Arizona

Arizona's Instrument to Measure Standards

2016
Technical Report

Submitted to the
Arizona Department of Education
October 2016

ALWAYS LEARNING PEARSON

FOREWORD

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, 2014).

TABLE OF CONTENTS

FIGURES AND TABLES

PART 1: EXECUTIVE SUMMARY

This document provides information regarding processes and procedures implemented in the Spring 2016 Arizona's Instrument to Measure Standards (AIMS) assessments for the development of tests, analysis of data, calibration, scoring, and scaling. This document also describes the results of the Spring 2016 AIMS assessments. The technical information in this report is intended for those who evaluate tests, interpret scores, or use test results in making educational decisions.

This document also provides information relevant to the *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, National Council on Measurement in Education, 1999). The *Standards* were revised in 2014, *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, National Council on Measurement in Education, 2014). The beginning of each part of this technical report will list the different standards addressed in each edition. Part 1 (the Executive Summary) of the technical report addresses 1999 standards 2.7, 3.2, 3.3, 6.3, 6.4, 6.15, and 13.6, and 2014 standards 4.1, 4.2, 7.0, 7.2, 7.4, and 12.9.

Structure of AIMS Technical Report

The Spring 2016 AIMS assessments were designed and developed to provide fair and accurate ability scores that support appropriate, meaningful, and useful educational decisions. In addition to the evidence provided in Part 2 (Involvement of Arizona Educators), additional validity evidence may be found in the following parts as described: Part 3 (Test Design), Part 4 (Test Development), Part 5 (Test Administration), Part 6 (Classical Item Analysis), Part 7 (Calibration, Scaling and Equating), Part 8 (Reliability), and Part 10 (Classification). As the technical report progresses chapter by chapter, it moves through the phases of the testing cycle. Each part of the technical report details the procedures and processes applied in the creation of AIMS, as well as their results. Each part also highlights the meaning and significance of the procedures, processes, and results in terms of content and construct validity and the relationship to the *Standards.*

The Spring 2016 AIMS tests were administered in science to students in grade 4, 8, and high school. This was the sixth year that Grades 4, 8, and high school were administered science. Science tests remain mandatory for all students in these grades. Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them as eligible for an alternate assessment, AIMS A, are excluded from AIMS Science testing. The AIMS Science tests consist of multiple-choice items, which are written entirely by Arizona teachers.

The AIMS assessments are designed to measure Arizona students' performance on the Arizona content standards. All AIMS Science tests are written to Arizona content standards approved by the State Board on May 24, 2004, and updated on March 10, 2005.

Based on the input of Arizona educators' review of the content standards, a design was derived, developed, administered, and scored. The present technical report documents all aspects of the testing cycle in the subsequent chapters. A brief content summary of the report is provided below.

Involvement of Arizona Educators

- Part 2 of this report describes the involvement of Arizona educators in test development and the work they performed to help prepare the 2016 AIMS Science assessments.

Test Design and Development

- Part 3 of this report describes the test design and the item development process. It provides the content frameworks and the blueprints upon which all of the AIMS tests are based. This section also includes descriptions and the structure of each AIMS test administered in the 2015-2016 academic year.
- Part 4 of this report provides a chronological description of the passage, stimulus, and item development process including modification of specifications, committee passage/stimulus reviews, item content and sensitivity reviews, data analysis and item selection committees, and customer and contractor reviews to guarantee a quality, error-free product.

Administration

- Part 5 briefly describes test administration, security, and the written procedures available to all test administrations and school personnel and the accommodations that were available to eligible students while testing on Spring 2016 AIMS Science. This section also describes instituted proceedures to ensure the security and standardization of test administrations.

Data for Operational Analysis

- Part 6 describes the data used for calibration and scaling of the Spring 2016 AIMS Science and presents classical test statistics and item analysis statistics. This section includes steps taken to ensure the valid calibration and scaling of these tests as well as the resulting measures of internal consistency.

Calibration, Scaling, and Equating

- Part 7 reviews calibration, equating, scoring methods, and calibration results. This section includes considerations for the evaluation of the calibration results and anchor items. It also presents the relationships between raw scores and scale score through scoring tables and scaling results including the standard error of measurement.

Test Results

- Part 8 summarizes information about the results of the Spring 2016 AIMS Science administration. The test results for different ethnic backgrounds and special program membership status are provided. Students in cohorts 2018 and 2019 are included separately in the high school science results. The results presented include frequency distributions and longitudinal comparisons of scale scores.

Validity Evidence

- Part 9 reviews the main validity issues discussed in all prior chapters and provides additional validity evidence supporting the AIMS Science tests. The evidence presented includes the results of an analysis of differential item functioning.

Classification

- Part 10 provides information regarding classification consistency and accuracy when students were classified into proficiency categories. These analyses used cut scores that were determined during standard setting and adopted by the State Board of Education.

PART 2: INVOLVEMENT OF ARIZONA EDUCATORS AT ALL LEVELS

Part 2 of the technical report addresses the involvement of Arizona educators in test development. This part of the technical report addresses standard 3.5 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), and standard 4.6 in the 2014 edition.

Typically several committees met in preparation for AIMS Science assessments. These committees included teachers, curriculum specialists, and administrators from across the state and were an integral part of both the AIMS test development processes and AIMS results interpretation. However, starting spring 2015, because ADE had developed a sufficient number and quality of items in the Science item bank, they chose to change their process for the development of the spring 2016 test.

The Spring 2016 AIMS Science called for administering one operational test for grade 4, 8, and high school in science. All items available for placement on an operational test had been previously field-tested, and passed through multiple educator committees, including development, bias and content, and data analysis meetings, prior to the start of the development of these tests. The AIMS Science tests for the spring 2016 administration were built by trained ADE staff, most of whom also held Arizona teacher certificates, in the summer of 2014 to match the blueprint, difficulty distribution, and include as many higher Depth-of-Knowledge (DOK) items as possible.

PART 3: TEST DESIGN

Part 3 of the technical report provides information regarding test design. The following AERA/APA/NCME *Standards* from the 1999 edition are addressed: 1.2, 1.6, 3.1, 3.2, 3.3, 3.11, 6.4, 6.15, 13.3, and 13.5. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this part of the technical report are 1.1, 1.11, 4.0, 4.1, 4.2, 4.12, 7.0, 7.2, 12.4, and 12.8.

3.1 Content Standards

The AIMS assessments are designed to measure performance on the Arizona content standards adopted in March 2005 for science. These standards are organized by strand, concept, and performance objective. The AIMS Science test blueprints are based on the concepts and strands of the Arizona content standards, presented in Figures 3.1.1 through 3.1.3.

Figure 3.1.1
Arizona Science Concepts and Strands – Grade 4

Strand 1: Inquiry Process
- **Concept 1: Observations, Questions, and Hypotheses**
- **Concept 2: Scientific Testing (Investigating and Modeling)**
- **Concept 3: Analysis and Conclusions**
- **Concept 4: Communication**

Strand 2: History and Nature of Science
- **Concept 1: History of Science as a Human Endeavor**
- **Concept 2: Nature of Scientific Knowledge**

Strand 3: Science in Personal and Social Perspectives
- **Concept 1: Changes in Environments**
- **Concept 2: Science and Technology in Society**

Strand 4: Life Science
- **Concept 1: Characteristics of Organisms**
- **Concept 2: Life Cycles**
- **Concept 3: Organisms and Environments**
- **Concept 4: Diversity, Adaptation, and Behavior**

Strand 5: Physical Science
- **Concept 1: Properties of Objects and Materials**
- **Concept 2: Position and Motion of Objects**
- **Concept 3: Energy and Magnetism**

Strand 6: Earth and Space Science
- **Concept 1: Properties of Earth Materials**
- **Concept 2: Earth's Processes and Systems**
- **Concept 3: Changes in the Earth and Sky**

Figure 3.1.2
Arizona Science Concepts and Strands – Grade 8

Strand 1: Inquiry Process

- **Concept 1: Observations, Questions, and Hypotheses**
- **Concept 2: Scientific Testing (Investigating and Modeling)**
- **Concept 3: Analysis and Conclusions**
- **Concept 4: Communication**

Strand 2: History and Nature of Science

- **Concept 1: History of Science as a Human Endeavor**
- **Concept 2: Nature of Scientific Knowledge**

Strand 3: Science in Personal and Social Perspectives

- **Concept 1: Changes in Environments**
- **Concept 2: Science and Technology in Society**

Strand 4: Life Science

- **Concept 1: Structure and Function in Living Systems**
- **Concept 2: Reproduction and Heredity**
- **Concept 3: Populations of Organisms in an Ecosystem**
- **Concept 4: Diversity, Adaptation, and Behavior**

Strand 5: Physical Science

- **Concept 1: Properties and Changes of Properties in Matter**
- **Concept 2: Motion and Forces**
- **Concept 3: Transfer of Energy**

Strand 6: Earth and Space Science

- **Concept 1: Structure of the Earth**
- **Concept 2: Earth's Processes and Systems**
- **Concept 3: Earth in the Solar System**

Figure 3.1.3
Arizona Science Concepts and Strands – High School

Strand 1: Inquiry Process

- **Concept 1: Observations, Questions, and Hypotheses**
- **Concept 2: Scientific Testing (Investigating and Modeling)**
- **Concept 3: Analysis, Conclusions, and Refinements**
- **Concept 4: Communication**

Strand 2: History and Nature of Science

- **Concept 1: History of Science as a Human Endeavor**
- **Concept 2: Nature of Scientific Knowledge**

Strand 3: Science in Personal and Social Perspectives

- **Concept 1: Changes in Environments**
- **Concept 2: Science and Technology in Society**
- **Concept 3: Human Population Characteristics**

Strand 4: Life Science

- **Concept 1: The Cell**
- **Concept 2: Molecular Basis of Heredity**
- **Concept 3: Interdependence of Organisms**
- **Concept 4: Biological Evolution**
- **Concept 5: Matter, Energy, and Organization in Living Systems (Including Human Systems)**

Strand 5: Physical Science

- **Concept 1: Structure and Properties of Matter**
- **Concept 2: Motions and Forces**
- **Concept 3: Conservation of Energy and Increase in Disorder**
- **Concept 4: Chemical Reactions**
- **Concept 5: Interactions of Energy and Matter**

Strand 6: Earth and Space Science

- **Concept 1: Geochemical Cycles**
- **Concept 2: Energy in the Earth System (Both Internal and External)**
- **Concept 3: Origin and Evolution of the Earth System**
- **Concept 4: Origin and Evolution of the Universe**

3.2 Test Blueprints

A test blueprint designates the percentage of items that should measure each strand and concept. AIMS assessments in science were designed in accordance with the following blueprints in Tables 3.2.1 through 3.2.3. Further discussion of item selection to match the blueprints is included in Part 4 of this report.

Table 3.2.1
AIMS Blueprint for Science Grade 4

AIMS Science
Grade 4 Test Blueprint

Strand/Concept	% of Test	# of Items
Strand 1: Inquiry Process	**33.3%**	
Concept 1: Observations, Questions, and Hypotheses	11.1%	6
Concept 2: Scientific Testing (Investigating and Modeling)	11.1%	6
Concept 3: Analysis and Conclusions	11.1%	6
Concept 4: Communications		
Strand 2: History and Nature of Science	**11.1%**	
Concept 1: History of Science as a Human Endeavor	11.1%	6
Concept 2: Nature of Scientific Knowledge		
Strand 3: Science in Personal and Social Perspectives	**11.1%**	
Concept 1: Changes in Environments	11.1%	6
Concept 2: Science and Technology in Society		
Strand 4: Life Science	**11.1%**	
Concept 1: Characteristics of Organisms	11.1%	6
Concept 3: Organisms and Environments		
Concept 4: Diversity, Adaptations, and Behavior		
Strand 5: Physical Science	**11.1%**	
Concept 3: Energy and Magnetism	11.1%	6
Strand 6: Earth and Space Science	**22.2%**	
Concept 2: Earth's Processes and Systems	11.1%	6
Concept 3: Changes in the Earth and Sky	11.1%	6
According to the Science Standard, the following Strands and Concepts do not have Performance Objectives for Grade 4: **Strand 4: Life Science, Concept 2** (Life Cycles); **Strand 5: Physical Science, Concept 1** (Properties of Objects and Materials) and **Concept 2** (Position and Motion of Objects); **Strand 6: Earth and Space Science, Concept 1** (Properties of Earth Materials).		**54**

Source: http://www.azed.gov/assessment/files/2014/06/science-blueprint-with-item-counts-11-10-09.pdf

Table 3.2.2
AIMS Blueprint for Science Grade 8

AIMS Science
Grade 8 Test Blueprint

Strand/Concept	% of Test	# of Items
Strand 1: Inquiry Process	**34.5%**	
Concept 1: Observations, Questions, and Hypotheses	10.3%	**6**
Concept 2: Scientific Testing (Investigating and Modeling)	6.9%	**4**
Concept 3: Analysis and Conclusions	10.3%	**6**
Concept 4: Communications	6.9%	**4**
Strand 2: History and Nature of Science	**10.3%**	
Concept 1: History of Science as a Human Endeavor Concept 2: Nature of Scientific Knowledge	10.3%	**6**
Strand 3: Science in Personal and Social Perspectives	**10.3%**	
Concept 1: Changes in Environments Concept 2: Science and Technology in Society	10.3%	**6**
Strand 4: Life Science	**13.8%**	
Concept 2: Reproduction and Heredity Concept 4: Diversity, Adaptations, and Behavior	13.8%	**8**
Strand 5: Physical Science	**31.0%**	
Concept 1: Properties and Changes of Properties in Matter	17.2%	**10**
Concept 2: Motion and Forces	13.8%	**8**
According to the Science Standard, the following Strands and Concepts do not have Performance Objectives for Grade 8: **Strand 4: Life Science**, **Concept 1** (Structure and Function in Living Organisms) and **Concept 3** (Populations of Organisms in an Ecosystem); **Strand 5: Physical Science, Concept 3** (Transfer of Energy).		**58**

Source: http://www.azed.gov/assessment/files/2014/06/science-blueprint-with-item-counts-11-10-09.pdf

Table 3.2.3
AIMS Blueprint for Science High School

AIMS Science

High School Test Blueprint

Strand/Concept	% of Test	# of Items
Strand 1: Inquiry Process	**33.8%**	
Concept 1: Observations, Questions, and Hypotheses	9.2%	**6**
Concept 2: Scientific Testing (Investigating and Modeling)	9.2%	**6**
Concept 3: Analysis, Conclusions, and Refinements	9.2%	**6**
Concept 4: Communications	6.2%	**4**
Strand 2: History and Nature of Science	**9.2%**	
Concept 1: History of Science as a Human Endeavor	9.2%	**6**
Concept 2: Nature of Scientific Knowledge		
Strand 3: Science in Personal and Social Perspectives	**10.8%**	
Concept 1: Changes in Environments	10.8%	**7**
Concept 2: Science and Technology in Society		
Concept 3: Human Population Characteristics		
Strand 4: Life Science	**46.2%**	
Concept 1: The Cell	9.2%	**6**
Concept 2: Molecular Basis of Heredity	9.2%	**6**
Concept 3: Interdependence of Organisms	9.2%	**6**
Concept 4: Biological Evolution	9.2%	**6**
Concept 5: Matter, Energy, and Organization in Living Systems (Including Human Systems)	9.2%	**6**
		65

Source: http://www.azed.gov/assessment/files/2014/06/science-blueprint-with-item-counts-11-10-09.pdf

3.3 Description of 2016 AIM Tests

The test blueprints were used with the processes described in detail in Part 4 to develop all AIMS tests administered in 2016. The resulting test configurations are as follows.

3.3.1 Science for Grades 4, 8, and High School

The 2016 AIMS Science tests consisted of one operational form with 54 multiple-choice items on the grade 4 test, 58 multiple-choice items on the grade 8 test, and 65 multiple-choice items on the high school test. All multiple-choice items were developed by Arizona teachers. The scale scores for each test range from 200 to 800 and all items on each test reported to a criterion-referenced score. No norm-referenced items were included on any of the science tests. Table 3.3.1.1 displays the structure of the science tests.

Table 3.3.1.1
Spring 2016 AIMS Test Structure of Science

Grade	SC FT	SC OP	TOTAL ITEMS ON TEST	Anchor
4	N/A	54	54	21
8	N/A	58	58	23
HS	N/A	65	65	18

*Grades 4, 8, and HS science each had no field test items on either the spring 2015 or 2016 tests.

3.3.2 AIMS Score Ranges

Raw score and scale score ranges of 2016 AIMS Science in grades 4, 8, and high school are presented in Table 3.3.2.1.

Table 3.3.2.1
Raw Score and Scale Score ranges of 2016 AIMS Assessments

Content	Grade	Raw Score Range	Scale Score range
Science	4	0-54	200-800
	8	0-58	200-800
	HS	0-65	200-800

PART 4: TEST DEVELOPMENT

Part 4 of the technical report provides a summary of the development activities that occurred for the Spring 2016 AIMS Science tests. Information is provided relating to the following topics as they pertain to AIMS:

- a discussion of the AIMS test development and editing process;
- a description of the use of previously created AIMS item specifications;
- a description of the AIMS item editing procedures;
- a description of the data analysis committee procedures;
- a description of the AIMS item selection committee meetings; and

A comprehensive, multi-segment development process guides the development of assessment materials. The following section outlines this process in general terms. The remainder of Part 4 provides details of how these processes were implemented in Arizona. This section of the technical report addresses the following AERA/APA/NCME *Standards* from the 1999 edition: 1.6, 3.1, 3.5, 3.6, 3.7, 3.9, 3.11, 3.16, 6.4, 6.15, 7.3, 7.4, 7.7, 13.3, and 13.5, and Standards 1.11, 3.2, 3.6, 4.0, 4.6, 4.7, 4.8, 4.10, 4.12, 7.0, 7.2, 12.4, 12.8 in the new edition of *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

4.1 AIMS Test Development and Editing Process

4.1.1 Test Development Process

Test development for the 2016 test administration began with the planning meeting held in Phoenix, January 16-18, 2013. During this meeting, the project deliverables were defined, such as number of forms, answer documents, test administration manuals, test coordinator manuals, test interpretation guides, and materials to support special accommodations, including Braille and large print books. The actual test form design was unchanged from the previous year. The ancillary materials were modified and all modifications were discussed and shared among all team members to ensure understanding.

4.1.2 Documents and Materials Development

Following definition of project deliverables, Pearson's entire test development team reviewed the blueprints, item specifications, and the *ADE Style Guide* to ensure that the 2016 assessment would meet all of the required, previously-developed criteria.

4.1.3 Item Writing for Science

The no new items were developed for field testing in the Spring 2016 AIMS Science assessments since there were sufficient items of sufficient quality in the AIMS item bank.

4.1.4 Quality Reviews

ADE and Pearson personnel implemented a series of quality review checks at various stages of production to ensure all AIMS materials were error free.

ADE first reviewed each component at a relatively early stage of forms production. Items were

compared to the way they were presented to the content/bias review committee to be sure no unauthorized changes had been introduced. Answer keys were checked. All changes were approved in writing by ADE.

A smooth AIMS test administration requires that all test materials, including test books, answer documents, and directions to students and test coordinators align with each other. Therefore, Pearson and ADE conducted a review of all materials as the second quality check.

Prior to creation of proofs (blueline stage), Pearson performed a Final Forms review. The purpose of the Final Forms review was to ensure that all publishable products met ADE's high quality standards and expectations.

After Pearson conducted their Final Forms review, all test forms were again submitted to ADE for review. All final forms and documents were reviewed and approved by ADE content specialists.

4.2 Pool of Items Used for Test Construction

4.2.1 Item Specifications

The item specifications were developed by Pearson and ADE in May 2009. The item specifications provide a definition of what is tested by each Performance Objective (PO) and, where needed, provide clarification of the PO statements, the content limits, and the stimulus and response attribute descriptions. Taken together, these help to inform instruction by explaining in detail what each PO means at each grade level and by describing how each PO is to be tested.

4.2.2 Data Analysis

The most recent AIMS Data Analysis workshop was conducted for Science in June 2014. Primary responsibility for conducting this workshop rested with ADE. The primary purpose of the Data Analysis meeting was to examine the item data generated for field tested items within the Spring 2014 AIMS Science test. Each item was assigned a status code to be included with the item information in the item bank, and determine each item's eligibility for possible selection as an operational item starting in spring 2015.

ADE staff were trained on how to interpret basic statistical concepts related to item data including *p*-values, Rasch values, infit/outfit, point biserial correlations, response distributions and race/ethnicity and gender differential item functioning (DIF) flags, omit rates, and population counts.

Items that measured the content they were intended to measure and whose statistics were within acceptable limits were assigned Item Available (IA) status. These items were eligible for selection as operational items. Throughout the meeting, content was stressed as the deciding factor over statistics for items to attain IA status. Across all grades in Science, approximately 87% of the items received IA status.

Items whose statistics indicated a fixable problem and that defined where the items could be improved were assigned Re-Field Test (RFT) status. These items would be revised during future item writing workshops and would be re-field tested in future assessments. None of items reviewed was coded RFT.

Items whose statistics indicated they would not function fairly and reliably were rejected and assigned Do Not Use (DNU) status. These items were removed from consideration as operational items. Across the content and grade levels, about 13% of the items were assigned DNU status.

Table 4.1 shows the number and portion of items classified into each category during the June 2014 Data Analysis workshop by grade level.

Table 4.1
Items Given Special Codes

Content Area	Grade	Items Reviewed	Items Assigned IA * Status		Items Assigned RFT* Status		Items Assigned DNU* Status	
Science	4	40	36	90%	0	0%	4	10%
	8	40	34	85%	0	0%	6	15%
	HS	40	34	85%	0	0%	6	15%
Science Total		**120**	**104**	**87%**	**0**	**0%**	**16**	**13%**

Note:* Item Available (IA) - Re-field Test (RFT) - Do Not Use (DNU)

4.2.3 AIMS Item Selection

The Item Selection meeting for Spring 2016 AIMS Science was conducted by ADE staff in July 2014. The purpose of the Item Selection meeting was to select items to place on test forms that would produce valid and reliable scores using items from previous test administrations as well as items from the 2014 field test administration that had been designated as "item accepted" (IA). Two sets of criteria primarily guided the selection of AIMS items: content representation and statistical requirements. In addition, the committee members were encouraged to select items with high-level DOKs that most reflect the expectation of skills represented within the Arizona Science Standard.

All of the items in the item bank that were available and eligible for selection as operational items in spring 2014 were displayed in grade level and content area item pool tables. With minor exceptions, the pool consisted of items field tested in 2008 through 2013. The items field tested in spring 2014 were also available in the data analysis materials. The item pool tables for the science committee were arranged by Performance Objective. All tables could also be sorted according to any of the columns, making them extremely useful tools for searching for items with specific characteristics. These items formed the pool for item selection. Item images could be viewed electronically via the item bank. The meeting room was equipped with a laptop with access to the item bank and a projection screen so that the entire group could view items at the same time.

Each entry on the table contained identification numbers, content alignment information (Strand, Concept, Performance Objective), the most recent test administration, and the most current statistical information about that item (*p*-value, Rasch values, point biserial, differential item functioning summary flags, Rasch model fit statistics, and the percent of students who omitted the item). Participants were given training to interpret these statistics and statistical guidelines for test selection. These guidelines included a target difficulty level for each test. Specifically, a target mean and range of selected item *p*-values, as well as a suggested distribution for the item *p*-values was provided for each grade/subject combination. Careful adherence to the specified distribution of *p*-values guaranteed students a reasonable opportunity to do well on a test that would be neither too easy nor too hard.

In addition to selecting items within specific *p*-values ranges, committee members were also asked to select items with item discriminations that indicate that getting the item correct is reasonably correlated with performance on the entire test (i.e., preferably item correlations greater than 0.3) and do not exhibit the potential for item bias (i.e., the items should not be flagged using various differential item functioning statistics).

Content considerations were addressed by the test blueprints. Careful adherence to the blueprints guaranteed the tests would validly measure the construct of science as represented in the Arizona Science Standard, maintain consistency, link to instruction, and allow for selection of items from

different performance objectives within each concept. Substantial variance from the test blueprint could alter the test alignment and thus the validity of the scores being reported. Items were selected to represent the significant content categories specified in the test blueprint in the same proportion as the content categories represented in the test blueprint.

Prior to the Item Selection Committee meeting, ADE selected an anchor set of items upon which the operational forms would be constructed. The anchor set consisted of items that had been operational at least the previous year (during the spring 2014 test administration). Regardless of the grade, each anchor set was carefully selected to meet statistical criteria and to proportionally represent the blueprint. Anchor sets were finalized by ADE prior to the item selection workshop.

To facilitate the selection process and to guarantee that the proper number and proportion of items would be selected, participants were provided with item pool tables and item replacement tables. Table 4.2 shows a sample of an item pool table and the available data considered by the Item Selection Committee in its selection of replacement items. An analysis of differential item functioning is performed for every administration. The latest values are included in the item pool tables for each grade/content area and provided to participants in the Item Selection Committee. Table 4.3 is a sample portion of the Item Replacement Table used by the participants to note their replacement requirements for grade 4 Science and to capture proposed items to be used on the spring 2016 assessment. This sample table shows the portion relevant to Strand 1 Concept 1 only. The entire table included all strands and concepts. This sample table shows the portion of columns relevant to spring 2015 and spring 2016. The information in the first column shows the blueprint requirements for Strand 1, Concept 1 – six of the 54 operational items that should be covered by items from Strand 1, Concept 1 in the grade 4 Science test.

The set of columns labeled Spring 2016 New Operational Items include all of the AZ items covering Strand 1 Concept 1 that were in the spring 2015 test. The set of columns labeled Spring 2016 New Operational Items show the items that were retained from the spring 2015 or prior administrations (highlighted in blue). These retained items were designated as anchor items. During item selection for spring 2016, the participants' tasks were to retain anchor items, if possible, and select items to fill in any gaps in blueprint coverage. As the participants considered each option based on content and difficulty, they could refer to the Item Pool Table to determine if the statistical considerations were being met and to the item bank to see the actual items.

As selections were made, they were recorded on item replacement tables. These tables were loaded onto computers and projected for group discussion. These tables provided a running record of the selections and further helped to guarantee blueprint coverage. Table 4.4 shows a sample of the *p*-value target distribution table and graph used by the committees. Note that this table and graph are displayed as if items were in the process of being selected. These tables were completed for all selections and were subject to approval by both ADE and Pearson's content and psychometric departments.

Table 4.5 shows the numbers of AIMS Science items that were selected for each grade. All selections were approved by Pearson content and psychometric staff and ADE staff.

Table 4.2
Sample Grade 4 Science Item Pool Table

Page 1

Row	AZID	Subject	Grade	Status	Stimulus Title	Strand	Concept	Perf. Obj.	DOK	Year 2006	Year 2007	Year 2008	Year 2009	Year 2010	Year 2011	Year 2012	Year 2013	Year 2014	Recent Year	Item No.
1	44144025	Science	4	New	Circuit Study	5	3	2	2									FT	2014	7
2	44144005	Science	4	New	Soil Erosion	6	2	3	2									FT	2014	59
3	44144047	Science	4	New		1	1	1	2									FT	2014	6
4	44144049	Science	4	New		1	1	2	2									FT	2014	6
5	44144051	Science	4	New		1	1	2	2									FT	2014	6
6	44144055	Science	4	New		1	1	2	2									FT	2014	7
7	44144041	Science	4	New		1	1	2	4											
8	44144054	Science	4	New		1	1	3	2											
9	44144043	Science	4	New		1	1	3	2									FT	2014	7
10	44144046	Science	4	New		1	1	3	2									FT	2014	7

Page 2

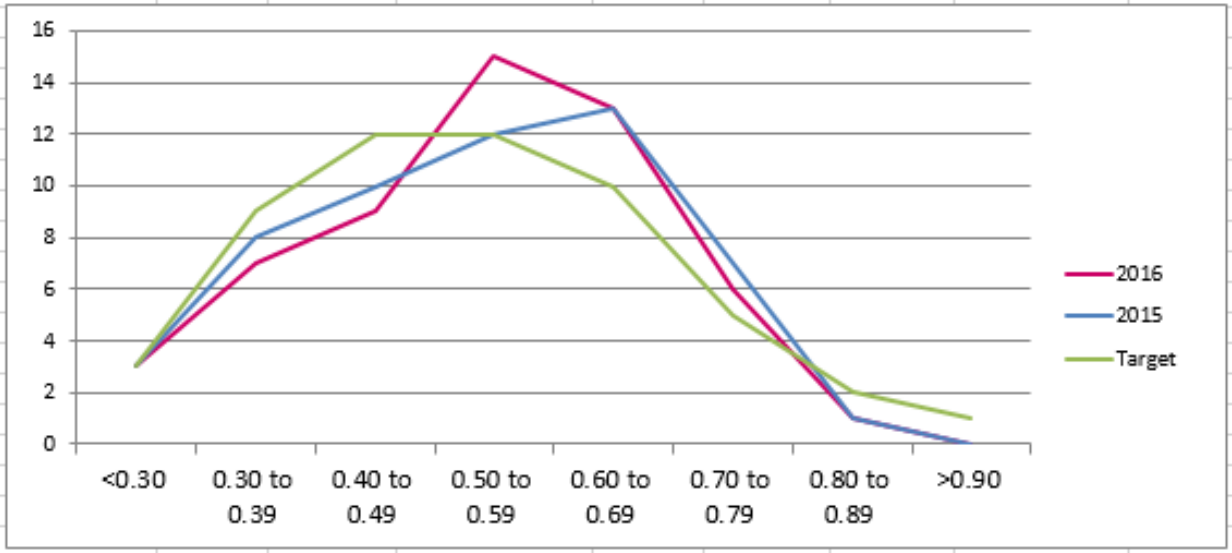
Row	N Count	Rasch	PVal	Flag PVal	PT Bis.	Flag PTBIS	Male vs Female Bias Flag	Non-Hispanic vs Hispanic Bias Flag	White vs Black Bias Flag	White vs Hispanic Bias Flag	White vs AmIn Bias Flag	White vs Asian Bias Flag	White vs Hawi/PaclsIr Bias Flag	White vs Multiracial Bias Flag	Dist A	Dist B	Dist C	Dist D	Omit
1	20638	1.116	0.46		0.24	*	A	A	A		A	A	A	A	38.3	45.8	9.5	6.4	0.0
2	20339	-0.411	0.76		0.50		A	A	A		A	A	A	A	76.1	7.5	7.4	8.8	0.1
3	20500	1.850	0.30	*	0.28	*	A	A	A		A	A	A	A	13.5	12.4	29.7	44.4	0.0
4	20340	1.759	0.33		0.21	*	A	A	A		A	A	A	A	15.6	34.2	16.9	33.3	0.0
5	20638	2.455	0.21	*	0.13	*	A	A	A		A	A	A	A	12.2	7.8	58.9	21.1	0.0
6	20339	1.205	0.45		0.26	*	A	A	A		A	A	A	A	15.2	14.7	25.1	44.9	0.1
7																			
8																			
9	20500	-0.283	0.72		0.44		A	A	A		A	A	A	A	4.9	5.4	72.4	17.3	0.0
10	20340	-0.268	0.74		0.52		A	A	A		A	A	A	A	74.0	14.8	4.5	6.6	0.0

Table 4.3
Sample Grade 4 Science Item Replacement Table

AZ AIMS Grade 4 Spring 16 Operational Item Replacement Plan for Science																	
# of Items Required per Blueprint	Strand	Concept	Spring 15 - New Operational Items Selections								Spring 16 - New Operational Items Selections							
			Actual # of Items	PO	AZID	Passg ID	P-VALUE	Rasch	PtBis	DOK	Actual # of Items	PO	AZID	Passg ID	P-VALUE	Rasch	PtBis	DOK
6	1	1	6	1.1.1	3514444	0	0.399	1.3943	0.373	1	6	1.1.1	3514444	0	0.399	1.3943	0.373	1
	1	1		1.1.3	3514583	0	0.62	0.5167	0.402	3		1.1.3	3514583	0	0.62	0.5167	0.402	3
	1	1		1.1.1	3514504	0	0.519	0.816	0.354	1		1.1.1	3514504	0	0.519	0.816	0.354	1
	1	1		1.1.3	44114434	Electricity and Magnetism	0.674	0.1057	0.532	2		1.1.3	44114434	Electricity and Magnetism	0.674	0.1057	0.532	2
	1	1		1.1.2	44114447	Volcanoes	0.736	-0.2494	0.545	1		1.1.2	44114447	Volcanoes	0.736	-0.2494	0.545	1
	1	1		1.1.3	44104325	0	0.575	0.624	0.451	3		1.1.2	44114318	0	0.445	1.2724	0.335	2

Table 4.4
Sample *P*-Value Target Table and Graph

		<0.30	0.30 to 0.39	0.40 to 0.49	0.50 to 0.59	0.60 to 0.69	0.70 to 0.79	0.80 to 0.89	>0.90	Total Number of Items
Pct of items for target		6%	17%	22%	22%	19%	9%	4%	2%	
Target Totals		**3**	**9**	**12**	**12**	**10**	**5**	**2**	**1**	**54**
Actual - Anchors		0	1	3	6	7	1	0	0	**18**
Actual - new selections		3	6	6	9	6	5	1	0	**36**
NRT/Dual Purpose		0	0	0	0	0	0	0	0	**0**
TOTAL	2016	3	7	9	15	13	6	1	0	**54**
Actual	2015	3	8	10	12	13	7	1	0	**54**



	2016	
DOK	Target	Actual
Level	N	N
1	5	10
2	36	36
3	13	8
4	0	0
Total	54	54

DOK	2015
1	10
2	33
3	11
4	0
Total	54

Table 4.5
Number of Science Items Selected by Committee

Content Area	Grade	Total Items	Anchor Items		Total Selected	
Science	4	54	22	41%	32	59%
	8	58	24	41%	34	59%
	HS	65	22	34%	43	66%
Science Total		**177**	**68**	**38%**	**109**	**62%**

4.3 Customer Approvals

Approvals from ADE staff were obtained during several phases of development: during selection of the items, after forms were created, at the completion of the QA reviews, and when pre-press test books were available. Each is described below.

4.3.1 Item Selection Approval

ADE staff members were given the item replacement tables. Approval was verbal. The item selection tables were then reviewed by Pearson's research scientist. Psychometric evaluation of the test selection was the main focus of this review. Recommended changes were discussed with and approved by ADE.

4.3.2 Test Book Approvals

At the test book phase of development, items had been arranged into test book format. That is, they were no longer treated as individual items, but appeared in page layouts as they would appear in the final, printed test books. By this point, all content issues were resolved. The focus of this approval was on format and presentation issues, rather than on content issues. Formal approval was given. Desired changes were communicated via PDF markup and the Development Tracking Form, which included a description of the change, a justification, and space for ADE to grant or deny approval. Formal sign-off of test books by ADE was achieved via the use of signed electronic Final Proof Approval Forms.

4.3.3 FTP Site

A secure FTP site had been established by ADE for transfer of electronic documents (annotated test books, test book reviews, etc.) that need to be reviewed by ADE staff. After careful review by ADE staff, corrections and edits were transmitted via this site to Pearson for inclusion/revision of the test documents.

4.3.4 Final Forms Review (Pearson)

The Final Forms review provided an opportunity for Pearson staff members who had not previously seen the test materials to review them. This review helped assure that test books, answer documents, and test administration directions all work in concert. In addition, this review helped in detecting errors, inconsistencies, cosmetic errors, and key verifications. Items with problems identified during the Final Forms review were annotated. Pearson staff resolved all

comments and made necessary corrections prior to releasing the materials.

4.3.5 ADE Quality Review

After Pearson reviewed and edited test documents, ADE staff conducted a final review of forms to determine if all edits had been accomplished properly.

4.3.6 Final Sign-off

A final, formal approval (blueline stage) was given as test books became available for printing. A copy of the test book was sent for ADE to review and to provide formal approval.

PART 5: TEST ADMINISTRATION

Part 5 of the technical report describes administration procedures, including accommodations, security, and written procedures available to test administrators and school personnel for the Spring 2016 AIMS Science testing. The following 1999 AERA/APA/NCME *Standards* (AERA, APA, NCME, 1999) are addressed: 1.13, 3.3, 3.19, 3.20, 3.21, 3.24, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 6.11, 6.15, 9.1, 10.1, and 10.2. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this part of the technical report are 1.10, 3.1, 3.9, 4.2, 4.5, 4.15, 4.16, 4.21, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 7.0, 7.8.

5.1 Accommodations

Accommodations were made available for the Spring 2016 AIMS Science grades 4, 8, and high school tests. All of the AIMS tests allow some of the same accommodations but exclude others if there is evidence that the accommodation changes the construct that is being assessed. All statistics include students who have received accommodations.

Arizona statutes (A.R.S. §15-741 and §15-755), the Individuals with Disabilities Education Act (IDEA) (300.160), and the Elementary and Secondary Education Act (ESEA) (§1111) mandate that all students who are educated with public funds must participate in state assessment, including all students with disabilities and all students identified as English Language Learners.

For the purposes of assessment, a Special Education student is eligible to receive services under the Individuals with Disabilities Education Act and has an Individualized Education Program (IEP); and a 504 student is eligible under Section 504 of the Rehabilitation Act of 1973 and has a 504 Accommodation Plan.

Students with disabilities who have an IEP, or who have a 504 plan, may be considered for both universal test administration conditions and standard accommodations (described in section 5.1.1). Also, students identified as English Language Learner (ELL) and students who have been identified as Fluent English Proficient (FEP) for no more than two years may be considered for universal test administration conditions and standard accommodations.

Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them as eligible for an alternate assessment, AIMS A, are excluded from AIMS testing.

The Arizona English Language Learner Assessment (AZELLA), a language proficiency assessment, is given to determine a student's proficiency in English and respective instructional placement. An English Language Learner (ELL) is a student whose primary home language is other than English, who scores below the proficient level on the AZELLA. Fluent English Proficient (FEP) is a term that is used to refer to a former ELL student who has scored at the proficient level of the AZELLA.

For detailed information on testing accommodations, please see *AIMS Testing Accommodations: Guidelines for School Year 2014-2016* on the Arizona Department of Education website at the following location: http://www.azed.gov/assessment/files/2014/08/testing-accommodations-2014-2016.pdf .

5.1.1 Overview of Accommodations

Accommodations are specific practices and procedures that provide students with equitable access during instruction and assessment. Accommodations are made in order to provide a student equal access to learning and equal opportunity to demonstrate what is known. They are intended to reduce or even eliminate the effects of a student's disability.

Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. There should be a direct connection between a student's disability, special education need or language need and the accommodation(s) provided to the student during educational activities, including assessment.

Students should receive the same accommodations for classroom instruction, classroom assessments, district assessments, and state assessments. No accommodations should be provided during assessments that are not also provided during instruction. However, not all accommodations appropriate for instruction are appropriate for use during a standardized state assessment. The accommodations available to students while testing on AIMS Science are limited to those listed in later sections of this document.

Accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, prompt, or multiple-choice option. Additionally, accommodations provided for one student may not impede or impact other students in the testing room. It is the responsibility of the testing administrator to see that each student, who qualifies for testing accommodations, receives appropriate accommodations while also ensuring that other students, who do not receive accommodations, are not affected.

5.1.2 Descriptions of Universal and Standard Accommodations

Arizona offers two levels of accommodations to students participating in state assessments: universal test administration conditions and standard accommodations.

Universal Test Administration Conditions are specific testing situations and conditions that may be offered to **any** student in order to provide him/her a comfortable and distraction-free testing environment. Universal test administration conditions may be included in a student's IEP or 504 plan as a required "accommodation"; however, for Arizona state testing purposes, these are not considered testing accommodations and are not limited to only students with IEPs or 504 plans.

Standard Accommodations are provisions made in how a student accesses and demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. For students with disabilities, standard accommodations are intended to reduce or even eliminate the effects of a student's disability. For ELLs and FEP Year 1 and Year 2 students, standard accommodations are intended to allow students the opportunity to demonstrate their content knowledge even though the student is not functioning at grade level in English.

During the assessment, all accommodations for assessment identified in a student's IEP or 504 plan must be made available. However, students may choose not to use the accommodation(s).

5.1.3 Determining if a Student Needs a Testing Accommodation

When students need accommodations in how they learn or demonstrate learning, they are likely to need accommodations in how they are assessed. Conversely, if students do not need accommodations in how they learn or demonstrate learning, they will not need accommodations in how they are assessed. Therefore, no accommodation can be put in place for an assessment that is not already used regularly in the classroom.

To determine if a student will need testing accommodations to participate in state assessments, the following questions were asked:

- Does the student use accommodations during daily instruction?
- If the student uses accommodations during daily instruction, does the student need accommodations in order to participate in the state assessment?
- If so, which testing accommodations are necessary and appropriate for the student?

It is important to annually re-consider the types of accommodations used for students, particularly as they gain more skills. The following is a list of the specific testing accommodations available to students while participating in a state assessment.

Universal Test Administration Conditions

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel
- Being seated in a specific location within the testing room or being seated at special furniture
- Having the test administered by a familiar test administrator
- Using a special pencil or pencil grip
- Using devices that allow the student to see the test: glasses, contacts, magnification, special lighting, and color overlays
- Using devices that allow the student to hear the test directions: hearing aids and amplification
- Wearing noise buffers after the scripted directions have been read
- Having the scripted directions included in the *Test Administration Directions* repeated (at student request) and having questions about the scripted directions or the directions that students read on their own answered.

Standard Accommodations

Injury

For students who were eligible to receive a standard accommodation due to an injury.

- Have answers transferred from a test book into an answer document
- Record or dictate multiple-choice responses to a scribe

ELL/FEP

For students who were eligible to receive a standard accommodation due to their classification as an ELL student or as a FEP (Year 1 or Year 2) student.

- More breaks and/or several shorter sessions
- Simplified language for the scripted directions in English
- Read aloud in English the writing prompt, mathematics test items, or science test items, as needed upon student request
- Provide a word-for-word published, paper translation dictionary
- Exact oral translation of the scripted directions or the directions that students read on their own as needed upon student request

IEP/504

For students who were eligible to receive a standard accommodation due to their IEP or 504 plan.

- Place marker used
- More breaks and/or several shorter sessions
- Test at a different time of day
- Simplify language for the scripted directions in English
- Read aloud or sign the directions that students read on their own
- Read aloud in English or sign the science test items
- Large print edition of test
- Have answers transferred from the test book into an answer document
- Record or dictate multiple-choice responses to a scribe
- Use of a Braille edition of the test

5.1.4 Reporting Results of Assessments Taken with Accommodations

The use of standard accommodations results in scores that are considered valid for comparison and accountability purposes. Students who received standard accommodations on AIMS Science assessments will count as having tested for accountability purposes. Their AIMS results will be included in aggregate results at the school, district, and state level on reports provided by the testing contractor.

Students who receive standard testing accommodations while participating in AIMS Science assessments must have their accommodations appropriately identified on their answer document as directed in the corresponding *Test Administration Directions*. It is not necessary to identify students who received universal test administration conditions while participating in AIMS Science assessments.

5.2 Test Security

All AIMS tests were administered under secure testing conditions. Figure 5.2.1 includes the security agreement signed by the superintendent/charter representative and district test coordinator involved with the testing administration. Figure 5.2.2 includes the security agreement signed by personnel involved with the testing administration.

District test coordinators are responsible for establishing and enforcing test security procedures that comply with the Test Security Agreement, the State Board of Education Rule regarding test security, and Test Security guidance provided in the Pre-Test Workshop package and included in the *AIMS Test Administration Directions.*

Figure 5.2.1
Spring 2016 AIMS Test security agreement for Superintendents/Charter Representatives and District Test Coordinators

Achievement Tests
(AIMS Science and AzMERIT)
School Year 2015-2016 Test Security Agreement
For Superintendents/Charter Representatives and District Test Coordinators

As Superintendent/Charter Representative or District Test Coordinator, I acknowledge that Achievement Tests are secure tests and agree to the following conditions of use to ensure the security of the tests. For this document Achievement Tests refers to AIMS Science, AzMERIT 3-8, and AzMERIT EOC.

1. Superintendents and Charter Representatives are responsible for all testing activities within their district/charter. Superintendents and Charter Representatives are allowed to designate a District Test Coordinator to act on their behalf.
   a. An accurate Test Coordinator Information Sheet for school year 2015-2016 must be on file with the Assessment Section of the Arizona Department of Education (ADE).
   b. The designated Achievement Testing District Test Coordinator must complete all pre-test trainings provided by ADE for each of the test administrations in which that the district will be participating.
2. All necessary security precautions shall be in place to safeguard test materials.
   a. Access to test books and answer documents shall be restricted.
   b. The names of all persons having access to the test books and answer documents shall be kept on file by the designated district test coordinator.
   c. A list of students must be kept on file with the test administrator(s) and test proctors(s) that were in the test room during the administration of the test.
   d. All persons having access to the Achievement Test materials, other than students to whom the tests are administered, shall sign a School Year 2015-2016 Achievement Test Security Agreement. Signed test security agreements shall be kept on file for 6 years.
      i. Building administrators shall maintain the agreements signed by building staff.
      ii. Superintendents/charter representatives shall maintain the agreements signed by building administrators.
      iii. The Assessment Section of ADE shall maintain the agreements signed by superintendents and charter representatives.
   e. All test books and answer documents shall be kept under lock and key except during actual test times.
      i. Test books and answer documents shall be delivered to test administrators no sooner than the date of testing.
      ii. Test books and answer documents shall be kept secure until they are distributed to students.
      iii. Students shall not be permitted to remove test material from the testing room except under supervision of staff.
   f. Achievement Tests shall not be examined, read, or reviewed.
      i. No content of the test shall be disclosed nor allowed to be disclosed.
      ii. No test item shall be discussed at any time.
      iii. No student test booklet or answer document shall be examined, read, or reviewed.
      iv. No student response or notations (including stray marks) on a student test booklet or answer document can be changed or erased.
      v. Student test booklets or answer documents will be submitted for scoring exactly as completed by the student.
   g. Upon completion of testing, all AzMERIT test materials shall be returned to the designated district test coordinator.
3. The district superintendent or charter representative shall develop, distribute, and enforce disciplinary procedures for the violation of test security by staff.
4. *Test Preparation and Administration Practices*, the guidelines approved by the State Board of Education in January 2003 and updated December 2007, shall be followed.
5. **All instructions in the *Test Coordinator's Manual* and the *Test Administration Directions*, which include reading the directions to students exactly as scripted in the *Test Administration Directions*, shall be followed.**

By signing my name to this document, I am assuring the Arizona Department of Education that I will abide by the above conditions and that anyone I supervise, who will have access to the Achievement tests for school year 15-16, will also sign a Test Security Agreement.
Superintendent/Charter Representative Signature: _____ Date: _____

Printed Name: _____ Title: _____

District Test Coordinator Signature: _____ Date:

Printed Name: _____

District/Charter: _____ Entity #: _____

Address: _____

City, State, Zip: _____

Fax: 602-542-5467 or Email: marypat.wood@azed.gov **Due to ADE by January 29, 2016**

Figure 5.2.2
Spring 2016 AIMS Test security agreement for all school/district/charter personnel

Achievement Tests
(AIMS Science and AzMERIT)
School Year 2015-2016 Test Security Agreement

I acknowledge that all Achievement Tests are secure tests and agree to the following conditions of use to ensure the security of the test. For this document Achievement Tests refers to AIMS Science, AzMERIT 3-8, and AzMERIT EOC.

1. I shall take necessary precautions to safeguard test materials.
   a. I shall sign an Achievement Test Security Agreement for School Year 2015-2016.
   b. Access to test books and answer documents is restricted. I shall not attempt to gain access to test materials beyond that which is granted to me by my school/district test coordinator, superintendent, or charter representative.
   c. If test books and answer documents are distributed to me, I shall keep them under lock and key except during actual test times.
   d. I shall not permit students to remove test material from the testing room except under the supervision of staff.
   e. I shall not examine, read, or review the Achievement Tests.
      i. I shall not disclose, nor allow to be disclosed, the content of the test.
      ii. I shall not discuss any test item at any time.
      iii. I shall not examine, read, or review any student responses.
   f. I shall not erase or change any student responses or any marks (including stray marks) on a scorable test booklet or answer document.
   g. If test books and answer documents are distributed to me, I shall return all AzMERIT test materials to the school/district test coordinator immediately upon the completion of testing.
2. I understand that the district superintendent or charter representative will develop, distribute, and enforce disciplinary procedures for the violation of test security by staff.

Individuals who will administer or proctor Achievement Tests for school year 2015-2016 must also agree to the following conditions to ensure the correct administration of the tests.

3. I shall participate in training activities prior to administering the tests.
4. I shall follow *Test Preparation and Administration Practices*, the guidelines approved by the State Board of Education in January 2003 and updated in December 2007.
5. I shall review the appropriate Test Administration Directions prior to administering the test.
6. I shall follow all instructions in the appropriate Test Administration Directions including **reading the directions to students exactly as scripted**.

By signing my name to this document, I am assuring my district/charter and the Arizona Department of Education that I will abide by the above conditions and that anyone I supervise, who will have access to the Achievement Tests, will also sign a Test Security Agreement.

Signed By: _____ Date: _____

Printed Name: _____

Title: _____ School: _____

Please return signed copy as per instructions from your school/district test coordinator.
Signed copies will be maintained by school/district administrators for 6 years.

5.3 Test Administration

In order to ensure a standardized testing administration for all students, a *Test Coordinator's Manual* was made available to all test coordinators for the spring 2016 administration. The manual included the following topics:

- Responsibilities of the District Test Coordinator
  - Before Testing
  - During Testing
  - After Testing
- Procedures for Test Administration
  - Students to Be Tested
  - Test Administration Schedules
  - Required Test Materials
  - Test Security
  - Student Identification Information
  - Arrangements Prior to Test Administration
- Procedures for Handling Test Materials (before, during, and after testing)
  - Receiving Test Materials
  - Inventorying Test Materials
  - Precautions
  - Inspecting and Organizing Test Materials
  - Assembling Scorable Test Materials
  - Assembling Nonscorable Test Materials
  - Materials Retrieval
- State Board of Education Rule
- Important Dates for Spring 2016 Testing

Test Administration Directions were made available to all test administrators for the spring 2016 assessments. The *Test Administration Directions* included the following topics:

- Overview for the Administration of AIMS
  - Test Administrator Responsibilities
  - Students to Be Tested
  - Test Administration Schedule
  - Test Materials
  - Precautions
- Before Testing Guidelines
  - Training and Test Security
  - Preparing the Room for Testing
- During Testing Guidelines
  - Reading the Scripted Directions

  - Student Identification Information
  - Monitoring Testing
  - Use of Resources
  - Use of Unacceptable Resources
  - Disruptive Students and Students Who Leave During Testing
  - Detailed Scripts for Administration of Each Part of Each Test
- After Testing
  - Inspecting Test Materials
  - Completing Student Identification Information
  - Transferring Student Responses
  - Returning Test Materials to the Test Coordinator

For specific information related to test administration, refer to the *Test Coordinator's Manual* and/or the *Test Administration Directions*.

Pre-Test Workshops were conducted online prior to the spring test administration Every district test coordinator is required to view a 3-session online Pre-Test Workshop. The Pre-Test Workshop encompasses training related to test administration which includes test security, accommodations, test coordinator responsibility, and test schedule. Materials handling is included in these online workshops, covering ordering, receiving, preparing for retrieval, and the retrieval of test materials.

PART 6: CLASSICAL ITEM ANALYSIS

Part 6 presents classical test statistics and item analysis statistics for the AIMS Science grade 4, 8, and high school tests computed from the data used for calibration and scaling. Addressed in this part of the technical report are the following 1999 AERA/APA/NCME *Standards*: 1.5, 1.13, 2.4, 2.8, 3.18, 6.5, and 7.1. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this chapter are: 1.8, 1.10, 2.19, 3.6, 4.14, and 7.4.

6.1 Data

Arizona had one test window for operational testing in spring 2016. The AIMS Science tests for grade 4 and 8, and high school were administered between March 16 and April 24.

6.2 Descriptive Statistics by Test

Table 6.2.1 presents descriptive statistics by grade level which are computed with the calibration samples. The table shows the number of students (N), the maximum obtained raw score (Max RS), the raw score mean (RS M), the raw score standard deviation (RS SD), the average *p*-value (P-Value M), the average item-to-total correlation (rpb M) and the estimate of internal consistency. Cronbach's alpha is the measure of internal consistency used for the AIMS Science tests. The item-to-total correlation is computed as a point biserial correlation. The point biserial correlation reported is the correlation of the item scores and the total test score.

Table 6.2.1
Spring 2016 AIMS Classical Test Analysis Statistics

	Content	Grade	Prompt	N	Max RS Obtained	RS M	RS SD	P-value M	rpb M	Internal Consistency
CRT										
	Science									
		4		85477	54	30.30	10.15	0.56	0.35	0.90
		8		81978	58	33.29	10.79	0.57	0.36	0.91
		HS		79976	65	31.72	11.65	0.49	0.33	0.90

Note: CRT = Criterion-referenced test.

6.3 Classical Item Analysis

Classical item analysis was conducted for each Science test. Tables 6.3.1-6.3.3 present item statistics for the spring science tests. The tables show the number of students (N), the item difficulty (P-Value), point biserial correlation (rpb) and biserial correlation (rbi), percentage of students who omitted the item (% Omit), and the percentage of students responding to and point biserial for each response option. The keyed response has a percent responding that matches the *p*-value and a positive point biserial correlation while the incorrect response options have a negative point biserial correlation. The point biserial correlation (rpb) reported is the correlation between student performance on an item and the total score on a test. The biserial correlation (rbi) is an adjusted point-biserial correlation intended to estimate the value of the correlation between the item and total score as if the item scores were normally distributed rather than binary.

Table 6.3.1
Spring 2016 AIMS Classical Item Analysis
Science Grade 4

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
1	85477	0.78	0.2633538	0.37	0.06	4.32	-0.1356	7.41	-0.194892	10.15	-0.1001868	78.06	0.2633538
2	85477	0.71	0.2977921	0.39	0.07	70.6	0.2977921	13.56	-0.1297157	7.33	-0.2182138	8.45	-0.1233163
3	85477	0.76	0.3668955	0.5	0.15	8.37	-0.2372443	3.97	-0.1941827	11.17	-0.1649364	76.35	0.3668955
4	85477	0.59	0.439784	0.54	0.14	25.67	-0.2807017	59.18	0.439784	8.9	-0.1760837	6.12	-0.1803997
5	85477	0.67	0.3578755	0.46	0.13	10.04	-0.2518306	6.62	-0.1692832	66.68	0.3578755	16.52	-0.1364852
6	85477	0.5	0.2996179	0.37	0.1	19.68	-0.0474239	49.93	0.2996179	23.33	-0.141446	6.95	-0.279451
7	85477	0.63	0.3629244	0.46	0.1	13.94	-0.2231842	10.04	-0.0650714	62.84	0.3629244	13.09	-0.2324452
8	85477	0.63	0.3962762	0.5	0.1	9.14	-0.2335505	8.53	-0.2316686	62.58	0.3962762	19.65	-0.1500344
9	85477	0.45	0.1878649	0.23	0.12	28.3	0.0185994	8.56	-0.1856387	44.91	0.1878649	18.12	-0.1294239
10	85477	0.76	0.4297118	0.58	0.09	5.48	-0.2675764	75.73	0.4297118	4.01	-0.2003869	14.7	-0.2366374
11	85477	0.51	0.3532826	0.43	0.05	11.02	-0.1608319	15.05	-0.1055523	50.76	0.3532826	23.13	-0.2098475
12	85477	0.51	0.4005181	0.49	0.09	16.12	-0.1281191	15.35	-0.1624267	17.11	-0.2508399	51.34	0.4005181
13	85477	0.4	0.2731292	0.34	0.12	22.71	-0.0856479	28.58	-0.0943496	8.67	-0.1961783	39.92	0.2731292
14	85477	0.4	0.3159584	0.39	0.1	40.25	0.3159584	47.8	-0.1176709	6.84	-0.2165061	5.01	-0.1902537
15	85477	0.63	0.5122923	0.64	0.1	15.1	-0.2671993	7.64	-0.2254344	13.95	-0.263521	63.21	0.5122923
16	85477	0.35	0.1771762	0.23	0.13	18.64	-0.2242224	39.28	0.1370124	35.32	0.1771762	6.63	-0.2581907
17	85477	0.36	0.2119562	0.27	0.14	28.93	-0.085025	16.85	-0.0429309	18.07	-0.1224245	36.01	0.2119562
18	85477	0.61	0.3246453	0.41	0.11	8.68	-0.2252863	61.42	0.3246453	9.52	-0.2326526	20.26	-0.0651255
19	85477	0.52	0.4448305	0.55	0.15	20.96	-0.2933384	12.84	-0.2141791	52.2	0.4448305	13.85	-0.0897395
20	85477	0.74	0.4221626	0.56	0.22	11.89	-0.1911156	6.03	-0.2490225	7.9	-0.2356528	73.96	0.4221626
21	85477	0.69	0.4686604	0.6	0.2	11.58	-0.2574886	69.26	0.4686604	14.34	-0.249482	4.63	-0.2188381
22	85477	0.55	0.3356623	0.42	0.26	11.24	-0.1453454	14.56	-0.1268745	18.96	-0.1941746	54.98	0.3356623
23	85477	0.75	0.512034	0.68	0.18	75.19	0.512034	8.94	-0.2407992	5.53	-0.2625346	10.16	-0.3040075
24	85477	0.43	0.2925875	0.35	1.42	16.53	-0.1358629	21.17	-0.0822897	43.09	0.2925875	17.79	-0.1575308
25	85477	0.36	0.1665136	0.21	0.12	17.55	-0.2009144	8.62	-0.141549	37.94	0.0749363	35.77	0.1665136
26	85477	0.43	0.3070692	0.38	0.18	26.6	-0.0414556	42.81	0.3070692	10.74	-0.2721208	19.67	-0.1240261
27	85477	0.71	0.439534	0.57	0.35	6.79	-0.2108062	70.89	0.439534	6.78	-0.2435881	15.19	-0.2356367
28	85477	0.61	0.3039554	0.38	0.04	3.02	-0.1989291	5.48	-0.2314182	30.05	-0.13354	61.42	0.3039554
29	85477	0.7	0.5176121	0.67	0.1	9.78	-0.2917193	12.76	-0.2460666	6.85	-0.2651992	70.49	0.5176121
30	85477	0.46	0.277373	0.34	0.18	8.01	-0.2447835	5.93	-0.2067824	39.85	-0.0466844	46.02	0.277373
31	85477	0.31	0.1540205	0.2	0.15	29.13	-0.0463486	27.79	-0.0241636	30.87	0.1540205	12.06	-0.1205463
32	85477	0.37	0.3423958	0.43	0.16	17.31	-0.2054564	33.46	-0.09472	37.5	0.3423958	11.57	-0.1353027
33	85477	0.44	0.4255925	0.52	0.19	9.81	-0.1503586	43.98	0.4255925	40.64	-0.233972	5.37	-0.2285704
34	85477	0.64	0.4505616	0.57	0.17	64.17	0.4505616	10.22	-0.2472443	18.33	-0.2019459	7.11	-0.2439912
35	85477	0.62	0.4379745	0.55	0.18	8.37	-0.2865227	7.81	-0.244457	22.09	-0.1635219	61.55	0.4379745
36	85477	0.6	0.212435	0.27	0.2	17.67	-0.0044914	59.74	0.212435	5.74	-0.1708093	16.65	-0.1680604
37	85477	0.36	0.2125858	0.27	0.18	36.36	0.2125858	39.65	-0.0745438	6.33	-0.2073296	17.48	-0.0402906
38	85477	0.45	0.3192255	0.39	0.21	25.68	-0.2150987	12.44	0.013673	45.48	0.3192255	16.19	-0.1884471
39	85477	0.42	0.3303285	0.41	0.07	41.88	0.3303285	20.17	-0.1406264	14.97	-0.16102	22.9	-0.1167479
40	85477	0.74	0.2952091	0.4	0.12	8.59	-0.0966613	11.5	-0.1886814	74.28	0.2952091	5.51	-0.1820253

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

(table continues)

Table 6.3.1 (continued)
Spring 2016 AIMS Classical Item Analysis
Science Grade 4 (continued)

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
41	85477	0.54	0.3443753	0.43	0.23	54.48	0.3443753	10.3	-0.2069445	19.78	-0.1787462	15.21	-0.1035386
42	85477	0.64	0.3631898	0.46	0.17	8.66	-0.2439764	9.35	-0.19334	18.1	-0.1284725	63.72	0.3631898
43	85477	0.61	0.4240631	0.53	0.15	10.34	-0.2815874	6.29	-0.2082731	22.33	-0.1690852	60.89	0.4240631
44	85477	0.23	0.1682673	0.23	0.18	22.76	0.1682673	14.16	-0.0291579	14.76	-0.1648702	48.14	-0.0038044
45	85477	0.59	0.4967985	0.61	0.22	9.35	-0.2611306	7.83	-0.2831008	59.33	0.4967985	23.27	-0.2168439
46	85477	0.82	0.4531038	0.65	0.18	6.58	-0.2660618	81.56	0.4531038	6.36	-0.2510692	5.33	-0.2129344
47	85477	0.5	0.3942734	0.48	0.24	23	-0.1726779	18.96	-0.1679251	7.97	-0.2158554	49.83	0.3942734
48	85477	0.59	0.399403	0.5	0.21	58.68	0.399403	9.15	-0.2677403	11.43	-0.201664	20.53	-0.1363191
49	85477	0.72	0.5280778	0.69	0.23	10.15	-0.3015767	7.7	-0.2780728	10.01	-0.2380421	71.9	0.5280778
50	85477	0.54	0.3821554	0.47	0.24	17.47	-0.074655	54.3	0.3821554	16.1	-0.2443969	11.89	-0.2222961
51	85477	0.52	0.3915532	0.48	0.23	22.1	-0.1175735	14.92	-0.2024154	10.41	-0.2438408	52.35	0.3915532
52	85477	0.56	0.4412834	0.54	0.41	20.22	-0.1994497	9.88	-0.2269187	56.07	0.4412834	13.41	-0.2074176
53	85477	0.6	0.3000683	0.37	0.58	9.89	-0.0961311	60.15	0.3000683	10.73	-0.1817996	18.65	-0.1575655
54	85477	0.67	0.4264245	0.54	0.79	67.24	0.4264245	8.69	-0.228785	8.55	-0.2270892	14.74	-0.1993461

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

Table 6.3.2
Spring 2016 AIMS Classical Item Analysis
Science Grade 8

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
1	81978	0.84	0.2273089	0.34	0.02	9.62	-0.1409296	83.92	0.2273089	2.89	-0.1332781	3.54	-0.1057778
2	81978	0.72	0.3277043	0.43	0.08	72.35	0.3277043	7.06	-0.1407743	12.29	-0.2231177	8.22	-0.1350942
3	81978	0.61	0.1541497	0.19	0.05	12.29	-0.0997042	21.92	-0.0350539	61.02	0.1541497	4.73	-0.1315243
4	81978	0.75	0.349486	0.47	0.05	74.62	0.349486	3.86	-0.1932086	17.49	-0.2048546	3.99	-0.1889041
5	81978	0.49	0.5013827	0.61	0.03	49.47	0.5013827	24.89	-0.2303918	22.78	-0.3036215	2.83	-0.142822
6	81978	0.57	0.2318186	0.29	0.06	23.33	-0.0987846	7.51	-0.1721438	56.68	0.2318186	12.41	-0.0839335
7	81978	0.8	0.3468802	0.49	0.05	4.21	-0.2189739	5.38	-0.1651137	10.06	-0.1881302	80.29	0.3468802
8	81978	0.88	0.3798604	0.6	0.06	2.25	-0.1545009	6.94	-0.2669767	87.74	0.3798604	3.02	-0.1963543
9	81978	0.63	0.2639092	0.33	0.08	16.28	-0.0699996	11.38	-0.2229662	62.91	0.2639092	9.35	-0.1056474
10	81978	0.61	0.3476323	0.43	0.04	16.94	-0.120458	10.22	-0.1919759	60.76	0.3476323	12.04	-0.2039096
11	81978	0.77	0.3031046	0.42	0.03	12.55	-0.1453608	77.3	0.3031046	5.61	-0.1954703	4.53	-0.1626324
12	81978	0.69	0.383121	0.49	0.04	16.53	-0.1498876	4.55	-0.2114583	9.61	-0.2610033	69.26	0.383121
13	81978	0.42	0.2422539	0.3	0.07	7.94	-0.177406	32.33	-0.0840674	42.23	0.2422539	17.43	-0.0852479
14	81978	0.47	0.3334068	0.41	0.07	29.04	-0.0803069	13.83	-0.172135	9.64	-0.239115	47.42	0.3334068
15	81978	0.29	0.1077549	0.14	0.08	22.27	-0.0390572	21.11	-0.0552242	27.12	-0.0232101	29.43	0.1077549
16	81978	0.63	0.3855314	0.48	0.05	12.36	-0.1836605	9.74	-0.2074597	63.09	0.3855314	14.76	-0.1804701
17	81978	0.59	0.4000705	0.5	0.08	18.99	-0.0969072	7.03	-0.2467682	14.58	-0.2700936	59.32	0.4000705
18	81978	0.73	0.4289345	0.56	0.06	10.17	-0.2060675	7.58	-0.2522853	72.62	0.4289345	9.57	-0.2109446
19	81978	0.57	0.3852496	0.48	0.1	56.57	0.3852496	15.26	-0.2151823	5.92	-0.2626296	22.14	-0.1239419
20	81978	0.29	0.3573362	0.47	0.1	28.86	0.3573362	17.53	-0.0990356	28.95	-0.2172718	24.56	-0.0597279
21	81978	0.51	0.3670659	0.45	0.11	6.56	-0.2855609	10.71	-0.2729394	50.7	0.3670659	31.93	-0.0607832
22	81978	0.68	0.3265059	0.42	0.11	21.14	-0.1386612	67.87	0.3265059	6.9	-0.237243	3.98	-0.1818981
23	81978	0.51	0.4290979	0.53	0.14	16.93	-0.2039533	50.66	0.4290979	27.5	-0.2009048	4.78	-0.2260757
24	81978	0.36	0.3912148	0.49	0.1	4.25	-0.1475333	26.2	-0.0367241	33.82	-0.2989454	35.63	0.3912148
25	81978	0.41	0.4462987	0.55	0.06	24.6	-0.2263877	15.94	-0.257968	41.12	0.4462987	18.28	-0.07151
26	81978	0.44	0.3357989	0.42	0.09	28.51	-0.0805294	44.11	0.3357989	11.71	-0.2060588	15.59	-0.1766732
27	81978	0.21	0.1483442	0.21	0.13	20.58	0.1483442	12.71	-0.1770699	17.62	-0.0825885	48.96	0.0609851
28	81978	0.46	0.32848	0.41	0.1	21.17	-0.111656	45.82	0.32848	17.3	-0.1944831	15.61	-0.1224224
29	81978	0.47	0.4086153	0.5	0.08	14.43	-0.1764161	47.48	0.4086153	16.89	-0.2041284	21.11	-0.1605039
30	81978	0.7	0.4900467	0.63	0.03	11.35	-0.2898966	6.74	-0.2623468	12.32	-0.2059944	69.57	0.4900467
31	81978	0.5	0.3488875	0.43	0.08	20.42	-0.2634746	50.19	0.3488875	18.35	-0.0903477	10.96	-0.1063121
32	81978	0.81	0.4482	0.64	0.06	4.44	-0.2450018	81.44	0.4482	10.07	-0.2693065	3.99	-0.217284
33	81978	0.71	0.4470334	0.58	0.07	5.24	-0.2030012	9.95	-0.2256401	71.5	0.4470334	13.23	-0.2621448
34	81978	0.45	0.358357	0.44	0.14	44.68	0.358357	8.65	-0.1742742	8.25	-0.1874592	38.28	-0.159522
35	81978	0.49	0.3031559	0.37	0.06	24.42	-0.0495664	17.64	-0.2100149	48.9	0.3031559	8.97	-0.1755062
36	81978	0.46	0.3357739	0.41	0.1	10.45	-0.2430261	46.28	0.3357739	21.75	-0.0623261	21.42	-0.1640401
37	81978	0.66	0.4366301	0.55	0.09	7.03	-0.256843	11.47	-0.2872239	65.98	0.4366301	15.42	-0.1369851
38	81978	0.58	0.3350175	0.42	0.09	10.01	-0.2645486	7.04	-0.2747705	24.49	-0.035718	58.37	0.3350175
39	81978	0.48	0.3594509	0.44	0.14	39.37	-0.1632177	47.87	0.3594509	5.6	-0.2573177	7.02	-0.15882
40	81978	0.62	0.4586465	0.57	0.12	11.16	-0.2208369	22.44	-0.2734499	62.3	0.4586465	3.98	-0.1967992

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

(table continues)

Table 6.3.2 (continued)
Spring 2016 AIMS Classical Item Analysis
Science Grade 8

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
41	81978	0.73	0.4479424	0.59	0.02	4.64	-0.256705	14.09	-0.2392361	72.75	0.4479424	8.5	-0.2229544
42	81978	0.78	0.4646449	0.64	0.05	5.86	-0.2650559	8.44	-0.2711806	78.23	0.4646449	7.43	-0.2058505
43	81978	0.51	0.4071545	0.5	0.12	14.02	-0.2668665	10.37	-0.2678598	50.57	0.4071545	24.92	-0.0673653
44	81978	0.79	0.5091001	0.71	0.08	7	-0.2674996	7.67	-0.2762187	5.94	-0.2714615	79.31	0.5091001
45	81978	0.46	0.2415091	0.3	0.09	40.2	-0.0136991	6.76	-0.2292208	45.99	0.2415091	6.97	-0.2202423
46	81978	0.52	0.3996721	0.49	0.08	7.63	-0.2440524	51.54	0.3996721	7.65	-0.2616307	33.1	-0.1388665
47	81978	0.53	0.3307626	0.41	0.09	9.15	-0.2601764	26.32	-0.0627004	52.83	0.3307626	11.61	-0.1948332
48	81978	0.59	0.4879053	0.6	0.1	18.25	-0.2849259	8.31	-0.1957254	14.15	-0.2167772	59.19	0.4879053
49	81978	0.72	0.4327115	0.57	0.12	5.05	-0.2312955	5.54	-0.2673079	72.19	0.4327115	17.1	-0.2172764
50	81978	0.32	0.2769323	0.36	0.12	44.22	-0.0940404	32.09	0.2769323	11.86	-0.0980382	11.71	-0.1581689
51	81978	0.58	0.3475723	0.43	0.1	21.24	-0.0494986	10.62	-0.2823937	58.45	0.3475723	9.58	-0.2171066
52	81978	0.56	0.3997785	0.49	0.15	9.84	-0.2219828	10.85	-0.3097119	55.64	0.3997785	23.52	-0.0849299
53	81978	0.77	0.5037628	0.69	0.17	77.44	0.5037628	8.73	-0.2832539	7.17	-0.2580807	6.49	-0.2574546
54	81978	0.55	0.4118982	0.51	0.2	10.01	-0.2865042	12.89	-0.2329498	21.44	-0.0985911	55.47	0.4118982
55	81978	0.32	0.2721542	0.35	0.1	27.57	-0.0069236	16.44	-0.140979	23.72	-0.1687014	32.16	0.2721542
56	81978	0.67	0.3914727	0.5	0.12	5.73	-0.2397943	21.25	-0.1753981	5.48	-0.2451119	67.42	0.3914727
57	81978	0.49	0.2079621	0.26	0.14	14	-0.1796382	49.2	0.2079621	27.59	0.0322575	9.05	-0.1950861
58	81978	0.52	0.3200541	0.4	0.14	51.94	0.3200541	9.67	-0.250248	29.07	-0.0257435	9.18	-0.2568108

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

Table 6.3.3
Spring 2016 AIMS Classical Item Analysis
Science Grade 10

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
1	79976	0.55	0.3130497	0.39	0.14	26.35	-0.1142269	9.55	-0.1975984	9.18	-0.1638076	54.78	0.3130497
2	79976	0.57	0.4719246	0.58	0.04	5.89	-0.0712097	27.63	-0.3813927	9.1	-0.1600798	57.34	0.4719246
3	79976	0.44	0.3129711	0.39	0.09	34.74	-0.0871156	14.06	-0.2220085	7.13	-0.1424758	43.98	0.3129711
4	79976	0.4	0.2268172	0.28	0.13	14.94	-0.1207347	29.54	-0.0211076	40.25	0.2268172	15.14	-0.1632776
5	79976	0.39	0.1815204	0.23	0.12	15.09	-0.1567957	37.53	0.0216653	7.99	-0.1585043	39.28	0.1815204
6	79976	0.4	0.2640124	0.33	0.1	6.51	-0.1541008	39.57	0.2640124	3.14	-0.135578	50.69	-0.134893
7	79976	0.26	0.3850562	0.51	0.14	22.01	0.0254853	32.22	-0.2318965	25.84	0.3850562	19.79	-0.1775923
8	79976	0.58	0.3405391	0.42	0.09	8.93	-0.0943857	12.54	-0.1565364	20.22	-0.2218549	58.22	0.3405391
9	79976	0.65	0.3168875	0.4	0.13	19.06	-0.1246641	5.88	-0.1907884	65.21	0.3168875	9.71	-0.192166
10	79976	0.57	0.3931372	0.49	0.17	21.91	-0.181272	5.15	-0.1664732	56.98	0.3931372	15.78	-0.2267828
11	79976	0.56	0.3297878	0.41	0.05	15.01	-0.1485914	55.81	0.3297878	8.67	-0.2223874	20.46	-0.1191846
12	79976	0.3	0.2465936	0.32	0.13	13.16	-0.0926418	23.02	0.0368364	34.12	-0.2040271	29.58	0.2465936
13	79976	0.73	0.3921135	0.52	0.04	13.47	-0.2162328	73.38	0.3921135	3.34	-0.1749294	9.77	-0.2288929
14	79976	0.55	0.1755111	0.22	0.08	8.37	-0.183174	30.14	0.0221673	6.48	-0.1899148	54.94	0.1755111
15	79976	0.74	0.2652316	0.36	0.07	74.16	0.2652316	14.54	-0.1048415	9.12	-0.19468	2.12	-0.1598013
16	79976	0.64	0.3541757	0.45	0.1	63.55	0.3541757	5.12	-0.1661992	25.05	-0.1909212	6.18	-0.2116619
17	79976	0.52	0.3772118	0.47	0.06	6.53	-0.165989	52.29	0.3772118	16.61	-0.2827418	24.51	-0.0979503
18	79976	0.52	0.4622766	0.57	0.09	12.69	-0.2177181	19.57	-0.2611051	15.63	-0.1508649	52.01	0.4622766
19	79976	0.33	0.2731647	0.35	0.21	26.08	-0.1424777	32.75	0.2731647	25.95	-0.0473584	15	-0.1256106
20	79976	0.84	0.3648661	0.55	0.1	4.56	-0.2063309	5.75	-0.1989693	84.34	0.3648661	5.26	-0.1922524
21	79976	0.71	0.2830482	0.37	0.08	12.87	-0.2262824	5.75	-0.1125699	10.39	-0.0867018	70.91	0.2830482
22	79976	0.49	0.2812044	0.35	0.13	18.58	-0.0590164	25.41	-0.2055427	48.61	0.2812044	7.27	-0.1080066
23	79976	0.66	0.4355455	0.56	0.13	14.95	-0.2356082	7.33	-0.2514212	66.49	0.4355455	11.09	-0.1776029
24	79976	0.57	0.3457248	0.43	0.16	12.22	-0.1497981	10.33	-0.2220414	20.5	-0.1349357	56.8	0.3457248
25	79976	0.42	0.3818093	0.47	0.04	20.16	-0.1585022	16.64	-0.1735675	41.81	0.3818093	21.35	-0.1466061
26	79976	0.31	0.2471347	0.32	0.11	31.15	0.2471347	15.02	-0.2241313	15.74	-0.2888952	37.98	0.1460132
27	79976	0.55	0.3395847	0.42	0.15	19.15	-0.1543164	10.58	-0.1380397	14.74	-0.184756	55.38	0.3395847
28	79976	0.46	0.3524463	0.44	0.12	13.85	-0.1740426	45.78	0.3524463	28.66	-0.1324263	11.59	-0.173471
29	79976	0.39	0.3537423	0.44	0.2	34.7	-0.1139655	38.94	0.3537423	19.6	-0.1984192	6.56	-0.1592338
30	79976	0.44	0.255347	0.32	0.11	43.97	0.255347	21.04	-0.1382745	27.93	-0.0570801	6.96	-0.175838
31	79976	0.46	0.3669092	0.45	0.08	4.98	-0.1403588	33.07	-0.2698607	45.56	0.3669092	16.3	-0.068213
32	79976	0.5	0.3391228	0.42	0.1	31.73	-0.1216207	10.72	-0.1906911	49.51	0.3391228	7.94	-0.1993589
33	79976	0.44	0.3715794	0.46	0.19	44.29	0.3715794	22.35	-0.1552227	19.17	-0.193372	13.99	-0.1259355
34	79976	0.28	0.2996926	0.39	0.13	24.67	-0.015647	27.12	-0.1426791	19.8	-0.1625708	28.28	0.2996926
35	79976	0.62	0.4799357	0.6	0.14	7.03	-0.2216828	16.56	-0.2786395	14.3	-0.2071323	61.98	0.4799357
36	79976	0.75	0.4378193	0.58	0.09	14.23	-0.2905232	74.53	0.4378193	5.44	-0.2092822	5.7	-0.1790953
37	79976	0.61	0.419464	0.53	0.18	11.69	-0.197072	12.4	-0.2215109	61.13	0.419464	14.6	-0.1922446
38	79976	0.5	0.4008461	0.49	0.15	10.99	-0.2258556	12.45	-0.1857436	26.88	-0.1541222	49.53	0.4008461
39	79976	0.43	0.3367391	0.42	0.06	17.55	-0.166939	42.55	0.3367391	17.01	-0.1496561	22.84	-0.1113383
40	79976	0.56	0.4048422	0.5	0.19	16.04	-0.1896285	14.67	-0.2446397	56.1	0.4048422	13.01	-0.1324851

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

(table continues)

Table 6.3.3 (continued)
Spring 2016 AIMS Classical Item Analysis
Science Grade 10

Item	N	P-Value	rpb	rbi	% Omit	Option A		Option B		Option C		Option D	
						%	rpb	%	rpb	%	rpb	%	rpb
41	79976	0.32	0.1968173	0.25	0.18	32.28	-0.0088464	32.38	0.1968173	22.64	-0.0614989	12.52	-0.1879355
42	79976	0.34	0.2709406	0.35	0.15	32.59	-0.0414405	22.23	-0.1944553	33.59	0.2709406	11.44	-0.0869174
43	79976	0.38	0.3014941	0.38	0.15	26.91	-0.0139084	14.19	-0.2235257	38.46	0.3014941	20.28	-0.1553495
44	79976	0.44	0.3127636	0.39	0.14	12.84	-0.1244827	32.27	-0.0691548	11.18	-0.2571889	43.57	0.3127636
45	79976	0.31	0.3005461	0.39	0.12	30.32	0.0854933	11.96	-0.2079773	26.19	-0.2530895	31.4	0.3005461
46	79976	0.47	0.4678196	0.58	0.13	46.87	0.4678196	9.24	-0.1690183	22.6	-0.2648879	21.16	-0.1802421
47	79976	0.53	0.3892941	0.48	0.13	22.06	-0.1529328	19.02	-0.2276423	53.39	0.3892941	5.4	-0.1828336
48	79976	0.26	0.1847578	0.25	0.17	14.67	-0.2054987	21.35	-0.1336307	25.81	0.1847578	38.01	0.096074
49	79976	0.44	0.3544517	0.44	0.12	44.35	0.3544517	24.62	-0.1795264	18.82	-0.1242831	12.09	-0.1537126
50	79976	0.6	0.4500747	0.56	0.18	59.72	0.4500747	8.01	-0.2192278	20.97	-0.2833742	11.12	-0.1451004
51	79976	0.51	0.365531	0.45	0.17	5.75	-0.2342871	31.69	-0.1237145	11.35	-0.2223143	51.05	0.365531
52	79976	0.56	0.4068802	0.5	0.16	14.72	-0.1526867	14.76	-0.2099273	55.79	0.4068802	14.58	-0.2078416
53	79976	0.51	0.3579597	0.44	0.09	51.11	0.3579597	18.19	-0.2194656	21.36	-0.1513727	9.26	-0.1110174
54	79976	0.56	0.5188473	0.64	0.11	56.12	0.5188473	9.33	-0.2374131	27.7	-0.3342422	6.74	-0.154344
55	79976	0.34	0.2068293	0.27	0.15	34.29	0.2068293	12.96	-0.2386423	6.84	-0.2366902	45.76	0.0838806
56	79976	0.41	0.2671998	0.33	0.16	13.21	-0.1186086	22.16	-0.0388238	40.94	0.2671998	23.53	-0.1769523
57	79976	0.46	0.3078486	0.38	0.17	45.61	0.3078486	26.98	-0.0609194	19.65	-0.1786338	7.59	-0.2085396
58	79976	0.39	0.3812993	0.48	0.14	38.82	0.3812993	13.11	-0.2127021	21.86	-0.1499666	26.07	-0.1184776
59	79976	0.44	0.3190518	0.4	0.15	43.55	0.3190518	15.72	-0.2031472	12.56	-0.2503839	28.03	-0.002703
60	79976	0.52	0.373318	0.46	0.17	10.86	-0.2241261	9.7	-0.2118118	52.36	0.373318	26.91	-0.1214903
61	79976	0.7	0.3958884	0.52	0.14	10.97	-0.2171357	70.23	0.3958884	13.18	-0.2235065	5.48	-0.1637564
62	79976	0.46	0.3151016	0.39	0.16	45.7	0.3151016	26.5	-0.1281563	10.03	-0.205826	17.62	-0.1010763
63	79976	0.27	0.2085069	0.28	0.14	44.37	-0.0597114	18.88	-0.0931046	26.99	0.2085069	9.62	-0.0897549
64	79976	0.35	0.1465731	0.19	0.12	9.57	-0.1740907	35.38	0.1465731	33.11	-0.0793557	21.82	0.0447515
65	79976	0.47	0.3626572	0.45	0.13	21.63	-0.1350525	47.28	0.3626572	16.95	-0.1712262	14.01	-0.1761319

Note. This test included multiple-choice items only. The statistics presented in this table are based on a calibration sample, which was near census for this administration.

PART 7: CALIBRATION, SCALING AND EQUATING

Part 7 of the technical report describes calibration and scaling procedures and results for the Spring 2016 AIMS Science assessments. Each grade level was calibrated and scaled with calibration samples that typically consisted close to the entire student population. Part 7 of this report addresses the following AERA/APA/NCME *Standards* from the 1999 edition: 1.13, 2.1, 2.2, 2.14, 4.1, 4.2, 4.3, 6.4, 6.5, and 13.6. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this chapter are: 1.10, 2.3, 2.13, 2.14, 5.1, 5.2, 5.3, 7.2, 7.4, and 12.9.

7.1 Ensuring Valid Records in Calibration Sample

In order to ensure valid calibration results, several data cleaning steps occurred upon receipt of raw data from the scanning and scoring processes. These steps allowed for calibration to be conducted on valid student responses at the targeted grade level.

The cleaning process removed the following records from the calibration datasets for each content area and grade level:

- records with invalid tests noted by a special invalidation code obtained from ADE and marked on the answer document;
- records with non-valid attempts noted by less than one response in any of the test sessions;
- records for Bureau of Indian Affairs schools, juvenile corrections centers, state hospital schools, private schools, and home schooled students;
- records for students in cohorts other than 2018 or 2019 (high school tests only);
- records which indicated the student took a test other than their grade level test; and
- duplicate records (score sheets were double scanned or students indicated as taking the test more than one time).

7.2 Calibration Methods

Item Response Theory (IRT) models were used in the item calibration for all AIMS Science tests. Each grade-level test was calibrated separately. . All calibration activities were replicated by ADE staff as an added quality control check.

7.2.1 Calibration Model

The AIMS Science assessments are composed of multiple-choice items. Historically, the AIMS Science tests have been developed and calibrated using the Rasch Model. The Rasch model (Rasch, 1960; Wright, 1977) can be conceptualized as a one-parameter IRT (1PL) model in which item difficulty and student ability are estimated on the same scale. The Rasch model defines a multiple-choice item in terms of one parameter: item difficulty. In the Rasch model, the probability that a student with an ability estimate (θ) responds correctly to item i is:

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the item difficulty.

7.2.2 Calibration Software

Parameter estimation for items on the science tests in grade 4, 8, and high school was implemented using WINSTEPS 3.71 (Linacre, 2011). WINSTEPS uses joint maximum likelihood estimation (JMLE) as described by Wright and Masters (1982).

7.3 Calibration Results

7.3.1 IRT Item Statistics

Item statistics resulting from calibration of the AIMS science tests for grades 4, 8, and high school are presented in Tables 7.3.1.1 through 7.3.1.3. These tables contain each item's Rasch difficulty, standard error of the difficulty (SE), weighted mean-square (MNSQ infit), and unweighted mean-square (MNSQ outfit).

All items for all AIMS tests converged during calibration using typical procedures for WINSTEPS software. Typically in IRT, Rasch difficulty values range from -3.00 to +3.00 with positive values indicating that the item is relatively difficult and negative values indicating that it is relatively easy. Standard error of estimates for the Rasch difficulty measures indicated that the parameters were well estimated. Model-to-item data fit was monitored using MNSQ infit and MNSQ outfit statistics, which indicate the degree of accuracy and predictability with which the data fits the model (Linacre, 2002). The MNSQ infit is sensitive to unexpected responses at or near the item's calibrated level; whereas, MNSQ outfit is sensitive to unexpected responses away from the item's calibrated level. Typically, values less than 0.6 and greater than 1.4 for MNSQ infit infit indicate misfit, and values greater than 1.4 for MNSQ outfit indicate misfit (Wright & Linacre, 1994). No item was flagged as having misfit as indicated by either MNSQ infit or MNSQ outfit.

Table 7.3.1.1
Spring 2016 AIMS IRT Item Statistics
Science Grade 4

Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit	Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit
1	-0.57	0.01	1.03	1.08	28	0.42	0.01	1.05	1.04
2	-0.20	0.01	1.06	1.10	29	-0.11	0.01	0.84	0.74
3	-0.45	0.01	0.95	0.90	30	1.16	0.01	1.09	1.12
4	0.48	0.01	0.93	0.89	31	1.90	0.01	1.17	1.33
5	0.10	0.01	0.99	0.98	32	1.55	0.01	1.00	1.05
6	0.89	0.01	1.07	1.09	33	1.28	0.01	0.95	0.95
7	0.52	0.01	0.97	0.97	34	0.38	0.01	0.89	0.84
8	0.27	0.01	0.97	0.95	35	0.34	0.01	0.93	0.90
9	1.18	0.01	1.18	1.22	36	0.25	0.01	1.18	1.26
10	-0.41	0.01	0.90	0.82	37	1.61	0.01	1.15	1.20
11	0.89	0.01	1.01	1.01	38	1.15	0.01	1.05	1.08
12	0.87	0.01	0.97	0.96	39	1.39	0.01	1.03	1.08
13	1.18	0.01	1.07	1.10	40	-0.33	0.01	1.02	1.07
14	1.41	0.01	1.04	1.07	41	0.71	0.01	1.02	1.05
15	0.28	0.01	0.86	0.79	42	0.15	0.01	1.01	1.02
16	1.66	0.01	1.17	1.26	43	0.67	0.01	0.93	0.89
17	1.62	0.01	1.13	1.22	44	2.39	0.01	1.12	1.37
18	0.42	0.01	1.02	1.06	45	0.48	0.01	0.87	0.83
19	0.82	0.01	0.93	0.90	46	-0.83	0.01	0.87	0.70
20	-0.31	0.01	0.91	0.85	47	0.94	0.01	0.98	0.98
21	-0.04	0.01	0.88	0.83	48	0.51	0.01	0.96	0.97
22	0.82	0.01	1.03	1.03	49	-0.19	0.01	0.83	0.71
23	-0.38	0.01	0.83	0.70	50	0.69	0.01	0.99	0.97
24	1.13	0.01	1.07	1.08	51	0.68	0.01	0.99	0.97
25	1.64	0.01	1.17	1.27	52	0.64	0.01	0.93	0.90
26	1.28	0.01	1.06	1.08	53	0.48	0.01	1.05	1.05
27	-0.13	0.01	0.91	0.83	54	0.13	0.01	0.92	0.87

Table 7.3.1.2
Spring 2016 AIMS IRT Item Statistics
Science Grade 8

Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit	Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit
1	-1.12	0.01	1.03	1.18	30	-0.09	0.01	0.84	0.78
2	-0.33	0.01	1.01	1.05	31	0.91	0.01	1.03	1.04
3	0.28	0.01	1.20	1.38	32	-1.06	0.01	0.94	0.79
4	-0.46	0.01	0.98	0.98	33	-0.28	0.01	0.90	0.83
5	0.85	0.01	0.88	0.85	34	1.08	0.01	1.02	1.03
6	0.50	0.01	1.14	1.21	35	0.87	0.01	1.08	1.12
7	-0.78	0.01	0.93	0.90	36	1.00	0.01	1.04	1.06
8	-1.29	0.01	0.79	0.65	37	0.02	0.01	0.92	0.89
9	0.21	0.01	1.09	1.17	38	0.37	0.01	1.04	1.11
10	0.29	0.01	1.02	1.03	39	0.93	0.01	1.01	1.02
11	-0.63	0.01	1.00	1.05	40	0.22	0.01	0.91	0.87
12	-0.15	0.01	0.97	0.94	41	-0.35	0.01	0.90	0.84
13	1.20	0.01	1.13	1.19	42	-0.70	0.01	0.87	0.72
14	1.05	0.01	1.05	1.07	43	0.67	0.01	0.98	0.97
15	1.88	0.01	1.23	1.46	44	-0.97	0.01	0.93	0.71
16	0.04	0.01	1.01	1.01	45	1.02	0.01	1.13	1.18
17	0.23	0.01	0.99	0.99	46	0.75	0.01	0.98	0.97
18	-0.38	0.01	0.93	0.87	47	0.68	0.01	1.05	1.06
19	0.50	0.01	0.99	0.96	48	0.37	0.01	0.89	0.86
20	1.99	0.01	1.01	1.06	49	-0.32	0.01	0.91	0.84
21	0.79	0.01	1.01	1.01	50	1.73	0.01	1.06	1.19
22	-0.08	0.01	1.02	1.06	51	0.53	0.01	1.02	1.03
23	0.71	0.01	0.95	0.94	52	0.46	0.01	0.99	0.98
24	1.64	0.01	0.99	1.04	53	-0.64	0.01	0.83	0.68
25	1.27	0.01	0.92	0.93	54	0.56	0.01	0.97	0.94
26	1.13	0.01	1.04	1.07	55	1.73	0.01	1.08	1.16
27	2.45	0.01	1.13	1.54	56	-0.05	0.01	0.96	0.94
28	1.17	0.01	1.06	1.11	57	0.86	0.01	1.17	1.21
29	0.98	0.01	0.97	0.97	58	0.73	0.01	1.06	1.07

Table 7.3.1.3
Spring 2016 AIMS IRT Item Statistics
Science Grade HS

Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit	Item	Rasch Difficulty	SE	MNSQ Infit	MNSQ Outfit
1	0.20	0.01	1.02	1.06	34	1.56	0.01	1.02	1.08
2	0.20	0.01	0.87	0.85	35	-0.10	0.01	0.86	0.82
3	0.76	0.01	1.03	1.04	36	-0.73	0.01	0.86	0.76
4	0.93	0.01	1.11	1.14	37	-0.06	0.01	0.92	0.87
5	0.98	0.01	1.15	1.19	38	0.49	0.01	0.95	0.94
6	1.03	0.01	1.09	1.12	39	0.82	0.01	1.01	1.01
7	1.70	0.01	0.94	0.97	40	0.18	0.01	0.94	0.91
8	0.08	0.01	0.99	0.99	41	1.33	0.01	1.12	1.20
9	-0.26	0.01	0.99	1.04	42	1.27	0.01	1.06	1.10
10	0.14	0.01	0.95	0.92	43	1.02	0.01	1.04	1.06
11	0.19	0.01	1.00	1.00	44	0.77	0.01	1.03	1.05
12	1.48	0.01	1.09	1.12	45	1.38	0.01	1.04	1.07
13	-0.74	0.01	0.92	0.88	46	0.61	0.01	0.89	0.87
14	0.23	0.01	1.14	1.19	47	0.31	0.01	0.96	0.94
15	-0.74	0.01	0.99	1.14	48	1.71	0.01	1.13	1.24
16	-0.18	0.01	0.97	0.97	49	0.73	0.01	0.99	0.99
17	0.36	0.01	0.97	0.96	50	-0.24	0.01	0.95	0.91
18	0.37	0.01	0.90	0.87	51	0.42	0.01	0.98	0.98
19	1.36	0.01	1.08	1.12	52	0.19	0.01	0.94	0.92
20	-1.43	0.01	0.90	0.73	53	0.41	0.01	0.99	0.98
21	-0.56	0.01	0.99	1.06	54	0.30	0.01	0.84	0.80
22	0.45	0.01	1.05	1.10	55	1.23	0.01	1.12	1.19
23	-0.49	0.01	0.94	0.89	56	0.89	0.01	1.07	1.09
24	0.23	0.01	0.98	0.98	57	0.78	0.01	1.05	1.06
25	0.85	0.01	0.97	0.97	58	1.09	0.01	0.99	1.01
26	1.40	0.01	1.08	1.13	59	0.77	0.01	1.02	1.03
27	0.21	0.01	0.99	1.02	60	0.35	0.01	0.97	0.96
28	0.65	0.01	0.99	1.00	61	-0.43	0.01	0.89	0.83
29	0.99	0.01	0.99	1.01	62	0.62	0.01	1.03	1.02
30	0.75	0.01	1.08	1.10	63	1.63	0.01	1.10	1.21
31	0.67	0.01	0.98	0.98	64	1.17	0.01	1.18	1.25
32	0.48	0.01	1.00	1.00	65	0.46	0.01	0.98	0.98
33	0.86	0.01	1.00	1.00					

7.4 Scaling Methods

7.4.1 Science

A scale of measurement was determined for science using spring 2008 operational test results and cut scores were determined during standard setting meetings. A detailed description concerning the development of the scale of measurement can be found in Appendix B of the *2008 AIMS Technical Report* which can be obtained from the Arizona Department of Education. A report detailing the procedures used to set performance standards on the science tests is available at http://www.azed.gov/assessment/files/2014/05/aims2008sciencerevisedstandardsettingtechnicalreport.pdf. The AIMS science scales for grades 4, 8, and high school ranged from 200 to 800. The science scales are not on a vertical scale. Each grade has its own unique scale so that the scale scores for different grades can NOT be compared.

7.5 Equating

7.5.1 Science

The 2016 AIMS Science tests were equated and placed on the operational AIMS scale using a common-item, non-equivalent groups design. A set of anchor items was selected from the 2015 and previous operational assessments before the item selection workshop. The anchor items were selected with two principles in mind. First, the subset of anchor items should represent the content covered by the full AIMS assessment. Second, the subset of anchor items should be representative of the distribution of item difficulties for the full assessment. Table 7.5.1.1 presents the number of anchor/common items for each grade/subject area. Table 7.5.1.2 show the content representation for the 2016 anchor items compared to the 2016 operational form. Table 7.5.1.3 presents descriptive statistics for the 2016 anchor/common item difficulties and the 2016 operational form.

Table 7.5.1.1
Spring 2016 AIMS Anchor Items

Content	Grade	CRT Total	Anchor
Science	4	54	22
	8	58	23
	HS	65	22

Table 7.5.1.2
Representation of Content by 2016 Anchor Sets, Science

		Strand																		
		1				2		3		4					5			6		Total
		Concept				Concept		Concept		Concept					Concept			Concept		
		1	2	3	4	1	2	1	2	1	2	3	4	5	1	2	3	2	3	
SC04 All	N	6	6	5	1	3	3	5	1	3	0	0	3		0	0	6	6	6	54
	Pct	11.11	11.11	9.26	1.85	5.56	5.56	9.26	1.85	5.56	0	0	5.56		0	0	11.11	11.11	11.11	100
Anchor	N	3	2	2		1	2	2	1	2					0	0	3	2	2	22
	Pct	13.64	9.09	9.09		4.55	9.09	9.09	4.55	9.09					0	0	13.64	9.09	9.09	100
SC08 All	N	6	4	6	4	4	2	2	4	0	3	0	5		10	8				58
	Pct	10.34	6.9	10.34	6.9	6.9	3.45	3.45	6.9	0	5.17	0	8.62		17.24	13.79				100
Anchor	N	2	2	2	1	1	2	0	3	0	0	0	3		4	3				23
	Pct	8.7	8.7	8.7	4.35	4.35	8.7	0	13.04	0	0	0	13.04		17.39	13.04				100
SCHS All	N	6	6	6	4	4	2	7		6	6	6	6	6						65
	Pct	9.23	9.23	9.23	6.15	6.15	3.08	10.77		9.23	9.23	9.23	9.23	9.23						100
Anchor	N	2	2	2	1	1	1	3		2	2	2	2	2						22
	Pct	9.09	9.09	9.09	4.55	4.55	4.55	13.64		9.09	9.09	9.09	9.09	9.09						100

Table 7.5.1.3
Representation of Difficulty by 2016 Anchor Sets, Science

			Difficulty Parameter		P-Value	
Content	Grade	Statistic	Entire 2016 Test	All Anchor Items	Entire 2016 Test	All Anchor Items
SC	4	N	54	22	54	22
		Mean	0.6200	0.5600	0.5600	0.5700
		Std Dev	0.7100	0.5200	0.1400	0.1000
		Min	-0.8300	-0.8300	0.2300	0.4000
		Max	2.3900	1.3900	0.8200	0.8200
SC	8	N	58	23	58	23
		Mean	0.4300	0.4000	0.5700	0.5800
		Std Dev	0.8300	0.8600	0.1500	0.1500
		Min	-1.2900	-1.2900	0.2100	0.2900
		Max	2.4500	1.9900	0.8800	0.8800
SC	HS	N	65	22	65	22
		Mean	0.5200	0.3500	0.4900	0.5200
		Std Dev	0.6600	0.5800	0.1300	0.1100
		Min	-1.4300	-0.7400	0.2600	0.3300
		Max	1.7100	1.3600	0.8400	0.7500

A fixed-parameter equating was implemented within WINSTEPS in order to link the 2016 science tests to the operational reporting scale. This is implemented by constraining the 2016 parameter estimates for the common anchor items to equal the final parameter estimates obtained in the most recent AIMS calibration analyses. The displacement statistic, which estimates the difference between the fixed parameter and the estimate had the item parameter not been constrained, was evaluated for each anchor item. Displacement statistic greater than 0.5 or less than -0.5 are considered significant in the Rasch literature and caused the anchor item to be removed from the anchor set. The following procedure was used to examine anchor item performance and determine whether to remove anchor items that exhibited significant displacement statistics from the annual equating:

1. All anchor items with displacement statistics greater than 0.3 or less than -0.3 were flagged. Any anchor item with displacement statistic greater than 0.3 or less than -0.3 was dropped from the anchor item set. If more than one item was observed with a displacement statistic greater than 0.3 or less than -0.3, then only the first item with the largest displacement value was dropped from the anchor set. The displacement values of the remaining anchor items were re-estimated by equating the test again using the remaining anchor items. This process of equating and dropping the anchor item with the largest displacement greater than 0.3 or less than -0.3 was repeated until all displacements were acceptable. All items with displacement values greater than 0.3 or less than -0.3 were noted to be carried over for removal from the anchor set for next year.

2. Whenever an anchor item was removed, content and difficulty representativeness of the remaining anchor set was examined. In instances where more than one anchor item was considered for removal for a given content and grade, the content strand and difficulty level of the item was

considered to prevent removal of more than one item from the same content strand and difficulty level.

3. If more than one item was removed from the same content strand, a note was made to address the problem in the setup of anchors for the succeeding year's assessment.

This procedure resulted in removing only one item from the anchor sets for grade 4, 8, and high school. This one item was within the grade 8 test leaving 3 items aligned to the same strand and concept as anchors, 38% of the operational items so aligned.

7.5.2 Scoring and Standard Error of Measurement

Item response theory makes available two types of scoring: number-correct and item-pattern. With number-correct scoring, the value of theta corresponding to each number-correct score (or raw score) is converted to a scale score. Item-pattern scoring produces a scale score, taking into account not only how many items were answered correctly but also which items and the characteristics of those items. For groups of 25 or more students, the two methods produce tau-equivalent results (Yen, 1984.) Tau-equivalent means that examinees are expected to receive the same score on average between the two methods. Number-correct scoring was used to derive scales scores for the AIMS tests.

Typically, a test score is obtained from a single observation of performance and represents an estimate of the trait being measured. As an estimate, an observed test score contains some measurement error and does not perfectly reflect an individual's true score. The degree of measurement error in a test score can be estimated using a statistic called the standard error of measurement (SEM).

A student's exact true score cannot be known. The true score is defined as the average test score that would result if the test could be administered repeatedly without the effects of practice or fatigue. The standard error of measurement is an estimate of the standard deviation of an individual's observed scores from these repeated administrations. For practical purposes, this statistic can be used to obtain a range within which a student's true score is likely to fall. Using item response theory, the standard error of measurement can be calculated for every possible scale score.

Tables 7.5.2.1 through 7.5.2.3 present raw score to scale score conversion tables and IRT conditional standard errors of measurement for Science grades 4, 8, and high school assessments. The values in bold represent the scale score with the smallest value greater than or equal to the established cut score for each grade level and content area. The "greater than" rule is evoked when the actual scale score is not observed in any given table.

Table 7.5.2.1
Spring 2016 AIMS Raw Score to Scale Score Table
Science Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	200	70	**28**	**502**	**14**
1	293	50	29	506	14
2	328	36	30	511	14
3	350	29	31	515	14
4	365	26	32	519	14
5	377	23	33	523	14
6	388	22	34	527	15
7	397	20	35	532	15
8	405	19	36	536	15
9	412	18	37	541	15
10	419	18	38	545	15
11	425	17	**39**	**550**	**16**
12	431	17	40	555	16
13	436	16	41	561	16
14	442	16	42	566	17
15	447	16	43	572	17
16	451	15	44	578	18
17	456	15	45	585	19
18	461	15	46	592	19
19	**465**	**15**	47	600	20
20	469	15	48	609	22
21	474	14	49	620	24
22	478	14	50	632	26
23	482	14	51	648	30
24	486	14	52	669	36
25	490	14	53	705	50
26	494	14	54	800	70
27	498	14			

Note. SEM is the standard error of measurement for the scale score.
Note. Cut scores for Approaches the Standard, Meets the Standard, and Exceeds the Standard are in boldface. The complete list of AIMS Science score cuts and ranges is presented in Table 10.1.1.

Table 7.5.2.2
Spring 2016 AIMS Raw Score to Scale Score Table
Science Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	200	68	30	497	14
1	283	49	**31**	**500**	**14**
2	318	35	32	504	14
3	339	29	33	508	14
4	354	25	34	512	14
5	367	23	35	516	14
6	377	21	36	520	14
7	386	20	37	524	14
8	394	19	38	528	14
9	401	18	**39**	**532**	**14**
10	408	17	40	537	15
11	414	17	41	541	15
12	419	16	42	546	15
13	425	16	43	551	15
14	430	16	44	556	16
15	435	15	45	561	16
16	440	15	46	566	16
17	444	15	47	572	17
18	449	15	48	578	17
19	453	14	49	585	18
20	457	14	50	592	19
21	462	14	51	600	20
22	466	14	52	609	21
23	470	14	53	619	23
24	**474**	**14**	54	631	26
25	477	14	55	647	29
26	481	14	56	668	35
27	485	14	57	702	49
28	489	14	58	800	69
29	493	14			

Note. SEM is the standard error of measurement for the scale score.
Note. Cut scores for Approaches the Standard, Meets the Standard, and Exceeds the Standard are in boldface. The complete list of AIMS Science score cuts and ranges is presented in Table 10.1.1.

Table 7.5.2.3
Spring 2016 AIMS Raw Score to Scale Score Table
Science High School

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	200	74	33	493	14
1	263	53	34	497	14
2	300	38	**35**	**500**	**14**
3	323	31	36	504	14
4	339	27	37	508	14
5	352	25	38	511	14
6	363	23	39	515	14
7	372	21	40	518	14
8	381	20	41	522	14
9	388	19	42	526	14
10	395	18	43	530	14
11	401	18	44	534	14
12	407	17	**45**	**538**	**15**
13	413	17	46	542	15
14	418	16	47	546	15
15	423	16	48	550	15
16	428	16	49	555	15
17	432	15	50	560	16
18	437	15	51	565	16
19	441	15	52	570	17
20	445	15	53	575	17
21	449	14	54	581	18
22	453	14	55	587	18
23	457	14	56	594	19
24	461	14	57	601	20
25	465	14	58	609	21
26	468	14	59	619	23
27	472	14	60	629	25
28	**476**	**14**	61	642	27
29	479	14	62	658	31
30	483	14	63	680	38
31	486	14	64	718	53
32	490	14	65	800	74

Note. SEM is the standard error of measurement for the scale score.
Note. Cut scores for Approaches the Standard, Meets the Standard, and Exceeds the Standard are in boldface. The complete list of AIMS Science score cuts and ranges is presented in Table 10.1.1.

PART 8: TEST RESULTS

8.1 Data

Part 8 of this technical report contains information about the results of Spring 2016AIMS Science in grades 4, 8, and high school. The 1999 AERA/APA/NCME *Standards* addressed in Part 8 include: 1.5, 4.3, 4.5, 4.6, 4.7, 6.5, 7.1, 7.10, 13.15, and 13.19. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this chapter are: 1.10, 5.1, 5.2, 5.3, 5.8, 5.9, 7.2, 7.4, and 12.9. Please note that the corresponding information for the Fall 2014 AIMS administration can be found in Appendix A.

Results presented below are based on population data contained within the final electronic data files and gone through the same clean-up process as calibration data in Part 7. The results presented in this part of the technical report may differ slightly from final testing results presented on the Arizona Department of Education website due to slight differences in the application of exclusion rules. Official final results typically use more detailed school-level information than is used to conduct research analyses. The results in the following tables are presented as evidence of reliability and validity of the AIMS assessments and should not be used for state accountability purposes.

8.1.1 AIMS State Test Results

The AIMS test results for Science for grades 4, 8, and high school are not on a vertical scale and therefore the scale scores across grades can not be compared. For each grade, the lowest obtainable scale score (LOSS) on the science tests is 200, and the highest obtainable scale score (HOSS) is 800.

Test results are presented in Tables 8.1.1.2 through 8.1.1.6. For each grade, scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state as a whole and disaggregated into various demographic groups.

In addition to the descriptive statistics presented in Tables 8.1.1.2 through 8.1.1.6, scale score frequency distributions are displayed in Tables 8.1.1.7 through 8.1.1.13. The information for each grade is contained within a separate table. These tables show the scale score, frequency (Freq), cumulative frequency (Cum Freq), percentage (%), and cumulative percentage (Cum %).

Results for AIMS assessments for high school are reported by graduating cohort for Science. Cohort 19 is defined as the group of students that expect to graduate in 2019 and typically includes grade 9 students. Cohort 18 is defined as the group of students that expect to graduate in 2018 and typically includes 10th grade students.

Table 8.1.1.2
Spring 2016 AIMS State Test Results
Science Grades 4 and 8

		Scale Score		% at Performance Level			
	N	M	SD	FFBS	AS	MS	ES
Grade 4							
Total	85675	514.49	47.56	15	25	34	25
Hispanic	38981	499.24	42.36	21	33	33	14
Non Hispanic	46248	527.43	47.85	10	19	36	35
Race							
White	67687	516.87	47.37	14	24	35	27
Black or African American	5144	496.74	43.78	24	32	31	13
Asian	2529	541.12	47.81	6	13	35	46
American Indian or Alaskan Native	5648	488.63	38.24	27	37	28	7
Native Hawaiian or Other Pacific Islander	427	510.79	42.66	15	27	38	20
Multiple Indication	3468	521.82	46.57	11	23	36	30
Special Program Membership							
English Learner Program	7919	464.96	28.71	51	38	10	1
Special Education	10781	485.41	43.29	35	33	22	10
Low SES	44836	499.25	42.38	21	32	33	14
Migrant	465	478.67	38.55	40	33	21	6
Grade 8							
Total	82258	512.64	48.83	22	19	25	35
Hispanic	36492	497.47	42.03	29	24	26	22
Non Hispanic	45511	524.92	50.44	16	15	24	46
Race							
White	66016	515.00	48.32	20	18	25	37
Black or African American	4960	496.84	43.98	31	23	25	22
Asian	2499	548.66	55.11	8	9	20	63
American Indian or Alaskan Native	5578	485.15	38.92	40	26	22	13
Native Hawaiian or Other Pacific Islander	307	504.18	44.04	27	18	29	26
Multiple Indication	2364	514.59	47.34	20	18	25	36
Special Program Membership							
English Learner Program	2852	450.43	26.01	82	13	4	1
Special Education	8564	469.74	38.8	60	20	12	8
Low SES	39908	496.97	42.17	30	24	25	21
Migrant	488	482.22	39.91	45	19	21	14

Note. FFBS= Falls Far Below the Standard; AS= Approaches the Standard; MS= Meets the Standard; ES= Exceeds the Standard. Students with no valid attempt, invalidation, or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections centers, and students attending state hospital schools are not included in this summary. These results are not final results and are presented here for purposes of addressing reliability and validity. These results should not be used for accountability purposes. Science results are not on a vertical scale.

Table 8.1.1.3
Spring 2016 AIMS State Test Results
Science High School

		Scale Score		% at Performance Level			
	N	M	SD	FFBS	AS	MS	ES
Cohort 18							
Total	46242	482.21	44.16	48	20	19	12
Hispanic	21462	470.23	37.33	60	20	15	6
Non Hispanic	24313	492.80	46.91	38	20	23	18
Race							
White	36292	484.35	44.20	46	20	21	13
Black or African American	2917	469.99	37.96	60	19	15	6
Asian	1166	512.29	56.19	26	18	24	32
American Indian or Alaskan Native	3333	462.47	33.73	68	18	10	4
Native Hawaiian or Other Pacific Islander	166	474.83	42.28	58	16	17	9
Multiple Indication	1366	485.54	43.18	44	21	23	12
Special Program Membership							
English Learner Program	1556	436.84	21.06	95	4	1	0
Special Education	5068	452.67	33.20	80	10	7	3
Low SES	22909	470.84	38.04	59	20	15	6
Migrant	495	452.46	29.22	78	15	6	1
Cohort 19							
Total	33782	499.41	48.36	33	20	25	22
Hispanic	13243	482.85	41.55	47	21	21	11
Non Hispanic	19512	510.99	49.31	24	19	27	30
Race							
White	25922	501.71	48.12	31	20	26	23
Black or African American	1746	481.65	41.68	47	23	20	10
Asian	1336	532.70	51.69	13	13	27	46
American Indian or Alaskan Native	1104	472.21	39.04	58	19	16	7
Native Hawaiian or Other Pacific Islander	156	498.27	44.92	28	28	23	22
Multiple Indication	894	501.07	48.09	32	20	24	23
Special Program Membership							
English Learner Program	333	439.97	24.37	92	5	2	1
Special Education	2179	456.37	37.64	77	11	7	5
Low SES	13826	481.55	41.4	48	22	20	11
Migrant	27	442.89	29.02	78	19	4	0

Note. FFBS= Falls Far Below the Standard; AS= Approaches the Standard; MS= Meets the Standard; ES= Exceeds the Standard. Students with no valid attempt, invalidation, or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections centers, and students attending state hospital schools are not included in this summary. These results are not final results and are presented here for purposes of addressing reliability and validity. These results should not be used for accountability purposes. Science results are not on a vertical scale.

Table 8.1.1.4
Spring 2016 AIMS Frequency Distribution
Science CRT Grade 4

Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %	Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %
0	200	0	0	0	0	**28**	**502**	**2574**	**3**	**37467**	**43.61**
1	293	1	0	1	0	29	506	2640	3.07	40107	46.68
2	328	0	0	1	0	30	511	2704	3.15	42811	49.83
3	350	1	0	2	0	31	515	2644	3.08	45455	52.91
4	365	2	0	4	0	32	519	2647	3.08	48102	55.99
5	377	7	0.01	11	0.01	33	523	2738	3.19	50840	59.17
6	388	28	0.03	39	0.05	34	527	2753	3.2	53593	62.38
7	397	60	0.07	99	0.12	35	532	2735	3.18	56328	65.56
8	405	146	0.17	245	0.29	36	536	2745	3.19	59073	68.76
9	412	218	0.25	463	0.54	37	541	2667	3.1	61740	71.86
10	419	432	0.5	895	1.04	38	545	2693	3.13	64433	74.99
11	425	694	0.81	1589	1.85	**39**	**550**	**2650**	**3.08**	**67083**	**78.08**
12	431	837	0.97	2426	2.82	40	555	2546	2.96	69629	81.04
13	436	1191	1.39	3617	4.21	41	561	2479	2.89	72108	83.93
14	442	1431	1.67	5048	5.88	42	566	2339	2.72	74447	86.65
15	447	1719	2	6767	7.88	43	572	2081	2.42	76528	89.07
16	451	1970	2.29	8737	10.17	44	578	2103	2.45	78631	91.52
17	456	2174	2.53	10911	12.7	45	585	1793	2.09	80424	93.61
18	461	2192	2.55	13103	15.25	46	592	1499	1.74	81923	95.35
19	**465**	**2222**	**2.59**	**15325**	**17.84**	47	600	1266	1.47	83189	96.82
20	469	2408	2.8	17733	20.64	48	609	1050	1.22	84239	98.05
21	474	2306	2.68	20039	23.32	49	620	761	0.89	85000	98.93
22	478	2404	2.8	22443	26.12	50	632	455	0.53	85455	99.46
23	482	2472	2.88	24915	29	51	648	303	0.35	85758	99.81
24	486	2359	2.75	27274	31.74	52	669	108	0.13	85866	99.94
25	490	2522	2.94	29796	34.68	53	705	40	0.05	85906	99.99
26	494	2493	2.9	32289	37.58	54	800	11	0.01	85917	100
27	498	2604	3.03	34893	40.61						

Note: Freq. = Frequency, Cum = Cumulative. Students with no valid attempt, invalidation or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections facilities, and students attending hospital schools are not included in this summary.

Table 8.1.1.5
Spring 2016 AIMS Frequency Distribution
Science CRT Grade 8

Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %	Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %
0	200	0	0	0	0	30	497	2415	2.93	33586	40.72
1	283	0	0	0	0	**31**	**500**	**2466**	**2.99**	**36052**	**43.71**
2	318	0	0	0	0	32	504	2531	3.07	38583	46.78
3	339	0	0	0	0	33	508	2437	2.95	41020	49.74
4	354	0	0	0	0	34	512	2595	3.15	43615	52.88
5	367	3	0	3	0	35	516	2518	3.05	46133	55.94
6	377	7	0.01	10	0.01	36	520	2562	3.11	48695	59.04
7	386	20	0.02	30	0.04	37	524	2549	3.09	51244	62.13
8	394	65	0.08	95	0.12	38	528	2550	3.09	53794	65.22
9	401	117	0.14	212	0.26	**39**	**532**	**2564**	**3.11**	**56358**	**68.33**
10	408	191	0.23	403	0.49	40	537	2491	3.02	58849	71.35
11	414	296	0.36	699	0.85	41	541	2402	2.91	61251	74.27
12	419	506	0.61	1205	1.46	42	546	2401	2.91	63652	77.18
13	425	667	0.81	1872	2.27	43	551	2283	2.77	65935	79.95
14	430	902	1.09	2774	3.36	44	556	2266	2.75	68201	82.69
15	435	1028	1.25	3802	4.61	45	561	2057	2.49	70258	85.19
16	440	1279	1.55	5081	6.16	46	566	1934	2.34	72192	87.53
17	444	1500	1.82	6581	7.98	47	572	1741	2.11	73933	89.64
18	449	1686	2.04	8267	10.02	48	578	1637	1.98	75570	91.63
19	453	1754	2.13	10021	12.15	49	585	1485	1.8	77055	93.43
20	457	1854	2.25	11875	14.4	50	592	1312	1.59	78367	95.02
21	462	1919	2.33	13794	16.73	51	600	1131	1.37	79498	96.39
22	466	2023	2.45	15817	19.18	52	609	904	1.1	80402	97.49
23	470	2095	2.54	17912	21.72	53	619	756	0.92	81158	98.4
24	**474**	**2148**	**2.6**	**20060**	**24.32**	54	631	541	0.66	81699	99.06
25	477	2130	2.58	22190	26.91	55	647	400	0.48	82099	99.54
26	481	2222	2.69	24412	29.6	56	668	217	0.26	82316	99.81
27	485	2162	2.62	26574	32.22	57	702	126	0.15	82442	99.96
28	489	2230	2.7	28804	34.92	58	800	33	0.04	82475	100
29	493	2367	2.87	31171	37.79						

Note: Freq. = Frequency, Cum = Cumulative. Students with no valid attempt, invalidation or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections facilities, and students attending hospital schools are not included in this summary.

Table 8.1.1.6
Spring 2016 AIMS Frequency Distribution
Science CRT High School Cohort 18

Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %	Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %
0	200	0	0	0	0	33	493	1130	2.43	30537	65.77
1	263	0	0	0	0	34	497	1193	2.57	31730	68.34
2	300	2	0	2	0	**35**	**500**	**1072**	**2.31**	**32802**	**70.65**
3	323	2	0	4	0.01	36	504	1109	2.39	33911	73.04
4	339	1	0	5	0.01	37	508	1022	2.2	34933	75.24
5	352	4	0.01	9	0.02	38	511	949	2.04	35882	77.29
6	363	6	0.01	15	0.03	39	515	941	2.03	36823	79.31
7	372	13	0.03	28	0.06	40	518	885	1.91	37708	81.22
8	381	23	0.05	51	0.11	41	522	812	1.75	38520	82.97
9	388	47	0.1	98	0.21	42	526	779	1.68	39299	84.65
10	395	122	0.26	220	0.47	43	530	737	1.59	40036	86.23
11	401	213	0.46	433	0.93	44	534	713	1.54	40749	87.77
12	407	325	0.7	758	1.63	**45**	**538**	**688**	**1.48**	**41437**	**89.25**
13	413	541	1.17	1299	2.8	46	542	594	1.28	42031	90.53
14	418	754	1.62	2053	4.42	47	546	560	1.21	42591	91.74
15	423	955	2.06	3008	6.48	48	550	522	1.12	43113	92.86
16	428	1188	2.56	4196	9.04	49	555	473	1.02	43586	93.88
17	432	1381	2.97	5577	12.01	50	560	430	0.93	44016	94.81
18	437	1520	3.27	7097	15.29	51	565	371	0.8	44387	95.61
19	441	1609	3.47	8706	18.75	52	570	358	0.77	44745	96.38
20	445	1772	3.82	10478	22.57	53	575	301	0.65	45046	97.03
21	449	1823	3.93	12301	26.5	54	581	285	0.61	45331	97.64
22	453	1815	3.91	14116	30.4	55	587	242	0.52	45573	98.16
23	457	1732	3.73	15848	34.14	56	594	209	0.45	45782	98.61
24	461	1739	3.75	17587	37.88	57	601	156	0.34	45938	98.95
25	465	1630	3.51	19217	41.39	58	609	153	0.33	46091	99.28
26	468	1661	3.58	20878	44.97	59	619	123	0.26	46214	99.54
27	472	1565	3.37	22443	48.34	60	629	77	0.17	46291	99.71
28	**476**	**1579**	**3.4**	**24022**	**51.74**	61	642	62	0.13	46353	99.84
29	479	1423	3.07	25445	54.81	62	658	39	0.08	46392	99.92
30	483	1357	2.92	26802	57.73	63	680	21	0.05	46413	99.97
31	486	1317	2.84	28119	60.57	64	718	11	0.02	46424	99.99
32	490	1288	2.77	29407	63.34	65	800	3	0.01	46427	100

Note: Freq. = Frequency, Cum = Cumulative. Students with no valid attempt, invalidation or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections facilities, and students attending hospital schools are not included in this summary.

Table 8.1.1.7
Spring 2016 AIMS Frequency Distribution
Science CRT High School Cohort 19

Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %	Raw Score	Scale Score	Freq	%	Cum. Freq.	Cum. %
0	200	0	0	0	0	33	493	908	2.68	17034	50.22
1	263	0	0	0	0	34	497	965	2.84	17999	53.06
2	300	2	0.01	2	0.01	**35**	**500**	**946**	**2.79**	**18945**	**55.85**
3	323	2	0.01	4	0.01	36	504	936	2.76	19881	58.61
4	339	2	0.01	6	0.02	37	508	869	2.56	20750	61.17
5	352	0	0	6	0.02	38	511	848	2.5	21598	63.67
6	363	2	0.01	8	0.02	39	515	815	2.4	22413	66.07
7	372	6	0.02	14	0.04	40	518	830	2.45	23243	68.52
8	381	11	0.03	25	0.07	41	522	860	2.54	24103	71.05
9	388	22	0.06	47	0.14	42	526	762	2.25	24865	73.3
10	395	43	0.13	90	0.27	43	530	775	2.28	25640	75.59
11	401	90	0.27	180	0.53	44	534	781	2.3	26421	77.89
12	407	116	0.34	296	0.87	**45**	**538**	**682**	**2.01**	**27103**	**79.9**
13	413	208	0.61	504	1.49	46	542	676	1.99	27779	81.89
14	418	305	0.9	809	2.38	47	546	660	1.95	28439	83.84
15	423	381	1.12	1190	3.51	48	550	624	1.84	29063	85.68
16	428	524	1.54	1714	5.05	49	555	645	1.9	29708	87.58
17	432	594	1.75	2308	6.8	50	560	584	1.72	30292	89.3
18	437	728	2.15	3036	8.95	51	565	467	1.38	30759	90.68
19	441	749	2.21	3785	11.16	52	570	522	1.54	31281	92.21
20	445	859	2.53	4644	13.69	53	575	459	1.35	31740	93.57
21	449	944	2.78	5588	16.47	54	581	404	1.19	32144	94.76
22	453	891	2.63	6479	19.1	55	587	371	1.09	32515	95.85
23	457	984	2.9	7463	22	56	594	289	0.85	32804	96.7
24	461	994	2.93	8457	24.93	57	601	275	0.81	33079	97.51
25	465	989	2.92	9446	27.85	58	609	261	0.77	33340	98.28
26	468	973	2.87	10419	30.71	59	619	203	0.6	33543	98.88
27	472	935	2.76	11354	33.47	60	629	166	0.49	33709	99.37
28	**476**	**928**	**2.74**	**12282**	**36.21**	61	642	92	0.27	33801	99.64
29	479	937	2.76	13219	38.97	62	658	74	0.22	33875	99.86
30	483	974	2.87	14193	41.84	63	680	35	0.1	33910	99.96
31	486	959	2.83	15152	44.67	64	718	10	0.03	33920	99.99
32	490	974	2.87	16126	47.54	65	800	2	0.01	33922	100

Note: Freq. = Frequency, Cum = Cumulative. Students with no valid attempt, invalidation or off-grade are not included in this summary. In addition, home-schooled students, students attending Bureau of Indian Affairs schools, students attending juvenile corrections facilities, and students attending hospital schools are not included in this summary.

8.2 Longitudinal Data

The spring 2008 administration represents the baseline year for the AIMS Science assessment. In this section, the spring 2016 results are presented along with results back to 2008 to provide longitudinal information. Tables 8.2.1 and 8.2.2 include scale score descriptive statistics and performance level distributions for the AIMS Science administrations.

Tables 8.2.1 - 8.2.2 include scale score descriptive statistics (mean scale score (M) and standard deviation (SD), as well as the scale score values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentile ranking (P10 – P90) and the percentage of students scoring within each performance level for the AIMS Science administration from each year. Caution should be taken when interpreting year-to-year or grade-to-grade comparisons, as slight differences in exclusion rules, changes in the manner in which accommodations were identified, and changes in the manner in which high school results were separated may result in different student population characteristics reported in these tables.

Table 8.2.1
Longitudinal Comparison of Scale Scores in Science

			Scale Score		Percentiles				
Grade	Year	N	M	SD	P10	P25	P50	P75	P90
4	2008	80296	501.8	50.2	436	466	503	536	567
	2009	81724	508.2	50.5	443	475	508	540	567
	2010	80982	513.8	52.7	446	478	515	547	583
	2011	81934	534.8	61.7	455	492	536	575	615
	2012	81892	518.9	57.6	448	478	514	554	589
	2013	83028	513.4	51.9	445	477	511	549	581
	2014	83408	513.5	46.6	457	480	510	546	574
	2015	84113	513.8	46.5	452	479	512	547	573
	2016	85917	514.4	47.6	451	478	515	550	578
8	2008	79482	500.6	50.0	435	463	498	534	568
	2009	78703	506.4	50.0	439	471	506	539	571
	2010	79293	510.4	51.5	446	473	508	545	578
	2011	79409	517.7	47.6	454	484	521	551	578
	2012	80019	519.3	47.9	456	487	521	553	581
	2013	81485	516.7	43.1	459	486	518	544	571
	2014	82470	516.7	45.7	459	483	516	546	573
	2015	82248	513.0	48.1	454	479	509	547	573
	2016	82475	512.6	48.8	449	477	512	546	578
HS	2008 (Cohort 10)	45286	477.3	50.1	414	440	475	510	543
	2009 (Cohort 11)	51195	475.8	49.7	410	439	477	508	541
	2010(Cohort 12)	53671	479.1	51.8	414	442	474	512	545
	2011(Cohort 13)	54610	484.6	58.3	407	443	484	524	559
	2011(Cohort 14)	19392	523.7	58.8	446	488	524	559	596
	2012(Cohort 14)	53344	487.0	62.6	403	441	487	528	569
	2012(Cohort 15)	21142	526.3	65.4	441	487	528	569	603
	2013(Cohort 15)	52650	485.7	56.0	414	442	482	521	562
	2013(Cohort 16)	24094	517.3	59.0	438	475	517	556	591
	2014(Cohort 16)	50096	487.2	52.9	421	448	484	522	555
	2014(Cohort 17)	26254	514.5	53.0	445	477	514	550	582
	2015(Cohort 17)	50975	484.2	44.7	432	453	479	514	546
	2015(Cohort 18)	29063	504.2	49.3	441	468	500	537	569
	2016(Cohort 18)	46427	482.2	44.2	432	449	476	508	542
	2016(Cohort 19)	33922	499.4	48.4	441	465	493	530	565

Note: Students without a valid attempt, invalidation, off-grade, a non-standard accommodation (not in 2008), home-schooled students, attending Bureau of Indian Affairs schools, attending juvenile corrections centers (not in 2005), and attending state hospital schools (not in 2005) are not included in this summary. These results are not final results and are presented here for purposes of addressing reliability and validity. Caution should be used when interpreting results across years, as exclusion rules differ slightly and high school identification of grade versus cohort may result in different student population characteristics.

Table 8.2.2
Longitudinal Comparison of Performance Level Distribution in Science

			% at Performance Level			
Grade	Year	N	FFBS	AS	MS	ES
4	2008	80296	22	25	35	18
	2009	81724	17	26	36	21
	2010	80982	17	22	33	28
	2011	81934	12	17	29	43
	2012	81892	16	21	31	32
	2013	83028	17	25	32	26
	2014	83408	12	29	36	22
	2015	84113	13	29	32	26
	2016	85917	15	25	34	25
8	2008	79482	31	20	22	28
	2009	78703	26	19	23	32
	2010	79293	23	18	25	34
	2011	79409	17	17	27	39
	2012	80019	18	15	28	40
	2013	81485	16	18	29	37
	2014	82470	18	20	24	38
	2015	82248	22	20	24	34
	2016	82475	22	19	25	35
HS	2008 (Cohort 10)	45286	49	19	20	12
	2009 (Cohort 11)	51195	50	18	22	11
	2010 (Cohort 12)	53671	50	16	21	14
	2011 (Cohort 13)	54610	43	15	23	18
	2011 (Cohort 14)	19392	19	12	27	41
	2012(Cohort 14)	53344	41	17	21	21
	2012(Cohort 15)	21142	20	14	23	43
	2013(Cohort 15)	52650	44	17	21	18
	2013(Cohort 16)	24094	23	15	25	36
	2014(Cohort 16)	50096	44	17	21	18
	2014(Cohort 17)	26254	24	16	27	33
	2015(Cohort 17)	50975	45	20	21	14
	2015(Cohort 18)	29063	29	19	26	26
	2016(Cohort 18)	46427	48	20	19	12
	2016(Cohort 19)	33922	33	20	25	22

Note: Students without a valid attempt, invalidation, off-grade, a non-standard accommodation (not in 2008), home-schooled students, attending Bureau of Indian Affairs schools, attending juvenile corrections centers (not in 2005), and attending state hospital schools (not in 2005) are not included in this summary. These results are not final results and are presented here for purposes of addressing reliability and validity. Caution should be used when interpreting results across years, as exclusion rules differ slightly and high school identification of grade versus cohort may result in different student population characteristics.

PART 9: VALIDITY EVIDENCE

Part 9 of the technical report provides evidence supporting the reliability and validity of the 2016 AIMS Science assessments in grades 4, 8, and high school. All data presented in this section were computed using population test data available in the final electronic data files gone through the same clean-up process as the calibration data in Part 7. The following AERA/APA/NCME *Standards* from the 1999 edition are addressed: 1.5, 1.7, 2.1, 2.4, 2.10, 2.13, 3.16, 4.15, 6.5, 7.1, 7.3, and 7.10. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this chapter are: 1.8, 1.9, 2.3, 2.7, 2.8, 2.19, 3.3, 3.6, 4.4, 5.19 and 7.4.

9.1 Reliability

AERA/APA/NCME *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) refer to reliability as the "consistency of [a measure] when the testing procedure is repeated on a population of individuals or groups." The 2014 edition of AERA/APA/NCME *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) indicates that "The term *reliability* has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across relications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficient, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency)".

A reliable test produces stable scores; that is, very similar score distributions would result if the test were administered repeatedly under similar conditions to the same students without memory or fatigue affecting the scores. Reliability of the Spring 2016 AIMS Science assessments is an estimate of its internal consistency.

9.1.1 Measures of Internal Consistency

For tests consisting of only constructed response or multiple-choice items, such as AIMS Science tests, Cronbach's alpha is a frequently used measure of internal consistency. Cronbach's alpha is computed as (Crocker & Algina, 1986)

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right),$$

where k = number of items, σ_X^2 = the total score variance, and σ_i^2 = the variance of item i.

Reliability estimates (Alpha) for the the Spring 2016 AIMS Science assessments, for all students as well as for the various subgroups, are presented in Table 9.1.1.1. Note that a high degree of internal consistency is evident for all three tests.

Table 9.1.1.1
Spring 2016 AIMS Internal Consistency for Science

Subgroup	Value	N	Alpha
Grade 4			
All Students		85675	0.90
Ethnicity	Hispanic	38981	0.87
Ethnicity	Non-Hispanic	46248	0.90
Race	White	67687	0.90
Race	Black/African American	5144	0.88
Race	Asian	2529	0.90
Race	Amerian Indian	5648	0.85
Race	Hawaii/Pacific Islander	427	0.88
Race	Multiple Indicators	3468	0.89
Gender	Female	42134	0.89
Gender	Male	43459	0.90
ELL	Yes	7919	0.72
SPED	Yes	10781	0.88
Low SES	Yes	44836	0.87
Migrant	Yes	465	0.85
Grade 8			
All Students		82258	0.91
Ethnicity	Hispanic	36492	0.88
Ethnicity	Non-Hispanic	45511	0.91
Race	White	66016	0.90
Race	Black/African American	4960	0.89
Race	Asian	2499	0.91
Race	Amerian Indian	5578	0.86
Race	Hawaii/Pacific Islander	307	0.89
Race	Multiple Indicators	2364	0.90
Gender	Female	40466	0.90
Gender	Male	41752	0.91
ELL	Yes	2852	0.67
SPED	Yes	8564	0.85
Low SES	Yes	39908	0.88
Migrant	Yes	488	0.87
HS			
All Students		80024	0.90
Ethnicity	Hispanic	34705	0.87
Ethnicity	Non-Hispanic	43825	0.91
Race	White	62214	0.90
Race	Black/African American	4663	0.87
Race	Asian	2502	0.92
Race	Amerian Indian	4437	0.84
Race	Hawaii/Pacific Islander	322	0.90
Race	Multiple Indicators	2260	0.90
Gender	Female	39981	0.90
Gender	Male	39781	0.91
ELL	Yes	1889	0.51
SPED	Yes	7247	0.82
Low SES	Yes	36735	0.87
Migrant	Yes	522	0.75

Presented in Tables 9.1.1.2 through 9.1.1.4 are number of items, mean and standard deviation (STD) of the raw scores, and the internal consistency reliability estimates (Alpha) at the science strand and concept level.

Table 9.1.1.2
Spring 2016 AIMS Strand/Concept Internal Consistency
Science Grade 4

Strand	Concept	Number of Items	N	Raw Score Mean	Raw Score STD	Alpha
1. Scientific Inquiry		18	85917	10.72	3.86	0.77
	Concept 1: Observations, Questions, and Hypotheses	6	85917	3.37	1.55	0.52
	Concept 2: Scientific Testing (Investigating and Modeling)	6	85917	3.87	1.49	0.50
	Concept 3/4: Analysis and Conclusions/Communication	6	85917	3.47	1.64	0.57
2. History and Nature of Science		6	85917	3.96	1.64	0.62
	Concept 1/2: History of Science as a Human Endeavor/Nature of Scientific Knowledge	6	85917	3.96	1.64	0.62
3. Science in Personal and Social Perspectives		6	85917	3.37	1.63	0.54
	Concept 1/2: Changes in Environments/Science and Technology in Society	6	85917	3.37	1.63	0.54
4. Life Science		6	85917	3.68	1.57	0.58
	Concept 1/3/4: Characteristics of Organisms/Organisms and Environments/Diversity, Adaptation, and Behavior	6	85917	3.68	1.57	0.58
5. Physical Science		6	85917	3.07	1.57	0.52
	Concept 3: Energy and Magnetism	6	85917	3.07	1.57	0.52
6. Earth and Space Science		12	85917	5.52	2.34	0.54
	Concept 2: Earth's Processes and Systems	6	85917	2.58	1.52	0.47
	Concept 3: Changes in the Earth and Sky	6	85917	2.94	1.35	0.29

Table 9.1.1.3
Spring 2016 AIMS Strand/Concept Internal Consistency
Science Grade 8

Strand	Concept	Number of Items	N	Raw Score Mean	Raw Score STD	Alpha
1. Scientific Inquiry		20	82475	12.04	3.91	0.75
	Concept 1: Observations, Questions, and Hypotheses	6	82475	3.66	1.58	0.54
	Concept 2: Scientific Testing (Investigating and Modeling)	4	82475	2.32	1.01	0.32
	Concept 3: Analysis, Conclusions, and Refinements	6	82475	3.21	1.51	0.45
	Concept 4: Communication	4	82475	2.85	1.07	0.40
2. History and Nature of Science		6	82475	3.08	1.63	0.54
	Concept 1/2: History of Science as a Human Endeavor/Nature of Scientific Knowledge	6	82475	3.08	1.63	0.54
3. Science in Personal and Social Perspectives		6	82475	3.53	1.63	0.56
	Concept 1/2: Changes in Environments/Science and Technology in Society	6	82475	3.53	1.63	0.56
4. Life Science		8	82475	4.79	2.06	0.65
	Concept 2/4: Reproduction and Heredity/Diversity, Adaptation, and Behavior	8	82475	4.79	2.06	0.65
5. Physical Science		18	82475	9.87	3.45	0.72
	Concept 1: Properties and Changes of Properties in Matter	10	82475	5.21	2.05	0.56
	Concept 2: Motion and Forces	8	82475	4.66	1.88	0.60

Table 9.1.1.4
Spring 2016 AIMS Strand/Concept Internal Consistency
Science High School

Strand	Concept	Number of Items	N	Raw Score Mean	Raw Score STD	Alpha
1. Scientific Inquiry		22	80349	11.23	4.59	0.79
	Concept 1: Observations, Questions, and Hypotheses	6	80349	3.20	1.65	0.56
	Concept 2: Scientific Testing (Investigating and Modeling)	6	80349	2.77	1.55	0.47
	Concept 3: Analysis, Conclusions, and Refinements	6	80349	3.22	1.56	0.50
	Concept 4: Communication	4	80349	2.05	1.11	0.37
2. History and Nature of Science		6	80349	3.25	1.49	0.45
	Concept 1/2: History of Science as a Human Endeavor/Nature of Scientific Knowledge	6	80349	3.25	1.49	0.45
3. Science in Personal and Social Perspectives		7	80349	3.01	1.66	0.49
	Concept 1/2/3: Changes in Environments/Science and Technology in Society/Human Population Characteristics	7	80349	3.01	1.66	0.49
4. Life Science		30	80349	14.21	5.60	0.81
	Concept 1: The Cell	6	80349	2.35	1.50	0.45
	Concept 2: Molecular Basis of Heredity	6	80349	3.07	1.61	0.54
	Concept 3: Interdependence of Organisms	6	80349	3.41	1.60	0.58
	Concept 4: Biological Evolution	6	80349	2.93	1.40	0.37
	Concept 5: Matter, Energy, and Organization in Living Systems (Including Human Systems)	6	80349	2.45	1.47	0.42

9.2 Validity

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (AERA/APA/NCME, 1999, 2014). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process.

The Spring 2016 AIMS tests were designed and developed to provide fair and accurate ability scores that support appropriate, meaningful, and useful educational decisions. In addition to the evidence provided in Part 2 (Involvement of Arizona Educators), additional validity evidence may be found in the following parts as described: Part 3 (Test Design), Part 4 (Test Development), Part 5 (Test Administration), Part 6 (Classical Item Analysis), Part 7 (Calibration, Scaling and Equating), Part 9.1 (Reliability), and Part 10 (Classification). As the technical report has progressed, chapter by chapter, it has moved through the phases of the testing cycle. Each part of the technical report detailed the procedures and processes applied in the creation of AIMS, as well as their results. Each part also highlights the meaning and significance of the procedures, processes, and results in terms of content and construct validity and the relationship to the Standards. Part 9.2 addresses two final issues in validity: the issues of bias and construct validity. The analyses presented here add to the perspectives provided in Chapters 2 through 10. Below is a brief review.

Part 2 of the technical report described the involvement of Arizona educators, ADE, and Pearson in the test development process. As indicated in Part 2, the test development process and the involvement of Arizona educators in that process formed an important part of the validity of the entire AIMS. The knowledge, expertise, and professional judgment offered by Arizona educators ultimately ensured that the content of AIMS formed an adequate and representative sample of appropriate content and that the content formed a legitimate basis upon which to validly derive conclusions about student achievement.

Parts 3 and 4 of the technical report addressed the issue of test form development. Part 3 provided a general discussion of test book creation and editing process, the process of selecting operational test items, the content distribution of embedded field test items, and the process of obtaining ADE approvals. The test design process and the participation of Arizona educators in the process of test selection, including item content and bias review, provide a solid rationale for having confidence in the content and design of AIMS as a tool from which to derive valid inferences about Arizona student performance.

Part 5 of the technical report described the process, procedures, and policies that guided the administration of the AIMS, including accommodations, security, and the written procedures provided to test administrators and school personnel.

Part 6 described classical data analysis of the Spring 2016 AIMS Science tests.

Part 7 of the technical report described the calibration, scaling and equating methods, as well as processes and procedures for deriving scale scores from students' raw scores and the data cleaning steps which ensure valid calibration and scaling. Some references to introductory and advanced discussions of IRT are provided.

Part 8 of the technical report dealt with the test results, longitudinal comparisons, score distributions and performance levels.

Part 9, above, dealt with Cronbach's alpha as a measure for internal consistency.

Part 9 below presents the results of an analysis of DIF (Differential Item Functioning). Complete tables of gender, ethnic, and race differential functioning of all operational items for the 2016 AIMS Science assessments are presented in Appendix B.

Part 10 of the technical report will describe a detailed analysis of classification consistency and classification accuracy.

Also note that further evidence in support of the AIMS assessment has been documented in previous AIMS technical reports and standard setting technical reports.

9.2.1 Differential Item Functioning

Because test scores can have many sources of variation, the test publishers' task is to develop assessments that measure the intended abilities and skills without introducing extraneous elements or construct irrelevant variance. When tests measure something other than what they are intended to measure, test scores will reflect these unintended skills and knowledge, as well as what is purportedly assessed by the test. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975). One of the factors that may render test scores to be biased is differing cultural and socioeconomic experiences.

The Spring 2016 AIMS tests were developed using procedures to minimize item and test bias and included reviews such as the Content and Sensitivity Reviews and Data Analysis Workshops after each item was field-tested as described in Part 4, Test Development. Expertise in this area is not, however, a substitute for statistical analyses of the items or the continued monitoring of the fairness of items. Thus, an empirical differential item functioning (DIF) approach was used to examine potential item bias on all operational items. DIF studies include systematic item analyses to determine if examinees with the same underlying level of ability have the same probability of correctly responding to the item. Items identified with DIF are further examined to determine if item performance differences between identifiable subgroups of the population are due to extraneous or construct irrelevant information which makes the items unfairly difficult for one of the subgroups.

DIF analyses of the Spring 2016 AIMS tests were conducted for ethnic/race subgroups and gender. In order to compute DIF, students must be matched on ability level using a conditioning variable. For these analyses, raw score on the test was used as the conditioning variable.

The Mantel-Haenszel chi-square statistic was used to identify DIF in multiple-choice items. The Mantel-Haenszel statistic was first recommended by Holland and Thayer (1988), is frequently used, and is efficient in terms of statistical power (Clauser & Mazor, 1998). The Mantel-Haenszel statistic is computed as (Zwick, Donoghue, & Grima, 1993):

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k \text{Var}(F_k)},$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable. Note that the Mantel-Haenszel statistic is sensitive to N such that larger sample sizes increase the value of chi square.

In addition to the Mantel-Haenszel chi-square statistic, the Mantel-Haenszel delta statistic (ΔMH DIF) was computed for all items. Educational Testing Service (ETS) first developed the ΔMH DIF statistic. To compute delta, alpha (the odds ratio) is first computed as:

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

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*. ΔMH DIF is then 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.

The Mantel-Haenszel chi-square statistic and the delta statistic were used in combination to identify the Spring 2016 AIMS items that exhibit strong, weak, or no DIF (Zieky, 1993). Table 9.2.1.1 indicates the criteria for each category used for the 2016 AIMS DIF analysis. An alpha level of .01 was used for all Mantel-Haenszel statistics. Note that the criteria are very lenient given very large sample sizes and the number of DIF statistics computed. In other words, a large number of items will be placed in categories B and C given the critical value. For reference, the critical value for the chi-square statistic to be significant at $p < 0.01$ is 6.635, at $p < 0.001$ the critical value is 10.827, and at $p < 0.0005$ the critical value is 12.116.

Table 9.2.1.1
Differential Item Functioning Flag Categories

Category	Description	Criterion
A	No DIF	Mantel-Haenzel chi-square not significantly different from zero
B	Weak DIF	Significant Mantel-Haenzel chi-square ($p<0.01$) and $1.0 \leq \lvert\Delta MH\rvert < 1.5$
C	Strong DIF	Significant Mantel-Haenzel chi-square ($p<0.01$) and $\lvert\Delta MH\rvert \geq 1.5$

Another measure, also use to analyze DIF for the Spring 2016 AIMS operational items, is the standardized mean difference (SMD; Zwick et al., 1993). The SMD is an effect size index of DIF, which is relatively easy to interpret. The SMD compares the means of the reference and focus groups, adjusting for the distribution of reference and focal group members on the conditioning variable, which for these analyses is the CRT raw score. SMD is computed as (Zwick et al., 1993):

$$SMD = \sum_{k} p_{Fk} \left(m_{Fk} - m_{Rk} \right),$$

where p_{Fk} is the proportion of the focal group members at the k^{th} level of the matching variable, m_{Fk} is the mean item response of the focal group at the k^{th} level and m_{Rk} is the mean item response of the reference group at the k^{th} level . A negative SMD value indicates an item on which the focal group

has a lower mean than the reference group. A positive SMD value indicates an item on which the reference group has a lower mean than the focal group.

Mantel-Haenszel chi-square statistic, MH-D DIF, SMD, and flag category results for all items in the Spring 2016 AIMS tests are presented in Appendix B. It is important to note that DIF analyses are also conducted on field test items prior to form construction. Very few AIMS items are identified as exhibiting strong DIF in field testing. All items exhibiting strong DIF are investigated for possible sources of differential functioning by Pearson and ADE staff and such items are avoided in form construction. Not surprisingly, the vast majority of items on the operational AIMS exhibit no DIF or weak DIF. The one item that was flagged for exhibiting strong DIF is summarized in Table 9.2.1.2 with the results for all items used in 2016 presented in Appendix B.

Table 9.2.1.2
DIF Statistics for Items Exhibiting Strong DIF

Content	Grade	Item	Item Type	In favor of/ Against	Group	MH χ2	ΔMH	SMD
Science	8	34	MC	Against	Female	1753.61	-1.51	-0.14

Note: MH χ2 = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference,

PART 10: CLASSIFICATION

Part 10 of this technical report provides information regarding classifying students into proficiency categories. The following AERA/APA/NCME *Standards* from the 1999 edition are covered in this part: 1.5, 1.7, 2.2, 2.14, 2.15, 4.9, 4.19, 4.20, 4.21, and 6.5. The 2014 AERA/APA/NCME *Standards* (AERA, APA, NCME, 2014) addressed by this chapter are: 1.8, 1.9, 2.13, 2.14, 2.16, 5.5, 5.21, 5.22, 5.23, and 7.4.

Scores from the Spring 2016 AIMS assessments are used to classify students into one of four performance categories: *Falls Far Below the Standard*, *Approaches the Standard*, *Meets the Standard*, and *Exceeds the Standard*. This part of the technical report provides information regarding classifying students into these four performance categories. Arizona educators made recommendations for cut scores for each category in the standard setting workshops. Analyses were conducted to examine the consistency and accuracy with which students were assigned to performance categories.

10.1 Standard Setting Technical Documentation

Standard setting for the AIMS Science tests was conducted in early June, 2008, using the bookmark standard setting procedure. All technical documentation regarding the standard setting is available in the bookmark standard setting technical report, available from the ADE at http://www.azed.gov.

The scale score ranges for each of the four performance level categories and their associated cut scores, along with the lowest possible and highest possible scale scores for the AIMS Science tests are presented below in Table 10.1.1.

Table 10.1.1
Spring 2016 AIMS Science
Final Scale Score Ranges by Performance Level

Grade	LOSS	FFBS	AS Cut	AS	MS Cut	MS	ES Cut	ES	HOSS
4	200	293-461	462	465-498	500	502-545	547	550-800	800
8	200	367-470	473	474-497	500	500-528	532	532-800	800
HS	200	300-472	475	476-497	500	500-534	537	538-800	800

Note: LOSS=Lowest Observable Scale Score, FFBS=Fall Far Below the Standard, AS=Approaches the Standard, MS=Meets the Standard, ES=Exceeds the Standard, HOSS=Highest Observable Scale Score.

10.2 Classification Consistency and Accuracy

This section describes the analyses conducted to estimate classification consistency and accuracy for the Spring 2016 AIMS Science adminstrations in grades 4, 8, and high school. Classification consistency can be defined as the agreement between examinees' performance category classification from two independent administrations of the same test (or two parallel forms of the test). Classification accuracy can be defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston & Lewis, 1995).

In conjunction with internal consistency, classification consistency is an important type of reliability and is particularly relevant to high-stakes tests. As a form of reliability, classification consistency represents how reliably students can be classified into performance categories. Please see Part 9 of this report for more information on the internal consistency of the AIMS Science assessments.

Classification consistency is most important for students whose ability is near each cut score. Students whose ability is far above or far below the established cut value are unlikely to be misclassified because repeated administration of the test will nearly always result in the same classification. Examinees whose true scores are close to the cut score are a more serious concern. These students' true scores will likely lie within the standard error of measurement of the cut score. For this reason, the measurement error at the cut scores should be considered when evaluating the classification consistency of a test. For convenience, the cut scores with their associated conditional standard error of measurement (CSEM) are presented in Table 10.2.2.1. The CSEMs around the Performance Level cuts were lower than those outside of the lowest and highest Performance Level cuts, indicating better measurement precision around the cuts.

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000) for the AIMS Science assessments. The following description of classification consistency and accuracy is based on the paper by Lee et al. (2002).

Table 10.2.1
Spring 2016 AIMS
Standard Error of Measurement at Cut Scores

Test	Grade	AS		MS		ES	
		Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
Science	4	462	15	500	14	547	16
Science	8	473	14	500	14	532	14
Science	HS	475	14	500	14	537	15

Note: AS = Approaches the Standard; MS = Meets the Standard; ES = Exceeds the Standard

10.2.1 Classification Consistency

Assume that θ is a single latent trait measured by a test and denote Φ as a latent random variable. When a test X consists of K items and its maximum number-correct score is N, the marginal probability of the number-correct (NC) score *x* is:

$$P(X = x) = \int P(X = x \mid \Phi = \theta) g(\theta) d\theta, \quad x = 0,1,..., N.,$$

where $g(\theta)$ is the density of θ.

In this report, the marginal distribution $P(X = x)$ is denoted as $f(x)$, and the conditional error distribution $P(X = x \mid \Phi = \theta)$ is denoted as $f(x \mid \theta)$. It is assumed that examinees are classified into one of H mutually exclusive categories on the basis of predetermined H-1 observed score cutoffs, $C_1, C_2, \ldots, C_{H-1}$. Let L_h represent the h^{th} category into which examinees with $C_{h-1} \le X \le C_h$ are

classified. $C_0 = 0$ and $C_H =$ the maximum number-correct score. Then, the conditional and marginal probabilities of each category classification are as follows:

$$P(X \in L_h \mid \theta) = \sum_{x=C_{h-1}}^{C_{h-1}} f(x \mid \theta), \quad h = 1, 2, \ldots, \text{H.}$$

$$P(X \in L_h) = \int \sum_{x=C_{h-1}}^{C_{h-1}} f(x \mid \theta) g(\theta) d\theta, \quad h = 1, 2, \ldots, \text{H.}$$

Because obtaining test scores from two independent administrations of AIMS was not feasible due to security, logistic, and cost constraints, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Based on the psychometric model, a symmetric H*H contingency table can be constructed. The elements of H*H contingency table consist of the joint probabilities of the row and column observed category classifications.

That two administrations are independent implies that if x_1 and x_2 represent the raw score random variables on the two administrations, then, conditioned on θ, x_1 and x_2 are independent and identically distributed. Consequently, the conditional bivariate distribution of x_1 and x_2 is:

$$f(x_1, \ x_2 \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta).$$

The marginal bivariate distribution of X_1 and X_2 can be expressed as follows:

$$f(x_1, \ x_2) = \int f(x_1, x_2 \mid \theta) f(\theta) d\theta.$$

Consistent classification means that both X_1 and X_2 fall in the same category. The conditional probability of falling in the same category on the two administrations is:

$$P(X_1 \in L_h, \ X_2 \in L_h \mid \theta) = \left[\sum_{x_1=C_{h-1}}^{C_{h-1}} f(x_1 \mid \theta) \right]^2,$$
$$h = 1, 2, \ldots, \text{H.}$$

The agreement index P, conditional on theta, is obtained by:

$$P(\theta) = \sum_{h=1}^{H} P(X_1 \in L_h, \ X_2 \in L_h \mid \theta).$$

The agreement index (classification consistency) can be computed as:

$$P = \int P(\theta) g(\theta) d(\theta).$$

The probability of consistent classification by chance, P_C, is the sum of squared marginal probabilities of each category classification:

$$P_C = \sum_{h=1}^{H} P(X_1 \in L_h)P(X_2 \in L_h) = \sum_{h=1}^{H} \left[P(X_1 \in L_h)\right]^2 .$$

Then, the coefficient kappa (Cohen, 1960) is:

$$k = \frac{P - P_C}{1 - P_C}$$

10.2.2 Classification Accuracy

Let Γ_w denote true category. When an examinee has an observed score, $x \in L_h$ (h =1, 2,…, H), and a latent score , $\theta \in \Gamma_w$(w=1, 2,…, H), an accurate classification is made when $h = w$. The conditional probability of accurate classification is

$$\Gamma(\theta) = P(X \in L_w \mid \theta),$$

where w is the category such that $\theta \in \Gamma_w$.

10.2.3 Classification Consistency and Accuracy Results

Table 10.2.3.1 presents results from the classification consistency and classification accuracy analyses. These results are for classifying students into AIMS' four performance levels. Included in the table for each grade are case counts (N), classification consistency (Agreement), classification inconsistency (Inconsistency), probability of consistent classification by chance (Chance), Cohen's Kappa (Kappa), and classification accuracy (Accuracy). Inconsistency is defined as 1-agreement.

The 2016 AIMS classification consistency and accuracy results are consistent with classification analyses from previous AIMS administrations. It is important to note that the classification results are dependent on the number of cut scores maintained in a testing program. Moreover, the acceptability of the classification results should be evaluated with respect to the associated stakes of the testing program. The results for the AIMS assessments are quite consistent with other testing programs with similar structure and purpose.

Table 10.2.3.1
Spring 2016 AIMS
Classification Consistency and Accuracy

Test	Grade	N	Agreement	Inconsistency	Chance	Kappa	Accuracy
Science	4	85917	0.69	0.31	0.26	0.58	0.77
Science	8	82475	0.69	0.31	0.27	0.58	0.77
Science	HS	80349	0.70	0.30	0.29	0.59	0.78

Note: High school results include students in all cohorts. Results were computed with the IRT method suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000).

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2104). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Warner (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Arizona Department of Education. (2015). *Arizona's Instrument to Measure Standards 2015 technical report*. Iowa City, IA: Pearson.

Arizona Department of Education. (2009). *Bookmark standard setting technical report 2008 for grades 4, 8, and high school science.* Monterey, CA: CTB/McGraw-Hill.

Brennan, R. L. (2004). BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy [Computer program]. Iowa City, IA: The University of Iowa Center for Advanced Studies in Measurement and Assessment.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group/Thompson Learning.

Green, D. R. (1975, December). *Procedures for assessing bias in achievement tests*. Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lee, W., Hanson, B. A., Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26,* 412-432.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16(2),* 878.

Linacre, J. M. (2005). WINSTEPS Rasch measurement [Computer software]. Chicago: WINSTEPS.com.

Linacre, J. M. (2011). A User's Guide to WINSTEPS Rasch-Model Computer Programs. WINSTEPS.com.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification consistency and accuracy based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F. M. (1980). *Applications of item response theory to practical testing programs.* Hillsdale, NJ: Lawrence Erlbaum.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2),* 97-116.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.

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

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 26,* 44-66.

APPENDIX A: COMMITTEE MEMBER SELECTION CRITERIA

AIMS Committee Participant Selection Criteria

ARIZONA DEPARTMENT OF EDUCATION

PROCEDURE FOR SELECTION OF EDUCATOR COMMITTEES
ARIZONA ASSESSMENT SECTION

The Assessment Section is always recruiting new teachers to serve on the committees, and have prevailed upon veteran teachers to become Ambassadors of the Assessment by encouraging their colleagues to apply.

Once Arizona educators are identified and entered into the database, the Assessment Section uses the following procedures for selecting membership for a committee:

- Identify the purpose/function of the committee
- Establish the date and time of the committee
- Determine the criteria for membership on the committee:
    - Content area of expertise
    - Grade level experience
    - Specific skill or knowledge expertise for committee function
    - Prior experience on ADE committees—a minimum 50% of each committee will have prior experience
    - Location of district/school
        - Rural/urban/suburban
        - Approximately 50% of committee members from Maricopa County when appropriate for purpose of committee
    - Ethnicity of school population or committee member
    - SES of school population
    - Number of committees served on recently—a committee member cannot serve on a series of committees used to develop items. Otherwise, they would be passing judgment on their own prior work.
- Review the database for educators that meet the criteria established
- Select committee members based on criteria for particular committee for primary and alternate list
- Invitations are sent to selected committee members
- After decline and accept emails are received by established deadline, additional invitations issued to members on alternate list

- Once the committee meeting is held, performance of participants is reviewed.

Recognition of existing AIMS committee participants is an important aspect of retaining our Ambassadors of the Assessment; therefore, after each committee meeting, each participant receives a letter recognizing their excellent contributions to the assessment program and to all Arizona students.

APPENDIX B: DIF RESULTS

Table B,1
Spring 2016 AIMS Differential Item Functioning
Science CRT Grade 4

	Reference: Male N= 43355				Reference: Hispanic N= 46567				Reference: White N= 67402				Reference: White N= 67402				Reference: White N= 67402				Reference: White N= 67402				Reference: White N= 67402			
	Focal: Female N= 41995				Focal: Non Hispanic N= 38910				Focal: Africa American N= 5151				Focal: Native American N= 5652				Focal: Asian N= 2517				Focal: Hawaii N= 429				Focal: Multiple Indicator N= 3462			
Item	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag
1	39.68	0.26	0.02	A	38.59	0.26	0.02	A	20.46	-0.36	-0.03	A	1.90	0.11	0.01	A	0.69	0.12	0.01	A	0.60	-0.22	-0.02	A	1.06	-0.11	-0.01	A
2	243.25	0.58	0.05	A	1.04	0.04	0.01	A	28.10	-0.39	-0.04	A	2.11	-0.10	-0.01	A	13.29	0.47	0.03	A	1.87	0.37	0.03	A	7.01	-0.25	-0.02	A
3	46.20	0.28	0.02	A	45.99	-0.28	-0.02	A	0.09	-0.02	0.00	A	29.38	-0.40	-0.03	A	4.32	-0.29	-0.01	A	2.31	-0.41	-0.03	A	1.51	-0.13	-0.01	A
4	11.61	-0.13	-0.01	A	39.20	-0.24	-0.02	A	9.02	0.23	0.02	A	40.21	-0.46	-0.04	A	21.04	0.58	0.04	A	0.06	-0.06	-0.01	A	1.70	0.13	0.01	A
5	3.62	-0.07	-0.01	A	20.18	-0.17	-0.02	A	9.14	0.23	0.02	A	38.27	-0.43	-0.04	A	21.92	0.59	0.04	A	0.57	-0.19	-0.02	A	3.27	-0.17	-0.01	A
6	1.02	-0.03	0.00	A	24.17	-0.17	-0.02	A	0.24	-0.04	0.00	A	9.71	-0.22	-0.02	A	1.40	-0.12	-0.01	A	0.74	-0.20	-0.02	A	4.93	0.19	0.02	A
7	225.30	0.54	0.05	A	1.23	0.04	0.01	A	2.13	0.11	0.01	A	11.11	-0.23	-0.02	A	1.11	0.12	0.01	A	0.14	0.10	0.01	A	0.03	0.02	0.00	A
8	126.43	0.41	0.03	A	31.66	-0.21	-0.02	A	4.38	0.16	0.01	A	2.89	-0.12	-0.01	A	0.23	-0.06	0.00	A	0.16	-0.10	-0.01	A	2.67	0.16	0.01	A
9	3.80	0.06	0.01	A	2.01	0.05	0.00	A	37.45	-0.44	-0.04	A	73.44	0.58	0.06	A	0.93	-0.10	-0.01	A	0.01	0.03	0.00	A	1.24	-0.09	-0.01	A
10	42.88	-0.28	-0.02	A	176.58	-0.57	-0.04	A	1.01	0.08	0.01	A	0.05	-0.02	0.00	A	4.07	-0.30	-0.01	A	0.01	0.03	0.00	A	20.85	0.53	0.03	A
11	44.25	0.23	0.02	A	32.90	-0.21	-0.01	A	10.84	-0.24	-0.02	A	5.62	-0.17	-0.01	A	0.07	0.03	0.00	A	3.90	-0.48	-0.04	A	0.70	-0.07	-0.01	A
12	120.77	0.39	0.03	A	57.44	-0.28	-0.02	A	0.17	-0.03	0.00	A	25.13	-0.36	-0.03	A	12.71	0.39	0.03	A	0.05	0.05	0.00	A	0.03	-0.01	0.00	A
13	159.84	0.44	0.04	A	22.42	0.17	0.01	A	1.98	0.11	0.01	A	0.03	-0.01	0.00	A	2.58	0.16	0.02	A	4.34	-0.51	-0.05	A	0.36	-0.05	0.00	A
14	0.06	0.01	0.00	A	147.14	-0.44	-0.03	A	15.77	-0.30	-0.03	A	2.79	0.12	0.01	A	12.80	-0.37	-0.03	A	1.64	-0.33	-0.03	A	2.80	0.15	0.01	A
15	248.69	0.63	0.04	A	41.73	-0.26	-0.02	A	7.48	-0.21	-0.02	A	6.31	-0.18	-0.01	A	0.21	0.06	0.00	A	0.02	0.04	0.00	A	0.67	-0.08	-0.01	A
16	468.59	0.75	0.07	A	18.53	0.16	0.02	A	22.39	0.35	0.03	A	1.92	-0.10	-0.01	A	12.10	0.34	0.03	A	0.01	-0.02	0.00	A	0.70	-0.07	-0.01	A
17	43.10	0.23	0.02	A	6.02	0.09	0.01	A	3.95	0.15	0.01	A	6.55	0.19	0.02	A	21.97	0.48	0.05	A	0.48	0.17	0.02	A	5.39	-0.21	-0.02	A
18	2.16	0.05	0.00	A	11.75	-0.13	-0.01	A	2.84	-0.12	-0.01	A	5.59	-0.16	-0.02	A	5.57	-0.25	-0.02	A	0.16	-0.10	-0.01	A	0.13	-0.03	0.00	A
19	223.91	-0.55	-0.05	A	85.48	-0.35	-0.03	A	0.86	-0.07	-0.01	A	19.78	-0.33	-0.03	A	26.91	0.61	0.04	A	1.94	0.35	0.03	A	1.34	0.11	0.01	A
20	192.07	0.57	0.04	A	0.04	0.01	0.00	A	0.57	0.06	0.00	A	0.00	0.00	0.00	A	3.37	0.26	0.01	A	0.82	0.27	0.02	A	0.53	-0.08	0.00	A
21	105.97	-0.42	-0.03	A	57.53	-0.31	-0.02	A	7.81	-0.22	-0.02	A	0.01	-0.01	0.00	A	18.17	0.62	0.03	A	1.71	0.39	0.03	A	1.99	0.15	0.01	A
22	119.24	0.38	0.03	A	14.75	0.14	0.01	A	1.14	-0.08	-0.01	A	4.07	-0.14	-0.01	A	0.01	0.01	0.00	A	0.04	0.05	0.00	A	1.00	-0.09	-0.01	A
23	0.77	0.04	0.00	A	61.32	-0.36	-0.02	A	0.18	-0.04	0.00	A	17.26	-0.32	-0.02	A	5.68	0.42	0.01	A	2.43	-0.47	-0.03	A	1.50	0.15	0.01	A
24	0.36	-0.02	0.00	A	48.77	-0.25	-0.02	A	1.25	0.08	0.01	A	8.99	0.21	0.02	A	18.63	0.45	0.04	A	1.18	-0.27	-0.02	A	0.11	-0.03	0.00	A
25	15.94	0.14	0.01	A	7.17	-0.10	-0.01	A	0.06	0.02	0.00	A	6.21	0.18	0.02	A	0.12	0.04	0.00	A	2.94	-0.43	-0.04	A	0.27	0.04	0.00	A
26	1.26	0.04	0.00	A	15.44	-0.14	-0.01	A	0.10	0.02	0.00	A	0.09	-0.02	0.00	A	0.00	0.01	0.00	A	0.91	-0.23	-0.02	A	0.42	0.06	0.01	A
27	39.44	-0.25	-0.02	A	4.70	-0.09	-0.01	A	9.88	-0.24	-0.02	A	0.46	0.05	0.01	A	0.14	0.05	0.00	A	0.00	0.01	0.00	A	0.26	0.05	0.00	A
28	15.67	0.14	0.01	A	14.95	0.14	0.02	A	5.46	-0.17	-0.02	A	1.52	-0.08	-0.01	A	4.35	0.24	0.02	A	3.02	-0.42	-0.04	A	8.56	-0.26	-0.02	A
29	34.26	0.25	0.02	A	11.72	0.15	0.01	A	8.75	-0.24	-0.02	A	8.30	-0.22	-0.02	A	5.08	0.35	0.01	A	6.76	0.79	0.05	A	3.69	-0.22	-0.01	A
30	18.57	0.15	0.01	A	5.53	-0.08	0.00	A	2.32	-0.11	-0.01	A	7.59	-0.19	-0.02	A	1.90	-0.14	-0.01	A	0.64	-0.19	-0.02	A	3.26	0.16	0.02	A

Note: African Am. = African American, Native Am. = Native American, MH χ2 = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors reference > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.

(table continues)

Table B.1 (continued)
Spring 2016 AIMS Differential Item Functioning
Science CRT Grade 4

	Reference: Male N= 43355 Focal: Female N= 41995				Reference: Hispanic N= 46567 Focal: Non Hispanic N= 38910				Reference: White N= 67402 Focal: Africa American N= 5151				Reference: White N= 67402 Focal: Native American N= 5652				Reference: White N= 67402 Focal: Asian N= 2517				Reference: White N= 67402 Focal: Hawaii N= 429				Reference: White N= 67402 Focal: Multiple Indicator N= 3462			
Item	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag
31	0.04	0.01	0.00	A	1.84	0.05	0.00	A	3.09	-0.14	-0.01	A	0.34	0.04	0.00	A	65.86	0.80	0.08	A	1.31	0.28	0.02	A	1.32	0.10	0.01	A
32	543.61	-0.84	-0.07	A	90.15	-0.36	-0.03	A	24.41	-0.39	-0.03	A	1.47	0.09	0.01	A	1.30	-0.12	-0.01	A	1.15	-0.27	-0.02	A	1.85	-0.12	-0.01	A
33	1507.90	-1.43	-0.12	B<	52.87	-0.27	-0.02	A	6.84	0.20	0.02	A	1.70	-0.10	-0.01	A	5.36	0.25	0.02	A	2.38	0.39	0.03	A	28.84	0.49	0.04	A
34	8.41	0.11	0.01	A	20.96	-0.18	-0.01	A	7.55	0.21	0.02	A	40.35	-0.46	-0.04	A	7.39	0.36	0.02	A	1.50	0.33	0.03	A	0.12	0.03	0.00	A
35	0.84	-0.03	0.00	A	33.25	-0.22	-0.02	A	2.09	-0.11	-0.01	A	1.56	0.09	0.01	A	1.15	0.13	0.01	A	0.49	-0.18	-0.01	A	0.68	0.08	0.01	A
36	47.77	0.23	0.02	A	1.82	-0.05	0.00	A	4.46	-0.15	-0.01	A	1.92	0.09	0.01	A	14.17	-0.38	-0.04	A	0.00	-0.01	0.00	A	0.04	0.02	0.00	A
37	33.61	-0.20	-0.02	A	39.28	0.23	0.02	A	0.03	-0.01	0.00	A	0.23	-0.04	0.00	A	1.03	-0.10	-0.01	A	0.00	-0.01	0.00	A	2.05	-0.12	-0.01	A
38	20.66	-0.16	-0.01	A	3.04	-0.06	-0.01	A	26.03	0.37	0.03	A	0.08	0.02	0.00	A	3.60	0.20	0.02	A	0.62	0.19	0.02	A	2.08	0.13	0.01	A
39	56.46	0.27	0.02	A	0.08	-0.01	0.00	A	0.36	-0.05	0.00	A	36.66	-0.45	-0.04	A	7.21	0.28	0.03	A	0.36	-0.15	-0.01	A	0.27	0.05	0.00	A
40	352.11	-0.73	-0.05	A	6.00	-0.10	-0.01	A	6.43	-0.20	-0.02	A	11.28	-0.24	-0.02	A	1.33	0.15	0.01	A	1.20	-0.30	-0.02	A	0.96	-0.10	-0.01	A
41	1.14	-0.04	0.00	A	20.48	0.16	0.01	A	3.67	-0.14	-0.01	A	31.26	0.40	0.04	A	4.11	0.21	0.02	A	0.29	0.13	0.01	A	1.56	0.11	0.01	A
42	4.85	-0.08	-0.01	A	9.09	-0.11	-0.01	A	6.34	-0.19	-0.02	A	8.27	-0.20	-0.02	A	0.30	0.06	0.00	A	0.14	0.10	0.01	A	1.33	-0.11	-0.01	A
43	244.20	-0.58	-0.05	A	30.91	-0.21	-0.02	A	20.88	-0.34	-0.03	A	0.01	0.01	0.00	A	0.02	0.02	0.00	A	0.00	0.00	0.00	A	1.81	-0.13	-0.01	A
44	66.33	-0.32	-0.02	A	14.52	-0.16	-0.01	A	14.96	-0.35	-0.02	A	1.32	0.10	0.01	A	10.56	-0.37	-0.03	A	1.03	-0.30	-0.02	A	8.74	-0.30	-0.02	A
45	233.34	-0.59	-0.04	A	122.31	-0.43	-0.03	A	0.15	-0.03	0.00	A	17.19	-0.31	-0.03	A	0.44	0.09	0.01	A	0.31	-0.14	-0.01	A	1.69	0.13	0.01	A
46	51.11	-0.34	-0.02	A	68.84	-0.40	-0.02	A	0.01	-0.01	0.00	A	0.03	0.01	0.00	A	1.38	0.22	0.01	A	0.26	-0.17	-0.01	A	6.90	0.35	0.01	A
47	376.95	-0.69	-0.06	A	34.63	-0.22	-0.02	A	0.91	-0.07	-0.01	A	4.02	-0.15	-0.01	A	0.38	-0.07	-0.01	A	1.06	-0.26	-0.02	A	1.67	0.12	0.01	A
48	4.46	0.08	0.01	A	2.32	0.06	0.01	A	0.02	-0.01	0.00	A	2.55	-0.11	-0.01	A	2.91	0.20	0.01	A	0.80	-0.22	-0.02	A	0.03	-0.02	0.00	A
49	45.78	0.29	0.02	A	0.02	0.01	0.00	A	1.76	-0.11	-0.01	A	1.81	0.11	0.01	A	0.64	0.13	0.01	A	2.66	0.50	0.03	A	1.84	0.16	0.01	A
50	76.88	-0.31	-0.03	A	14.94	-0.14	-0.01	A	5.72	-0.18	-0.02	A	1.36	0.08	0.01	A	6.11	-0.27	-0.02	A	1.64	-0.31	-0.03	A	1.10	-0.10	-0.01	A
51	26.70	0.18	0.02	A	0.50	0.03	0.00	A	1.34	-0.09	-0.01	A	7.94	0.20	0.02	A	0.04	0.02	0.00	A	2.02	0.36	0.03	A	1.07	-0.09	-0.01	A
52	6.40	0.09	0.01	A	3.57	-0.07	0.00	A	15.79	-0.30	-0.03	A	12.77	0.26	0.02	A	0.77	-0.10	-0.01	A	0.77	0.21	0.02	A	0.20	-0.04	0.00	A
53	119.52	0.38	0.03	A	45.89	-0.24	-0.02	A	14.62	-0.27	-0.03	A	11.22	-0.23	-0.02	A	11.91	-0.37	-0.03	A	0.07	0.06	0.01	A	0.42	-0.06	-0.01	A
54	146.24	0.47	0.04	A	77.39	0.35	0.03	A	5.52	0.18	0.01	A	0.14	0.03	0.00	A	22.17	0.64	0.04	A	3.41	0.50	0.04	A	0.58	-0.08	-0.01	A

Note: African Am. = African American, Native Am. = Native American, MH $\chi 2$ = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors reference > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.

Table B.2
Spring 2016 AIMS Differential Item Functioning
Science CRT Grade 8

	Reference: Male N= 41573 Focal: Female N= 40292				Reference: Hispanic N= 45666 Focal: Non Hispanic N= 36312				Reference: White N= 65677 Focal: Africa American N= 4950				Reference: White N= 65677 Focal: Native American N= 5567				Reference: White N= 65677 Focal: Asian N= 2495				Reference: White N= 65677 Focal: Hawaii N= 305				Reference: White N= 65677 Focal: Multiple Indicator N= 2354			
Item	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag
1	80.67	-0.42	-0.02	A	9.27	-0.14	-0.01	A	24.18	-0.44	-0.03	A	17.10	-0.34	-0.02	A	24.88	-0.71	-0.03	A	0.08	-0.11	-0.01	A	0.02	0.02	0.00	A
2	99.10	-0.39	-0.03	A	3.99	-0.08	-0.01	A	0.18	-0.03	0.00	A	1.26	-0.08	-0.01	A	10.48	-0.41	-0.02	A	10.36	1.10	0.08	B>	0.00	0.00	0.00	A
3	169.57	0.45	0.04	A	26.15	0.18	0.02	A	0.73	0.06	0.01	A	0.09	0.02	0.00	A	0.10	0.03	0.00	A	0.03	0.05	0.01	A	0.77	-0.09	-0.01	A
4	93.58	0.39	0.03	A	7.02	-0.11	-0.01	A	21.49	-0.37	-0.03	A	7.31	0.21	0.02	A	5.23	0.33	0.02	A	0.65	0.27	0.02	A	3.99	0.25	0.02	A
5	73.90	0.33	0.03	A	41.93	-0.26	-0.02	A	0.31	-0.04	0.00	A	54.97	-0.60	-0.04	A	20.20	0.56	0.04	A	1.20	0.34	0.03	A	0.23	-0.06	0.00	A
6	1.53	-0.04	0.00	A	48.74	-0.25	-0.03	A	20.40	-0.32	-0.03	A	5.41	-0.16	-0.02	A	24.37	-0.50	-0.05	A	4.38	-0.59	-0.06	A	0.04	-0.02	0.00	A
7	13.92	-0.17	-0.01	A	12.30	-0.16	-0.01	A	0.00	0.00	0.00	A	0.01	0.01	0.00	A	0.69	-0.13	-0.01	A	1.03	0.38	0.02	A	0.71	0.11	0.01	A
8	255.74	-0.89	-0.03	A	12.40	-0.20	-0.01	A	25.70	-0.52	-0.03	A	92.05	-0.84	-0.05	A	9.27	-0.64	-0.01	A	0.84	-0.38	-0.02	A	0.03	0.03	0.00	A
9	774.48	1.00	0.09	A	0.53	0.03	0.00	A	6.48	-0.19	-0.02	A	6.69	-0.18	-0.02	A	2.38	0.17	0.01	A	0.19	0.12	0.01	A	2.64	-0.17	-0.02	A
10	106.36	-0.37	-0.03	A	62.30	-0.29	-0.03	A	2.66	-0.12	-0.01	A	13.10	-0.26	-0.02	A	1.73	-0.15	-0.01	A	3.48	0.55	0.05	A	6.07	0.27	0.02	A
11	2.08	0.06	0.00	A	5.14	0.10	0.01	A	0.74	0.07	0.01	A	0.96	-0.07	-0.01	A	0.00	0.01	0.00	A	0.29	0.18	0.01	A	0.51	0.09	0.01	A
12	263.55	0.63	0.05	A	47.64	0.28	0.02	A	0.00	0.00	0.00	A	0.17	0.03	0.00	A	15.93	0.55	0.03	A	1.38	0.37	0.03	A	3.56	-0.22	-0.02	A
13	1.51	0.04	0.00	A	0.88	0.03	0.01	A	3.60	-0.14	-0.01	A	0.52	0.05	0.00	A	18.67	0.44	0.04	A	0.57	0.21	0.02	A	1.94	-0.15	-0.01	A
14	0.24	0.02	0.00	A	2.71	-0.06	-0.01	A	0.20	0.03	0.00	A	7.63	0.20	0.02	A	28.91	0.58	0.05	A	0.56	0.22	0.02	A	1.03	0.11	0.01	A
15	23.05	0.18	0.01	A	0.74	0.03	0.01	A	0.87	0.07	0.01	A	0.08	-0.02	0.00	A	14.16	0.39	0.04	A	1.16	-0.34	-0.03	A	0.28	0.06	0.01	A
16	27.91	0.20	0.02	A	2.26	0.06	0.01	A	18.45	-0.32	-0.03	A	0.34	0.04	0.00	A	2.86	0.21	0.01	A	0.22	0.14	0.01	A	0.93	-0.11	-0.01	A
17	115.69	0.40	0.03	A	4.94	-0.08	-0.01	A	0.11	-0.03	0.00	A	4.48	-0.15	-0.01	A	1.38	-0.13	-0.01	A	1.33	-0.32	-0.03	A	0.24	0.05	0.00	A
18	510.41	0.94	0.06	A	9.36	0.13	0.01	A	0.29	0.04	0.00	A	6.62	0.20	0.02	A	33.87	0.93	0.04	A	0.35	0.19	0.01	A	3.58	0.24	0.02	A
19	183.29	-0.49	-0.04	A	27.73	-0.20	-0.01	A	12.93	0.27	0.02	A	11.39	-0.24	-0.02	A	10.83	0.39	0.03	A	0.91	-0.28	-0.03	A	0.21	-0.05	0.00	A
20	153.75	-0.49	-0.04	A	57.53	-0.32	-0.02	A	14.25	-0.33	-0.02	A	28.31	-0.49	-0.03	A	24.90	0.54	0.05	A	0.27	-0.18	-0.01	A	2.52	-0.19	-0.01	A
21	31.79	-0.20	-0.02	A	0.07	-0.01	0.00	A	0.21	-0.03	0.00	A	3.06	-0.13	-0.01	A	10.08	0.36	0.03	A	3.14	-0.50	-0.05	A	0.10	0.03	0.00	A
22	94.91	0.37	0.03	A	5.58	0.09	0.01	A	14.29	-0.29	-0.03	A	0.34	0.04	0.00	A	0.13	-0.04	0.00	A	2.22	-0.45	-0.04	A	0.67	-0.09	-0.01	A
23	75.09	-0.32	-0.03	A	152.94	-0.47	-0.04	A	2.52	-0.12	-0.01	A	3.21	-0.13	-0.01	A	10.78	-0.38	-0.03	A	0.99	-0.29	-0.03	A	4.27	0.23	0.02	A
24	0.30	-0.02	0.00	A	79.82	-0.35	-0.02	A	1.11	0.08	0.01	A	18.26	-0.36	-0.02	A	28.26	0.58	0.05	A	0.90	-0.31	-0.02	A	1.76	-0.15	-0.01	A
25	1.80	-0.05	0.00	A	292.16	-0.67	-0.05	A	1.52	0.10	0.01	A	54.90	-0.61	-0.04	A	47.66	0.80	0.06	A	0.06	-0.08	-0.01	A	4.60	0.24	0.02	A
26	0.32	-0.02	0.00	A	103.83	-0.38	-0.03	A	1.26	0.08	0.01	A	14.36	-0.28	-0.02	A	0.14	0.04	0.00	A	0.00	0.01	0.00	A	0.61	0.08	0.01	A
27	98.12	-0.41	-0.03	A	57.14	-0.34	-0.02	A	0.00	0.00	0.00	A	7.67	-0.26	-0.01	A	1.79	0.15	0.01	A	0.11	-0.12	-0.01	A	5.93	-0.31	-0.02	A
28	81.88	-0.32	-0.03	A	18.16	-0.16	-0.01	A	24.61	-0.37	-0.03	A	0.37	-0.04	0.00	A	10.11	0.34	0.03	A	0.00	-0.02	0.00	A	1.89	0.15	0.01	A
29	461.46	-0.79	-0.07	A	300.45	-0.65	-0.05	A	33.42	-0.44	-0.04	A	26.28	-0.38	-0.03	A	2.81	-0.19	-0.01	A	5.62	-0.73	-0.06	A	0.92	0.10	0.01	A
30	422.58	0.86	0.06	A	64.92	0.34	0.03	A	6.59	-0.21	-0.02	A	9.68	0.24	0.02	A	3.04	0.26	0.01	A	0.72	0.28	0.02	A	0.08	-0.04	0.00	A

Note: African Am. = African American, Native Am. = Native American, MH χ2 = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.

(table continues)

Table B.2 (continued)
Spring 2016 AIMS Differential Item Functioning
Science CRT Grade 8

	Reference: Male N= 41573 Focal: Female N= 40292				Reference: Hispanic N= 45666 Focal: Non Hispanic N= 36312				Reference: White N= 65677 Focal: Africa American N= 4950				Reference: White N= 65677 Focal: Native American N= 5567				Reference: White N= 65677 Focal: Asian N= 2495				Reference: White N= 65677 Focal: Hawaii N= 305				Reference: White N= 65677 Focal: Multiple Indicator N= 2354			
Item	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag
31	54.03	-0.26	-0.02	A	2.52	-0.06	0.00	A	0.27	-0.04	0.00	A	0.40	-0.05	0.00	A	8.27	0.32	0.03	A	0.51	-0.20	-0.02	A	1.22	0.12	0.01	A
32	403.16	0.98	0.05	A	1.33	-0.06	0.00	A	4.26	-0.19	-0.01	A	0.48	0.06	0.00	A	0.15	-0.07	0.00	A	0.12	0.13	0.01	A	0.27	0.08	0.00	A
33	1.80	0.06	0.00	A	0.33	0.02	0.00	A	0.01	-0.01	0.00	A	0.18	0.03	0.00	A	15.52	0.63	0.03	A	0.04	-0.07	0.00	A	0.15	0.05	0.00	A
34	1753.61	-1.51	-0.14	C<	10.61	-0.12	-0.01	A	19.29	-0.34	-0.03	A	2.14	-0.11	-0.01	A	6.98	0.29	0.03	A	3.49	-0.57	-0.05	A	0.25	-0.05	0.00	A
35	29.24	0.19	0.02	A	126.05	-0.41	-0.04	A	15.27	0.29	0.03	A	0.49	-0.05	0.00	A	2.36	0.16	0.02	A	2.09	0.42	0.04	A	1.62	0.13	0.01	A
36	69.82	-0.30	-0.03	A	31.89	-0.21	-0.02	A	6.48	0.19	0.02	A	0.57	-0.05	0.00	A	28.83	0.58	0.05	A	0.00	-0.01	0.00	A	0.35	0.06	0.01	A
37	2.92	-0.07	-0.01	A	7.29	-0.11	-0.01	A	0.32	-0.04	0.00	A	0.94	-0.07	-0.01	A	7.92	-0.36	-0.02	A	2.39	-0.46	-0.04	A	2.30	0.18	0.01	A
38	201.00	0.51	0.05	A	43.34	0.25	0.02	A	0.41	-0.05	0.00	A	0.21	0.03	0.00	A	16.73	0.47	0.04	A	0.00	0.01	0.00	A	0.59	0.08	0.01	A
39	0.78	-0.03	0.00	A	4.86	-0.08	-0.01	A	0.98	0.07	0.01	A	13.35	0.26	0.02	A	2.80	0.19	0.02	A	3.31	0.51	0.05	A	0.21	0.05	0.00	A
40	11.44	-0.13	-0.01	A	13.06	0.14	0.01	A	7.14	-0.21	-0.02	A	9.17	-0.22	-0.02	A	27.45	0.72	0.04	A	1.70	-0.41	-0.03	A	0.18	-0.05	0.00	A
41	0.04	0.01	0.00	A	37.48	-0.26	-0.02	A	11.67	-0.28	-0.02	A	0.00	0.00	0.00	A	0.11	-0.05	0.00	A	0.29	0.19	0.01	A	0.94	0.12	0.01	A
42	540.74	1.08	0.06	B>	5.42	0.11	0.01	A	0.73	-0.08	-0.01	A	0.65	-0.06	0.00	A	8.93	0.55	0.02	A	1.23	-0.37	-0.02	A	0.91	0.14	0.01	A
43	1.57	-0.05	0.00	A	10.95	-0.12	-0.01	A	12.09	-0.26	-0.02	A	31.94	-0.42	-0.04	A	0.62	-0.09	-0.01	A	1.55	-0.37	-0.03	A	7.62	-0.30	-0.03	A
44	129.60	0.56	0.03	A	7.14	0.13	0.01	A	3.31	-0.17	-0.01	A	0.34	0.05	0.00	A	0.53	-0.14	0.00	A	0.07	-0.10	-0.01	A	0.46	-0.10	0.00	A
45	29.63	-0.19	-0.02	A	31.78	-0.20	-0.02	A	9.24	-0.22	-0.02	A	10.92	-0.23	-0.02	A	12.95	0.37	0.04	A	0.22	0.13	0.01	A	9.84	-0.32	-0.03	A
46	130.78	-0.42	-0.04	A	5.64	-0.09	-0.01	A	0.00	0.00	0.00	A	5.41	0.17	0.02	A	9.62	0.36	0.03	A	0.36	-0.18	-0.02	A	0.15	-0.04	0.00	A
47	56.14	0.26	0.03	A	5.38	0.08	0.01	A	0.00	0.00	0.00	A	8.66	-0.21	-0.02	A	13.77	0.41	0.03	A	0.17	0.12	0.01	A	3.50	-0.19	-0.02	A
48	13.88	-0.15	-0.01	A	60.63	-0.31	-0.03	A	0.28	0.04	0.00	A	15.05	-0.29	-0.02	A	0.45	0.09	0.01	A	0.05	-0.07	-0.01	A	3.53	0.22	0.02	A
49	0.00	0.00	0.00	A	0.60	-0.03	0.00	A	1.06	-0.08	-0.01	A	3.54	0.14	0.01	A	0.67	-0.12	-0.01	A	7.73	-0.84	-0.07	A	0.40	-0.08	-0.01	A
50	140.01	0.44	0.04	A	26.38	-0.20	-0.01	A	0.17	0.03	0.00	A	0.42	0.05	0.00	A	69.20	0.87	0.08	A	0.30	0.16	0.01	A	1.98	-0.16	-0.01	A
51	267.36	0.59	0.05	A	56.15	0.28	0.02	A	6.74	0.20	0.02	A	0.16	-0.03	0.00	A	7.90	0.32	0.03	A	0.29	-0.15	-0.01	A	1.27	-0.12	-0.01	A
52	26.44	-0.19	-0.02	A	69.00	-0.31	-0.02	A	0.62	0.06	0.01	A	1.52	-0.09	-0.01	A	0.33	0.07	0.00	A	0.06	0.07	0.01	A	1.99	0.15	0.01	A
53	36.80	0.28	0.01	A	13.90	0.18	0.01	A	5.18	-0.20	-0.01	A	2.31	-0.12	-0.01	A	1.78	0.25	0.01	A	4.65	-0.81	-0.05	A	0.10	0.05	0.00	A
54	30.49	-0.20	-0.02	A	38.98	-0.23	-0.02	A	0.53	-0.05	0.00	A	7.54	-0.20	-0.02	A	17.90	0.51	0.04	A	0.00	-0.01	0.00	A	1.51	0.14	0.01	A
55	0.60	0.03	0.00	A	9.41	-0.12	-0.01	A	0.00	0.00	0.00	A	2.61	-0.13	-0.01	A	53.25	0.76	0.07	A	0.17	-0.12	-0.01	A	0.18	0.05	0.00	A
56	32.10	0.22	0.02	A	7.34	0.11	0.01	A	13.17	-0.28	-0.02	A	4.63	0.16	0.01	A	4.24	0.27	0.02	A	2.34	-0.47	-0.04	A	7.49	-0.32	-0.02	A
57	115.51	-0.36	-0.04	A	19.26	-0.15	-0.01	A	6.59	-0.18	-0.02	A	0.32	-0.04	0.00	A	0.49	0.07	0.01	A	0.03	-0.05	0.00	A	0.02	-0.01	0.00	A
58	68.04	-0.29	-0.03	A	111.60	-0.38	-0.03	A	29.76	-0.40	-0.04	A	0.92	-0.07	-0.01	A	21.32	0.50	0.04	A	0.42	0.19	0.02	A	0.86	0.10	0.01	A

Note: African Am. = African American, Native Am. = Native American, MH χ2 = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.

Table B.3
Spring 2016 AIMS Differential Item Functioning
Science CRT High School

	Reference: Male N= 39746 Focal: Female N= 39925				Reference: Hispanic N= 45262 Focal: Non Hispanic N= 34714				Reference: White N= 62142 Focal: Africa American N= 4682				Reference: White N= 62142 Focal: Native American N= 4322				Reference: White N= 62142 Focal: Asian N= 2508				Reference: White N= 62142 Focal: Hawaii N= 320				Reference: White N= 62142 Focal: Multiple Indicator N= 2265			
Item	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag	MH χ2	ΔMH	SMD	Flag
1	198.21	0.50	0.05	A	8.24	0.11	0.01	A	35.94	-0.46	-0.04	A	1.63	-0.10	-0.01	A	10.39	-0.34	-0.03	A	0.03	0.05	0.00	A	0.22	-0.05	0.00	A
2	114.23	0.42	0.03	A	9.57	0.12	0.01	A	0.33	-0.05	0.00	A	9.80	-0.26	-0.02	A	22.21	0.60	0.04	A	1.59	0.37	0.03	A	1.48	-0.14	-0.01	A
3	61.97	-0.28	-0.03	A	7.89	-0.10	-0.01	A	1.08	0.08	0.01	A	0.01	-0.01	0.00	A	38.78	0.66	0.06	A	1.43	0.33	0.03	A	1.49	0.13	0.01	A
4	56.44	-0.26	-0.02	A	2.00	-0.05	0.00	A	3.72	-0.15	-0.01	A	4.42	0.17	0.02	A	84.05	0.95	0.09	A	0.01	-0.02	0.00	A	0.73	-0.09	-0.01	A
5	0.25	-0.02	0.00	A	88.25	0.34	0.04	A	4.04	0.15	0.01	A	55.76	0.58	0.06	A	114.78	1.08	0.11	B>	0.04	0.06	0.01	A	0.01	-0.01	0.00	A
6	2.16	-0.05	0.00	A	87.01	-0.35	-0.03	A	8.48	0.22	0.02	A	0.98	-0.08	-0.01	A	4.26	0.21	0.02	A	0.18	0.12	0.01	A	1.84	0.14	0.01	A
7	0.33	-0.02	0.00	A	176.99	-0.59	-0.03	A	4.12	0.19	0.01	A	23.79	0.47	0.03	A	162.16	1.41	0.12	B>	0.00	0.02	0.00	A	3.78	0.24	0.02	A
8	77.85	-0.32	-0.03	A	0.02	0.01	0.00	A	17.17	-0.31	-0.03	A	0.33	0.05	0.00	A	0.60	0.09	0.01	A	4.18	-0.56	-0.05	A	0.04	0.02	0.00	A
9	6.91	0.10	0.01	A	3.35	-0.07	-0.01	A	15.10	-0.30	-0.03	A	8.19	-0.23	-0.02	A	0.45	0.08	0.01	A	1.58	0.37	0.03	A	0.13	-0.04	0.00	A
10	80.51	0.33	0.03	A	0.89	0.04	0.00	A	0.12	0.03	0.00	A	0.22	-0.04	0.00	A	14.38	0.45	0.03	A	0.04	0.06	0.01	A	0.04	-0.02	0.00	A
11	2.04	0.05	0.01	A	0.96	-0.04	0.00	A	0.00	0.00	0.00	A	5.62	-0.19	-0.02	A	1.28	-0.12	-0.01	A	0.11	0.09	0.01	A	0.03	-0.02	0.00	A
12	20.83	-0.17	-0.01	A	29.62	-0.22	-0.02	A	27.16	0.43	0.03	A	6.06	0.22	0.02	A	10.32	0.33	0.03	A	1.91	0.40	0.03	A	0.58	-0.09	-0.01	A
13	5.19	0.10	0.01	A	56.07	-0.32	-0.03	A	2.05	-0.12	-0.01	A	40.20	-0.53	-0.04	A	0.30	-0.08	0.00	A	0.01	0.03	0.00	A	3.62	0.25	0.02	A
14	0.06	-0.01	0.00	A	27.14	0.19	0.02	A	2.26	0.11	0.01	A	3.47	0.14	0.01	A	3.53	0.19	0.02	A	0.07	-0.07	-0.01	A	4.49	0.22	0.02	A
15	154.30	0.50	0.04	A	26.08	-0.21	-0.01	A	0.00	0.00	0.00	A	4.67	-0.18	-0.02	A	7.94	-0.35	-0.02	A	2.10	-0.44	-0.03	A	1.93	0.17	0.01	A
16	253.39	-0.60	-0.05	A	118.23	-0.42	-0.04	A	59.45	-0.60	-0.05	A	32.40	-0.46	-0.04	A	12.57	-0.41	-0.03	A	2.32	-0.44	-0.04	A	1.58	-0.14	-0.01	A
17	20.37	-0.16	-0.02	A	11.76	-0.13	-0.01	A	0.82	0.07	0.01	A	45.66	-0.55	-0.05	A	32.66	0.66	0.05	A	0.34	0.16	0.02	A	0.17	-0.05	0.00	A
18	1446.10	-1.46	-0.12	B<	97.31	-0.38	-0.03	A	25.51	-0.41	-0.03	A	0.59	0.06	0.01	A	0.35	-0.07	0.00	A	0.37	-0.18	-0.02	A	2.77	0.19	0.02	A
19	47.21	-0.25	-0.02	A	15.98	-0.16	-0.01	A	4.66	0.17	0.01	A	0.96	-0.09	-0.01	A	70.08	0.86	0.08	A	1.08	0.30	0.03	A	1.49	0.14	0.01	A
20	162.66	-0.65	-0.03	A	2.64	-0.08	0.00	A	32.37	-0.55	-0.03	A	15.06	-0.37	-0.02	A	11.35	-0.62	-0.02	A	0.67	0.33	0.02	A	0.03	-0.03	0.00	A
21	357.98	-0.74	-0.06	A	40.36	-0.25	-0.02	A	20.32	-0.36	-0.03	A	0.04	-0.02	0.00	A	2.12	0.18	0.01	A	0.72	0.27	0.02	A	0.00	0.00	0.00	A
22	5.27	0.08	0.01	A	6.99	0.10	0.01	A	17.01	0.31	0.03	A	63.33	0.63	0.06	A	1.34	-0.12	-0.01	A	2.47	0.45	0.04	A	0.36	0.06	0.01	A
23	259.09	-0.65	-0.05	A	4.54	0.09	0.01	A	1.15	-0.09	-0.01	A	2.06	-0.12	-0.01	A	0.99	-0.13	-0.01	A	0.72	0.26	0.02	A	0.35	0.07	0.01	A
24	531.21	0.84	0.08	A	2.97	0.06	0.01	A	1.75	0.10	0.01	A	2.75	0.13	0.01	A	27.11	0.60	0.05	A	2.48	0.45	0.04	A	0.03	-0.02	0.00	A
25	401.01	0.74	0.06	A	81.87	-0.34	-0.02	A	42.09	-0.52	-0.04	A	103.37	-0.89	-0.07	A	9.22	-0.33	-0.03	A	13.56	-1.09	-0.09	B<	4.43	-0.23	-0.02	A
26	385.48	-0.73	-0.06	A	30.94	-0.22	-0.02	A	0.57	0.06	0.01	A	8.67	-0.26	-0.02	A	0.04	0.02	0.00	A	0.71	0.26	0.02	A	0.19	0.05	0.00	A
27	9.22	-0.11	-0.01	A	12.39	-0.13	-0.01	A	0.27	-0.04	0.00	A	0.66	-0.06	-0.01	A	8.48	0.32	0.03	A	0.60	0.22	0.02	A	0.46	-0.07	-0.01	A
28	25.42	0.18	0.02	A	0.47	0.03	0.00	A	6.55	0.20	0.02	A	0.16	0.03	0.00	A	9.26	0.33	0.03	A	4.98	0.62	0.06	A	0.96	-0.11	-0.01	A
29	17.34	-0.15	-0.01	A	3.39	-0.07	0.00	A	15.71	-0.32	-0.03	A	2.56	-0.14	-0.01	A	0.97	-0.11	-0.01	A	1.46	0.36	0.03	A	1.07	0.11	0.01	A
30	30.89	0.19	0.02	A	22.93	0.17	0.02	A	0.21	-0.03	0.00	A	2.04	-0.11	-0.01	A	56.77	0.79	0.07	A	4.10	-0.58	-0.05	A	0.58	0.08	0.01	A

Note: African Am. = African American, Native Am. = Native American, MH χ2 = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.

(table continues)

Table B.3 (continued)
Spring 2016 AIMS Differential Item Functioning
Science CRT High School

	Reference: Male N= 39746				Reference: Hispanic N= 45262				Reference: White N= 62142				Reference: White N= 62142				Reference: White N= 62142				Reference: White N= 62142				Reference: White N= 62142			
	Focal: Female N= 39925				Focal: Non Hispanic N= 34714				Focal: Africa American N= 4682				Focal: Native American N= 4322				Focal: Asian N= 2508				Focal: Hawaii N= 320				Focal: Multiple Indicator N= 2265			
Item	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag	MH $\chi 2$	ΔMH	SMD	Flag
31	0.04	0.01	0.00	A	0.02	0.00	0.00	A	6.59	0.20	0.02	A	9.41	-0.25	-0.02	A	3.35	0.20	0.02	A	0.12	0.10	0.01	A	0.15	0.04	0.00	A
32	190.51	0.49	0.05	A	13.24	0.13	0.01	A	0.00	0.00	0.00	A	0.02	-0.01	0.00	A	27.31	0.57	0.05	A	2.48	0.45	0.04	A	3.49	0.20	0.02	A
33	18.93	-0.16	-0.01	A	4.62	-0.08	0.00	A	0.84	-0.07	-0.01	A	2.01	-0.12	-0.01	A	18.94	0.47	0.04	A	0.90	0.28	0.02	A	0.07	0.03	0.00	A
34	73.04	-0.33	-0.03	A	3.45	-0.08	0.00	A	0.00	0.00	0.00	A	21.44	0.41	0.03	A	21.52	0.50	0.04	A	0.00	-0.01	0.00	A	0.02	-0.02	0.00	A
35	1.06	-0.04	0.00	A	0.00	0.00	0.00	A	2.04	-0.12	-0.01	A	34.99	-0.50	-0.04	A	0.62	-0.10	-0.01	A	1.03	-0.32	-0.02	A	5.84	-0.29	-0.02	A
36	40.28	0.28	0.02	A	0.19	-0.02	0.00	A	35.03	-0.51	-0.04	A	0.67	0.07	0.01	A	0.68	0.14	0.01	A	0.27	-0.19	-0.01	A	6.59	-0.34	-0.02	A
37	3.39	0.07	0.01	A	4.07	0.08	0.01	A	13.35	-0.29	-0.03	A	1.95	0.11	0.01	A	0.48	0.08	0.01	A	0.01	0.03	0.00	A	3.98	-0.23	-0.02	A
38	40.25	0.23	0.02	A	107.67	-0.39	-0.03	A	17.80	-0.32	-0.03	A	19.77	-0.36	-0.03	A	0.76	-0.09	-0.01	A	0.14	-0.10	-0.01	A	0.74	-0.09	-0.01	A
39	0.03	0.01	0.00	A	58.48	-0.29	-0.02	A	7.14	-0.21	-0.02	A	65.30	0.64	0.06	A	3.30	-0.19	-0.02	A	0.50	0.21	0.02	A	2.92	0.19	0.02	A
40	6.33	0.09	0.01	A	21.93	0.18	0.02	A	5.75	-0.19	-0.02	A	9.65	-0.25	-0.02	A	3.03	0.21	0.01	A	6.14	0.72	0.06	A	2.65	0.18	0.02	A
41	2.28	0.06	0.00	A	5.18	0.09	0.01	A	6.08	0.19	0.02	A	0.01	-0.01	0.00	A	78.89	0.91	0.09	A	1.38	-0.35	-0.03	A	1.29	-0.13	-0.01	A
42	36.31	-0.22	-0.02	A	78.92	-0.34	-0.03	A	0.10	-0.03	0.00	A	0.05	0.02	0.00	A	17.71	0.43	0.04	A	0.33	-0.17	-0.01	A	0.53	-0.08	-0.01	A
43	0.98	-0.04	0.00	A	24.98	-0.19	-0.01	A	2.84	-0.13	-0.01	A	0.04	-0.02	0.00	A	2.32	0.16	0.01	A	1.35	-0.34	-0.03	A	0.13	-0.04	0.00	A
44	47.67	-0.25	-0.02	A	6.12	-0.09	-0.01	A	78.19	-0.70	-0.06	A	0.19	0.04	0.00	A	27.70	-0.53	-0.05	A	0.20	-0.13	-0.01	A	0.00	0.01	0.00	A
45	9.11	0.11	0.01	A	15.85	-0.16	-0.01	A	0.84	0.08	0.01	A	10.52	-0.30	-0.02	A	0.21	0.05	0.00	A	0.01	0.03	0.00	A	8.20	0.33	0.03	A
46	198.54	-0.54	-0.04	A	454.95	-0.83	-0.06	A	0.00	0.00	0.00	A	2.93	-0.15	-0.01	A	0.01	0.01	0.00	A	0.10	-0.09	-0.01	A	0.72	0.10	0.01	A
47	248.50	0.58	0.05	A	143.67	-0.45	-0.04	A	1.30	-0.09	-0.01	A	1.67	0.10	0.01	A	11.09	-0.37	-0.03	A	1.10	0.30	0.03	A	0.03	0.02	0.00	A
48	157.70	-0.49	-0.04	A	8.92	-0.12	-0.01	A	0.50	-0.06	0.00	A	0.05	-0.02	0.00	A	0.78	-0.09	-0.01	A	1.74	-0.42	-0.03	A	0.01	0.01	0.00	A
49	41.33	0.23	0.02	A	25.09	-0.19	-0.01	A	1.29	-0.09	-0.01	A	71.38	-0.71	-0.06	A	0.03	-0.02	0.00	A	1.50	-0.36	-0.03	A	2.50	-0.17	-0.02	A
50	95.48	0.38	0.03	A	2.70	0.06	0.01	A	4.65	-0.17	-0.01	A	12.15	-0.28	-0.02	A	14.35	-0.46	-0.03	A	6.04	-0.77	-0.06	A	0.13	-0.04	0.00	A
51	278.36	0.61	0.05	A	35.38	-0.22	-0.02	A	3.04	0.14	0.01	A	12.70	-0.28	-0.02	A	34.80	0.65	0.05	A	0.05	-0.06	-0.01	A	2.55	0.18	0.02	A
52	106.15	0.38	0.03	A	0.17	-0.02	0.00	A	0.85	-0.07	-0.01	A	2.24	-0.12	-0.01	A	2.12	0.17	0.01	A	2.29	0.44	0.04	A	0.55	0.09	0.01	A
53	26.51	0.19	0.02	A	0.42	-0.02	0.00	A	0.50	-0.05	-0.01	A	1.90	-0.11	-0.01	A	49.55	0.81	0.06	A	1.43	0.34	0.03	A	0.71	-0.09	-0.01	A
54	254.08	-0.64	-0.05	A	209.70	-0.59	-0.04	A	23.15	-0.39	-0.03	A	22.13	-0.41	-0.03	A	0.01	0.01	0.00	A	2.14	-0.46	-0.03	A	1.81	0.16	0.01	A
55	184.08	-0.49	-0.04	A	1.95	0.05	0.01	A	2.23	-0.12	-0.01	A	10.11	-0.27	-0.02	A	5.57	0.24	0.02	A	0.92	-0.27	-0.02	A	0.00	0.00	0.00	A
56	3.22	0.06	0.01	A	10.12	-0.12	-0.01	A	1.87	0.10	0.01	A	0.61	0.06	0.01	A	0.34	0.06	0.01	A	7.73	0.79	0.07	A	0.01	-0.01	0.00	A
57	9.30	-0.11	-0.01	A	0.00	0.00	0.00	A	0.04	0.01	0.00	A	16.10	-0.32	-0.03	A	6.94	0.27	0.03	A	0.11	-0.10	-0.01	A	7.04	-0.28	-0.03	A
58	25.30	-0.19	-0.02	A	28.39	-0.21	-0.01	A	12.75	-0.29	-0.02	A	12.81	-0.31	-0.02	A	10.40	0.35	0.03	A	0.43	-0.20	-0.02	A	3.64	-0.22	-0.02	A
59	44.52	0.24	0.02	A	8.65	-0.11	-0.01	A	3.75	-0.15	-0.01	A	0.00	0.00	0.00	A	18.90	0.46	0.04	A	0.07	-0.08	-0.01	A	0.04	0.02	0.00	A
60	15.06	-0.14	-0.01	A	33.18	-0.21	-0.02	A	2.53	-0.12	-0.01	A	2.73	0.13	0.01	A	13.70	0.41	0.03	A	0.00	0.00	0.00	A	0.50	0.08	0.01	A
61	438.02	0.85	0.06	A	24.45	0.20	0.02	A	7.54	0.23	0.02	A	19.39	-0.36	-0.03	A	29.87	0.80	0.04	A	0.10	0.10	0.01	A	0.77	0.11	0.01	A
62	0.00	0.00	0.00	A	0.74	-0.03	0.00	A	2.21	0.11	0.01	A	13.32	-0.30	-0.03	A	6.51	0.27	0.02	A	0.45	-0.19	-0.02	A	0.08	-0.03	0.00	A
63	38.64	-0.24	-0.02	A	8.96	-0.12	-0.01	A	7.93	0.23	0.02	A	0.25	0.05	0.00	A	44.69	0.70	0.06	A	0.00	-0.02	0.00	A	1.11	0.12	0.01	A
64	0.55	-0.03	0.00	A	16.56	0.15	0.02	A	5.34	0.18	0.02	A	1.20	0.09	0.01	A	1.49	0.12	0.01	A	0.01	0.03	0.00	A	1.03	-0.11	-0.01	A
65	3.24	0.07	0.01	A	26.21	-0.19	-0.02	A	11.56	-0.26	-0.02	A	19.44	0.35	0.03	A	16.52	-0.42	-0.04	A	7.20	-0.77	-0.07	A	0.00	0.00	0.00	A

Note: African Am. = African American, Native Am. = Native American, MH $\chi 2$ = Mantel_Haenszel Chi-Square, ΔMH = Mantel-Haenszel Delta DIF, SMD = Standardized Mean Difference, A=No DIF, B=Weak DIF, C=Strong DIF, < favors > favors focal group. Item number does not indicate test booklet location due to field test items and NRT items.