selection of technical passages, creation of effective and authentic stimuli, and characteristics of the MISA item formats (selected response, multiple response, technology enhanced, and constructed response). Additionally, item writers were given style information, item delivery schedules, and content resource suggestions.

Once the items and stimuli were created and submitted, they were reviewed by Pearson assessment specialists. Items were reviewed according to well-defined criteria approved by MSDE. Item writers were provided feedback on items that were rejected due to not meeting the criteria for quality or in need of major revision and allowed to resubmit. Once accepted, items and stimuli were vetted through the internal Pearson item development process. The following diagram depicts the overall item and stimulus development workflow.

Figure 2.1. Maryland Science Stimulus Item Development Workflow



The 2019 operational MISA are comprised of three types of items: selected response (SR), technology enhanced (TE), and constructed response (CR). SR items require students to select a correct answer from four options. Each SR is scored dichotomously (i.e., 0 or 1). CR items require students to provide a short answer using words, numbers, and/or symbols. All CR items are scored using generic rubrics by maximum score point. CRs range 0-2, 0-3, and 0-4 based on concordant scores from two independent raters (see Chapter 5 for details).

MISA Blueprint

Table 4 presents the general blueprint used for MISA grades 5 and 8. Each core test is created such that a broad range of performance expectations within and across domains is represented (see Tables 2.1 and 2.2). The performance expectations on the 2019 MISA tests are presented in Appendix A by grade and core form.

Table 2.4. MISA Blueprint

	Core Sets*		Matrix Sets**	
Domain	**# SR/TEI**	**CR**	**# SR/TEI**	**CR**
Physical Science	15	3	0 or 5	0 or 1
Life Science	15	3	0 or 5	0 or 1
Earth and Space Science	15	3	0 or 5	0 or 1
Total	45	9	15	3

**Note: item sets that contain a 2-point CR are combined with a 2-point TE item in the same set.*
***Note: "0" denotes forms which have a field test matrix set as opposed to an operational set.*

Chapter 3: Test Construction

The 2019 operational MISA tests were created in line with the test design and blueprint presented earlier. The process of selecting items for the two core forms per grade was an iterative process primarily involving Pearson content experts, MSDE, and Pearson psychometricians.

Initial Build

Pearson content specialists and psychometricians worked jointly on the preliminary test build. The test development team selected the "best" items within an item set from a content perspective, to meet the MISA test construction guidelines. The general process to follow was:

- For each grade, item level statistics were used to select the "best" 6 of the 12 items within each item set for the Core sections
- At the item set level statistical analysis information was used to get a sense of the overall difficulty of the item sets
- First, Psychometrics recommended item set combinations to serve as linking items for equating purposes based on statistical and content match to base forms. The item sets that served as a common link back to the 2018 base forms and across cores 1 and 2 (one each of ESS, LS, and PS) were considered by Pearson content and finalized
- Next, the remaining item sets were chosen to fill out the units of each core such that each domain appears once per Unit
- Sets were reviewed and adjusted as needed based on
    - The numbers of 2-point and 3-point items
    - The coverage of performance expectations
    - Content considerations such as cluing, content "duplication", etc.
    - Level of student engagement
    - Spread of item difficulties across the full performance range (overall and by domain)

Selection of Item Sets for Linking Operational Core Forms

As noted, the 2019 MISA linking sets that were used to provide a statistical link to the 2018 base operational forms and also between the two core operational forms within the 2019 administration in order to provide a mechanism for placing the two core forms on a common scale. As such, care was taken to select only high-quality item sets to serve as links. Additionally, set combinations were selected to most closely match the overall statistical characteristics of the 2018 base forms.

The linking sets were intact MISA item sets and all item sets that were eligible to serve on an operational core form were eligible to be used as a linking set. This meant that the item sets could be selected from the pool of field-tested item sets.

Ideally, the linking sets should representative of the entire MISA test in terms of content coverage and difficulty. As such, three item sets were selected for linking the core forms, one from each of the Domains. The linking item sets appeared in the same location on both operational core forms and were spread throughout all sessions of the test.

Statistical Guidelines for Selection of Items and Sets

The purposes of statistical analyses and reviews were:

- First, to develop 2019 MISA operational test forms and item banks based on the results of the 2019 field test analyses and existing bank of items from 2018; and
- Second, to create statistical targets based on the 2018 operational administration that would be used to guide MISA operational test form development for 2019 and beyond.

With respect to the analyses of the field test results for constructing the 2019 operational forms, it is important to remember that MISA is a new kind of test. In particular, the entire MISA consists of stimuli (i.e., phenomena) with their associated sets of items. Because of this, consideration of different levels of analysis was important:

- Individual items
- Item sets

Individual Items

Several classical item statistics were used to evaluate the quality of individual items within item sets during the test construction process. These statistics include:

For dichotomously scored items

- p-value for item difficulty
- point-biserial correlation for item discrimination
- percent choosing each item option (i.e., distractor) for multiple choice items
- item option point-biserial for multiple choice items
- Mantel-Haenszel differential item functioning flags and levels

For polytomously scored items

- mean score for item difficulty
- item-total correlation for item discrimination
- item score distribution
- standardized mean difference (SMD) DIF statistics flags and levels

Items were flagged for further review when their field test statistics failed to meet certain statistical criteria. These included:

- Extremely high or low p-value, or item mean with respect to range:

  If greater than 0.90 or less than 0.20

- High omit rate:

  If greater than 5% omit rate

- Extremely low point-biserial or item-total correlation:

  If less than 0.10 (Note that items with point-biserial item-total correlations less than zero are extremely flawed and not acceptable for operational forms):

- Highly attractive multiple-choice item option (distractor):

  If an item option percentage greater than 40%

- Highly attractive multiple-choice item option (distractor):

  If an item option point-biserial is greater than the point-biserial.

- Item shows differential item functioning

  If the DIF index is significant.

Differential Item Functioning

Differential item functioning (DIF), is a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups. In examining 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 (a detailed description of DIF is presented in Chapter 6).

For the MISA DIF analyses, the reference groups were White for ethnicity, and male for gender. The focal groups were females, and African-American and Hispanic ethnic groups.

Items were flagged into one of three categories based on the magnitude of their DIF statistics:

- Category A: no or negligible DIF
- Category B: slight or moderate DIF, and
- Category C: moderate to large values of DIF. These items which exhibit significant DIF, are of primary concern.

All items exhibiting DIF underwent additional content review in order to determine the source and meaning of performance differences.

Item Sets

Additional information was available to the test development team regarding the performance of the item sets and field test forms. This included:

For item sets

- mean and standard deviation of item set total
- item set score distribution

Content Review

To better identify potentially "dependent" items (i.e., items in which the answer to one question may influence how students perform on another) during content review, the entire test was taken in one sitting to identify and address potential issues of cluing. It was important that all items selected for use within and across item sets do not clue one another.

Each item was reviewed for accuracy, clarity, and appropriateness of content. The test was also reviewed for coherency, diversity of content and flow. Additionally, the test development team verified the following:

- the accuracy of item-level content classifications
- the accuracy of scoring keys
- the representation of scoring keys (i.e., want 25% of each A-D)
- the appropriateness of the proposed item sequence (e.g., no more than 3 items with same key in row)
- diversity of subject matter within stimulus

When determining the order in which items should be presented several factors were considered:

- Location of linking items – Items that serve to link the two base forms should be in the same position on each form.
- Item keys – Several selected response items having the same key should not be presented adjacent to each other on a form.
- Similarity of passages and lab stimulus – To the extent possible the subject matter, length and reading difficulty was varied across the test.

After content review was completed, the content team determined whether the initial build needed to be revised. If not, the form was sent to psychometrics for review and then to MSDE for their review. This iterative process continued until content experts, MSDE, and psychometrics finalized and approve each respective core form.

Field Test Form Assembly

After operational forms were approved, the test development team assigned newly developed items to field test forms for field-testing. Factors that were considered in determining how to assign items to forms are outlined below. The number of items associated with a given item set varied slightly from one form to another in some instances. As noted previously, the field-test item locations are in Unit 4.

Several factors were considered when assigning items/passages to forms:

- *Cueing/Clueing.* Field-test items were evaluated against the given core form to ensure they did not clue the answer to other field test items on the form OR any of the operational items.
- *The type of items represented on each form.* Ideally a mix of item types appeared on each form. Similarly, multiple standards and objectives were represented.
- *The number of items associated with a given item set.* Item sets were field-tested with enough items to allow for attrition. Each field-tested item set was placed on two different field test forms with its own set of items. (For example, the same item set stimulus appeared on one form with six items and on a second form with another six items).
- *The distribution of keys and the number of items having the same key placed adjacent to one another.* Similar to operational forms – the key distribution and placement was considered when selecting/sequencing items.
- *Stimulus passage difficulty/reading load.* The mix of stimulus passage difficulties and lengths on a given form was considered.

In selecting field-test items for forms the team reviewed the accuracy and appropriateness of the proposed field-test forms. Specifically, they considered:

- Does every item clearly meet its identified objective?
- Is every item free from cueing?
- Is the content of every item clear and accurate?
- Are there a variety of item types on each form?
- Are there a variety of standards and objectives represented on each form?
- Is the key accurate and accurately represented in the test map?
- Is there only one correct response?
- Are the items/passages free of typographical, spelling, punctuation, or grammatical errors?
- Does the hardcopy test build match the test map provided to the customer (form number, item number, item UIN, key, passage title, and objective)?
- Are items ordered appropriately (limit the number of items having same key adjacent to one another)?

Chapter 4: Test Administration

Test Window

The overall test window for MISA is established by MSDE. Each Local Education Agency (LEA) sets a specific schedule for administration of MISA within the testing window for its district. Each LEA must submit a schedule of their paper testing dates to MSDE in advance for approval by the state. For each given grade level (grade 5 or 8), all testing takes place according to the state approved schedule established by each LEA.

The testing schedule allows for approximately 60 minutes for each testing Unit (excluding preparation time). Testing is scheduled to allow for the completion of applicable Units each day. Extra consideration is given to scheduling test administrations for students who receive the extended time accommodation to ensure enough time is available to complete the started Unit tested that day. MISA consists of four Units. Units can be tested over the course of four days, or multiple Units can be tested in a day; however, it is recommended that no more than two Units be tested on the same day.

For MISA, the testing schedule for 2019 was as follows:

- Test Materials arrive in Schools February 20-22, 2019
- Paper Test Window March 11-22, 2019
- Online Test Window March 11-29, 2019

Test Format

Each set of items, within a Unit, consists of Selected Response (SR), Constructed Response (CR) items, and Technology Enhanced (TE) items (online only), based on shared stimuli. The online version of the test also includes interactive stimuli and may also contain videos.

For the paper version of MISA, each student uses a test book containing all test items and response areas. Since the test books are scanned for scoring, students do not use a highlighter or make stray marks in any part of the book or tamper with the barcode on the label. In addition, for CR items, students can write their responses within the boxed area only. Responses written outside the boxed area are not scored.

Preprinted student ID labels are used for most students participating in the paper test administration. Students and staff cannot write on or tamper in any way with the student barcode label. The barcode on the labels contained encoding which links the Test Book to a specific student.

Testing Accommodations

Testing accommodations for students with disabilities (i.e., students having an Individualized Education Program [IEP] or a 504 Plan) or students who are English Learners (EL) (i.e., students who have an EL Plan) have to be approved and documented according to the procedures and requirements outlined in the document entitled Maryland Accommodations Manual: Selecting, Administering, and Evaluating the Use of Accommodations for Instruction and Assessment

(MAM). No accommodations can be made for students merely because they are members of an instructional group. Any accommodation has to be based on individual need documented in the student's IEP—not on a category of disability area, level of instruction, environment, or other group characteristics.

Large Print and Braille Test Books, and Transcription

MISA is administered to students requiring Large Print and Braille Test Books. For Large Print Test Books and Braille Test Books, student responses have to be transcribed into the standard-size Test Book or TestNav after testing. The student's name, date of birth, LEA number, and school number are to be written on the Large Print or Braille Test Book for proper transcribing into the standard-size Test Book. An eligible TA transcribes the student's responses into a standard-size Test Book or into TestNav exactly as given by the student. At least two persons are present during transcription of student responses.

Human Reader Accessibility Feature and Text-to-Speech Tests on PearsonAccess^Next

Students who receive this accessibility feature in regular instruction receive the same accessibility feature on the MISA. The accessibility feature is provided either by a human reader or through Text-to Speech in TestNav.

Online Human Reader Accessibility Feature

For those students that take MISA online and receive a Human Reader accessibility feature, this is provided in one of two ways; either (1) by using TestNav to access a Text-to-Speech form or (2) by a human reader, individually or in a group called a "Human Reader" Session.

For individual students who test online and receive an individual Human Reader accessibility feature by a human reader, the individual providing the reading sits next to the student and reads the text which appears on the computer screen. Students who test online and receive their Human Reader accessibility feature in a group are placed into a Human Reader session in PearsonAccess^Next. Placing students in the Human Reader session allows all students in that session to be assigned the same test form and allows the Test Administrator to receive a "Proctor Testing Ticket" Testing PearsonAccess^Next. The Proctor Testing Ticket allows the TA to log in to TestNav and view the same test as the students in the Human Reader session. The TA then reads the test aloud to the students.

Administrative Procedures for Students with IEP, 504 Plan, or EL Plan Permitting a Dictated Response or Use of a Word Processor

A student whose IEP, 504 Plan, or EL Plan permits a dictated response has his/her responses transcribed at the school level by an eligible TA into the student's Test Book or into TestNav. At least two persons are present during transcription of students responses. A student whose IEP, 504 Plan, or EL Plan permits the use of a word processor either takes the test online via TestNav or has his or her responses transcribed by hand exactly as the student enters the responses on the word processor. After the student's responses are transcribed, the memory of the word processor is cleared. The original word-processed printout is returned to Pearson with the nonscorable materials.

Extended Time Accommodation

The extended time accommodation is given in one continuous block of time. The extended time student is not told to stop testing at the end of the standard testing time, and brought back to that Unit at a later time to complete the extended time accommodation. Special attention must be considered when arranging testing groups to ensure that students without the extended time accommodation do not receive more than the specified testing time stated in the Test Administrator Manual (TAM) for each assessment.

Test Security

The following code of ethics conforms to the Standards for Educational and Psychological Testing developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

IT IS A BREACH OF PROFESSIONAL ETHICS FOR SCHOOL PERSONNEL TO PROVIDE VERBAL OR NONVERBAL CLUES OR ANSWERS, TEACH ITEMS ON THE TEST, SHARE WRITING PROMPTS, COACH, HINT, OR IN ANY WAY INFLUENCE A STUDENT'S PERFORMANCE DURING THE TESTING SITUATION. A BREACH OF ETHICS MAY RESULT IN INVALIDATION OF TEST RESULTS AND LOCAL EDUCATION AGENCY (LEA) OR MSDE DISCIPLINARY ACTION.

Online versions and Test Books for MISA are confidential and are kept secure at all times. Unauthorized use, duplication, or reproduction of any or all portions of the assessment is prohibited.

VIOLATION OF SECURITY CAN RESULT IN PROSECUTION AND/OR PENALTIES AS IMPOSED BY THE MARYLAND STATE BOARD OF EDUCATION AND/OR THE STATE SUPERINTENDENT OF SCHOOLS IN ACCORDANCE WITH COMAR 13A.03.04 AND 13A.12.05.

TAs and anyone else with access to test materials are aware of the consequences of test security violations and sign a Test Administration and Certification of Training Form and Non-Disclosure Agreement, which is kept on file. Anyone handling test materials solely for clerical purposes sign a Non-Disclosure Agreement, which is kept on file.

Administration Monitoring by MSDE

MSDE sends representatives to schools throughout the state to monitor and observe testing to ensure that standardized testing procedures are being followed. Schools are not notified in advance of a monitor's visit. All monitors follow local procedures for reporting to the school's main office and signing the school's visitor log. Monitors also sign Non-Disclosure forms as requested by the school and provide a copy of a memorandum from the Assistant Superintendent for Accountability and Assessment giving authorization to monitor testing. LEAs who permit central office personnel in making observations during Maryland State testing train personnel on proper test security procedures and have all personnel sign a Non-Disclosure Agreement.

Chapter 5: Scoring Procedures

Rangefinding

Rangefinding is the activity of identifying student responses to define the range of performance levels within each score point on a given scoring rubric. Ultimately, the purpose is to arrive at consensus scores according to the standards established by the rubric so that training sets can be built that accurately reflect those standards.

Pearson's scoring staff conduct rangefinding in Maryland, in the greater Baltimore metropolitan area. To help ensure that decisions remain consistent, there are three rangefinding committees, one for each grade, and one focused on grade 5 week one and grade 8 week two. Each grade-level committee is comprised of one MSDE scoring or content facilitator, four to five Maryland educators, and two Pearson scoring directors. MSDE and Pearson begins each week with a one hour Monday meeting with a brief review of the purpose of rangefinding and the rubric, as well as other documentation of standard evaluation criteria that facilitate a common understanding of the standards and intentions of MSDE.

Each rangefinding committee systematically reviews copies of student responses for the first item, determining and recording consensus scores. The goal is to reach consensus scores on a sufficient number of student responses to construct effective training materials for each item. These responses accurately represent the range of student performance levels described in the rubrics, as interpreted by the committee members and MSDE.

The general process for review of rangefinding materials was:

1. An item is introduced and the committee members are encouraged to create a short response. A brief discussion is held to gain further insight into the prompt and possible student responses.
2. The committee then reviews the Set A responses selected by Pearson and MSDE as "grounding papers." These responses reflect the entire range of scores and be representative so that they help the committee define the lines between score points. The first "grounding paper" reviewed with the committee are a highest score point response.
3. The first set of responses is then assigned to all the attendees to read individually. The committee members read each response and assign scores on their copies of the matrix. The scoring directors collect and record all committee members' scores on the consensus sheet/matrix before any discussion begins.
4. The committee discusses each response so that scoring directors can take adequate notes for training purposes, but discussion is more extensive on responses that do not have immediate consensus. The discussion always refers to the rubric and all scores are justified with the rubric in mind. A consensus score is reached by the teacher committee members. The scoring directors note any discussion points during the review of each response.
5. Upon the completion of the first item, the process is repeated for subsequent items.

MSDE and Scoring Services staff meet at the end of each day to:

1. Review and compare the scoring of items that measure the same objectives within and across grade levels to confirm the consistency of scoring.
2. Finalize consensus scores.
3. Discuss the committee work and any scoring issues from the day.
4. Sign and date the matrix (consensus sheet) to certify the scores are recorded accurately.

Scorer Training

Students' responses to MC and TE items are machine-scored, and their responses to CR items are individually read and scored by Pearson. Using MSDE-approved training materials, Pearson scoring directors and supervisors train readers to score the MISA. Scorers attend all training and prove they have internalized the project standards by qualifying on item-specific content. Only qualified readers are allowed to score the MISA.

All scorers complete training and qualifying in order to score the MISA. To maintain security of test items, student responses, data, and employees, the following safeguards are employed:

Pearson allows only controlled access to the facility.

- Scoring personnel sign a Confidentiality and Acknowledgement agreement when hired, as well as an MSDE non-disclosure form in which they agree not to use or divulge any information concerning test items, scoring guides, or individual student work.
- All staff display Pearson identification badges at all times while in the scoring facility.
- Pearson allows no recording or photographic equipment in the scoring area without the consent of Pearson or MSDE.

Supervisors and scorers for the MISA test are selected based on their ability to commit to the duration of the project and to the professional standards of scoring, including their willingness to complete the entire training program. Pearson strives to hire only scorers that have experience in elementary and/or middle school science. Regardless of previous experience or education, however, scorers are required to demonstrate an understanding of the scoring criteria and to meet the project's qualification standards (acceptable scores on qualifying sets).

The training includes the following information:

1. Overview of Pearson
2. Overview of Next Gen Science Standards
3. Overview of MISA
4. Reader Bias Training
5. Training goals and objectives
6. Item Training
7. Overview of how to use the ePEN2 scoring system

Supervisor Training

Prior to scorer training, scoring directors train supervisors on the items their teams scores. Content training for supervisors follows the same steps as scorer training. Pearson provides all qualifying statistics for supervisors to MSDE. Scoring supervisors do not complete training for all items in the upfront supervisor training window; however, supervisors are trained on each item prior to scorer training on the item. Supervisors receive training on backreading, providing feedback to scorers, scoring issue documentation, condition codes, resolution scoring, and scorer documentation. Supervisors also receive training on the supervisor tools in the image-based scoring system.

Scorer Training

Eight scoring directors train one item per scoring group for operational scoring. When scoring on an item is complete, scoring directors train scorers on a new item. Scorers are required to qualify on each new item. Each scoring group scores 4 - 6 items.

The training process for each item consists of the following materials:

1. Scoring Guide (which includes the MISA rubric, the item, item stimulus and/or technical passage [if applicable] for the constructed response items, the anchor set, and anchor annotations)

2. 2 practice sets

3. 3 qualifying sets

For both supervisor and scorer training, scorers begin by reviewing the Before You Score MISA on their first item, then the online training item level training material. Scorers then take the first practice set in the image-based scoring system, and assign scores to these sample responses. Scorer performance on practice set 1 is recorded in reports in the image-based scoring system. Once a scorer completes the set, he/she then reviews the true scores and annotations for the given practice set; if they have any questions about the scores or annotations in the practice set, the scoring director is available to answer those questions. The same process occurs for the second practice set. If scorer performance or discussion of practice sets indicates any need for review or retraining with the Scoring Director, it occurs at that time. When all scorers complete those practice sets, everyone moves on to qualification sets.

Finally, scorers complete the three qualification sets, each consisting of 10 sample student responses. Scoring directors and scoring supervisors monitor scorers’ progress on each qualification set through online reports. If scorer performance on qualification set 1 indicates any need for review and discussion with the Scoring Director, it occurs at that time. The scores achieved on these qualification sets determine if a trainee understands and can apply the scoring criteria. Table 5.1 below shows the qualification, provisional qualification, inter-rater reliability (IRR), and Validity thresholds.

Table 5.1. Qualification, Provisional Qualification, IRR, and Validity standards

Item Type	Qualification (average on 2 of 3 sets)	Provisional Qualification (average on 2 of 3 sets)	IRR	Validity
SP 0-4	70%/100%	65%/100%	65%	65%
SP 0-3	80%/100%	75%/100%	70%	70%
SP 0-2	80%/100%	75%/100%	80%	80%

Scorers who qualify outright but have 60% or lower on any one set of the three are considered low qualifiers. Low qualifiers are coached by the supervisory staff and heavily backread. Provisional scorers are allowed if the scorer meets the criteria above. MSDE has to approve all provisional scorers.

Qualified scorers receive training on how to identify responses (alerts and condition codes) that need to be sent to scoring directors or scoring supervisors, as well as how to navigate and use the image-based scoring system. Training on the types of responses that may receive condition codes occurs after scorer qualification. Scorers are trained to recognize these types of responses and to forward them to scoring directors, but scorers do not assign condition codes themselves aside from blanks.

Scoring directors are responsible for assigning condition codes. Where necessary, scoring directors assign selected scoring supervisors to assist in reviewing responses and assigning condition codes. During scoring, scoring directors escalate any new issues about condition codes as quickly as possible to MSDE. Scoring directors and project managers closely monitor the frequency distribution of condition codes and notify MSDE if any percentage of responses receiving condition codes is greater than anticipated.

Scoring and Monitoring

All scoring is computer-based, with a 10 percent second scoring for operational items. Automated scoring performed by Pearson’s Intelligent Essay Assessor (IEA) was the first score for all but two operational CR items on MISA in 2019. Field test scoring consists of approximately 2500 responses per item and is 100 percent second scored. Scorers begin scoring each item immediately after qualification. Scorers do not know whether a response has received a previous score or what that previous score is if so.

There are three generic rubrics used to score CR items based on the maximum points earnable; 0 to 2, 0 to 3, and 0 to 4. For responses scored by two scorers, the higher score is the score of record where scores are adjacent (one-point difference). Resolution reads are required where there is a two-point or greater difference between two readers. In such cases, the "expert" third reader score will override the scores of the previous two readers.

The following highlights the quality measures that scoring services staff takes to ensure accurate scoring of MISA. A sample of the PSC Quality Management Extended Guide is presented in Appendix B, which includes validity, IRR and frequency distribution results by item and are available on demand as well as cumulatively.

Backreading

Backreading is one of the primary responsibilities of scoring directors and scoring supervisors and starts at the beginning of scoring. It is an immediate source of information on scoring accuracy. It alerts scoring directors and scoring supervisors to misconceptions at the team level, allowing them to quickly calibrate or retrain scorers. Backreading continues throughout the scoring of the project. Approximately five percent of the scored responses will be reviewed through backreading. To help ensure that students receive accurate scores, scores assigned in the backreading queue will override scores assigned in the first or second scoring queue.

Findings from backreading result in any or all of the following:

- The supervisor clarifies the issue(s).
- Scorers review training materials.
- Supervisor backreads the scorers' work more extensively.
- Supervisory staff gives scorers further training.
- Supervisor monitors reports for improvement.

If a scorer's inter-rater reliability and/or validity statistics fall below the expected rate (see Table 5.1), scoring supervisors increase backreading on the scorer. If a scorer has low backreading agreement, an intervention log is opened for that scorer. This log provides documentation of the steps taken to retrain the scorer and is signed by the scorer. The scoring director determines whether the same issue or trend is being experienced by several scorers and determines the need for a calibration set.

General Calibration

Calibration sets are administered as project leadership deems necessary. Calibration provides a way to proactively promote accuracy by exploring project- or item-specific issues, score boundaries, or types of responses particularly challenging to score consistently. Scorers who miss two consecutive days must be retrained before they can return to scoring. Scorers who fall below acceptable standards are retrained a maximum of two times before being dismissed from the project. General calibration sets consist of 2-3 papers, address a single issue, and be administered online. General calibration responses are approved by MSDE. If an item spans the weekend during scoring a Monday calibration is given.

Validity

Pre-scored validation responses are used to verify that scorers are applying the same standards throughout the project, and we watch for early indications of reader drift from the standards. Validity is presented blind; scorers cannot distinguish them from live responses. Validity papers are prepared by item and administered on a regular schedule (at least 1 percent of responses). Validity papers are interspersed with and indistinguishable from unscored student responses.

Inter-Rater Reliability

This reliability statistic allows scoring leadership to monitor individual and group scoring agreement. The statistic reflects a level of agreement between two scorers' scores to the same student response. Monitoring allows scoring supervisory staff can target individuals for increased backreading, feedback, and—if necessary—retraining. Readers with less than expected IRR (see Table 5.1) are monitored closely and their work is backread at a higher rate.

Frequency Distribution

The number or percentage of scores assigned at each score point of a given rubric. This is calculated at the scorer and item levels. Anomalous scoring trends are evaluated in conjunction with validity and other statistics which allow for intervention as needed with the individuals involved to ensure that individual drift has not occurred. Frequency distribution reports are monitored and available to MSDE. Items not performing as anticipated can result in further investigation or intervention.

Validity Reports

Validity reports are used to identify struggling scorers (scorer below the validity requirement and/or significantly below the room average) or room drift (as a group, the scorers are scoring an item incorrectly or inconsistently). These reports are also used to determine whether a scorer is misunderstanding a particular issue. An extension of the validity process whereby select validity responses are annotated and used to provide feedback to scorers. If a validity response is scored incorrectly, it subsequently appears on the scorer's screen with the true score, the score they assigned, and an annotation explaining the true score. In this way, this quality monitoring tool serves an immediate, valuable secondary function: that of automated real-time feedback.

If struggling scorers or room drift is identified, scoring directors and scoring supervisors will follow the same procedure described in backreading. All reports are monitored daily by the scoring director(s), the content specialist, and the project manager.

Automated Scoring

As noted, automated scoring performed by Pearson's Intelligent Essay Assessor (IEA) was the default option for scoring the MISA CR items in spring 2019. For 10 percent of responses, a second "reliability" score was assigned. The purpose of the reliability score was to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. All reliability scoring was done by human scorers.

Continuous Flow

Continuous flow scoring results in an integrated connection between human scoring and automated scoring. It refers to a system of scoring where either an automated score, a human score, or both can be assigned based on a predetermined asynchronous operational flow.

Smart Routing

Smart routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores. Smart routing can be applied item by item to the extent needed to meet scoring quality criteria for automated scoring.

Confidence Level

When a response is to be smart routed to human scorers, the automated score is marked with a low confidence flag. Otherwise, the automated score is marked as high confidence.

Quality Criteria for Evaluating Automated Scoring

A variety of measures of inter-rater agreement for evaluating automated scoring have been proposed based on the research literature (Williamson, Xi, and Breyer, 2012). These measures are utilized in Pearson's automated scoring research and include Pearson correlation, quadratic-weighted kappa, exact agreement, and standardized mean difference. These measures are computed between pairs of human scores, as well as between IEA and humans, to evaluate how performance was the same or different. Criteria for evaluating the training of IEA given these measures include the following:

- Pearson correlation between IEA-human should be within 0.1 of human-human.
- Quadratic-weighted kappa between IEA-human should be within 0.1 of human-human.
- Standardized mean difference between IEA-human should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria and are noted below.

1. Primary Criteria –IEA-Human exact agreement is within 5.00 percent of Human-Human exact agreement for each trait score.
2. Secondary Criteria - Based on the training responses: With smart routing applied as needed, IEA-human differences on statistical measures for each trait score are within the Williamson et al. tolerances for subgroups with at least 50 responses.

Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the following hierarchy determines which score was reported operationally:

- The IEA score is reported if it is the only score assigned;
- If an IEA score and a human score are assigned, the human score is reported;
- If two human scores are assigned, the first human score is reported;
- If a back read score and human and/or IEA scores are assigned, the back read score is reported;
- If a resolution score is assigned and an adjudicated score is not assigned, the resolution score is reported (note that if nonadjacent scores are encountered, responses are automatically routed to resolution);

- If an adjudicated score is assigned, it is reported (note that if a resolution score is nonadjacent to the other scores assigned, responses are automated routed to adjudication).

Sampling Responses Used for Training IEA

The responses used for training IEA came from prompts administered online as part of the 2017 census field test and 2018 embedded field test administrations. IEA was trained and evaluated prior to the start of operational scoring in 2019.

Criteria for Evaluating IEA Performance

IEA performance on each prompt was evaluated based on IEA-human exact agreement and compared to agreement based on responses that were double-scored by humans. A portion of the data was held out for evaluating IEA-human exact agreement according to the following steps:

1. Calculate agreement of the IEA scores with the human scores.
2. Compare the IEA-human agreement with the human-human agreement.
3. If the IEA-human agreement is within 5.00 percent of the human-human agreement and 50 percent IEA-human agreement by score point (i.e., conditioned on the human score), IEA can be deployed operationally.

Chapter 6: Classical Item Analysis

This section describes the results of the classical item analysis conducted for data obtained from the MISA 2019 operational administration. A set of classical item statistics were computed for each item. The following statistics and associated flagging rules were used to identify items that were not performing as expected. Appendix C presents classical item analysis summaries for the MISA 2019 operational test.

Classical Item Difficulty Indices (P-Value and Average Item Score)

Item difficulty offers an index of how easy or hard a given test question is to answer correctly or to earn a given score point for items scored according to a rubric. Item difficulty statistics are used by test developers to help construct test forms that contain a range of items from easy to hard. For items that appear to be unexpectedly difficult, this may indicate students' lack of familiarity with the item type or students' limited opportunity to learn the content represented in the item and are worth further review.

For dichotomously scored items (items scored correct or incorrect), item difficulty is indicated by its p-value, which is the proportion of test takers who answered that item correctly. The range for p-values is from .00 to 1.00. Items with high p-values are easy items and those with low p-values are difficult items. Dichotomously scored items were flagged for further review if the p-value was above .90 (i.e., too easy) or below .20 (i.e., too difficult).

For polytomously scored items (items scored according to a rubric with multiple points awarded), difficulty is indicated by the average item score (AIS). The AIS can range from .00 to the maximum total possible points for an item. To facilitate interpretation, the AIS values for polytomously scored items are often expressed as percentages of the maximum possible score, which are equivalent to the p-values of dichotomously scored items. The desired p-value range for polytomously scored items is also .20 to .90; items with values outside this range were flagged for review.

Item-Total Score Correlation

This statistic describes the relationship between test takers' performance on a specific item and their performance on the total test. The item-total correlation is usually referred to as the item discrimination index. For MISA item analysis, the total score on the assessment was used as the total test score. The point-biserial correlation was calculated for both selected response items and constructed response items as an estimate of the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson, Drasgow, and Dorans, 1982). Item-total correlations can range from -1.00 to 1.00. Desired values are positive and larger than .10. Negative item-total correlations indicate that low ability test takers perform better on an item than high ability test takers, an indication that the item may be potentially flawed.

The Percentage of Students Choosing each Response Option or Earning each Score Point

Selected response items refer primarily to single-select multiple-choice items. These items require that the test taker select a single response from a number of answer options (four in the case of MISA). These statistics for single-select multiple-choice items indicate the percentage of students who select each of the answer options. Also included are the percentage of students that omit the item. These statistics give an indication of whether the items are functioning well as a whole. Anomalies can indicate problems with item functioning, such as multiple correct answers or non-functioning distractors.

Constructed response items are scored according to rubrics in determining the number of points to award a given response. For these items, the statistics indicate the percentage of students who earn each possible score point. The percentage of students omitting the items are also indicated.

Differential Item Functioning

Differential item functioning (DIF) analyses were conducted using the data obtained from the MISA operational tests. If an item performs differentially across identifiable subgroups (e.g., gender or ethnicity) when students are matched on ability, this may indicate an issue with fairness or that the item may be measuring something other than the intended construct (i.e., possible evidence of DIF). It is important, however, to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify potential biases. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

This section provides information about differential item functioning (DIF) analyses used for the 2019 MISA operational tests. The reference group was either male or Caucasian students, and the focal group was either female, African-American students, or Hispanic students. Appendix D presents DIF results for items appearing on the MISA 2019 tests.

The Mantel-Haenszel (MH) DIF statistic was calculated for selected-response items and for dichotomously-scored constructed-response items. The Mantel-Haenszel chi-square statistic is computed as

$$MH - \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k Var(F_k)},$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable (Zwick, Donoghue, & Grima 1993). Note that the MH statistic is sensitive to N such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the MH delta statistic (ΔMH) was computed. Educational Testing Service (ETS) first developed the ΔMH DIF statistic. To compute the ΔMH DIF, the MH alpha (the odds ratio) is first computed

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k},$$

Where N_{r1k} is the number of correct responses in the reference group at ability level k, N_{f0k} is the number of incorrect responses in the focal group at ability level k, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level k, and N_{r0k} is the number of incorrect responses in the reference group at ability level k. The ΔMH DIF is computed as

$$\Delta MH\ DIF = -2.35\ln(\alpha_{MH}).$$

Positive values of ΔMH DIF indicate items that favor the focal group whereas negative values of ΔMH DIF indicate items that favor the reference group.

For polytomously scored constructed-response items, the standardized mean difference (SMD) (Dorans & Schmitt, 1991; Zwick, Thayer & Mazzeo, 1997; Dorans, 2013), in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959), is used to identify items with DIF. This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the values of the matching variable.

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

where

$P_{Fk} = \frac{n_{F+k}}{n_{F++}}$, the proportion of the focal group members who are at the k^{th} level of the matching variable,

$m_{Fk} = \frac{1}{n_{F+k}} x(\sum_t y_t n_{Ftk})$, the mean item score of the focal group members at the k^{th} level, and

m_{Rk} = 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 for the reference group are applied to make the weighted number of the reference group students the same as in the focal group within the same ability.

Classification of DIF statistics

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C (Zieky, 1993). Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values. Positive values indicate that, conditional on the total score, the focal group has a higher mean item score than the reference group. In contrast, negative DIF values indicate that, conditional on the total test score, the focal group has a lower mean item score than the reference group. The flagging criteria for dichotomously scored items are presented in Table 6.1; the flagging criteria for polytomously scored constructed response items are provided in Table 6.2.

Table 6.1. DIF Categories for Dichotomous Selected Response and Constructed Response Items

DIF Category	Criteria
A (negligible)	The Mantel Chi-square is not significantly different from zero, or the absolute value of ΔMH DIF is less than one.
B (slight to moderate)	1. The Mantel Chi-square is significantly different from zero but not from one, and the absolute value of ΔMH DIF is at least one; OR 2. The Mantel Chi-square is significantly different from one, but the absolute value of ΔMH DIF is less than 1.5. Positive values are classified as “B+” and negative values as “B-”.
C (moderate to large)	The Mantel Chi-square is significantly different from one, and the absolute value of ΔMH DIF is at least 1.5. Positive values are classified as “C+” and negative values as “C-”.

Table 6.2. DIF Categories for Polytomous Constructed Response Item

DIF Category	Criteria
A (negligible)	Mantel Chi-square *p value* > 0.05 or $\lvert SMD/SD \rvert \le 0.17$
B (slight to moderate)	Mantel Chi-square *p value* < 0.05 and $\lvert SMD/SD \rvert > 0.17$
C (moderate to large)	Mantel Chi-square *p value* < 0.05 and $\lvert SMD/SD \rvert > 0.25$

Note: SMD =Standardized Mean Difference; SD – total group standard deviation of item score.

Flagging Items for DIF

Items are flagged into one of three categories based on the magnitude of their DIF statistics:

- Category A: no or negligible DIF
- Category B: slight or moderate DIF, and
- Category C: moderate to large values of DIF. These items which exhibit significant DIF, are of primary concern.

Chapter 7: Calibration, Scaling, and Equating

This section describes calibration, scaling, and equating procedures that took place for the Spring 2019 MISA operational administration. As this administration marks the second operational administration of MISA, these procedures include calibration of operational forms and equating to place 2019 onto the base 2018 scales.

Measurement Models

The Rasch model (Rasch, 1980) and its polytomous extension, the Partial Credit model (PCM) (Masters, 1982) are the item response theory models used to develop and calibrate the 2019 operational MISA assessments. These measurement models are regularly used to construct test forms, for scaling and equating, and to develop and maintain large item banks in large scale K-12 testing programs. The PCM reduces to the Rasch model for items with only two response categories, such as multiple-choice items. For an item involving m_i score categories, the general expression for the probability of scoring x on item i is given by:

$$P_{xi} = \exp\sum_{j=0}^{x}(\theta - D_{ij}) / \sum_{k=0}^{m_i}\left[\exp\sum_{j=0}^{k}(\theta - D_{ij})\right]$$

where $x = 0, 1, ..., m_i$, and by definition, $\sum_{j=0}^{0}(\theta - D_{ij}) = 0$.

The above equation gives the probability of scoring x on the i-th test item as a function of ability (θ) and the difficulty (D_{ij}) of the m_i steps of the task. According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between θ and D_{ij} of all the completed steps, divided by the sum of the differences of all the steps of a task.

Operational MISA items for each respective grade were calibrated according to the Rasch and PCM concurrently and can be found in Appendix E. The following information is provided:

- Item type
- Rasch item difficulty estimate (D_i)
- Conditional standard error of Rasch item difficulty estimate
- Mean-square infit
- Mean-square outfit
- Rasch step difficulty estimate (or structure calibration estimate, F_{ij})

The following formula shows how structure measure estimate (D_{ij}) is calculated from both D_i and F_{ij} directly obtained from a run of Winsteps:

$$D_{ij} = D_i + F_{ij},$$

where D_{ij} = structure measure estimate
D_i = item difficulty estimate,
F_{ij} = structure calibration estimate (i.e., step difficulty estimate).

Finally, the following formulas show how conditional standard error (SE) of item difficulty estimate (D_i) and structure measure estimate (F_{ij}) were driven (Wright & Masters, 1982):

$$\mathrm{SE}(D_i) = 1/\sqrt{\sum_{n=1}^{N}[\sum_{k}^{m_i} k^2 p_{nik} - (\sum_{k}^{m_i} kp_{nik})^2]}$$

$$\mathrm{SE}(F_{ij}) = 1/\sqrt{\sum_{n=1}^{N}(\sum_{k=0}^{j} p_{nik} - (\sum_{k=j+1}^{m_i} p_{nik})^2)}$$

where $P_{nix} = \exp\sum_{j=0}^{x}(\theta_n - D_{ij}) / \sum_{k=0}^{m_i}\left[\exp\sum_{j=0}^{k}(\theta_n - D_{ij})\right]$

$x = 0, 1, ..., m_i$, and

$k = 1, 2, …., m_i$.

Fit Statistics for the Rasch Model

Fit statistics are used for evaluating the goodness-of-fit of a model to the data. Fit statistics are calculated by comparing the observed and expected trace lines obtained for an item after parameter estimates are obtained using a particular model. *WINSTEPS* provides two kinds of fit statistics called mean-squares that show the size of the randomness or amount of distortion of the measurement system.

Outfit mean-squares are influenced by outliers and are usually easy to diagnose and remedy. Infit mean-squares, on the other hand, are influenced by response patterns and are harder to diagnose and remedy. 7.1 provides a guideline for evaluating mean-square fit statistics (Linacre & Wright, 2000).

In general, mean-squares near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmodeled noise, model underfit).

Table 7.1. Criteria to Evaluate Mean-Square Fit Statistics

Mean-Square	Interpretation
> 2.0	Distorts or degrades the measurement system
1.5 – 2.0	Unproductive for construction of measurement, but not degraded
0.5 – 1.5	Productive for measurement
< 0.5	Unproductive for measurement, but not degrading. May produce misleadingly good reliabilities and separations

Calibration

As noted, the Rasch family of item response theory models were used to establish the operational base scales for MISA at grades 5 and 8. Each of the two core forms per grade level share a set of common item sets (described in Chapter 2). This common linkage was used to conduct a concurrent calibration of all MISA operational items by grade using a single *WINSTEPS* run (WINSTEPS version 3.91; Linacre 2015). *WINSTEPS* uses joint maximum likelihood estimation (JMLE) as described by Wright and Masters (1982) for determining item parameter estimates.

Equating

The 2019 MISA assessments were placed on the 2018 base scales through a non-equivalent groups anchor test design (NEAT). The items that comprised the anchors in each instance were three sets of items from the spring 2018 operational administration – one per domain (Physical, Life, Earth and Space Science). This reflected one-third of the operational test. Sets were chosen to best reflect the make-up and content and statistical targets (average Rasch difficulty). Items were chosen to be best possible quality and were administered in the identical positions they were in 2018. These anchor sets served as linkages both to the 2018 base scale as well as within year, across both core forms.

The NEAT fixed parameter equating was implemented within *WINSTEPS* in order to link to the 2018 operational reporting scale. This was carried out by constraining the spring parameter estimates for the anchor items to equal the final parameter estimates obtained in the original 2018 MISA calibration analysis. As part of this process, a stability check of the anchor item functioning was conducted using a robust *z* procedure (Huynh and Meyer, 2010; Huynh and Rawls, 2009; South Carolina Department of Education, 2001).

The robust *z* statistic is calculated using the following formula for each anchor item:

$$z = \frac{[(b_{iF} - b_{iE}) - M_d]}{IQR(0.74)}$$

Where b_{iF} is the fixed Rasch difficulty value (banked 2018 value), b_{iE} is the estimated Rasch difficulty for that item (freely calibrated 2019 value), M_d is the median difference of all anchor items, and IQR is the interquartile range of the differences for all anchor items.

Values greater than the absolute value of the robust *z* statistic greater than 1.645 are flagged for consideration of removal from a final anchor set. If the ratio of the standard deviations of the Rasch difficulties is between 0.9 and 1.1 in addition to the correlation of the Rasch difficulties being greater than 0.95, the set is considered stable. Until this is met, flagged items are removed one at a time beginning with the largest robust *z* value.

The robust *z* procedure was conducted for MISA 2019 assessments. Given that the anchor sets contained constructed response items as well, this comparison was also applied to the step difficulties. It should be noted that while there were items flagged by the procedure for potential removal from the final anchor sets, the full sets met the *sd* ratio and correlation criteria in all instances and were kept intact. That is, the final NEAT equating was based on all anchor items.

Scaling

Creation of the MISA base 2018 reporting scale scores follow the scaling approach established for the Partnership for Assessment of Readiness for College and Careers (PARCC). For each test, scale scores are linear transformations of the underlying IRT-based (theta) metric where the lowest and highest obtainable scale scores (LOSS and HOSS) are set to 650 and 850 respectively. The following linear transformation was used for transforming the underlying Rasch theta scales to the final operational MISA scales:

$$SS_{MISA} = m(\theta - \theta_{met}) + b$$

where the slope (m) is set to 15.5, the intercept (b) is set to 750, θ is the person ability estimate and θ_{met} is the cut point for the *Met Expectations* MISA performance level on the ability metric (denoted in Table 7.2 as the 3|4 θ_{met}). Table 7.2 summarizes the scaling constants used for MISA grades 5 and 8 scale score reporting. Note that all scale scores are rounded to the nearest whole number. Table 7.2 also presents all of the MISA cut scores on the IRT (θ) metric as a result of the standard setting held in summer of 2018 (see Chapter 8).

Table 7.2. MISA Scaling Values

Grade	Slope (m)	Intercept (b)	LOSS	HOSS	Theta Cut Scores		
					2|3	3|4 (θmet)	4|5
5	15.5	750	650	850	-1.2729	0.0763	1.4870
8	15.5	750	650	850	-1.3236	0.1933	1.6716

Note: MISA reports only four performance levels in line with PARCC levels 2 through 5

Chapter 8: Student Scores, Performance Standards, and Student Performance

Score Interpretation

To help provide appropriate interpretation of the 2019 MISA operational test scores, two types of scores were created: scale scores and performance levels with descriptions. As presented in the previous chapter, it was decided that the MISA reporting scales would utilize a similar approach to reporting MISA scale scores as has been used on PARCC. That is, scale scores are reported on a scale ranging from 650 to 850, with 750 designating Met Expectations. Alignment to PARCC is seen as valuable due to familiarity by stakeholders with respect to score reporting and interpretation of mathematics and ELA results.

In addition to the use of scale scores for reporting results, MISA also reports on performance levels. These too utilize the same PARCC framework. However, it was decided that MISA would not include the Level 1 PARCC classification, *Did Not Yet Meet Expectations*. Instead, the lowest reporting category for MISA is defined as (Level 2) *Partially Met Expectations*. Thus, MISA will report on four performance levels labelled according to PARCC levels 2 through 5:

- Level 2: *Partially Met Expectations*
- Level 3: *Approaching Expectations*
- Level 4: *Met Expectations*
- Level 5: *Exceeded Expectations*

Scale Scores

As explained in the proceeding section, the 2019 MISA assessments yield scale scores that range between 650 and 850. As a result of calibration and scaling, the scale scores from the two base forms are comparable within the same grade, but not across grade levels. Generally, the only inferences that can be appropriately drawn from scale scores are that higher scale scores represent higher performance on the MISA tests.

Performance Levels and Descriptions

The MISA tests were designed as criterion referenced tests in that they offer indicators of student performance in relation to a set of performance descriptions premised on the Next Generation Science Standards. Performance level descriptions (PLDs) describe what students at each of the four levels generally know and can do. The determination of what MISA scale score values reflect each of the thresholds between performance levels was determined in the summer of 2018 as a result of standard setting. A description of this process can be found within a memo from the State Superintendent of Schools to the State Board of Education at

http://www.marylandpublicschools.org/stateboard/Documents/10232018/TabH-ScienceStandards.pdf.

Table 8.1 provides scale score ranges for each of the MISA performance levels by grade.

Table 8.1. MISA Scale Score Ranges by Performance Level and Grade

Grade	Performance Level	Scale Score Range
5	Partially Met Expectations	650-728
	Approaching Expectations	729-749
	Met Expectations	750-771
	Exceeded Expectations	772-850
8	Partially Met Expectations	650-725
	Approaching Expectations	726-749
	Met Expectations	750-772
	Exceeded Expectations	773-850

Dimensions of Science Indicators

Lastly, students receive information around mastery of several dimensions of science. These include the three content-based domains of Physical, Life, and Earth and Space Science. For each grade level two MISA Evidence Families (see Chapter 2) were also reported on per grade. For 2019 the two Evidence Families for grade 5 were *Data and Information* and *Model Components, Relationships, and Connections*. The two Evidence Families for grade 8 were *Reasoning* and *Model Components, Relationships, and Connections*. Performance for each of these dimensions is reported using categories rather than scale scores or performance levels. The three categories are: *Met or Exceeded Expectations*, *Approached Expectations*, and *Did Not Yet Meet Expectations*. Performance on these respective indicators is based on student performance on the subset of items associated with each dimension and using each respective derived Rasch parameter estimate described in Chapter 7.

Student Performance

Tables 8.2 and 8.3 present performance information for grades 5 and 8 of the 2019 operational MISA administration. Results are presented overall for mean scale score and percentage of students being classified into each of the performance levels. Additionally, results are also broken out by subgroup. Appendix F presents performance level results for each of the dimensions of science indicators noted in the previous section by grade.

Table 8.2. 2019 MISA Grade 5 Scale Score and Performance Level Summary Results

	Scale Scores			% Within Each Performance Level			
Group	**N**	**Mean**	**SD**	**PL2**	**PL3**	**PL4**	**PL5**
Overall	69859	739.55	17.18	28	43	27	3
Female	34104	740.20	16.72	26	45	27	3
Male	35755	738.94	17.59	30	41	27	2
Hispanic\Latino	12935	732.87	15.95	41	44	14	1
Not-Hispanic\Latino	56924	741.07	17.09	25	43	29	3
Asian	4748	750.90	16.46	10	34	48	9
American Indian or Alaskan Native	2592	732.70	13.96	40	49	11	0
Black or African American	24324	732.37	14.85	42	45	13	0
Native Hawaiian or Other Pacific Islander	296	736.99	15.71	30	48	21	1
Multiple Indication	4134	741.98	16.18	22	46	29	3
White	33765	743.39	17.05	20	42	34	4
Economic Disadvantage	31487	731.95	14.90	43	45	12	0
Students with Disability	10991	729.19	16.76	54	32	12	1

Note: PL2 = *Partially Met Expectations*; PL3 = *Approaching Expectations*; PL4 = *Met Expectations*; PL5 = *Exceeded Expectations*

Table 8.3. 2019 MISA Grade 8 Scale Score and Performance Level Summary Results

	Scale Scores			% Within Each Performance Level			
Group	**N**	**Mean**	**SD**	**PL2**	**PL3**	**PL4**	**PL5**
Overall	64357	743.08	18.62	19	43	33	5
Female	31650	744.14	18.17	16	45	34	6
Male	32707	742.05	18.98	21	42	32	5
Hispanic\Latino	11137	735.74	17.05	29	49	20	1
Not-Hispanic\Latino	53220	744.61	18.56	17	42	35	6
Asian	4588	756.16	17.61	5	27	50	17
American Indian or Alaskan Native	2537	736.62	14.84	23	57	19	1
Black or African American	21998	734.43	15.93	30	52	17	1
Native Hawaiian or Other Pacific Islander	486	737.16	16.73	26	50	22	1
Multiple Indication	3883	745.02	17.64	14	45	36	6
White	30865	747.68	18.16	13	38	42	7
Economic Disadvantage	26003	734.02	15.95	31	52	17	1
Students with Disability	10093	731.89	17.69	42	41	16	2

Note: PL2 = *Partially Met Expectations*; PL3 = *Approaching Expectations*; PL4 = *Met Expectations*; PL5 = *Exceeded Expectations*

Chapter 9: Reliability and Validity

Reliability

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 estimates of reliability reported here are internal consistency measures, which are derived from analysis of the consistency of the performance of individuals on items within a test (internal consistency reliability). Therefore, they apply only to the test form being analyzed.

Internal Consistency

The equation displayed below is the formula for the most common index of reliability, namely, Cronbach's coefficient alpha (α ; Cronbach, 1951). In this formula, the s_i^2's denote the variances for the *k* individual items; s_{sum}^2 denotes the variance for the sum of all items.

$$s_{sum}^2 = \frac{k}{k-1} \times \left(1 - \frac{\sum_{i=1}^{k} S_i^2}{S_{sum}^2} \right)$$

Standard Error of Measurement

The standard error of measurement (SEM) is commonly used in interpreting and reporting individual test scores and score differences on tests (Harvill, 1991). Classical test theory is based on the following assumptions (Andrich & Luo, 2004):

- Each person *v* has a true score on the construct, usually denoted by the variable T_v
- The best overall indicator of the person's true score is the sum of the scores on the items and is usually denoted by the variable X_v
- This observed score will have an error for each person which is usually denoted by E_v
- These errors are not correlated with the true score
- Across a population of people, the errors sum to 0 and they are normally distributed.

The SEM is calculated by the following formula:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_x}$$

Coefficient alpha and SEM were calculated by core form for grade 5 and grade 8 as shown in Table 9.1. Across all forms, the overall reliabilities for each respective core test were roughly .93. Reliabilities were also presented for the science domains and dimension of science indices. Domain score reliabilities were generally high, averaging roughly .80.

Table 9.1. MISA Coefficient Alpha and SEM Overall and by Domain and Evidence Family

Grade	Core	Domain/Evidence Family	N	Coefficient Alpha	SEM
05	1	Overall	29685	0.920	4.758
		Physical Science	29685	0.715	3.715
		Life Science	29685	0.845	3.559
		Earth and Space Science	29685	0.810	3.524
		Data and Information	29685	0.785	4.055
		Model Components, Relationships and Connections	29685	0.839	2.967
05	2	Overall	29551	0.911	4.574
		Physical Science	29551	0.677	3.567
		Life Science	29551	0.817	3.470
		Earth and Space Science	29551	0.808	3.432
		Data and Information	29551	0.788	3.785
		Model Components, Relationships and Connections	29551	0.745	3.513
08	1	Overall	29985	0.937	4.627
		Physical Science	29985	0.834	3.463
		Life Science	29985	0.847	3.490
		Earth and Space Science	29985	0.817	3.631
		Reasoning	29985	0.739	5.537
		Model Components, Relationships and Connections	29985	0.817	3.746
08	2	Overall	29789	0.932	4.678
		Physical Science	29789	0.832	3.544
		Life Science	29789	0.794	3.673
		Earth and Space Science	29789	0.832	3.442
		Reasoning	29789	0.716	5.887
		Model Components, Relationships and Connections	29789	0.777	4.082

Appendix G provides coefficient alpha and SEM breakdowns by core form and subgroup for total test as well as domain and dimensions of science by grade.

Reliability of Classifications Accuracy and Consistency

Reliability of classification estimates the proportion of students who are accurately classified into proficiency levels. There are two kinds of classification reliability statistics provided here: *decision accuracy* and *decision consistency*. The reliability of the classifications for MISA were determined using the computer program BB-CLASS (Brennan, 2004), which operationalizes a statistical method developed by Livingston and Lewis (1995). This approach uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficients) to estimate two kinds of statistics, *decision accuracy* and *decision consistency* (Livingston and Lewis, 1993, 1995). Decision accuracy refers to the extent to which the classifications of examinees based on their scores on a given form agree with the classifications made on the basis of the classifications that would be made if the test scores were perfectly reliable. Decision consistency refers to the agreement between these classifications based on two non-overlapping, but equally difficult forms of a test.

Decision consistency values are always lower than the corresponding decision accuracy values, because in decision consistency, both of the classifications of the student are based on scores that depend on which form of the test the student took. In decision accuracy, only one of the classifications is based on a score that can vary in this way. It is not possible to know which specific students are accurately classified. But it is possible to estimate the *proportion* of students who were accurately classified. Similarly, it is not possible to know which students are consistently classified if they were retested using another form, but it is possible to estimate the proportion of the students who would be consistently classified.

Table 9.2 provides information about the accuracy and the consistency of two classifications made on the basis of the scores on the grades 5 and 8 MISA assessments (by core form). The columns labeled as *Exact Level* provide the classification of the student into one of four MISA achievement levels. The columns labeled as *Level 4 or Higher vs. 3 or Lower* provide the classification of the student as being either in one of the upper two levels (Levels 4 and 5) or in one of the lower two levels (Levels 2 or 3).

The table shows that for classifying each student into one of the four achievement levels, the proportion accurately classified ranges from .83 to .85; the proportion who would be consistently classified on two different test forms ranges from .77 to .79. For classifying each student simply as being at Level 4 or higher vs. being at Level 3 or lower, the proportion accurately classified ranges from .92 to .93; the proportion who would be consistently classified on two different test forms ranges from .89 to .90.

Table 9.2. MISA Classification Accuracy and Consistency Results by Grade and Core Form

		Decision Accuracy		Decision Consistency	
Grade	**Core**	**Exact Level**	**Level 4 or Higher vs. 3 or Lower**	**Exact Level**	**Level 4 or Higher vs. 3 or Lower**
5	**1**	0.83	0.92	0.77	0.89
	2	0.84	0.92	0.77	0.89
8	**1**	0.85	0.93	0.79	0.90
	2	0.85	0.92	0.78	0.89

Tables 9.3 and 9.4 provide more detailed information about the accuracy and the consistency of the classification of students into the MISA proficiency levels at grades 5 and 8 respectively. Each cell in the 4-by-4 tables shows the estimated proportion of students who would be classified into a combination of proficiency levels. The sum of the five bold italicized values on the diagonal should equal the *Exact Level* of decision accuracy or consistency presented in Table 9.2. For *Level 4 and Higher vs. 3 and Lower* found in Table 9.2, the sum of the shaded values in Table 9.3 and 9.4 should equal the level of decision accuracy or consistency presented in Table 9.2. Note that the sums based on values in Table 9.3 and 9.4 may not match exactly to the values in Table 9.2 due to truncation and rounding.

Table 9.3. Grade 5 MISA Reliability of Classifications Across Levels

Core	Reliability	Scale Score Range	Level 2	Level 3	Level 4	Level 5	Category Total
1	Accuracy	650-728	**0.24**	0.03	0.00	0.00	0.28
		729-749	0.04	**0.36**	0.04	0.00	0.45
		750-771	0.00	0.03	**0.21**	0.01	0.26
		772-850	0.00	0.00	0.00	**0.02**	0.02
	Consistency	650-728	**0.23**	0.05	0.00	0.00	0.28
		729-749	0.05	**0.32**	0.05	0.00	0.43
		750-771	0.00	0.05	**0.20**	0.01	0.26
		772-850	0.00	0.00	0.01	**0.02**	0.03
2	Accuracy	650-728	**0.16**	0.03	0.00	0.00	0.19
		729-749	0.04	**0.41**	0.04	0.00	0.50
		750-771	0.00	0.04	**0.25**	0.01	0.29
		772-850	0.00	0.00	0.00	**0.02**	0.02
	Consistency	650-728	**0.15**	0.05	0.00	0.00	0.20
		729-749	0.05	**0.37**	0.06	0.00	0.48
		750-771	0.00	0.06	**0.23**	0.01	0.30
		772-850	0.00	0.00	0.01	**0.02**	0.02

Table 9.4. Grade 8 MISA Reliability of Classifications Across Levels

Core	Reliability	Scale Score Range	Level 2	Level 3	Level 4	Level 5	Category Total
1	Accuracy	650-725	**0.16**	0.02	0.00	0.00	0.18
		726-749	0.03	**0.38**	0.04	0.00	0.45
		750-772	0.00	0.03	**0.27**	0.01	0.32
		773-850	0.00	0.00	0.01	**0.04**	0.05
	Consistency	650-725	**0.15**	0.04	0.00	0.00	0.19
		726-749	0.04	**0.35**	0.06	0.00	0.44
		750-772	0.00	0.05	**0.25**	0.02	0.32
		773-850	0.00	0.00	0.02	**0.04**	0.06
2	Accuracy	650-725	**0.13**	0.02	0.00	0.00	0.15
		726-749	0.03	**0.37**	0.04	0.00	0.45
		750-772	0.00	0.04	**0.31**	0.01	0.36
		773-850	0.00	0.00	0.01	**0.04**	0.05
	Consistency	650-725	**0.12**	0.03	0.00	0.00	0.15
		726-749	0.04	**0.34**	0.06	0.00	0.44
		750-772	0.00	0.05	**0.29**	0.02	0.35
		773-850	0.00	0.00	0.02	**0.04**	0.06

Validity

The Standards for Educational and Psychological Testing, issued jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) reports:

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations (p. 11).

The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, item development, and psychometric characteristics. The 2019 MISA operational assessments provided an opportunity to gather evidence of validity based on both test content and on the internal structure of the tests.

Evidence Based on Test Content

Content validity evidence addresses whether a given assessment adequately samples from the full given domain. Where the assessment is determined to be representative in terms of the standards and in the manner intended, it is said to have high content validity. For the MISA assessments, they are designed to measure NGSS broadly and involve more complex content and synthesis of responses according to the content and three-dimensional nature of the standards.

For MISA, test design and blueprint specifications were developed in concert between Pearson and MSDE science experts well versed in NGSS. These specifications drive item and stimulus development targets intended to effectively support the intended purposes of the MISA assessment in relation to the NGSS. As noted, both the testing contractor and MSDE were directly involved in item and stimulus development. Both were developed based on the test specifications and were rigorously scrutinized during the various content reviews, which involves all members of the assessment team. These reviews examine the appropriateness of test items, difficulty, clarity, correctness of answer choices, plausibility of distractors, and fairness of the items and tasks. Then the items must be reviewed and approved by the content review committees, which assure that each item appropriately measures the intended content, is appropriate in difficult, contains only one correct (or best) answer for multiple-choice questions, and has an appropriate and complete scoring guideline for technology-enhanced items. Next, a bias and sensitivity committee must approve the items, which review the item for language, or content, that may be inappropriate or offensive to students, parents, or community members, or that contain stereotypical or biased references to gender, ethnicity, or culture. The process of the MISA test design, development, and test construction is described in chapters 2 and 3 of this report. As documented, MSDE, Pearson, and educator committees expend tremendous effort to ensure the MISA assessments are content-valid. Additionally, evidence of the content coverage is presented in Appendix A.

MSDE also developed performance level descriptors (PLDs) for MISA, which provide a description of typical grade-level performance for each level of achievement in relation to the NGSS. The PLDs are descriptions of the knowledge and skills demonstrated by students in each performance category. Higher scores translate to a greater level of knowledge and skills demonstrated. There is a link between the PLDs and the knowledge and skills required to meet proficiency according to the standards. PLDs are used to relate performance on MISA to the NGSS through the process of standard setting. Content experts and stakeholders participated in a standard setting for MISA in August of 2018. This committee set the cut scores that delineate the four levels of science achievement at grades 5 and 8 as reported in Maryland (Partially Met, Approaching, Met, and Exceeded Expectations). Evidence of these activities is presented in the context of student performance on MISA (Chapter 8) and includes a link to the MISA standard setting report.

Also important for content validity is the control of random measurement error. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Reliability and the standard error of measurement (SEM) are discussed earlier in this chapter. The section presents tables reporting the SEM and the coefficient alpha reliabilities by core form and grade overall and broken down by demographic groups. These measures show the MISA tests to be reliable.

Evidence Based on Response Process

Validity evidence based on response processes involves explicit assumptions about the cognitive processes engaged in by the test takers. “Depth of knowledge” (DOK), or cognitive complexity, refers to the cognitive demand associated with interacting with a given item. The level of cognitive demand focuses on the type and level of thinking and reasoning required of the student. Levels of cognitive complexity for MISA are based on Norman L. Webb’s (Webb, 1999) DOK levels.

A Level 1 (recall) item requires the recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula. A well-defined and straight algorithmic procedure is considered to be at this level. A Level 1 item specifies the operation or method of solution and the student is required to carry it out.

A Level 2 (skill/concept) item calls for the engagement of some mental processing beyond a habitual response, with students required to make some decisions as to how to approach a problem or activity. Interpreting information from a simple graph and requiring reading information from the graph is a Level 2. An item that requires students to choose the operation or method of solution and then solve the problem is a Level 2. Level 2 items are often similar to examples used in textbooks.

Level 3 (strategic thinking) items require students to reason, plan, or use evidence to solve the problem. In most instances, requiring students to explain their thinking is a Level 3. A Level 3 item may be solved using routine skills, but the student is not cued or prompted as to which skills to use.

Level 4 (extended thinking) items require complex reasoning, planning, developing, and thinking, most likely over an extended period of time. Level 4 items are best assessed in the classroom, where the constraints of standardized testing are not a factor.

In line with the nature of NGSS, items developed and appearing on the MISA assessments only address DOK levels 2 through 4 with most items characterized as levels 2 and 3.

Evidence Based on Internal Structure

Internal structure evidence shows the degree to which items and test components conform to the construct on which the proposed test score interpretations are based (AERA, APA, and the NCME, 2014). For example, MISA reports overall science scale scores for individual students and also reports scores based on the science domains and on two evidence families at each grade level. Internal structure validity evidence identifies the degree to which the item relationships conform to the overall scores and individual subscales.

While the NGSS are presented as reflective of several interwoven components that address multiple dimensions, MISA test questions and sets are designed around scientific phenomena and crafted to be reflective collectively of the standards. While individual items may each measure multiple elements of the standards and dimensions, they are crafted without dependencies on other items. As such, the tests are designed to be unidimensional and to measure overall NGSS primarily. Assuming this holds true then it is appropriate to apply a unidimensional IRT model for calibrating and scaling the MISA assessments.

One commonly used approach to evaluating this, factor analysis, was used to evaluate each core MISA form to determine the extent to which a they indicate a single dominant factor is present. To check the unidimensionality of the 2019 MISA assessments, we examined the relative sizes of the eigenvalues associated with a principal components analysis (PCA) of the items comprising each respective core MISA form. The first and the second principal component eigenvalues were compared without rotation.

A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 in this analysis because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis within an IRT framework (Loehlin, 1987; Orlando, 2004). Table 9.5 summarizes the results of the first and second principal component eigenvalues of the 2019 MISA assessments. Here, the first eigenvalue is substantially larger than the second in all instances and indicative of essential unidimensionality.

Table 9.5 MISA PCA First and Second Eigenvalues

Grade	Core	First Eigenvalue	Second Eigenvalue
5	1	11.09	1.78
	2	10.31	1.34
8	1	13.07	1.45
	2	12.37	1.48

Model-data fit based on the Rasch model calibrations are also indicators of unidimensionality. That is, the model assumes unidimensionality as a necessary condition supporting its application. To the extent that indicators of fit suggest data do *not* appropriately fit the model as applied may be the result of multidimensionality. Discussion of model fit is presented in Chapter 7 with Rasch Infit and Outfit statistics for all MISA operational items presented in Appendix E. These statistics support the overall fit of MISA items to the Rasch model.

Lastly, correlations among the total test scores, domains, and evidence family scores offer additional evidence of the internal structure of the 2019 MISA assessments. These correlations quantify the strength of the relationships across scores. For these, the overall scale scores were compared to each respective scale score used to derive the performance indicators for domain and evidence family scores (only overall scale scores are reported). Tables 9.6 and 9.7 respectively present these correlations for each grade and by core form. The domain and evidence family scores are moderately to highly related to one another and to the total test score. Additionally, they are generally consistent across core forms, suggesting the internal structure is comparable.

Table 9.6 Grade 5 MISA Correlations of Overall and Sub-Claim Scores

Core		Life	Earth & Space	Physical	Data and Info	Models	Total Test
1	**Life**	1.00					
	Earth and Space	0.81	1.00				
	Physical	0.75	0.74	1.00			
	Data and Info	0.72	0.97	0.72	1.00		
	Models	0.92	0.81	0.93	0.80	1.00	
	Total Test	0.94	0.93	089	0.91	0.97	1.00
2	**Life**	1.00					
	Earth and Space	0.80	1.00				
	Physical	0.73	0.73	1.00			
	Data and Info	0.79	0.97	0.73	1.00		
	Models	0.85	0.77	0.90	0.75	1.00	
	Total Test	0.93	0.93	0.88	0.92	0.91	1.00

Table 9.7 Grade 8 MISA Correlations of Overall and Sub-Claim Scores

Core		Life	Earth & Space	Physical	Reasoning	Models	Total Test
1	**Life**	1.00					
	Earth and Space	0.82	1.00				
	Physical	0.83	0.80	1.00			
	Reasoning	0.84	0.83	0.82	1.00		
	Models	0.90	0.83	0.89	0.78	1.00	
	Total Test	0.95	0.93	0.94	0.89	0.93	1.00
2	**Life**	1.00					
	Earth and Space	0.80	1.00				
	Physical	0.81	0.83	1.00			
	Reasoning	0.79	0.80	0.83	1.00		
	Models	0.85	0.85	0.86	0.71	1.00	
	Total Test	0.93	0.94	0.94	0.82	0.91	1.00

Evidence Based on Different Student Populations

In addition, internal structure evidence should show that individual items are functioning similarly for different demographic subgroups within the population being measured. MISA assessments are developed to assess NGSS and administered to all students irrespective of any particular demographic characteristic (as described in Chapters 2 and 4). Great care has been taken to ensure the items on MISA are fair and representative of the content domain expressed in the content standards. Special attention is given to find evidence that construct-irrelevant content has not been inadvertently included in the test, as such content could result in an unfair advantage for one group versus another.

This begins with item writers trained on how to avoid economic, regional, cultural, and ethnic biases when writing items. After items have been written, they are reviewed by a bias and sensitivity committee, which evaluates each item to identify language or content that might be inappropriate or offensive to students, parents, or other community members or that contain stereotypical or biased references to gender, ethnic, or cultural groups. The bias and sensitivity committee accepts, edits, or rejects each item for use prior to the items' administration.

Differential item functioning (DIF) analyses are conducted for the purpose of identifying items that are differentially difficult for different subpopulations of individuals. Chapter 6 details the methodology used to evaluate DIF for MISA items. Though DIF analyses flag items as being differentially difficult for one group as compared to another, it does not solely provide sufficient evidence for removing the item from use. Flagged items are re-examined post administration for any potentially overlooked biases attributable to the content of those item.

Chapter 10: Quality-Control Procedures

Quality control is a critically important element of every phase of MISA development, administration, and score reporting in ensuring the accuracy of student-, school- and district-level data. Pearson has developed and refined a set of quality procedures to help ensure that all MSDE's testing requirements are met or exceeded. These quality-control procedures are detailed in the paragraphs that follow. In general, Pearson's commitment to quality is incorporated in both task-specific quality standards applied to processing functions and services as well as a network of systems and procedures that coordinate quality steps *across* functions and services.

Quality Control for Test Construction

Following a legally sanctioned test development process (Smisko, Twing, & Denny, 2000), items are selected and placed on particular test forms that are as comparable as possible with respect to content and statistical characteristics. The process is an iterative process involving content experts, psychometricians, and MSDE. The goals are to create forms that meet blueprint and statistical targets using the highest quality items (in terms of content and statistical characteristics) that result in the most comparable test forms. Once an initial core is selected, all responsible parties evaluate and recommend improvements until final *best* core forms have been affirmed and moved to production. Throughout the process, standard checklists are used to ensure all steps are followed.

Quality Control for Printed Documents

Pearson follows a meticulous set of internal quality standards to ensure high-quality printed products. Specific areas of responsibility for staff involved in materials production include monitoring all materials-production schedules to meet contract commitments, overseeing the production of test materials, coordinating detailed printing and post-printing specifications, outlining specific quality control requirements for all materials, and conducting print reviews and quality checks. The quality production and printing processes follow printers' reviews and quality checks. Project management and print procurement staff work closely with the printers during the production phase. Press proofs are checked to ensure high-quality printing and to verify adherence to specifications. The printing staff randomly pull documents throughout the print run for additional quality control inspections.

Quality Control for Online Test Delivery Components

Each release of every Online Test Delivery goes through a complete testing cycle, including regression and performance testing. The system goes through User Acceptance Testing (UAT). During UAT, operational MISA tests that will be administered are used. In addition to the UAT, Production Validation (PV) testing occurs. Pearson publishes the MISA assessments in a production environment and recommends test scenarios. The tests are completed and scoring deliverables are generated during the PV period. The validation process includes confirmation of the tests published and the scoring deliverables. Approvals are required at the close of the PV period prior to the opening of the testing window.

For changes required during the testing window, a patch build is implemented. The release notes are provided that include fixes made and/or system upgrades. Any patches are tested and approved before being deployed to the field. Deployments are scheduled outside of the regular testing window timeframes.

Quality Control for Test-Form Equating

Test-form equating is the process that enables fair and equitable comparisons both across test forms within a single year and between test administrations across years. Pearson uses several quality-control procedures to ensure this equating is performed accurately.

1. Pearson performs a statistical “key check” analysis for the multiple-choice (MC) item type to ensure the appropriate scoring key is being used.
2. Pearson performs an “adjudication” analysis for the technology-enhanced (TE) item types. The adjudication process includes a check of all responses given by students in the current administration to ensure all possible responses are scores as intended.
3. For all assessments, a anchor item stability analysis is conducted in order to determine whether the Rasch item parameters have shifted over time. Items which have shifted are investigated and a resolution whether to keep or remove an item within an equating protocol is made.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Andrich, A., & Luo, G. (2004). *Modern measurement and analysis in social science*. Murdoch University, Perth, Western Australia.

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy [computer software] (Version 1.0*). Iowa City, IA: University of Iowa.

Cronbach, L. J., (1951). *Coefficient alpha and the internal structure of tests*. Psychometrika, 16, 297–334.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach. (ETS Research Report No. 91-47).* Princeton, NJ: Educational Testing Service.

Dorans, N. J. (2013). *ETS Contributions to the Quantitative Assessment of Item, Test and Score Fairness (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04).* Princeton, NJ: Educational Testing Service.

ESSA (2015). *Every Student Succeeds Act of 2015*, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Harvill, L. M. (1991). *Standard error of measurement.* Educational Measurement: Issues and Practice, 10, 181-189.

Huynh, H., & Meyer, P. (2010*). Use of robust z in detecting unstable items in item response theory models*. Practical Assessment, Research & Evaluation, 15(2). Available online: http://pareonline.net/getvn.asp?v=15&n=2.

Huynh, H., & Rawls, A. (2009). *A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model*. In Everett V. Smith Jr. & Greg E. Stone (Eds.) Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting. Maple Grove, MN: JAM Press.

Linacre, J. M., & Wright, B. D. (2000). *A user's guide to WINSTEPS: Rasch-model computer program.* Chicago, IL: MESA Press.

Livingston, S. A., & Lewis, C. (1995). *Estimating the Consistency and Accuracy of Classifications Based on Test Scores*. Journal of Educational Measurement, 32: 179-197.

Loehlin, J. C. (1987). *Latent variable models.* NJ: Lawrence Erlbaum Associates, Publishers.

Mantel, N. (1963). *Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure.* Journal of the American Statistical Association, *58,* 690-700.

Mantel, N., & Haenszel, W. (1959). S*tatistical aspects of the analysis of data from retrospective studies of disease.* Journal of the National Cancer Institute, *22,* 719-748.

Masters, G. N. (1982). *A Rasch model for partial credit scoring*. Psychometrika, *47*(2), 149-174.

National Research Council. (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Board on Science Education, Division of Behavioral and Social Sciences and Education, Committee on a Conceptual Framework for New K-12 Science Education Standards. Washington, DC: The National Academies Press.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982), *The Polyserial Correlation Coefficient*, Biometrika, 47, 337–347.

Orlando, M. (2004, June). *Critical issues to address when applying item response theory (IRT) models.* Paper presented at the The Drug Information Association, Bethesda, MD.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Smisko, A., Twing, J. S., & Denny, P. L. (2000). *The Texas model for content and curricular validity*. Applied Measurement in Education, 13(4), 333-342.

South Carolina Department of Education. (2001). *Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics (Technical Report)*. Columbia: South Carolina Department of Education.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). *A framework for evaluation and use of automated scoring*. Educational Measurement: Issues and Practices, 31, 2–13.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA PRESS.

Webb, N. L. (1999) *Alignment of science and mathematics standards and assessments in four states (Research Monograph No. 18).* Madison, WI: University of Wisconsin – Madison, National Institute for Science Education.

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337–348). Hillsdale, NJ: Erlbaum

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items (ETS Research Report RR-97-05).* Princeton, NJ: Educational Testing Service.

Appendix A: Performance Expectations by Grade and Core Form

Table A.1. Grade 5 MISA Domain and Performance Expectations Assessed by Item

Item	Unit	Form	Type	Domain	Performance Expectations
1	1	Common	MC	Life Science	5-LS2, 5-LS2-1
2	1	Common	MC	Life Science	5-LS2, 5-LS2-1
3	1	Common	TE	Life Science	5-LS2, 5-LS2-1
4	1	Common	MC	Life Science	5-LS1, 5-LS1-1
5	1	Common	MC	Life Science	5-LS1, 5-LS1-1
6	1	Common	CR	Life Science	5-LS2, 5-LS2-1
7	2	Common	MC	Earth & Space Science	4-ESS2, 4-ESS2-2
8	2	Common	MC	Earth & Space Science	4-ESS2, 4-ESS2-2
9	2	Common	TE	Earth & Space Science	4-ESS2, 4-ESS2-1
10	2	Common	MC	Earth & Space Science	4-ESS2, 4-ESS2-1
11	2	Common	MC	Earth & Space Science	4-ESS2, 4-ESS2-1
12	2	Common	CR	Earth & Space Science	4-ESS2, 4-ESS2-1
13	3	Common	MC	Physical Science	5-PS1, 5-PS1-1
14	3	Common	TE	Physical Science	5-PS1, 5-PS1-1
15	3	Common	MC	Physical Science	5-PS1, 5-PS1-1
16	3	Common	MC	Physical Science	5-PS1, 5-PS1-2
17	3	Common	MC	Physical Science	5-PS1, 5-PS1-2
18	3	Common	CR	Physical Science	5-PS1, 5-PS1-1
19	1	Core 1	MC	Earth & Space Science	5-ESS3, 5-ESS3-1
20	1	Core 1	MC	Earth & Space Science	4-ESS3, 4-ESS3-1
21	1	Core 1	TE	Earth & Space Science	5-ESS3, 5-ESS3-1
22	1	Core 1	TE	Earth & Space Science	4-ESS3, 4-ESS3-1
23	1	Core 1	TE	Earth & Space Science	5-ESS3, 5-ESS3-1
24	1	Core 1	CR	Earth & Space Science	5-ESS3, 5-ESS3-1
25	1	Core 1	TE	Physical Science	5-PS1, 5-PS1-1
26	1	Core 1	TE	Physical Science	5-PS1, 5-PS1-2
27	1	Core 1	TE	Physical Science	5-PS1, 5-PS1-1
28	1	Core 1	MC	Physical Science	5-PS1, 5-PS1-1
29	1	Core 1	TE	Physical Science	5-PS1, 5-PS1-2
30	1	Core 1	CR	Physical Science	5-PS1, 5-PS1-1
31	2	Core 1	MC	Physical Science	4-PS4, 4-PS4-1
32	2	Core 1	TE	Physical Science	4-PS4, 4-PS4-2
33	2	Core 1	MC	Physical Science	4-PS4, 4-PS4-2
34	2	Core 1	MC	Physical Science	4-PS4, 4-PS4-1
35	2	Core 1	TE	Physical Science	4-PS4, 4-PS4-1
36	2	Core 1	CR	Physical Science	4-PS4, 4-PS4-1

Item	Unit	Form	Type	Domain	Performance Expectations
37	2	Core 1	TE	Life Science	4-LS1, 4-LS1-1
38	2	Core 1	TE	Life Science	3-LS1, 3-LS1-1
39	2	Core 1	MC	Life Science	4-LS1, 4-LS1-2
40	2	Core 1	MC	Life Science	3-LS1, 3-LS1-1
41	2	Core 1	TE	Life Science	3-LS1, 3-LS1-1
42	2	Core 1	CR	Life Science	4-LS1, 4-LS1-1
43	3	Core 1	MC	Earth & Space Science	4-ESS3, 4-ESS3-1
44	3	Core 1	MC	Earth & Space Science	4-ESS3, 4-ESS3-1
45	3	Core 1	TE	Earth & Space Science	5-ESS3, 5-ESS3-1
46	3	Core 1	MC	Earth & Space Science	5-ESS3, 5-ESS3-1
47	3	Core 1	TE	Earth & Space Science	4-ESS3, 4-ESS3-1
48	3	Core 1	CR	Earth & Space Science	5-ESS3, 5-ESS3-1
49	3	Core 1	TE	Life Science	4-LS1, 4-LS1-2
50	3	Core 1	MC	Life Science	4-LS1, 4-LS1-1
51	3	Core 1	TE	Life Science	4-LS1, 4-LS1-1
52	3	Core 1	MC	Life Science	4-LS1, 4-LS1-2
53	3	Core 1	TE	Life Science	3-LS1, 3-LS1-1
54	3	Core 1	CR	Life Science	4-LS1, 4-LS1-2
55	1	Core 2	MC	Earth & Space Science	4-ESS3, 4-ESS3-1
56	1	Core 2	MC	Earth & Space Science	4-ESS3, 4-ESS3-1
57	1	Core 2	TE	Earth & Space Science	5-ESS3, 5-ESS3-1
58	1	Core 2	TE	Earth & Space Science	5-ESS3, 5-ESS3-1
59	1	Core 2	TE	Earth & Space Science	4-ESS3, 4-ESS3-1
60	1	Core 2	CR	Earth & Space Science	5-ESS3, 5-ESS3-1
61	1	Core 2	TE	Physical Science	4-PS4, 4-PS4-1
62	1	Core 2	MC	Physical Science	4-PS4, 4-PS4-1
63	1	Core 2	MC	Physical Science	4-PS4, 4-PS4-1
64	1	Core 2	MC	Physical Science	4-PS4, 4-PS4-1
65	1	Core 2	MC	Physical Science	4-PS4, 4-PS4-1
66	1	Core 2	CR	Physical Science	4-PS4, 4-PS4-1
67	2	Core 2	MC	Physical Science	4-PS3, 4-PS3-4
68	2	Core 2	MC	Physical Science	4-PS3, 4-PS3-2
69	2	Core 2	MC	Physical Science	4-PS3, 4-PS3-4
70	2	Core 2	MC	Physical Science	4-PS3, 4-PS3-2
71	2	Core 2	MC	Physical Science	4-PS3, 4-PS3-2
72	2	Core 2	CR	Physical Science	4-PS3, 4-PS3-4
73	2	Core 2	TE	Life Science	3-LS4, 3-LS4-3

Item	Unit	Form	Type	Domain	Performance Expectations
74	2	Core 2	MC	Life Science	3-LS2, 3-LS2-1
75	2	Core 2	MC	Life Science	3-LS4, 3-LS4-3
76	2	Core 2	MC	Life Science	3-LS2, 3-LS2-1
77	2	Core 2	TE	Life Science	3-LS4, 3-LS4-3
78	2	Core 2	CR	Life Science	3-LS4, 3-LS4-3
79	3	Core 2	MC	Earth & Space Science	3-ESS2, 3-ESS2-2
80	3	Core 2	TE	Earth & Space Science	3-ESS2, 3-ESS2-1
81	3	Core 2	MC	Earth & Space Science	3-ESS2, 3-ESS2-1
82	3	Core 2	TE	Earth & Space Science	3-ESS2, 3-ESS2-2
83	3	Core 2	TE	Earth & Space Science	3-ESS2, 3-ESS2-2
84	3	Core 2	CR	Earth & Space Science	3-ESS2, 3-ESS2-2
85	3	Core 2	TE	Life Science	3-LS1, 3-LS1-1
86	3	Core 2	TE	Life Science	3-LS1, 3-LS1-1
87	3	Core 2	TE	Life Science	4-LS1, 4-LS1-1
88	3	Core 2	MC	Life Science	4-LS1, 4-LS1-1
89	3	Core 2	TE	Life Science	3-LS1, 3-LS1-1
90	3	Core 2	CR	Life Science	3-LS1, 3-LS1-1
91	4	Matrix	MC	Life Science	3-LS4, 3-LS4-3
92	4	Matrix	MC	Life Science	3-LS4, 3-LS4-3
93	4	Matrix	MC	Life Science	3-LS4, 3-LS4-3
94	4	Matrix	TE	Life Science	3-LS4, 3-LS4-3
95	4	Matrix	MC	Life Science	3-LS2, 3-LS2-1
96	4	Matrix	CR	Life Science	3-LS2, 3-LS2-1
97	4	Matrix	TE	Physical Science	3-PS2, 3-PS2-1
98	4	Matrix	MC	Physical Science	3-PS2, 3-PS2-1
99	4	Matrix	MC	Physical Science	3-PS2, 3-PS2-2
100	4	Matrix	TE	Physical Science	3-PS2, 3-PS2-2
101	4	Matrix	TE	Physical Science	3-PS2, 3-PS2-1
102	4	Matrix	CR	Physical Science	3-PS2, 3-PS2-1
103	4	Matrix	TE	Earth & Space Science	4-ESS3, 4-ESS3-2
104	4	Matrix	TE	Earth & Space Science	4-ESS2, 4-ESS2-2
105	4	Matrix	TE	Earth & Space Science	4-ESS2, 4-ESS2-2
106	4	Matrix	TE	Earth & Space Science	4-ESS2, 4-ESS2-1
107	4	Matrix	TE	Earth & Space Science	4-ESS2, 4-ESS2-1
108	4	Matrix	CR	Earth & Space Science	4-ESS3, 4-ESS3-2

Table A.2. Grade 8 MISA Domain and Performance Expectations Assessed by Item

Item	Unit	Form	Type	Domain	Performance Expectations
1	1	Common	TE	Physical Science	MS-PS3, MS-PS3-2
2	1	Common	MC	Physical Science	MS-PS3, MS-PS3-2
3	1	Common	TE	Physical Science	MS-PS3, MS-PS3-1
4	1	Common	TE	Physical Science	MS-PS3, MS-PS3-1
5	1	Common	MC	Physical Science	MS-PS3, MS-PS3-2
6	1	Common	CR	Physical Science	MS-PS3, MS-PS3-2
7	2	Common	MC	Earth & Space Science	MS-ESS3, MS-ESS3-3
8	2	Common	TE	Earth & Space Science	MS-ESS3, MS-ESS3-2
9	2	Common	TE	Earth & Space Science	MS-ESS3, MS-ESS3-4
10	2	Common	TE	Earth & Space Science	MS-ESS3, MS-ESS3-4
11	2	Common	MC	Earth & Space Science	MS-ESS3, MS-ESS3-4
12	2	Common	CR	Earth & Space Science	MS-ESS3, MS-ESS3-3
13	3	Common	MC	Life Science	MS-LS1, MS-LS1-7
14	3	Common	TE	Life Science	MS-LS1, MS-LS1-7
15	3	Common	TE	Life Science	MS-LS1, MS-LS1-7
16	3	Common	TE	Life Science	MS-LS1, MS-LS1-6
17	3	Common	MC	Life Science	MS-LS1, MS-LS1-6
18	3	Common	CR	Life Science	MS-LS1, MS-LS1-6
19	1	Core 1	MC	Life Science	MS-LS3, MS-LS3-2
20	1	Core 1	MC	Life Science	MS-LS3, MS-LS3-2
21	1	Core 1	TE	Life Science	MS-LS3, MS-LS3-2
22	1	Core 1	TE	Life Science	MS-LS4, MS-LS4-4
23	1	Core 1	MC	Life Science	MS-LS4, MS-LS4-4
24	1	Core 1	CR	Life Science	MS-LS3, MS-LS3-2
25	1	Core 1	MC	Earth & Space Science	MS-ESS2, MS-ESS2-2
26	1	Core 1	MC	Earth & Space Science	MS-ESS2, MS-ESS2-2
27	1	Core 1	MC	Earth & Space Science	MS-ESS2, MS-ESS2-3
28	1	Core 1	TE	Earth & Space Science	MS-ESS1, MS-ESS1-4
29	1	Core 1	MC	Earth & Space Science	MS-ESS1, MS-ESS1-4
30	1	Core 1	CR	Earth & Space Science	MS-ESS1, MS-ESS1-4
31	2	Core 1	TE	Life Science	MS-LS2, MS-LS2-2
32	2	Core 1	TE	Life Science	MS-LS2, MS-LS2-2
33	2	Core 1	TE	Life Science	MS-LS2, MS-LS2-1
34	2	Core 1	TE	Life Science	MS-LS2, MS-LS2-1
35	2	Core 1	TE	Life Science	MS-LS2, MS-LS2-2
36	2	Core 1	CR	Life Science	MS-LS2, MS-LS2-1

Item	Unit	Form	Type	Domain	Performance Expectations
37	2	Core 1	TE	Physical Science	MS-PS3, MS-PS3-3
38	2	Core 1	TE	Physical Science	MS-PS3, MS-PS3-3
39	2	Core 1	MC	Physical Science	MS-PS3, MS-PS3-4
40	2	Core 1	TE	Physical Science	MS-PS3, MS-PS3-4
41	2	Core 1	TE	Physical Science	MS-PS3, MS-PS3-5
42	2	Core 1	CR	Physical Science	MS-PS3, MS-PS3-5
43	3	Core 1	TE	Earth & Space Science	MS-ESS2, MS-ESS2-6
44	3	Core 1	TE	Earth & Space Science	MS-ESS2, MS-ESS2-6
45	3	Core 1	TE	Earth & Space Science	MS-ESS2, MS-ESS2-5
46	3	Core 1	TE	Earth & Space Science	MS-ESS2, MS-ESS2-6
47	3	Core 1	TE	Earth & Space Science	MS-ESS2, MS-ESS2-5
48	3	Core 1	CR	Earth & Space Science	MS-ESS3, MS-ESS3-5
49	3	Core 1	TE	Physical Science	MS-PS1, MS-PS1-6
50	3	Core 1	MC	Physical Science	MS-PS1, MS-PS1-5
51	3	Core 1	MC	Physical Science	MS-PS1, MS-PS1-6
52	3	Core 1	TE	Physical Science	MS-PS1, MS-PS1-2
53	3	Core 1	TE	Physical Science	MS-PS1, MS-PS1-5
54	3	Core 1	CR	Physical Science	MS-PS1, MS-PS1-2
55	1	Core 2	TE	Life Science	MS-LS1, MS-LS1-5
56	1	Core 2	MC	Life Science	MS-LS1, MS-LS1-4
57	1	Core 2	TE	Life Science	MS-LS1, MS-LS1-4
58	1	Core 2	TE	Life Science	MS-LS1, MS-LS1-5
59	1	Core 2	TE	Life Science	MS-LS1, MS-LS1-5
60	1	Core 2	CR	Life Science	MS-LS1, MS-LS1-4
61	1	Core 2	MC	Earth & Space Science	MS-ESS2, MS-ESS2-4
62	1	Core 2	TE	Earth & Space Science	MS-ESS3, MS-ESS3-1
63	1	Core 2	TE	Earth & Space Science	MS-ESS2, MS-ESS2-4
64	1	Core 2	MC	Earth & Space Science	MS-ESS3, MS-ESS3-1
65	1	Core 2	TE	Earth & Space Science	MS-ESS3, MS-ESS3-1
66	1	Core 2	CR	Earth & Space Science	MS-ESS3, MS-ESS3-1
67	2	Core 2	TE	Physical Science	MS-PS2, MS-PS2-4
68	2	Core 2	TE	Physical Science	MS-PS2, MS-PS2-1
69	2	Core 2	TE	Physical Science	MS-PS2, MS-PS2-4
70	2	Core 2	TE	Physical Science	MS-PS2, MS-PS2-4
71	2	Core 2	MC	Physical Science	MS-PS2, MS-PS2-1
72	2	Core 2	CR	Physical Science	MS-PS2, MS-PS2-1
73	2	Core 2	TE	Life Science	MS-LS2, MS-LS2-2

Item	Unit	Form	Type	Domain	Performance Expectations
74	2	Core 2	MC	Life Science	MS-LS2, MS-LS2-3
75	2	Core 2	TE	Life Science	MS-LS2, MS-LS2-1
76	2	Core 2	MC	Life Science	MS-LS2, MS-LS2-1
77	2	Core 2	TE	Life Science	MS-LS2, MS-LS2-2
78	2	Core 2	CR	Life Science	MS-LS2, MS-LS2-3
79	3	Core 2	MC	Physical Science	MS-PS2, MS-PS2-5
80	3	Core 2	TE	Physical Science	MS-PS2, MS-PS2-3
81	3	Core 2	MC	Physical Science	MS-PS2, MS-PS2-3
82	3	Core 2	MC	Physical Science	MS-PS2, MS-PS2-3
83	3	Core 2	MC	Physical Science	MS-PS2, MS-PS2-5
84	3	Core 2	CR	Physical Science	MS-PS2, MS-PS2-5
85	3	Core 2	TE	Earth & Space Science	MS-ESS2, MS-ESS2-4
86	3	Core 2	TE	Earth & Space Science	MS-ESS2, MS-ESS2-4
87	3	Core 2	MC	Earth & Space Science	MS-ESS2, MS-ESS2-4
88	3	Core 2	TE	Earth & Space Science	MS-ESS3, MS-ESS3-1
89	3	Core 2	MC	Earth & Space Science	MS-ESS3, MS-ESS3-1
90	3	Core 2	CR	Earth & Space Science	MS-ESS3, MS-ESS3-1
91	4	Matrix	MC	Earth & Space Science	MS-ESS1, MS-ESS1-4
92	4	Matrix	MC	Earth & Space Science	MS-ESS1, MS-ESS1-4
93	4	Matrix	MC	Earth & Space Science	MS-ESS2, MS-ESS2-3
94	4	Matrix	TE	Earth & Space Science	MS-ESS2, MS-ESS2-2
95	4	Matrix	TE	Earth & Space Science	MS-ESS2, MS-ESS2-3
96	4	Matrix	CR	Earth & Space Science	MS-ESS2, MS-ESS2-2
97	4	Matrix	MC	Physical Science	MS-PS3, MS-PS3-4
98	4	Matrix	TE	Physical Science	MS-PS3, MS-PS3-4
99	4	Matrix	TE	Physical Science	MS-PS3, MS-PS3-5
100	4	Matrix	TE	Physical Science	MS-PS3, MS-PS3-5
101	4	Matrix	MC	Physical Science	MS-PS3, MS-PS3-4
102	4	Matrix	CR	Physical Science	MS-PS3, MS-PS3-3
103	4	Matrix	MC	Life Science	MS-LS2, MS-LS2-3
104	4	Matrix	TE	Life Science	MS-LS2, MS-LS2-1
105	4	Matrix	TE	Life Science	MS-LS2, MS-LS2-2
106	4	Matrix	MC	Life Science	MS-LS2, MS-LS2-3
107	4	Matrix	TE	Life Science	MS-LS2, MS-LS2-2
108	4	Matrix	CR	Life Science	MS-LS2, MS-LS2-1, MS-LS2-2, MS-LS2-3

Appendix B: Sample PSC Quality Monitoring Report

Table B.1. Sample PSC Quality Management Report Extended Guide

Run Date: 06/06/2018
Run Time: 01:26:40 PM CDT
Report Date: 06/06/2018

PSC Quality Management Report Extended Guide
MD MISA SPR 2018 OPFT/814-375 for 06/06/2018

Report #: PSC4
Page 1 of 6

					Validity %				Reliability %								
					Perfect Agreement		Total Non-Adj		Perfect Agreement		Total Non-Adj		# of Resolutions				
Grade	Site	RIB ID	Item/Trait	Item Description	Daily	Cum	Daily	Cum	Daily	Cum	Daily	Cum	TBR*	Comp**	Distribution Current	Mean	Percent Complete
05	Inte		51603_12	Different Types of Water on Earth	0.0	92.9	0.0	0.0	0.0	80.7		0.2	0	7	**BL OI 0 1 2** 9.1 3.0 19.0 56.0 12.9	0.8	0.0%
			17MD05SC51_12			*95*		*0*		*88*		*0*			*3 1 17 68 12*	*0.915*	
05	Inte		51604_11	Prevent Land Loss	0.0	90.7	0.0	0.3	0.0	77.8		1.6	0	47	**BL OI 0 1 2 3** 2.1 1.1 36.1 42.9 15.4 2.4	0.8	0.0%
			17MD05SC32_06			*90*		*0*		*85*		*0*			*1 1 38 47 11 1*	*0.73*	
05	Inte		51605_12	Results of the Ladybug Investigation	0.0	89.8	0.0	0.0	0.0	78.9		9.4	0	367	**BL OI 0 1 2 3** 3.5 7.4 40.4 41.7 6.4 0.6	0.6	0.0%
			17MD05SC41_06			*87*		*0*		*82*		*1*			*2 5 24 55 12 2*	*0.843*	
05	Inte		51607_12	Effect of Fur Coloring	0.0	92.9	0.0	0.0	0.0	78.5		0.5	0	38	**BL OI 0 1 2 3** 4.9 2.3 30.5 46.5 14.0 1.7	0.8	0.0%
			17MD05SC34_06			*91*		*0*		*79*		*0*			*1 1 27 52 16 3*	*0.927*	
05	Inte		51609_06	Need for Multiple Trials	0.0	91.5	0.0	0.0	0.0	85.8		0.3	0	8	**BL OI 0 1 2** 5.5 2.1 58.4 29.6 4.4	0.4	0.0%
			17MD05SC33_06			*90*		*0*		*91*		*0*			*3 2 62 30 3*	*0.353*	
05	Inte		51613_06	Evidence of New Substance	0.0	98.2	0.0	0.0	0.0	78.3		0.7	0	29	**BL OI 0 1 2** 2.6 2.6 44.1 38.6 12.1	0.6	0.0%
			17MD05SC32_12			*95*		*0*		*91*		*0*			*3 2 44 44 7*	*0.586*	

Numbers in blue italics are previous final numbers

***TBR - To Be Resolved**

****Comp - Resolutions Completed**

Appendix C: Classical Item Statistics

Table C.1. Grade 5 MISA Classical Item Statistics

Item	Item Type	N-Count	P-value	Pbis Total	Prop A or 0pts	Prop B or 1pts	Prop C or 2pts	Prop D or 3pts	Prop 4pts	Prop Omitting
1	MC	69777	0.763	0.313	0.125	0.038	0.072	0.763		0.003
2	MC	69777	0.418	0.500	0.226	0.152	0.200	0.418		0.004
3	TE	69777	0.299	0.545	0.696	0.299				0.006
4	MC	69777	0.806	0.388	0.051	0.090	0.051	0.806		0.003
5	MC	69777	0.645	0.445	0.074	0.645	0.203	0.074		0.003
6	CR	69777	0.255	0.630	0.421	0.418	0.136	0.025		0.000
7	MC	59233	0.374	0.486	0.623	0.374				0.003
8	MC	59233	0.807	0.454	0.807	0.095	0.052	0.045		0.001
9	TE	59233	0.128	0.262	0.869	0.128				0.003
10	MC	59233	0.357	0.439	0.418	0.116	0.106	0.357		0.003
11	MC	59233	0.336	0.206	0.222	0.336	0.194	0.244		0.003
12	CR	59233	0.271	0.626	0.368	0.464	0.157	0.012		0.000
13	MC	59233	0.380	0.300	0.134	0.126	0.380	0.358		0.002
14	TE	59233	0.809	0.398	0.190	0.809				0.001
15	MC	59233	0.279	0.048	0.081	0.356	0.282	0.279		0.002
16	MC	59233	0.496	0.309	0.078	0.294	0.131	0.496		0.002
17	MC	59233	0.315	0.118	0.368	0.315	0.102	0.211		0.003
18	CR	59233	0.234	0.505	0.337	0.623	0.039	0.000		0.000
19	MC	29684	0.554	0.249	0.064	0.554	0.199	0.173		0.009
20	MC	29684	0.260	0.387	0.223	0.260	0.224	0.278		0.014
21	TE	29684	0.287	0.401	0.700	0.287				0.013
22	TE	29684	0.225	0.439	0.760	0.225				0.015
23	TE	29684	0.357	0.516	0.625	0.357				0.019
24	CR	29684	0.324	0.445	0.148	0.737	0.110	0.005		0.000
25	TE	29684	0.161	0.304	0.791	0.161				0.049
26	TE	29684	0.209	0.524	0.725	0.209				0.067
27	TE	29684	0.384	0.428	0.345	0.379	0.194			0.082
28	MC	29684	0.313	0.260	0.184	0.155	0.313	0.250		0.098
29	TE	29684	0.291	0.321	0.602	0.291				0.107
30	CR	29684	0.294	0.469	0.445	0.524	0.031			0.000
31	MC	29684	0.606	0.488	0.109	0.185	0.096	0.606		0.004
32	TE	29684	0.387	0.428	0.609	0.387				0.004
33	MC	29684	0.501	0.439	0.214	0.113	0.166	0.501		0.006
34	MC	29684	0.823	0.339	0.823	0.042	0.040	0.089		0.005
35	TE	29684	0.252	0.309	0.740	0.252				0.008
36	CR	29684	0.223	0.557	0.389	0.554	0.057	0.000		0.000
37	TE	29684	0.415	0.457	0.568	0.415				0.018
38	TE	29684	0.573	0.558	0.171	0.471	0.337			0.020
39	MC	29684	0.448	0.525	0.122	0.265	0.136	0.448		0.029
40	MC	29684	0.268	0.464	0.701	0.268				0.031
41	TE	29684	0.546	0.533	0.418	0.546				0.035
42	CR	29684	0.227	0.494	0.614	0.319	0.067			0.000
43	MC	40228	0.533	0.383	0.267	0.533	0.114	0.082		0.004
44	MC	40228	0.447	0.349	0.285	0.118	0.447	0.147		0.004
45	TE	39641	0.409	0.665	0.587	0.409				0.005
46	MC	29684	0.342	0.421	0.141	0.169	0.344	0.342		0.004
47	TE	40228	0.426	0.625	0.314	0.508	0.172			0.005
48	CR	40228	0.507	0.713	0.381	0.224	0.395			0.000
49	TE	29684	0.605	0.566	0.384	0.605				0.011
50	MC	29684	0.840	0.482	0.020	0.058	0.070	0.840		0.013
51	TE	29684	0.479	0.561	0.507	0.479				0.014
52	MC	29684	0.857	0.477	0.857	0.048	0.049	0.031		0.016
53	TE	29684	0.362	0.534	0.620	0.362				0.017
54	CR	29684	0.302	0.516	0.168	0.761	0.069	0.002		0.000
55	MC	29549	0.436	0.395	0.195	0.132	0.228	0.436		0.008
56	MC	29549	0.576	0.359	0.102	0.576	0.072	0.239		0.012

Item	Item Type	N-Count	P-value	Pbis Total	Prop A or 0pts	Prop B or 1pts	Prop C or 2pts	Prop D or 3pts	Prop 4pts	Prop Omitting
57	TE	29549	0.267	0.459	0.719	0.267				0.015
58	TE	29549	0.503	0.549	0.479	0.503				0.018
59	TE	29549	0.192	0.429	0.785	0.192				0.023
60	CR	29549	0.284	0.674	0.342	0.476	0.168	0.013		0.000
61	TE	29549	0.416	0.508	0.344	0.389	0.221			0.046
62	MC	29549	0.353	0.142	0.353	0.255	0.229	0.108		0.055
63	MC	29549	0.497	0.317	0.150	0.114	0.497	0.180		0.060
64	MC	29549	0.460	0.469	0.460	0.300	0.113	0.062		0.065
65	MC	29549	0.476	0.393	0.476	0.170	0.212	0.073		0.069
66	CR	29549	0.240	0.569	0.537	0.447	0.016			0.000
67	MC	29549	0.425	0.219	0.265	0.425	0.040	0.267		0.004
68	MC	29549	0.535	0.320	0.254	0.535	0.156	0.049		0.005
69	MC	29549	0.414	0.419	0.246	0.237	0.414	0.097		0.006
70	MC	29549	0.583	0.434	0.093	0.202	0.583	0.116		0.007
71	MC	29549	0.296	0.095	0.081	0.446	0.296	0.168		0.009
72	CR	29549	0.236	0.543	0.326	0.641	0.033	0.000		0.000
73	TE	29549	0.143	0.120	0.833	0.143				0.024
74	MC	29549	0.532	0.428	0.146	0.532	0.198	0.096		0.028
75	MC	29549	0.435	0.468	0.435	0.174	0.241	0.117		0.034
76	MC	29549	0.612	0.484	0.612	0.116	0.126	0.110		0.038
77	TE	29549	0.635	0.436	0.091	0.469	0.400			0.039
78	CR	29549	0.281	0.579	0.503	0.434	0.063			0.000
79	MC	29549	0.421	0.281	0.299	0.137	0.421	0.141		0.002
80	TE	29549	0.504	0.435	0.493	0.504				0.003
81	MC	29549	0.601	0.314	0.094	0.601	0.101	0.201		0.003
82	TE	29549	0.282	0.296	0.713	0.282				0.004
83	TE	29549	0.385	0.619	0.431	0.357	0.206			0.005
84	CR	29549	0.355	0.604	0.419	0.453	0.128			0.000
85	TE	29549	0.425	0.432	0.563	0.425				0.011
86	TE	29549	0.250	0.453	0.735	0.250				0.015
87	TE	29549	0.777	0.505	0.207	0.777				0.017
88	MC	29549	0.505	0.441	0.175	0.152	0.149	0.505		0.020
89	TE	29549	0.473	0.611	0.506	0.473				0.021
90	CR	29549	0.282	0.513	0.313	0.527	0.159	0.001		0.000
91	MC	12094	0.660	0.309	0.043	0.242	0.054	0.660		0.001
92	MC	12094	0.271	0.281	0.728	0.271				0.002
93	MC	12141	0.757	0.464	0.091	0.079	0.757	0.071		0.001
94	TE	12094	0.360	0.373	0.638	0.360				0.002
95	MC	12094	0.776	0.521	0.072	0.058	0.093	0.776		0.002
96	CR	12094	0.299	0.558	0.147	0.570	0.231	0.040	0.011	0.000
97	TE	11851	0.097	0.048	0.901	0.097				0.002
98	MC	11851	0.177	-.0320	0.106	0.177	0.368	0.346		0.003
99	MC	11851	0.215	0.287	0.265	0.287	0.229	0.215		0.003
100	TE	11851	0.080	0.173	0.917	0.080				0.003
101	TE	11851	0.217	0.151	0.779	0.217				0.005
102	CR	11851	0.255	0.561	0.253	0.505	0.212	0.030	0.001	0.000
103	TE	11624	0.457	0.540	0.542	0.457				0.002
104	TE	11624	0.358	0.417	0.637	0.358				0.005
105	TE	11624	0.348	0.412	0.650	0.348				0.002
106	TE	11624	0.086	0.202	0.912	0.086				0.003
107	TE	11624	0.026	0.133	0.971	0.026				0.003
108	CR	11624	0.302	0.555	0.143	0.567	0.237	0.045	0.007	0.000

Table C.2. Grade 8 MISA Classical Item Statistics

Item	Item Type	N-Count	P-value	Pbis Total	Prop A or 0pts	Prop B or 1pts	Prop C or 2pts	Prop D or 3pts	Prop 4pts	Prop Omitting
1	TE	59754	0.578	0.374	0.405	0.578				0.017
2	MC	59754	0.478	0.461	0.249	0.478	0.134	0.120		0.019
3	TE	59754	0.524	0.512	0.445	0.524				0.032
4	TE	59754	0.524	0.467	0.448	0.524				0.028
5	MC	59754	0.657	0.429	0.657	0.127	0.131	0.053		0.032
6	CR	59754	0.361	0.700	0.233	0.471	0.278	0.019		0.000
7	MC	64082	0.462	0.124	0.093	0.147	0.462	0.297		0.001
8	TE	64082	0.703	0.604	0.294	0.703				0.003
9	TE	64082	0.651	0.581	0.347	0.651				0.002
10	TE	64082	0.807	0.378	0.190	0.807				0.003
11	MC	64237	0.534	0.476	0.534	0.217	0.145	0.101		0.003
12	CR	64237	0.306	0.717	0.326	0.476	0.152	0.046		0.000
13	MC	64237	0.604	0.522	0.110	0.127	0.151	0.604		0.007
14	TE	64082	0.238	0.432	0.752	0.238				0.010
15	TE	64082	0.184	0.231	0.804	0.184				0.012
16	TE	64082	0.401	0.416	0.588	0.401				0.011
17	MC	64237	0.422	0.357	0.310	0.157	0.422	0.098		0.013
18	CR	64237	0.271	0.663	0.287	0.619	0.090	0.004		0.000
19	MC	29976	0.564	0.572	0.090	0.190	0.153	0.564		0.002
20	MC	34459	0.628	0.504	0.155	0.628	0.120	0.093		0.004
21	TE	29976	0.356	0.507	0.500	0.283	0.215			0.002
22	TE	34304	0.538	0.550	0.457	0.538				0.005
23	MC	34459	0.650	0.545	0.650	0.109	0.118	0.118		0.005
24	CR	34459	0.404	0.718	0.377	0.439	0.184			0.000
25	MC	29976	0.402	0.273	0.402	0.346	0.141	0.105		0.006
26	MC	34459	0.699	0.414	0.699	0.138	0.073	0.083		0.006
27	MC	34459	0.480	0.377	0.143	0.214	0.155	0.480		0.008
28	TE	34304	0.355	0.290	0.636	0.355				0.009
29	MC	34459	0.233	0.288	0.255	0.293	0.207	0.233		0.012
30	CR	34459	0.290	0.746	0.328	0.507	0.133	0.032		0.000
31	TE	29976	0.273	0.297	0.724	0.273				0.003
32	TE	29976	0.428	0.461	0.569	0.428				0.004
33	TE	29976	0.553	0.591	0.443	0.553				0.004
34	TE	29976	0.457	0.492	0.256	0.565	0.174			0.005
35	TE	29976	0.493	0.365	0.502	0.493				0.005
36	CR	29976	0.444	0.716	0.323	0.466	0.211			0.000
37	TE	29976	0.245	0.315	0.747	0.245				0.008
38	TE	29976	0.559	0.526	0.433	0.559				0.008
39	MC	29976	0.579	0.396	0.579	0.126	0.264	0.021		0.009
40	TE	29976	0.371	0.630	0.618	0.371				0.011
41	TE	29976	0.328	0.458	0.660	0.328				0.012
42	CR	29976	0.271	0.757	0.408	0.394	0.175	0.023		0.000
43	TE	29976	0.620	0.569	0.379	0.620				0.002
44	TE	29976	0.607	0.472	0.391	0.607				0.001
45	TE	29976	0.305	0.452	0.693	0.305				0.002
46	TE	29976	0.651	0.424	0.347	0.651				0.002
47	TE	29976	0.501	0.245	0.497	0.501				0.002
48	CR	29976	0.325	0.692	0.218	0.597	0.174	0.010		0.000
49	TE	29976	0.238	0.331	0.759	0.238				0.003
50	MC	29976	0.392	0.320	0.392	0.245	0.201	0.159		0.004
51	MC	29976	0.683	0.509	0.064	0.683	0.172	0.076		0.004
52	TE	29976	0.278	0.225	0.718	0.278				0.004
53	TE	29976	0.375	0.431	0.401	0.438	0.156			0.005
54	CR	29976	0.405	0.609	0.351	0.487	0.161			0.000
55	TE	29778	0.409	0.335	0.590	0.409				0.001
56	MC	29778	0.583	0.319	0.044	0.121	0.249	0.583		0.003

Item	Item Type	N-Count	P-value	Pbis Total	Prop A or 0pts	Prop B or 1pts	Prop C or 2pts	Prop D or 3pts	Prop 4pts	Prop Omitting
57	TE	29778	0.443	0.478	0.546	0.443				0.010
58	TE	29778	0.714	0.385	0.284	0.714				0.002
59	TE	29778	0.391	0.474	0.606	0.391				0.003
60	CR	29778	0.316	0.647	0.305	0.449	0.241	0.005		0.000
61	MC	34261	0.506	0.430	0.287	0.052	0.149	0.506		0.006
62	TE	34106	0.493	0.545	0.288	0.426	0.280			0.007
63	TE	34261	0.378	0.540	0.614	0.378				0.007
64	MC	29778	0.573	0.406	0.091	0.065	0.573	0.264		0.007
65	TE	29778	0.644	0.561	0.348	0.644				0.008
66	CR	34261	0.415	0.635	0.291	0.590	0.120			0.000
67	TE	34106	0.425	0.445	0.571	0.425				0.004
68	TE	29778	0.270	0.462	0.514	0.423	0.058			0.004
69	TE	29778	0.423	0.490	0.572	0.423				0.005
70	TE	34106	0.552	0.516	0.438	0.552				0.009
71	MC	29778	0.390	0.091	0.390	0.089	0.372	0.143		0.006
72	CR	29778	0.230	0.612	0.558	0.425	0.017			0.000
73	TE	29778	0.437	0.564	0.552	0.437				0.011
74	MC	29778	0.388	0.264	0.122	0.190	0.388	0.287		0.013
75	TE	29778	0.677	0.280	0.308	0.677				0.014
76	MC	29778	0.599	0.407	0.113	0.198	0.599	0.075		0.014
77	TE	29778	0.175	0.351	0.806	0.175				0.019
78	CR	29778	0.168	0.605	0.574	0.351	0.073	0.003		0.000
79	MC	29778	0.769	0.526	0.060	0.040	0.129	0.769		0.001
80	TE	29778	0.479	0.566	0.520	0.479				0.001
81	MC	29778	0.744	0.469	0.095	0.744	0.056	0.104		0.001
82	MC	29778	0.589	0.436	0.124	0.589	0.098	0.188		0.002
83	MC	29778	0.425	0.367	0.425	0.138	0.177	0.258		0.002
84	CR	29778	0.288	0.646	0.304	0.537	0.151	0.009		0.000
85	TE	29778	0.556	0.513	0.441	0.556				0.003
86	TE	29778	0.350	0.483	0.350	0.595	0.053			0.003
87	MC	29778	0.784	0.417	0.075	0.067	0.784	0.070		0.003
88	TE	29778	0.677	0.589	0.320	0.677				0.003
89	MC	29778	0.493	0.266	0.165	0.493	0.194	0.145		0.003
90	CR	29778	0.377	0.527	0.414	0.419	0.167			0.000
91	MC	11874	0.377	0.136	0.256	0.131	0.377	0.233		0.003
92	MC	11874	0.547	0.362	0.547	0.281	0.110	0.058		0.003
93	MC	11874	0.749	0.401	0.090	0.046	0.749	0.112		0.003
94	TE	11874	0.542	0.404	0.454	0.542				0.004
95	TE	11874	0.610	0.517	0.385	0.610				0.004
96	CR	11874	0.291	0.701	0.201	0.481	0.276	0.041	0.001	0.000
97	MC	12041	0.575	0.371	0.126	0.194	0.575	0.103		0.002
98	TE	12041	0.694	0.551	0.304	0.694				0.002
99	TE	12041	0.619	0.485	0.378	0.619				0.003
100	TE	12041	0.582	0.613	0.415	0.582				0.003
101	MC	12041	0.799	0.436	0.087	0.799	0.043	0.068		0.003
102	CR	12041	0.324	0.716	0.203	0.420	0.260	0.112	0.005	0.000
103	MC	12128	0.619	0.402	0.159	0.619	0.161	0.059		0.002
104	TE	12059	0.344	0.299	0.654	0.344				0.003
105	TE	12059	0.276	0.259	0.721	0.276				0.003
106	MC	12128	0.618	0.578	0.070	0.151	0.158	0.618		0.003
107	TE	12059	0.381	0.460	0.616	0.381				0.003
108	CR	12128	0.369	0.750	0.185	0.335	0.304	0.174	0.002	0.000

Appendix D: Differential Item Functioning

Table D.1. MISA Grade 5 Male vs Female Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	MC	35713	34064	A
2	MC	35713	34064	A
3	TE	35713	34064	A
4	MC	35713	34064	A
5	MC	35713	34064	A
6	CR	35713	34064	A
7	MC	29598	29635	A
8	MC	29598	29635	A
9	TE	29598	29635	A
10	MC	29598	29635	A
11	MC	29598	29635	A
12	CR	29598	29635	A
13	MC	29598	29635	A
14	TE	29598	29635	A
15	MC	29598	29635	A
16	MC	29598	29635	A
17	MC	29598	29635	A
18	CR	29598	29635	+B
19	MC	14880	14804	A
20	MC	14880	14804	-B
21	TE	14880	14804	A
22	TE	14880	14804	A
23	TE	14880	14804	A
24	CR	14880	14804	A
25	TE	14880	14804	A
26	TE	14880	14804	A
27	TE	14880	14804	A
28	MC	14880	14804	A
29	TE	14880	14804	A
30	CR	14880	14804	A
31	MC	14880	14804	A
32	TE	14880	14804	A
33	MC	14880	14804	A
34	MC	14880	14804	A
35	TE	14880	14804	A
36	CR	14880	14804	+B
37	TE	14880	14804	A
38	TE	14880	14804	A
39	MC	14880	14804	A
40	MC	14880	14804	A
41	TE	14880	14804	A
42	CR	14880	14804	+B

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
43	MC	20995	19233	A
44	MC	20995	19233	A
45	TE	20613	19028	A
46	MC	14880	14804	A
47	TE	20995	19233	A
48	CR	20995	19233	+B
49	TE	14880	14804	A
50	MC	14880	14804	-B
51	TE	14880	14804	A
52	MC	14880	14804	A
53	TE	14880	14804	A
54	CR	14880	14804	A
55	MC	14718	14831	A
56	MC	14718	14831	A
57	TE	14718	14831	A
58	TE	14718	14831	-B
59	TE	14718	14831	-B
60	CR	14718	14831	+B
61	TE	14718	14831	A
62	MC	14718	14831	A
63	MC	14718	14831	A
64	MC	14718	14831	A
65	MC	14718	14831	A
66	CR	14718	14831	A
67	MC	14718	14831	A
68	MC	14718	14831	A
69	MC	14718	14831	A
70	MC	14718	14831	A
71	MC	14718	14831	A
72	CR	14718	14831	+B
73	TE	14718	14831	A
74	MC	14718	14831	A
75	MC	14718	14831	A
76	MC	14718	14831	A
77	TE	14718	14831	A
78	CR	14718	14831	A
79	MC	14718	14831	A
80	TE	14718	14831	A
81	MC	14718	14831	A
82	TE	14718	14831	A
83	TE	14718	14831	A
84	CR	14718	14831	A
85	TE	14718	14831	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
86	TE	14718	14831	A
87	TE	14718	14831	A
88	MC	14718	14831	A
89	TE	14718	14831	A
90	CR	14718	14831	+B
91	MC	6145	5949	A
92	MC	6145	5949	A
93	MC	6172	5969	A
94	TE	6145	5949	A
95	MC	6145	5949	A
96	CR	6145	5949	+B
97	TE	5875	5976	A
98	MC	5875	5976	A
99	MC	5875	5976	A
100	TE	5875	5976	A
101	TE	5875	5976	A
102	CR	5875	5976	A
103	TE	5780	5844	A
104	TE	5780	5844	A
105	TE	5780	5844	A
106	TE	5780	5844	A
107	TE	5780	5844	A
108	CR	5780	5844	+B

Table D.2. MISA Grade 5 White vs. Black/African-American Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	MC	24886	23633	A
2	MC	24886	23633	A
3	TE	24886	23633	A
4	MC	24886	23633	A
5	MC	24886	23633	A
6	CR	24886	23633	A
7	MC	22594	19987	A
8	MC	22594	19987	A
9	TE	22594	19987	A
10	MC	22594	19987	A
11	MC	22594	19987	A
12	CR	22594	19987	A
13	MC	22594	19987	A
14	TE	22594	19987	A
15	MC	22594	19987	A
16	MC	22594	19987	A
17	MC	22594	19987	A
18	CR	22594	19987	A
19	MC	11261	10032	A
20	MC	11261	10032	A
21	TE	11261	10032	A
22	TE	11261	10032	A
23	TE	11261	10032	A
24	CR	11261	10032	A
25	TE	11261	10032	A
26	TE	11261	10032	A
27	TE	11261	10032	A
28	MC	11261	10032	A
29	TE	11261	10032	A
30	CR	11261	10032	A
31	MC	11261	10032	A
32	TE	11261	10032	A
33	MC	11261	10032	A
34	MC	11261	10032	A
35	TE	11261	10032	A
36	CR	11261	10032	A
37	TE	11261	10032	A
38	TE	11261	10032	A
39	MC	11261	10032	A
40	MC	11261	10032	A
41	TE	11261	10032	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
42	CR	11261	10032	A
43	MC	13553	13678	A
44	MC	13553	13678	A
45	TE	13405	13479	A
46	MC	11261	10032	A
47	TE	13553	13678	A
48	CR	13553	13678	A
49	TE	11261	10032	A
50	MC	11261	10032	A
51	TE	11261	10032	A
52	MC	11261	10032	A
53	TE	11261	10032	A
54	CR	11261	10032	A
55	MC	11333	9955	A
56	MC	11333	9955	A
57	TE	11333	9955	A
58	TE	11333	9955	A
59	TE	11333	9955	A
60	CR	11333	9955	A
61	TE	11333	9955	A
62	MC	11333	9955	A
63	MC	11333	9955	A
64	MC	11333	9955	A
65	MC	11333	9955	A
66	CR	11333	9955	A
67	MC	11333	9955	A
68	MC	11333	9955	A
69	MC	11333	9955	A
70	MC	11333	9955	A
71	MC	11333	9955	A
72	CR	11333	9955	A
73	TE	11333	9955	A
74	MC	11333	9955	A
75	MC	11333	9955	A
76	MC	11333	9955	A
77	TE	11333	9955	A
78	CR	11333	9955	A
79	MC	11333	9955	A
80	TE	11333	9955	A
81	MC	11333	9955	A
82	TE	11333	9955	A
83	TE	11333	9955	A
84	CR	11333	9955	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
85	TE	11333	9955	A
86	TE	11333	9955	A
87	TE	11333	9955	A
88	MC	11333	9955	A
89	TE	11333	9955	A
90	CR	11333	9955	A
91	MC	4561	4111	A
92	MC	4561	4111	A
93	MC	4584	4118	A
94	TE	4561	4111	A
95	MC	4561	4111	A
96	CR	4561	4111	A
97	TE	4512	3973	A
98	MC	4512	3973	A
99	MC	4512	3973	A
100	TE	4512	3973	A
101	TE	4512	3973	A
102	CR	4512	3973	A
103	TE	4421	3949	A
104	TE	4421	3949	A
105	TE	4421	3949	-B
106	TE	4421	3949	A
107	TE	4421	3949	A
108	CR	4421	3949	A

Table D.3. MISA Grade 5 White vs. Hispanic/Latino Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	MC	24886	12875	A
2	MC	24886	12875	A
3	TE	24886	12875	A
4	MC	24886	12875	A
5	MC	24886	12875	A
6	CR	24886	12875	A
7	MC	22594	8982	A
8	MC	22594	8982	A
9	TE	22594	8982	A
10	MC	22594	8982	A
11	MC	22594	8982	A
12	CR	22594	8982	A
13	MC	22594	8982	A
14	TE	22594	8982	A
15	MC	22594	8982	A
16	MC	22594	8982	A
17	MC	22594	8982	A
18	CR	22594	8982	A
19	MC	11261	4552	A
20	MC	11261	4552	A
21	TE	11261	4552	A
22	TE	11261	4552	A
23	TE	11261	4552	A
24	CR	11261	4552	A
25	TE	11261	4552	A
26	TE	11261	4552	A
27	TE	11261	4552	A
28	MC	11261	4552	A
29	TE	11261	4552	A
30	CR	11261	4552	A
31	MC	11261	4552	A
32	TE	11261	4552	A
33	MC	11261	4552	A
34	MC	11261	4552	A
35	TE	11261	4552	A
36	CR	11261	4552	A
37	TE	11261	4552	A
38	TE	11261	4552	A
39	MC	11261	4552	A
40	MC	11261	4552	A
41	TE	11261	4552	A
42	CR	11261	4552	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
43	MC	13553	8445	A
44	MC	13553	8445	A
45	TE	13405	8253	A
46	MC	11261	4552	A
47	TE	13553	8445	A
48	CR	13553	8445	A
49	TE	11261	4552	A
50	MC	11261	4552	A
51	TE	11261	4552	A
52	MC	11261	4552	A
53	TE	11261	4552	A
54	CR	11261	4552	A
55	MC	11333	4430	A
56	MC	11333	4430	A
57	TE	11333	4430	A
58	TE	11333	4430	A
59	TE	11333	4430	A
60	CR	11333	4430	A
61	TE	11333	4430	A
62	MC	11333	4430	A
63	MC	11333	4430	A
64	MC	11333	4430	A
65	MC	11333	4430	A
66	CR	11333	4430	A
67	MC	11333	4430	A
68	MC	11333	4430	A
69	MC	11333	4430	A
70	MC	11333	4430	A
71	MC	11333	4430	A
72	CR	11333	4430	A
73	TE	11333	4430	A
74	MC	11333	4430	A
75	MC	11333	4430	A
76	MC	11333	4430	A
77	TE	11333	4430	A
78	CR	11333	4430	A
79	MC	11333	4430	A
80	TE	11333	4430	A
81	MC	11333	4430	A
82	TE	11333	4430	A
83	TE	11333	4430	A
84	CR	11333	4430	A
85	TE	11333	4430	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
86	TE	11333	4430	A
87	TE	11333	4430	A
88	MC	11333	4430	A
89	TE	11333	4430	A
90	CR	11333	4430	A
91	MC	4561	1869	A
92	MC	4561	1869	A
93	MC	4584	1874	A
94	TE	4561	1869	A
95	MC	4561	1869	A
96	CR	4561	1869	A
97	TE	4512	1810	A
98	MC	4512	1810	A
99	MC	4512	1810	A
100	TE	4512	1810	A
101	TE	4512	1810	A
102	CR	4512	1810	A
103	TE	4421	1744	A
104	TE	4421	1744	A
105	TE	4421	1744	A
106	TE	4421	1744	A
107	TE	4421	1744	A
108	CR	4421	1744	A

Table D.4. MISA Grade 8 Male vs Female Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	TE	29865	29889	A
2	MC	29865	29889	A
3	TE	29865	29889	A
4	TE	29865	29889	A
5	MC	29865	29889	A
6	CR	29865	29889	A
7	MC	32527	31555	A
8	TE	32527	31555	A
9	TE	32527	31555	A
10	TE	32527	31555	A
11	MC	32631	31606	A
12	CR	32631	31606	+B
13	MC	32631	31606	A
14	TE	32527	31555	A
15	TE	32527	31555	A
16	TE	32527	31555	A
17	MC	32631	31606	A
18	CR	32631	31606	+B
19	MC	15103	14873	A
20	MC	17869	16590	A
21	TE	15103	14873	A
22	TE	17765	16539	A
23	MC	17869	16590	A
24	CR	17869	16590	+B
25	MC	15103	14873	A
26	MC	17869	16590	A
27	MC	17869	16590	A
28	TE	17765	16539	A
29	MC	17869	16590	A
30	CR	17869	16590	A
31	TE	15103	14873	A
32	TE	15103	14873	A
33	TE	15103	14873	A
34	TE	15103	14873	A
35	TE	15103	14873	A
36	CR	15103	14873	A
37	TE	15103	14873	A
38	TE	15103	14873	A
39	MC	15103	14873	A
40	TE	15103	14873	A
41	TE	15103	14873	A
42	CR	15103	14873	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
43	TE	15103	14873	-B
44	TE	15103	14873	A
45	TE	15103	14873	A
46	TE	15103	14873	A
47	TE	15103	14873	A
48	CR	15103	14873	A
49	TE	15103	14873	A
50	MC	15103	14873	A
51	MC	15103	14873	A
52	TE	15103	14873	A
53	TE	15103	14873	A
54	CR	15103	14873	A
55	TE	14762	15016	A
56	MC	14762	15016	A
57	TE	14762	15016	A
58	TE	14762	15016	A
59	TE	14762	15016	A
60	CR	14762	15016	+C
61	MC	17528	16733	A
62	TE	17424	16682	A
63	TE	17528	16733	A
64	MC	14762	15016	A
65	TE	14762	15016	A
66	CR	17528	16733	A
67	TE	17424	16682	A
68	TE	14762	15016	A
69	TE	14762	15016	A
70	TE	17424	16682	-B
71	MC	14762	15016	A
72	CR	14762	15016	A
73	TE	14762	15016	A
74	MC	14762	15016	A
75	TE	14762	15016	A
76	MC	14762	15016	A
77	TE	14762	15016	A
78	CR	14762	15016	A
79	MC	14762	15016	-B
80	TE	14762	15016	A
81	MC	14762	15016	A
82	MC	14762	15016	A
83	MC	14762	15016	A
84	CR	14762	15016	+B
85	TE	14762	15016	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
86	TE	14762	15016	A
87	MC	14762	15016	A
88	TE	14762	15016	A
89	MC	14762	15016	A
90	CR	14762	15016	A
91	MC	5958	5916	A
92	MC	5958	5916	A
93	MC	5958	5916	A
94	TE	5958	5916	A
95	TE	5958	5916	A
96	CR	5958	5916	+B
97	MC	5916	6125	A
98	TE	5916	6125	A
99	TE	5916	6125	A
100	TE	5916	6125	A
101	MC	5916	6125	A
102	CR	5916	6125	+B
103	MC	6110	6018	-B
104	TE	6075	5984	A
105	TE	6075	5984	A
106	MC	6110	6018	A
107	TE	6075	5984	A
108	CR	6110	6018	+B

Table D.5. MISA Grade 8 White vs. Black/African-American Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	TE	23061	19965	A
2	MC	23061	19965	A
3	TE	23061	19965	A
4	TE	23061	19965	A
5	MC	23061	19965	A
6	CR	23061	19965	A
7	MC	24143	21403	A
8	TE	24143	21403	A
9	TE	24143	21403	A
10	TE	24143	21403	A
11	MC	24188	21479	A
12	CR	24188	21479	A
13	MC	24188	21479	A
14	TE	24143	21403	A
15	TE	24143	21403	A
16	TE	24143	21403	A
17	MC	24188	21479	A
18	CR	24188	21479	A
19	MC	11679	9925	A
20	MC	12806	11439	A
21	TE	11679	9925	A
22	TE	12761	11363	A
23	MC	12806	11439	A
24	CR	12806	11439	A
25	MC	11679	9925	A
26	MC	12806	11439	A
27	MC	12806	11439	A
28	TE	12761	11363	A
29	MC	12806	11439	A
30	CR	12806	11439	A
31	TE	11679	9925	A
32	TE	11679	9925	A
33	TE	11679	9925	A
34	TE	11679	9925	A
35	TE	11679	9925	A
36	CR	11679	9925	A
37	TE	11679	9925	A
38	TE	11679	9925	A
39	MC	11679	9925	A
40	TE	11679	9925	A
41	TE	11679	9925	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
42	CR	11679	9925	A
43	TE	11679	9925	A
44	TE	11679	9925	A
45	TE	11679	9925	A
46	TE	11679	9925	A
47	TE	11679	9925	A
48	CR	11679	9925	A
49	TE	11679	9925	A
50	MC	11679	9925	A
51	MC	11679	9925	A
52	TE	11679	9925	A
53	TE	11679	9925	A
54	CR	11679	9925	A
55	TE	11382	10040	A
56	MC	11382	10040	A
57	TE	11382	10040	A
58	TE	11382	10040	A
59	TE	11382	10040	A
60	CR	11382	10040	A
61	MC	12509	11554	A
62	TE	12464	11478	A
63	TE	12509	11554	A
64	MC	11382	10040	A
65	TE	11382	10040	A
66	CR	12509	11554	A
67	TE	12464	11478	A
68	TE	11382	10040	A
69	TE	11382	10040	A
70	TE	12464	11478	A
71	MC	11382	10040	A
72	CR	11382	10040	A
73	TE	11382	10040	A
74	MC	11382	10040	A
75	TE	11382	10040	A
76	MC	11382	10040	A
77	TE	11382	10040	A
78	CR	11382	10040	A
79	MC	11382	10040	A
80	TE	11382	10040	A
81	MC	11382	10040	A
82	MC	11382	10040	A
83	MC	11382	10040	A
84	CR	11382	10040	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
85	TE	11382	10040	A
86	TE	11382	10040	A
87	MC	11382	10040	A
88	TE	11382	10040	A
89	MC	11382	10040	A
90	CR	11382	10040	A
91	MC	4523	4024	A
92	MC	4523	4024	A
93	MC	4523	4024	A
94	TE	4523	4024	A
95	TE	4523	4024	A
96	CR	4523	4024	A
97	MC	4726	3984	A
98	TE	4726	3984	A
99	TE	4726	3984	A
100	TE	4726	3984	A
101	MC	4726	3984	A
102	CR	4726	3984	A
103	MC	4663	4074	A
104	TE	4621	4063	A
105	TE	4621	4063	A
106	MC	4663	4074	A
107	TE	4621	4063	A
108	CR	4663	4074	A

Table D.6. MISA Grade 8 White vs. Hispanic/Latino Differential Item Functioning Statistics

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
1	TE	23061	9518	A
2	MC	23061	9518	A
3	TE	23061	9518	A
4	TE	23061	9518	A
5	MC	23061	9518	A
6	CR	23061	9518	A
7	MC	24143	11044	A
8	TE	24143	11044	A
9	TE	24143	11044	A
10	TE	24143	11044	A
11	MC	24188	11070	A
12	CR	24188	11070	A
13	MC	24188	11070	A
14	TE	24143	11044	A
15	TE	24143	11044	A
16	TE	24143	11044	A
17	MC	24188	11070	A
18	CR	24188	11070	A
19	MC	11679	4758	A
20	MC	12806	6310	A
21	TE	11679	4758	A
22	TE	12761	6284	A
23	MC	12806	6310	A
24	CR	12806	6310	A
25	MC	11679	4758	A
26	MC	12806	6310	A
27	MC	12806	6310	A
28	TE	12761	6284	A
29	MC	12806	6310	A
30	CR	12806	6310	A
31	TE	11679	4758	A
32	TE	11679	4758	A
33	TE	11679	4758	A
34	TE	11679	4758	A
35	TE	11679	4758	A
36	CR	11679	4758	A
37	TE	11679	4758	A
38	TE	11679	4758	A
39	MC	11679	4758	A
40	TE	11679	4758	A
41	TE	11679	4758	A
42	CR	11679	4758	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
43	TE	11679	4758	A
44	TE	11679	4758	A
45	TE	11679	4758	A
46	TE	11679	4758	A
47	TE	11679	4758	A
48	CR	11679	4758	A
49	TE	11679	4758	A
50	MC	11679	4758	A
51	MC	11679	4758	A
52	TE	11679	4758	A
53	TE	11679	4758	A
54	CR	11679	4758	A
55	TE	11382	4760	A
56	MC	11382	4760	A
57	TE	11382	4760	A
58	TE	11382	4760	A
59	TE	11382	4760	A
60	CR	11382	4760	A
61	MC	12509	6312	A
62	TE	12464	6286	A
63	TE	12509	6312	A
64	MC	11382	4760	A
65	TE	11382	4760	A
66	CR	12509	6312	A
67	TE	12464	6286	A
68	TE	11382	4760	A
69	TE	11382	4760	A
70	TE	12464	6286	A
71	MC	11382	4760	A
72	CR	11382	4760	A
73	TE	11382	4760	A
74	MC	11382	4760	A
75	TE	11382	4760	A
76	MC	11382	4760	A
77	TE	11382	4760	A
78	CR	11382	4760	A
79	MC	11382	4760	A
80	TE	11382	4760	A
81	MC	11382	4760	A
82	MC	11382	4760	A
83	MC	11382	4760	A
84	CR	11382	4760	A
85	TE	11382	4760	A

Item	Item Type	N-Count (Reference)	N-Count (Focal)	Flag
86	TE	11382	4760	A
87	MC	11382	4760	A
88	TE	11382	4760	A
89	MC	11382	4760	A
90	CR	11382	4760	A
91	MC	4523	1902	A
92	MC	4523	1902	A
93	MC	4523	1902	A
94	TE	4523	1902	A
95	TE	4523	1902	A
96	CR	4523	1902	A
97	MC	4726	1909	A
98	TE	4726	1909	A
99	TE	4726	1909	A
100	TE	4726	1909	A
101	MC	4726	1909	A
102	CR	4726	1909	A
103	MC	4663	1914	A
104	TE	4621	1907	A
105	TE	4621	1907	A
106	MC	4663	1914	A
107	TE	4621	1907	A
108	CR	4663	1914	A

Appendix E: Item Level Rasch Statistics

Table E.1. Grade 5 MISA Rasch Difficulties, SEs, Steps, and Fit Statistics

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
1	-2.0865	0.0111	1.10	1.27				
2	-0.1370	0.0091	0.93	0.93				
3	0.4921	0.0096	0.89	0.82				
4	-2.3339	0.0118	0.94	0.93				
5	-1.3961	0.0098	0.99	0.97				
6	0.8795	0.0063	0.92	0.91	-1.8066	0.2517	1.5549	
7	0.3611	0.0095	0.96	0.91				
8	-2.1144	0.0112	0.88	0.75				
9	1.9947	0.0135	1.12	1.26				
10	0.2989	0.0094	0.97	1.00				
11	0.4373	0.0096	1.20	1.41				
12	1.0319	0.0064	0.80	0.79	-1.7628	0.146	1.6168	
13	0.2526	0.0094	1.13	1.24				
14	-2.2126	0.0115	0.99	0.89				
15	0.7230	0.0099	1.32	1.66				
16	-0.4043	0.0091	1.13	1.15				
17	0.7331	0.0100	1.35	1.70				
18	2.2218	0.0082	0.86	0.86	-3.4711	0.3301	3.141	
19	-0.6860	0.0130	1.20	1.31				
20	0.8739	0.0144	0.98	1.27				
21	0.7115	0.0141	1.01	0.97				
22	1.0997	0.0151	0.92	1.06				
23	0.3198	0.0134	0.90	0.84				
24	1.1756	0.0124	1.00	1.00	-3.647	0.804	2.843	
25	1.5868	0.0169	1.02	1.19				
26	1.2123	0.0154	0.85	0.67				
27	0.1755	0.0093	1.23	1.30	-0.6743	0.6743		
28	0.5623	0.0138	1.15	1.37				
29	0.6886	0.0140	1.08	1.25				
30	1.2735	0.0119	1.01	1.00	-1.9293	1.9293		
31	-0.9528	0.0132	0.94	0.92				
32	0.1608	0.0132	1.01	0.97				
33	-0.4192	0.0129	1.00	1.01				
34	-2.2975	0.0165	0.99	1.09				
35	0.9209	0.0146	1.07	1.26				
36	2.8432	0.0114	0.91	0.90	-3.7605	-0.2502	4.0107	
37	0.0175	0.0131	0.97	0.99				
38	-0.8233	0.0097	0.98	0.99	-1.0416	1.0416		
39	-0.1539	0.0130	0.91	0.89				
40	0.8243	0.0143	0.94	0.84				
41	-0.6439	0.0130	0.90	0.88				
42	1.1578	0.0109	0.99	1.01	-0.8539	0.8539		

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
43	-0.8420	0.0131	1.07	1.07				
44	-0.3289	0.0129	1.08	1.13				
45	-0.4211	0.0129	0.77	0.72				
46	0.3971	0.0135	1.00	1.05				
47	-0.3648	0.0101	0.90	0.90	-1.3099	1.3099		
48	-0.8912	0.0086	0.83	0.78	0.0629	-0.0629		
49	-0.9444	0.0132	0.85	0.81				
50	-2.4383	0.0171	0.84	0.65				
51	-0.3096	0.0129	0.86	0.83				
52	-2.5864	0.0178	0.84	0.59				
53	0.2888	0.0133	0.88	0.82				
54	1.6036	0.0133	0.89	0.88	-3.9326	0.9404	2.9923	
55	-0.0869	0.0129	1.02	1.04				
56	-0.7673	0.0129	1.05	1.07				
57	0.8164	0.0143	0.92	0.91				
58	-0.4122	0.0128	0.87	0.83				
59	1.3063	0.0158	0.92	0.88				
60	1.1362	0.0093	0.79	0.79	-2.1156	-0.0219	2.1375	
61	0.0086	0.0091	1.06	1.07	-0.6344	0.6344		
62	0.3357	0.0133	1.26	1.40				
63	-0.3837	0.0128	1.11	1.14				
64	-0.2049	0.0128	0.95	0.94				
65	-0.2838	0.0128	1.03	1.04				
66	1.7647	0.0122	0.86	0.83	-1.9623	1.9623		
67	-0.0311	0.0129	1.20	1.28				
68	-0.5690	0.0128	1.10	1.13				
69	0.0198	0.0130	0.99	1.01				
70	-0.8048	0.0129	0.98	0.96				
71	0.6462	0.0139	1.27	1.58				
72	3.0389	0.0123	0.87	0.86	-4.2839	0.1524	4.1315	
73	1.7099	0.0176	1.15	1.60				
74	-0.5555	0.0128	0.99	1.01				
75	-0.0806	0.0129	0.95	0.97				
76	-0.9471	0.0131	0.92	0.90				
77	-1.1572	0.0100	1.10	1.10	-1.0843	1.0843		
78	0.9721	0.0108	0.87	0.85	-1.264	1.264		
79	-0.0121	0.0129	1.13	1.20				
80	-0.4169	0.0128	0.98	0.98				
81	-0.8935	0.0130	1.08	1.16				

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
82	0.7222	0.0140	1.09	1.18				
83	0.1416	0.0091	0.89	0.86	-0.526	0.526		
84	0.4103	0.0099	0.87	0.84	-1.0394	1.0394		
85	-0.0337	0.0129	0.98	0.95				
86	0.9151	0.0145	0.93	0.85				
87	-1.8855	0.0150	0.85	0.70				
88	-0.4207	0.0128	0.97	0.97				
89	-0.2676	0.0128	0.80	0.75				
90	2.1882	0.0100	0.99	0.97	-3.3664	-0.904	4.2703	
91	-1.2295	0.0212	1.08	1.09				
92	0.7936	0.0223	1.08	1.06				
93	-1.8050	0.0232	0.88	0.77				
94	0.2911	0.0208	1.01	1.06				
95	-1.9243	0.0237	0.80	0.65				
96	0.7663	0.0141	0.93	0.94	-3.0834	-0.0225	1.5034	1.6026
97	2.1978	0.0325	1.13	1.89				
98	1.4162	0.0257	1.27	2.09				
99	1.1368	0.0241	1.01	1.23				
100	2.4266	0.0353	1.02	1.40				
101	1.1275	0.0240	1.16	1.38				
102	1.7890	0.0143	0.91	0.90	-3.2692	-0.9764	0.759	3.4866
103	-0.1808	0.0206	0.85	0.81				
104	0.3259	0.0213	0.97	0.94				
105	0.3787	0.0214	0.97	0.94				
106	2.3878	0.0347	1.02	1.44				
107	3.6962	0.0590	1.01	1.04				
108	0.8849	0.0146	0.95	0.95	-3.2203	-0.1592	1.3302	2.0492

Table E.2. Grade 8 MISA Rasch Difficulties, SEs, Steps, and Fit Statistics

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
1	-0.5193	0.0093	1.10	1.14				
2	0.0080	0.0092	1.00	1.07				
3	-0.3528	0.0093	0.95	0.94				
4	-0.2492	0.0092	1.00	1.00				
5	-1.0080	0.0097	1.03	1.07				
6	1.0426	0.0067	0.84	0.83	-2.5528	-0.2325	2.7852	
7	-0.0784	0.0092	1.41	1.59				
8	-1.3284	0.0101	0.77	0.67				
9	-1.0049	0.0097	0.83	0.76				
10	-2.0015	0.0115	0.97	1.05				
11	-0.4646	0.0093	1.00	0.99				
12	0.9881	0.0064	0.84	0.82	-2.0089	0.4812	1.5277	
13	-0.7798	0.0095	0.92	0.89				
14	1.1952	0.0104	0.94	0.99				
15	1.5876	0.0112	1.11	1.48				
16	0.3096	0.0094	1.06	1.10				
17	0.0292	0.0092	1.13	1.18				
18	2.0911	0.0084	0.91	0.89	-3.5132	0.5741	2.9391	
19	-0.5021	0.0132	0.87	0.82				
20	-1.0360	0.0137	0.93	0.86				
21	0.5251	0.0091	1.19	1.33	-0.2914	0.2914		
22	-0.6031	0.0132	0.92	0.88				
23	-1.2560	0.0141	0.90	0.84				
24	0.1386	0.0098	0.79	0.78	-1.0840	1.0840		
25	0.3423	0.0133	1.22	1.36				
26	-1.4071	0.0144	1.02	1.07				
27	-0.1865	0.0131	1.12	1.17				
28	0.5077	0.0134	1.20	1.36				
29	1.2667	0.0149	1.11	1.45				
30	0.9932	0.0095	0.75	0.73	-2.2610	0.6415	1.6195	
31	1.0755	0.0144	1.12	1.42				
32	0.2049	0.0132	1.01	0.99				
33	-0.4471	0.0131	0.85	0.79				
34	0.1178	0.0105	1.09	1.09	-1.4756	1.4756		
35	-0.1341	0.0131	1.13	1.17				
36	0.1494	0.0098	0.76	0.75	-1.0574	1.0574		
37	1.2545	0.0148	1.11	1.26				
38	-0.4752	0.0132	0.93	0.89				

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
39	-0.5824	0.0132	1.09	1.11				
40	0.5082	0.0134	0.79	0.73				
41	0.7475	0.0138	0.98	1.03				
42	1.3764	0.0091	0.71	0.67	-1.7789	-0.1035	1.8824	
43	-0.7999	0.0134	0.86	0.81				
44	-0.7322	0.0133	0.98	0.99				
45	0.8823	0.0140	0.99	0.94				
46	-0.9729	0.0136	1.04	1.02				
47	-0.1752	0.0131	1.27	1.37				
48	1.3478	0.0105	0.78	0.77	-3.0057	0.2392	2.7666	
49	1.3012	0.0150	1.10	1.14				
50	0.3958	0.0133	1.16	1.30				
51	-1.1560	0.0139	0.92	0.87				
52	1.0449	0.0143	1.24	1.44				
53	0.5505	0.0099	1.23	1.30	-1.0303	1.0303		
54	0.4031	0.0101	0.93	0.92	-1.2000	1.2000		
55	0.3320	0.0132	1.14	1.20				
56	-0.5621	0.0132	1.17	1.26				
57	0.1539	0.0131	0.98	0.96				
58	-1.2947	0.0143	1.03	1.18				
59	0.4290	0.0133	0.98	0.94				
60	1.6823	0.0095	0.89	0.88	-2.6044	-0.6982	3.3026	
61	-0.3229	0.0131	1.06	1.08				
62	-0.2485	0.0095	1.10	1.12	-0.8791	0.8791		
63	0.2996	0.0132	0.93	0.89				
64	-0.5107	0.0132	1.07	1.09				
65	-0.8916	0.0136	0.86	0.82				
66	0.2663	0.0111	0.87	0.87	-1.7149	1.7149		
67	0.1055	0.0131	1.03	1.04				
68	1.4000	0.0112	1.08	1.08	-1.3618	1.3618		
69	0.2598	0.0132	0.96	0.95				
70	-0.5975	0.0133	0.96	0.95				
71	0.4341	0.0133	1.42	1.64				
72	2.1549	0.0124	0.83	0.78	-1.9701	1.9701		
73	0.1874	0.0131	0.87	0.86				
74	0.4420	0.0133	1.21	1.39				
75	-1.0788	0.0139	1.19	1.37				
76	-0.6500	0.0133	1.07	1.07				
77	1.7817	0.0165	1.01	1.07				

Item	*b*	SE	Infit	Outfit	Step 1	Step 2	Step 3	Step 4
78	2.5262	0.0107	0.86	0.81	-2.1552	-0.1741	2.3293	
79	-1.6517	0.0152	0.85	0.74				
80	-0.0311	0.0131	0.87	0.83				
81	-1.4828	0.0147	0.95	0.86				
82	-0.5954	0.0133	1.03	1.03				
83	0.2511	0.0132	1.10	1.19				
84	1.6046	0.0101	0.85	0.85	-2.6527	0.1255	2.5272	
85	-0.4236	0.0131	0.94	0.93				
86	1.1067	0.0119	1.02	1.02	-1.9726	1.9726		
87	-1.7586	0.0155	0.99	0.92				
88	-1.0785	0.0139	0.82	0.76				
89	-0.1006	0.0131	1.23	1.36				
90	0.5441	0.0098	1.08	1.07	-0.9140	0.9140		
91	0.4771	0.0213	1.32	1.65				
92	-0.4097	0.0208	1.08	1.14				
93	-1.5512	0.0235	0.98	0.97				
94	-0.3831	0.0208	1.03	1.08				
95	-0.7413	0.0212	0.89	0.85				
96	1.8301	0.0145	0.76	0.76	-3.4685	-1.1299	1.0585	3.5398
97	-0.5364	0.0209	1.08	1.11				
98	-1.1950	0.0222	0.83	0.72				
99	-0.7753	0.0212	0.93	0.88				
100	-0.5736	0.0209	0.77	0.71				
101	-1.8948	0.0251	0.92	0.80				
102	1.3296	0.0128	0.77	0.76	-2.8709	-0.7993	0.3478	3.3225
103	-0.7728	0.0211	1.05	1.02				
104	0.6732	0.0215	1.13	1.23				
105	1.0685	0.0226	1.16	1.20				
106	-0.7724	0.0211	0.82	0.76				
107	0.4709	0.0211	0.94	1.03				
108	1.3188	0.0123	0.73	0.72	-2.8507	-1.3462	-0.1077	4.3046

Appendix F: Performance Breakdowns on Indicators of Science

Table F.1. Grade 5 Performance Level Results for Earth & Space Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	26	43	31
Female	24	45	30
Male	27	42	31
Hispanic\Latino	36	47	17
Not-Hispanic\Latino	24	43	34
Asian	9	33	58
American Indian or Alaskan Native	35	51	14
Black or African American	39	47	15
Native Hawaiian or Other Pacific Islander	26	50	24
Multiple Indication	22	44	35
White	19	42	39
Economic Disadvantage	39	47	14
Students with Disability	46	40	14

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.2. Grade 5 Performance Level Results for Life Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	22	46	32
Female	19	47	33
Male	24	45	31
Hispanic\Latino	31	50	19
Not-Hispanic\Latino	19	46	35
Asian	8	34	57
American Indian or Alaskan Native	31	52	16
Black or African American	32	51	17
Native Hawaiian or Other Pacific Islander	21	56	23
Multiple Indication	17	48	35
White	16	44	40
Economic Disadvantage	32	51	16
Students with Disability	43	42	15

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.3. Grade 5 Performance Level Results for Physical Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	30	36	34
Female	28	37	35
Male	32	35	34
Hispanic\Latino	43	37	19
Not-Hispanic\Latino	27	36	38
Asian	12	29	59
American Indian or Alaskan Native	43	41	16
Black or African American	44	38	18
Native Hawaiian or Other Pacific Islander	36	36	28
Multiple Indication	23	38	39
White	22	35	43
Economic Disadvantage	44	38	17
Students with Disability	55	28	17

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.4. Grade 5 Performance Level Results for Data and Information

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	29	40	31
Female	28	42	30
Male	31	38	31
Hispanic\Latino	41	41	18
Not-Hispanic\Latino	27	39	34
Asian	11	31	58
American Indian or Alaskan Native	40	45	15
Black or African American	43	42	16
Native Hawaiian or Other Pacific Islander	31	44	25
Multiple Indication	25	41	34
White	22	39	39
Economic Disadvantage	44	42	14
Students with Disability	53	32	14

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.5. Grade 5 Performance Level Results for Model Components

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	19	47	35
Female	17	48	35
Male	20	45	35
Hispanic\Latino	25	52	23
Not-Hispanic\Latino	17	46	37
Asian	6	35	59
American Indian or Alaskan Native	26	56	18
Black or African American	31	52	17
Native Hawaiian or Other Pacific Islander	20	55	26
Multiple Indication	15	47	38
White	11	44	45
Economic Disadvantage	30	53	18
Students with Disability	33	44	22

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.6. Grade 8 Performance Level Results for Earth & Space Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	17	42	40
Female	15	43	42
Male	20	42	39
Hispanic\Latino	26	48	26
Not-Hispanic\Latino	16	41	43
Asian	5	27	68
American Indian or Alaskan Native	21	54	25
Black or African American	29	51	21
Native Hawaiian or Other Pacific Islander	22	51	28
Multiple Indication	13	43	44
White	12	37	51
Economic Disadvantage	29	50	21
Students with Disability	37	43	20

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.7. Grade 8 Performance Level Results for Life Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	20	39	41
Female	18	40	42
Male	22	38	40
Hispanic\Latino	32	43	25
Not-Hispanic\Latino	18	37	45
Asian	6	24	70
American Indian or Alaskan Native	27	48	24
Black or African American	32	46	22
Native Hawaiian or Other Pacific Islander	29	44	27
Multiple Indication	16	39	45
White	14	34	52
Economic Disadvantage	33	46	21
Students with Disability	41	39	20

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.8. Grade 8 Performance Level Results for Physical Science

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	21	38	41
Female	19	39	43
Male	24	37	39
Hispanic\Latino	31	44	25
Not-Hispanic\Latino	19	37	44
Asian	7	25	68
American Indian or Alaskan Native	28	48	25
Black or African American	33	45	22
Native Hawaiian or Other Pacific Islander	31	43	26
Multiple Indication	17	39	44
White	15	34	51
Economic Disadvantage	34	45	21
Students with Disability	42	39	19

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.9. Grade 8 Performance Level Results for Reasoning

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	21	35	43
Female	18	35	47
Male	25	36	40
Hispanic\Latino	32	40	28
Not-Hispanic\Latino	19	34	46
Asian	6	23	71
American Indian or Alaskan Native	27	43	29
Black or African American	33	42	24
Native Hawaiian or Other Pacific Islander	32	38	30
Multiple Indication	17	37	46
White	15	32	54
Economic Disadvantage	34	42	24
Students with Disability	44	35	21

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Table F.10. Grade 8 Performance Level Results for Model Components

	% Within Each Performance Level		
Group	**PL1**	**PL2**	**PL3**
Overall	18	42	40
Female	15	44	41
Male	20	40	40
Hispanic\Latino	28	48	24
Not-Hispanic\Latino	16	41	44
Asian	5	27	68
American Indian or Alaskan Native	23	54	23
Black or African American	27	51	22
Native Hawaiian or Other Pacific Islander	25	46	29
Multiple Indication	15	42	43
White	12	37	51
Economic Disadvantage	29	51	21
Students with Disability	37	43	20

Note: PL1 = Did Not Yet Meet Expectations; PL2 = Approached Expectations; PL3 =Met or Exceeded Expectations

Appendix G: Reliability Breakdowns by Grade, Core Form, and Subgroup

Table G.1. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Total Test for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.917	4.751
Male	14880	0.924	4.748
Hispanic\Latino	4552	0.903	4.788
Not-Hispanic\Latino	25133	0.921	4.755
Asian	2226	0.905	4.688
American Indian or Alaskan Native	931	0.880	4.724
Black or African American	10309	0.899	4.835
Native Hawaiian or Other Pacific Islander	121	0.904	4.739
Multiple Indication	1826	0.916	4.704
White	14272	0.911	4.706
Economic Disadvantage	12289	0.895	4.843
Students with Disability	2955	0.929	4.834

Table G.2. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Total Test for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.905	4.583
Male	14719	0.917	4.551
Hispanic\Latino	4430	0.887	4.626
Not-Hispanic\Latino	25121	0.913	4.569
Asian	2178	0.902	4.533
American Indian or Alaskan Native	985	0.867	4.544
Black or African American	10230	0.879	4.635
Native Hawaiian or Other Pacific Islander	129	0.904	4.573
Multiple Indication	1829	0.903	4.547
White	14200	0.903	4.529
Economic Disadvantage	12008	0.878	4.618
Students with Disability	2984	0.917	4.582

Table G.3. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Physical Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.704	3.671
Male	14880	0.726	3.744
Hispanic\Latino	4552	0.670	3.771
Not-Hispanic\Latino	25133	0.718	3.708
Asian	2226	0.697	3.605
American Indian or Alaskan Native	931	0.612	3.760
Black or African American	10309	0.652	3.877
Native Hawaiian or Other Pacific Islander	121	0.641	3.795
Multiple Indication	1826	0.702	3.672
White	14272	0.690	3.635
Economic Disadvantage	12289	0.648	3.883
Students with Disability	2955	0.740	3.891

Table G.4. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Physical Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.663	3.553
Male	14719	0.691	3.571
Hispanic\Latino	4430	0.615	3.643
Not-Hispanic\Latino	25121	0.681	3.556
Asian	2178	0.687	3.462
American Indian or Alaskan Native	985	0.593	3.657
Black or African American	10230	0.595	3.700
Native Hawaiian or Other Pacific Islander	129	0.607	3.526
Multiple Indication	1829	0.661	3.511
White	14200	0.654	3.485
Economic Disadvantage	12008	0.591	3.688
Students with Disability	2984	0.683	3.647

Table G.5. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Life Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.840	3.512
Male	14880	0.849	3.587
Hispanic\Latino	4552	0.819	3.593
Not-Hispanic\Latino	25133	0.846	3.557
Asian	2226	0.815	3.583
American Indian or Alaskan Native	931	0.789	3.549
Black or African American	10309	0.818	3.608
Native Hawaiian or Other Pacific Islander	121	0.828	3.427
Multiple Indication	1826	0.833	3.576
White	14272	0.825	3.526
Economic Disadvantage	12289	0.814	3.629
Students with Disability	2955	0.863	3.632

Table G.6. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Life Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.808	3.465
Male	14719	0.826	3.465
Hispanic\Latino	4430	0.785	3.443
Not-Hispanic\Latino	25121	0.818	3.481
Asian	2178	0.782	3.769
American Indian or Alaskan Native	985	0.751	3.556
Black or African American	10230	0.773	3.507
Native Hawaiian or Other Pacific Islander	129	0.838	2.953
Multiple Indication	1829	0.798	3.545
White	14200	0.801	3.436
Economic Disadvantage	12008	0.771	3.485
Students with Disability	2984	0.831	3.429

Table G.7. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Earth & Space Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.804	3.528
Male	14880	0.818	3.50
Hispanic\Latino	4552	0.770	3.535
Not-Hispanic\Latino	25133	0.813	3.532
Asian	2226	0.780	4.128
American Indian or Alaskan Native	931	0.724	3.513
Black or African American	10309	0.761	3.511
Native Hawaiian or Other Pacific Islander	121	0.782	3.230
Multiple Indication	1826	0.804	3.546
White	14272	0.794	3.549
Economic Disadvantage	12289	0.748	3.577
Students with Disability	2955	0.824	3.520

Table G.8. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Earth & Space Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.795	3.457
Male	14719	0.820	3.395
Hispanic\Latino	4430	0.757	3.440
Not-Hispanic\Latino	25121	0.811	3.444
Asian	2178	0.791	3.962
American Indian or Alaskan Native	985	0.723	3.577
Black or African American	10230	0.741	3.462
Native Hawaiian or Other Pacific Islander	129	0.758	3.563
Multiple Indication	1829	0.799	3.413
White	14200	0.796	3.462
Economic Disadvantage	12008	0.740	3.463
Students with Disability	2984	0.817	3.342

Table G.9. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Data & Information for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.777	4.070
Male	14880	0.795	4.012
Hispanic\Latino	4552	0.748	3.961
Not-Hispanic\Latino	25133	0.788	4.074
Asian	2226	0.749	4.857
American Indian or Alaskan Native	931	0.703	3.892
Black or African American	10309	0.742	3.876
Native Hawaiian or Other Pacific Islander	121	0.748	3.805
Multiple Indication	1826	0.778	4.077
White	14272	0.763	4.189
Economic Disadvantage	12289	0.725	3.974
Students with Disability	2955	0.797	4.001

Table G.10. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Data & Information for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.774	3.824
Male	14719	0.803	3.732
Hispanic\Latino	4430	0.741	3.706
Not-Hispanic\Latino	25121	0.792	3.807
Asian	2178	0.771	4.499
American Indian or Alaskan Native	985	0.711	3.752
Black or African American	10230	0.725	3.692
Native Hawaiian or Other Pacific Islander	129	0.742	4.003
Multiple Indication	1829	0.780	3.761
White	14200	0.775	3.884
Economic Disadvantage	12008	0.723	3.688
Students with Disability	2984	0.796	3.636

Table G.11. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Model Components, Relationships and Connections for Core 1

Group	N	Coefficient Alpha	SEM
Female	14805	0.833	2.971
Male	14880	0.844	2.956
Hispanic\Latino	4552	0.813	2.874
Not-Hispanic\Latino	25133	0.840	2.991
Asian	2226	0.815	3.778
American Indian or Alaskan Native	931	0.774	2.782
Black or African American	10309	0.806	2.798
Native Hawaiian or Other Pacific Islander	121	0.815	2.707
Multiple Indication	1826	0.827	3.035
White	14272	0.818	3.128
Economic Disadvantage	12289	0.803	2.805
Students with Disability	2955	0.860	2.778

Table G.12. MISA Grade 5 Coefficient Alpha and SEM by Subgroup in Model Components, Relationships and Connections for Core 2

Group	N	Coefficient Alpha	SEM
Female	14832	0.735	3.494
Male	14719	0.756	3.523
Hispanic\Latino	4430	0.704	3.353
Not-Hispanic\Latino	25121	0.747	3.549
Asian	2178	0.736	4.059
American Indian or Alaskan Native	985	0.673	3.332
Black or African American	10230	0.683	3.371
Native Hawaiian or Other Pacific Islander	129	0.717	4.231
Multiple Indication	1829	0.727	3.509
White	14200	0.718	3.688
Economic Disadvantage	12008	0.683	3.337
Students with Disability	2984	0.764	3.363

Table G.13. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Total Test for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.933	4.652
Male	15106	0.940	4.577
Hispanic\Latino	4762	0.922	4.628
Not-Hispanic\Latino	25223	0.937	4.626
Asian	2220	0.927	4.613
American Indian or Alaskan Native	1111	0.900	4.541
Black or African American	10160	0.914	4.637
Native Hawaiian or Other Pacific Islander	227	0.921	4.611
Multiple Indication	1799	0.930	4.584
White	14468	0.931	4.603
Economic Disadvantage	11623	0.915	4.645
Students with Disability	3635	0.938	4.625

Table G.14. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Total Test for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.928	4.709
Male	14768	0.936	4.619
Hispanic\Latino	4762	0.920	4.693
Not-Hispanic\Latino	25027	0.932	4.672
Asian	2223	0.917	4.797
American Indian or Alaskan Native	1153	0.900	4.626
Black or African American	10243	0.913	4.630
Native Hawaiian or Other Pacific Islander	210	0.925	4.588
Multiple Indication	1816	0.923	4.625
White	14144	0.924	4.676
Economic Disadvantage	11502	0.911	4.649
Students with Disability	3594	0.934	4.615

Table G.15. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Physical Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.828	3.498
Male	15106	0.841	3.413
Hispanic\Latino	4762	0.796	3.523
Not-Hispanic\Latino	25223	0.835	3.454
Asian	2220	0.818	3.391
American Indian or Alaskan Native	1111	0.740	3.527
Black or African American	10160	0.778	3.538
Native Hawaiian or Other Pacific Islander	227	0.799	3.508
Multiple Indication	1799	0.823	3.393
White	14468	0.824	3.419
Economic Disadvantage	11623	0.775	3.542
Students with Disability	3635	0.820	3.479

Table G.16. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Physical Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.821	3.572
Male	14768	0.844	3.492
Hispanic\Latino	4762	0.808	3.557
Not-Hispanic\Latino	25027	0.832	3.538
Asian	2223	0.805	3.605
American Indian or Alaskan Native	1153	0.775	3.426
Black or African American	10243	0.788	3.530
Native Hawaiian or Other Pacific Islander	210	0.834	3.609
Multiple Indication	1816	0.814	3.504
White	14144	0.815	3.542
Economic Disadvantage	11502	0.783	3.536
Students with Disability	3594	0.833	3.523

Table G.17. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Life Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.840	3.533
Male	15106	0.855	3.431
Hispanic\Latino	4762	0.814	3.460
Not-Hispanic\Latino	25223	0.848	3.489
Asian	2220	0.834	3.628
American Indian or Alaskan Native	1111	0.777	3.332
Black or African American	10160	0.798	3.462
Native Hawaiian or Other Pacific Islander	227	0.806	3.560
Multiple Indication	1799	0.827	3.515
White	14468	0.838	3.485
Economic Disadvantage	11623	0.799	3.430
Students with Disability	3635	0.850	3.315

Table G.18. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Life Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.785	3.659
Male	14768	0.803	3.678
Hispanic\Latino	4762	0.754	3.753
Not-Hispanic\Latino	25027	0.796	3.657
Asian	2223	0.778	3.787
American Indian or Alaskan Native	1153	0.702	3.638
Black or African American	10243	0.738	3.673
Native Hawaiian or Other Pacific Islander	210	0.786	3.842
Multiple Indication	1816	0.781	3.555
White	14144	0.777	3.623
Economic Disadvantage	11502	0.730	3.698
Students with Disability	3594	0.791	3.711

Table G.19. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Earth & Space Science for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.809	3.695
Male	15106	0.826	3.551
Hispanic\Latino	4762	0.794	3.364
Not-Hispanic\Latino	25223	0.817	3.671
Asian	2220	0.781	4.396
American Indian or Alaskan Native	1111	0.754	3.236
Black or African American	10160	0.778	3.354
Native Hawaiian or Other Pacific Islander	227	0.793	3.422
Multiple Indication	1799	0.802	3.669
White	14468	0.796	3.835
Economic Disadvantage	11623	0.783	3.311
Students with Disability	3635	0.836	3.217

Table G.20. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Earth & Space Science for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.824	3.459
Male	14768	0.842	3.409
Hispanic\Latino	4762	0.817	3.235
Not-Hispanic\Latino	25027	0.831	3.485
Asian	2223	0.783	4.558
American Indian or Alaskan Native	1153	0.776	3.267
Black or African American	10243	0.806	3.162
Native Hawaiian or Other Pacific Islander	210	0.799	3.640
Multiple Indication	1816	0.809	3.414
White	14144	0.813	3.569
Economic Disadvantage	11502	0.803	3.149
Students with Disability	3594	0.847	3.141

Table G.21. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Reasoning for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.736	5.548
Male	15106	0.744	5.501
Hispanic\Latino	4762	0.693	5.314
Not-Hispanic\Latino	25223	0.741	5.591
Asian	2220	0.711	6.764
American Indian or Alaskan Native	1111	0.650	5.001
Black or African American	10160	0.671	5.284
Native Hawaiian or Other Pacific Islander	227	0.692	5.297
Multiple Indication	1799	0.708	5.690
White	14468	0.717	5.829
Economic Disadvantage	11623	0.674	5.246
Students with Disability	3635	0.727	5.364

Table G.22. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Reasoning for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.709	5.821
Male	14768	0.720	5.984
Hispanic\Latino	4762	0.672	5.775
Not-Hispanic\Latino	25027	0.718	5.928
Asian	2223	0.688	7.166
American Indian or Alaskan Native	1153	0.624	5.508
Black or African American	10243	0.650	5.718
Native Hawaiian or Other Pacific Islander	210	0.727	5.726
Multiple Indication	1816	0.700	5.738
White	14144	0.691	6.119
Economic Disadvantage	11502	0.655	5.602
Students with Disability	3594	0.706	5.908

Table G.23. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Model Components for Core 1

Group	N	Coefficient Alpha	SEM
Female	14879	0.805	3.863
Male	15106	0.828	3.618
Hispanic\Latino	4762	0.779	3.408
Not-Hispanic\Latino	25223	0.818	3.812
Asian	2220	0.794	4.883
American Indian or Alaskan Native	1111	0.726	3.289
Black or African American	10160	0.763	3.366
Native Hawaiian or Other Pacific Islander	227	0.789	3.388
Multiple Indication	1799	0.801	3.839
White	14468	0.807	3.967
Economic Disadvantage	11623	0.760	3.393
Students with Disability	3635	0.812	3.360

Table G.24. MISA Grade 8 Coefficient Alpha and SEM by Subgroup in Model Components for Core 2

Group	N	Coefficient Alpha	SEM
Female	15021	0.761	4.150
Male	14768	0.793	4.002
Hispanic\Latino	4762	0.744	3.759
Not-Hispanic\Latino	25027	0.777	4.159
Asian	2223	0.753	5.334
American Indian or Alaskan Native	1153	0.702	3.666
Black or African American	10243	0.720	3.675
Native Hawaiian or Other Pacific Islander	210	0.766	4.011
Multiple Indication	1816	0.764	4.016
White	14144	0.761	4.323
Economic Disadvantage	11502	0.714	3.657
Students with Disability	3594	0.771	3.800