



Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Math

2016–2017 School Year

September 2017

ARIZONA STATEWIDE ASSESSMENT

**ARIZONA'S MEASUREMENT OF EDUCATIONAL READINESS TO INFORM
TEACHING (AzMERIT)**

ENGLISH LANGUAGE ARTS GRADES 3–11

MATH GRADES 3–8, ALGEBRA I, GEOMETRY, AND ALGEBRA II

2016–2017 ANNUAL TECHNICAL REPORT

SEPTEMBER 2017

Prepared by American Institutes for Research (AIR) in collaboration with the
Arizona Department of Education

TABLE OF CONTENTS

1.	Introduction: The Validity of AzMERIT Test Score Interpretations	1
1.1	Overview	1
1.2	Validity Evidence	2
1.3	Evidence Based on Test Content	9
1.4	Evidence for Interpretation of Performance Standards.....	12
1.5	Evidence Based on Internal Structure	15
1.5.1	ELA Content Model	16
1.5.2	ELA Depth of Knowledge.....	17
1.5.3	Math Content Model.....	18
1.5.4	Math Depth of Knowledge	19
1.6	Evidence for Relationships with Conceptually Related Constructs.....	20
1.7	Measurement Invariance Across Subgroups.....	22
1.8	Differential Mode Effects Across Subgroups.....	23
1.9	Evidence for Student Growth – Overall and by Subgroups.....	25
1.10	Day, Week, and Time of Day Effects on Performance.....	29
1.11	Arizona Glossary Study.....	31
1.12	Summary of Validity of Test Score Interpretations	35
2.	Background of Arizona Statewide Assessments.....	36
2.1	Development of Arizona College and Career Ready Standards	37
2.2	AzMERIT Test Design	37
3.	Summary of Summer 2016 and Fall 2016 Operational Test Administration.....	39
3.1	Student Population and Participation	39
3.2	Summary of Overall Student Performance	41
3.3	Student Performance by Subgroup	42
3.4	Reliability.....	48
3.4.1	Internal Consistency	48

3.4.2	Standard Error of Measurement	49
3.4.3	Student Classification Reliability	50
3.4.4	Classification Accuracy	50
3.4.5	Classification Consistency	51
3.4.6	Classification Reliability Estimates	51
3.4.7	Reliability for Subgroups in the Population.....	52
3.4.8	Subscale Reliability.....	54
3.5	Subscale Intercorrelations.....	56
4.	Summary of Spring 2017 Operational Test Administration	58
4.1	Student Population and Participation	58
4.2	Classical Item Analysis	59
4.3	Item Response Theory Analysis.....	61
4.4	Summary of Overall Student Performance	63
4.5	Student Performance by Subgroup.....	67
4.6	Reliability.....	73
4.6.1	Internal Consistency.....	74
4.6.2	Standard Error of Measurement	74
4.6.3	Student Classification Reliability	75
4.6.4	Classification Accuracy	76
4.6.5	Classification Consistency	76
4.6.6	Classification Accuracy and Consistency Estimates	77
4.6.7	Reliability for Subgroups in the Population.....	82
4.6.8	Subscale Reliability.....	84
4.7	Subscale Intercorrelations.....	86
4.8	Handscoring agreement rate.....	88
5.	Item Development and Test Construction	90
5.1	Item Development Process.....	91

5.1.1	Item Writing.....	91
5.1.2	Machine-Scored Constructed-Response Item Development Tools	94
5.1.3	Item Types	95
5.2	Item Review.....	96
5.3	Field Testing	98
5.4	Item Statistics.....	99
5.4.1	Classical Statistics.....	99
5.4.2	IRT Statistics.....	100
5.4.3	Analysis of Differential Item Functioning	100
5.5	Test Construction	102
5.5.1	Operational Form Construction	102
5.5.2	Assembling Test Forms	104
6.	Test Administration	106
6.1	Eligibility	106
6.2	Administration Procedures	106
6.2.1	Managing Testing.....	109
6.3	Testing Conditions, Tools, and Accommodations	109
6.3.1	Universal Test Administration Conditions.....	109
6.3.2	Universal Testing Tools for Computer-Based Testers	110
6.3.3	Subject Area Tools for CBT and PBT	111
6.3.4	Accommodations	112
6.4	System Security	116
6.4.1	Secure System Design	116
6.4.2	System Security Components.....	116
6.5	Test Security	117
6.6	Data Forensics Program	119
6.6.1	Changes in Student Performance.....	120

6.6.2	Item Response Latency	121
6.6.3	Inconsistent Item Response Pattern (Person Fit)	121
6.6.4	Response Change and Response Similarity	122
7.	Reporting and Interpreting AzMERIT Scores	125
7.1	Appropriate Uses for Scores and Reports	125
7.2	Reports Provided	126
7.2.1	Family Reports	126
7.2.2	Online Reporting System for Educators	127
7.3	Interpretation of Scores	131
8.	Performance Standards	133
8.1	Standard Setting Procedures	133
8.1.1	Performance-Level Descriptors	134
8.2	Recommended Performance Standards	134
9.	Scaling And Equating	138
9.1.1	Item Response Theory Procedures	139
9.1.2	Calibration of AzMERIT Item Banks	139
9.1.3	Estimating Student Ability Using Maximum Likelihood Estimation	140
9.2	Establishing a Vertical Scale in ELA and Math	141
9.2.1	Linking Items	141
9.2.2	Linking Analysis	142
9.3	AzMERITAzMerit Reporting Scale (Scale Scores)	149
9.4	Linking paper and Online Test Scores (Mode Comparability)	150
9.4.1	Mode Linking	150
9.4.2	School Performance	154
9.5	Linking the AzMERIT to Other Scales for Performance Comparison	154
9.5.1	Establishing Linkages to AIMS, SAGE, Smarter Balanced, PISA	154
9.5.2	Identifying the Location of the ACT College-Ready Cut on AzMERIT	155

10. Constructed-Response Scoring.....	158
10.1 Machine Scoring.....	158
10.1.1 Explicit Rubrics	158
10.1.2 Essay Autoscoring.....	158
10.2 HandScoring	165
10.2.1 Handscoring Process	165
10.2.2 HandScoring Quality Control.....	166
10.2.3 HandScoring Reliability and Validity	167
10.2.4 Machine-Scoring verification	168
11. Quality Assurance Procedures.....	169
11.1 Quality Assurance in Test Construction	169
11.2 Quality Assurance in Paper-Delivered Test Production	170
11.3 Quality Assurance in Computer-Delivered Test Production.....	171
11.3.1 Production of Content.....	171
11.3.2 Web Approval of Content During Development	172
11.3.3 Approval of Final Forms	172
11.3.4 Packaging	172
11.3.5 Platform Review	172
11.3.6 User Acceptance Testing and Final Review.....	173
11.3.7 Functionality and Configuration.....	173
11.4 Quality Assurance in Document Processing.....	173
11.4.1 Scanning Accuracy.....	173
11.4.2 Quality Assurance in Editing and Data Input.....	174
11.5 Quality Assurance in Data Preparation	175
11.6 Quality Assurance in Test Form Equating.....	175
11.7 Quality Assurance in Scoring and Reporting	176
11.7.1 Quality Assurance in HandScoring	176

11.7.2	Test Scoring	177
11.7.3	Reporting.....	180
12.	References.....	181

APPENDICES

Appendix A. AzMERIT Calculator Guidelines	A-1
Appendix B. AzMERIT ELA and Mathematics Test Blueprints.....	B-1
Appendix C. Measurement Invariance Testing by Subgroups	C-1
Appendix D. Differential Growth Analysis Across Subgroups – From Spring 2016 to Spring 2017.....	D-1
Appendix E. Equations and Formula for Estimating Reliability.....	E-1
Appendix F. Student Participation by Demographic Subgroup – Spring 2017 Administration	F-1
Appendix G. Operational Item Parameter Estimates – Spring 2017 Administration.....	G-1
Appendix H. Data Review Training Slides	H-1
Appendix I. Test Characteristic Curves – Spring 2017 Administration	I-1

1. INTRODUCTION: THE VALIDITY OF AZMERIT TEST SCORE INTERPRETATIONS

1.1 OVERVIEW

The purpose of this technical report is to document the evidence supporting the claims made for how Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) test scores may be interpreted. Evidence for the validity of test score interpretations is central to claims that AzMERIT test scores can be used to evaluate the effectiveness with which Arizona districts and schools teach students the Arizona College and Career Ready Standards (ACCRS) and whether individual students have achieved those standards by the end of each school year. Thus, the report begins with a review of validity evidence evaluated to date. Evidence for the validity of test score interpretations is expected to accrue over time, so this section will be expanded as further evidence is gained.

Chapter 2 describes the design and development of the AzMERIT assessment system, including the ACCRS, which define the content domain to be assessed by AzMERIT; the development of test specifications, including blueprints, that ensure that the breadth and depth of the content domain is adequately sampled by the assessments; and test development procedures that ensure alignment of test forms with the blueprint specifications.

Chapters 3 and 4 provide summaries of the AzMERIT test administrations. Chapter 3 presents results of the summer 2016 and fall 2016 administrations of the high school end-of-course (EOC) assessments, and Chapter 4 presents results of the spring 2017 administration of the full AzMERIT assessment system, including end-of-year (EOC) assessments in ELA and math for grades 3–8 and high school. These chapters provide summaries of the test-taking student population and their performance on the assessments. In addition, these chapters describe administration-specific evidence for the reliability of the AzMERIT assessments, including internal consistency reliability, standard errors of measurement, and the reliability of performance-level classifications.

The remaining chapters document technical details of the test development, administration, scoring, and reporting activities.

Chapter 5 describes the item development process, especially the sequence of reviews that each item must pass through before being eligible for AzMERIT test administration. This chapter also describes the procedures for constructing test forms from items successfully passing through the review process. Chapter 6 documents the test administration procedures, including eligibility of participation in the AzMERIT assessments; testing conditions, including accessibility tools and accommodations; systems security for assessments administered online; as well as test security procedures for all test administrations. Chapter 7 provides a description of the score reporting system and the interpretation of test scores. Chapter 8 describes the procedures that the Arizona Department of Education (ADE) uses to identify and adopt performance standards for AzMERIT assessments. Chapter 9 describes the procedures used to scale and equate the AzMERIT assessments for scoring and reporting. Chapter 10 describes the procedures for scoring constructed-response items, both machine-scored and handscored items, and it provides summary rater agreement results. Chapter 11 provides an overview of the quality assurance processes described throughout that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.2 VALIDITY EVIDENCE

Validity refers to the degree to which test score interpretations are supported by evidence, especially regarding the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the *Standards* describe the range of evidence supporting the validity of test score interpretations.

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the standards make explicit that validity is not an attribute of tests but rather of test score interpretations. Some test score interpretations are supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Central to evaluating the validity of test score interpretations is determining whether the test measures the intended construct. Such an evaluation in turn requires a clear definition of the measurement construct. For the AzMERIT, the definition of the measurement construct is provided by the Arizona College and Career Ready Standards (ACCRS).

In 2010, Arizona adopted new academic content standards in English language arts (ELA) and math. The ACCRS are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate.¹ In spring 2015, the ADE administered Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) to assess proficiency on the new ACCRS for the first time. The AzMERIT measures English language arts and math in grades 3–8 and following completion of high school coursework in ELA grades 9–11, Algebra I, Geometry, and Algebra II.

Because directly measuring student achievement against each benchmark in the ACCRS would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the ACCRS.² To ensure that each student is assessed on the intended breadth and depth of the ACCRS, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark.³ Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards, in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content

¹ Standard 1.1 – The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

² Standard 4.0 – Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

³ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

standards is critical. ADE has published the AzMERIT ELA and math test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels. The ELA and math blueprints are also provided in Appendix B.

While the blueprints ensure that the full range of the intended measurement construct is represented in each test administration, tests may also inadvertently measure attributes that are not relevant to the construct of interest. For example, when a high level of English language proficiency is necessary to access content in other subject-area assessments such as math or science, language proficiency may unnecessarily limit the student's ability to demonstrate achievement in those subject areas. Thus, while such tests may measure achievement of relevant subject-area content standards, they may also measure construct-irrelevant variation in language proficiency, limiting the generalizability of test score interpretations for some student populations.

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement.⁴ Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

- Inclusive assessment population
- Precisely-defined constructs
- Accessible, non-biased items
- Amenability to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including items and accompanying stimuli. In the review process, adherence to the principles of universal design is verified.

In addition, the AzMERIT test delivery system provides a range of accessibility tools and accommodations for reducing construct-irrelevant barriers to accessing test content for virtually all students.⁵ The range of accommodations, provided in the online testing environment, far exceeds the typical accommodations available in paper-based test administrations. Exhibits 1.2.1–1.2.5 list the accommodations and accessibility supports currently available for students taking the AzMERIT assessments online. Paper-based test forms are available as an accommodation for students testing in online schools should the accommodations provided online not be

⁴ Standard 3.0 – All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

⁵ Standard 3.1 – Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.

Standard 3.2 – Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

Standard 12.3 – Those responsible for the development and use of educational assessments should design all relevant steps of the testing process to promote access to the construct for all individuals and subgroups for whom the assessment is intended.

sufficient to remove barriers to accessing test content. These include both large print and Braille forms. Section 6.3 describes available testing tools and accommodations for students testing online and on paper.

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment. Universal test administration conditions are available for both paper-based test (PBT) and computer-based test (CBT) modes. Universal test administration conditions include the following:

- Testing in a small group, testing one-on-one, or 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 a place holder,
- Using devices that allow the student to see the test, such as eyeglasses, contact lenses, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions, such as hearing aids and amplification tools,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- Extended time. (Testing session must be completed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan (IEP) as an accommodation, for AzMERIT testing purposes, these are not considered testing accommodations and are available to any student who needs them, not just to students with IEPs.

Exhibit 1.2.1 summarizes the universal testing tools that are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 1.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	The student may click anywhere on the selected response text or button for multiple-choice options.
Expand/Collapse Passage	The student may expand a passage for easier readability. Expanded passages can also be collapsed.
Help	The student may view the on-screen <i>Test Instructions and Help</i> .
Highlighter	The student may highlight text in a passage or item.
Line Reader	The student may track the line he or she is reading.
Mark (Flag) for Review	The student may mark an item for review so that it can be easily found later.
Notes/Comments	The student may open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	The student may pause the session at any time and restart the test if taken over a one-day period. For test security purposes, visibility of past items is not allowed when the test is paused longer than 20 minutes.
Review Test	The student may review the test before ending it.
Strikethrough	The student may cross out answer options for multiple-choice and multi-select items.
System Settings	The student may adjust the audio volume during the test.
Text-to-Speech for Instructions	The student may listen to test instructions.
Tutorial	The student may view a short video about each item type and how to respond.
Writing Tools	The student may use these editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italics) for extended-response items.
Zoom In/Zoom Out	The student may “zoom in” to enlarge the font and images in the test and may “zoom out” to return the font and images in the test to original size.

AzMERIT testing requires specific subject-area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 1.2.2.

Exhibit 1.2.2 Subject-Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	CBT – Students may access the dictionary/thesaurus tool, or students may use a published, paper dictionary or thesaurus. PBT – Students may use published, paper dictionaries and thesauruses. Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned off.
Writing Guide	Writing	CBT – Students may access the writing guide tool. PBT – The writing guide is included within the test booklet.
Scratch Paper	Writing and Math	CBT – Schools must provide scratch paper (plain, lined, or graph) to students. PBT – Schools must provide scratch paper (plain, lined, or graph) to students.
Calculator Grades 7–8 (Part 1 only): specific scientific calculators are acceptable EOC (entire test): specific graphing calculators are acceptable	Math	CBT – Students may access the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted. PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.

Note: The details of the AzMERIT calculator guidance are presented in Appendix A.

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. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Learner (EL) or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education need, or language need and the accommodation(s) that are provided to the student during educational activities, including assessment. Test administrators are instructed to make accommodation decisions based on individual needs and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation that is not already used regularly in the classroom may be put in place for an AzMERIT test.

Testing accommodations may not violate the construct of a test item. Testing 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, writing prompt, or answer option. The

accommodations available to students while testing on AzMERIT are generally limited to those listed in the *AzMERIT Testing Conditions, Tools, and Accommodations Guidance* manual and summarized in this section. The ADE takes care to ensure that allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student’s individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact the ADE for guidance.

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. There are no specific CBT tools to support these accommodations.

Exhibit 1.2.3 Accommodations for Students with an Injury

Adult Transcription	If a student with an injury is testing at a CBT school and cannot enter his or her own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student’s responses exactly as provided, orally or by gestures, into the paper booklet and then into the DEI, or directly into the DEI. If a student with an injury at a PBT school cannot write his or her own responses in a booklet, an adult must transfer the student's responses exactly as provided orally or by gestures.
Assistive Technology	Assistive technology may be used for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires adult transcription (see above for rules on adult transcription).
Rest/Breaks	Student may take breaks during testing sessions to rest.

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the accommodations in Exhibit 1.2.4. This includes English Learner (EL) students and students withdrawn from English language services at parent request. Reclassified Fluent English Proficient (FEP) students are monitored for two school years. These FEP Year 1 and FEP Year 2 students also may use, as appropriate, any of the universal test administration conditions and any of the accommodations below.

The *upon student request* accommodations are required to be administered in a setting that does not disturb other students, such as in a one-on-one or very small group setting.

Exhibit 1.2.4 summarizes accommodations that may be provided for EL and FEP students.

Exhibit 1.2.4 Allowable Accommodations for EL and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test maybe be provided upon student request.</p> <p>Reading aloud the content of the Reading portion of the ELA test is prohibited.</p>
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Provide exact oral translation, in the student’s native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word published, paper translation dictionary. Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned off.

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 1.2.5, as designated in their IEP or Section 504 Plan.

Exhibit 1.2.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	If a student testing at a CBT school has an IEP indicating that he or she cannot enter their own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided orally or by gestures, into the paper booklet and then into the DEI or directly into the DEI. If a student testing at a PBT school has an IEP indicating Adult Transcription, an adult must transfer the student's responses exactly as provided orally or by gestures into the paper booklet.
Assistive Technology	This is the use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. PBT – Provide a Large Print test booklet. PBT – Large Print test booklet requires Adult Transcription into the DEI (see above for rules on Adult Transcription).
Paper Test Booklet	CBT – Student's IEP must indicate that student cannot enter his or her own responses on the computer and requires a paper-based test or Adult Transcription. The school will provide a Special Paper Version booklet for the student. The student's responses must be transcribed into the paper booklet, and then entered into the DEI, or entered directly into the DEI (see above for rules on Adult Transcription).

1.3 EVIDENCE BASED ON TEST CONTENT

Because the AzMERIT assessments are designed to measure student progress toward achievement of the Arizona College and Career Ready Standards (ACCRS), the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the academic standards.⁶

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards and refers to those standards in a highly iterative test development process that includes the ADE, test developers, and educator committees. Since spring 2016, the items used to develop operational test forms were drawn from custom Arizona item development and AIR's AIRCore pool of items. Both custom Arizona

⁶ Standard 12.4 – When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.

items and AIRCore items used in Arizona were developed to align with the ACCRS. These items were all reviewed by the Arizona Department of Education, Arizona content experts and educators, and Arizona community members prior to field testing in spring 2016 and subsequent operational test administration in spring 2017. Only items that were found to align well with the ACCRS were used. To supplement the AIRCore pool of items, a few previously-developed Arizona items that also aligned to the ACCRS were used. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning that the ADE and Arizona educator committees will act as reviewers throughout the item development cycle.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards will be covered in each test administration.⁷ Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet the blueprints. Developing test forms is difficult because test blueprints can be highly complex, specifying not only the range of items and points for each strand and standard but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. In addition to meeting complex blueprint requirements, test developers must work to meet psychometric goals so that alternate test forms measure equivalently across the range of ability.

Following a standard item-review process, item reviews proceeded through a series of internal reviews before items were eligible for external review by the ADE's staff and educator committees. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passed through four internal review steps before it was eligible for external review. Those steps include the following:

- Preliminary review, conducted by a group of AIR content-area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert.

At every stage of the item review process beginning with preliminary review, AIR's test developers analyze each item to ensure the following:

- The item is well-aligned with the intended content standard.
- The item conforms to the item specifications for the target being assessed.
- The item is based on a quality idea (i.e., it assesses something worthwhile in a reasonable way).
- The item is properly aligned to a depth of knowledge (DOK) level.
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward.

⁷ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

- Any accompanying graphic and stimulus materials are necessary to answer the question.
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as *no*, *not*, *none*, *never*, *unless absolutely necessary*), and ends with a question.
- For selected-response items, the set of response options are succinct; parallel in structure, grammar, length, and content; and sufficiently distinct from one another. All plausible, non-keyed response options are unambiguously incorrect.
- There is no obvious or subtle cluing within the item.
- The score points for constructed-response items are clearly defined.
- For machine-scored constructed-response (MSCR) items, item responses yield the intended score points based on the rubric.
- For human-scored constructed-response items, the scoring rubric clearly explains what characterizes responses at each possible level of achievement.

Based on review of each item, the test developer may accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to the ADE for review. At this stage, items may be further revised based on any edits or changes requested by the ADE, or they may be rejected outright. Items passing through the ADE's review must then pass through a stakeholder review in which a committee of educators reviews each item's accuracy, alignment to the intended standard and DOK level, and item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- aligned to the content standards;
- appropriate for the grade level;
- accurate;
- presented online in a way that is clear and appropriate; and
- free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics.

Items were also passed through to a parent/community sensitivity review committee to ensure that test content did not violate community standards. Items successfully passing through both the educator and parent/community review process were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is therefore an important step in constructing valid and equivalent operational test forms.

In addition, rubric-scored items, both machine-scored and human-scored, are validated following field test administration. Machine-scored items go through a rubric validation process wherein samples of student responses are reviewed, along with resulting scores, to ensure that rubrics are enacted as intended. This process is described in Section 10.1.1. Human-scored items go through a range-finding process prior to scoring where samples of item responses are used to create scorer training materials and ensure that the scoring rubric is appropriate, as described in Section 10.1.2.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass a three-stage review to be included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct, and there are no other obvious problems with the items.

ADE content and psychometric staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, the ADE determined that certain flagged items must be rejected or deemed the item eligible for inclusion in operational test administrations.

1.4 EVIDENCE FOR INTERPRETATION OF PERFORMANCE STANDARDS

Alignment of test content to the Arizona College and Career Ready Standards (ACCRS) ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the ACCRS. However, the interpretation of AzMERIT test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the Arizona standards. AzMERIT test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the ACCRS. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade-level expectations for achievement of the Arizona standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standards for the AzMERIT assessments are therefore central to the validity of test score interpretations.⁸

Following the first operational administration of the AzMERIT in spring 2015, a standard-setting workshop was conducted to recommend to the Arizona State Board of Education a set of performance standards for reporting student achievement of the ACCRS. Arizona educators, serving as standard-setting panelists, followed a standardized and rigorous procedure to recommend performance-level cut scores. The workshops employed the Bookmark procedure, a widely used method in which standard-setting panelists used their expert knowledge of the ACCRS and student achievement to map the performance-level descriptors adopted by Arizona onto an ordered item book comprising the spring 2015 operational test form and augmented with items administered in the embedded field test slots to minimize information gaps in the operational test form.⁹

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. For each assessment, panelists were provided the approximate location of performance standards for other important assessment systems. Panelists recommending performance standards for the high

⁸ Standard 4.22 – Test developers should specify the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.

⁹ Standard 1.18 – When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3–8 summative assessments were provided with the approximate location of relevant performance standards for the National Assessment of Educational Progress (NAEP) at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3–8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous performance standards for Arizona's Instrument to Measure Standards (AIMS). They were asked to consider the location of these benchmarks when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, the validity of test score interpretation is bolstered.

In addition, panelists were provided with feedback about the vertical articulation of their recommended performance standards so that they could view the relationship between the locations of recommended cut scores for each grade-level assessment to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and this further reinforced the interpretation of test scores as indicating not only achievement of current grade-level standards but also preparedness to benefit from instruction in the subsequent grade level.

Following recommendation of final performance standards, the recommended cut scores were presented to the Arizona State Board of Education for review and adoption. The Board adopted the recommended performance standards in August 2015.

Based on the adopted performance standards, Exhibit 1.4.1 shows the estimated percentage of students meeting the AzMERIT proficient standard for each assessment in spring 2015. Exhibit 1.4.1 also shows the approximate percentage of Arizona students expected to meet the ACT college-ready standards and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. It also presents the expected proficient rate for the Smarter Balanced assessments, system-wide, based on the spring 2014 field test administration. As indicated, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college-ready standards, and NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 1.4.1 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Test	Percentage of Students Meeting Standard			
	AzMERIT Proficient	Arizona ACT College-Ready	Arizona NAEP Proficient	Projected SBAC
<i>ELA</i>				
Grade 3	41%			38%
Grade 4	38%		28%	41%
Grade 5	30%			44%
Grade 6	34%			41%
Grade 7	33%			38%
Grade 8	32%		28%	41%
Grade 9	27%			

Grade 10	30%		
Grade 11	25%	34%	41%
Math			
Grade 3	42%		39%
Grade 4	42%	42%	38%
Grade 5	40%		33%
Grade 6	32%		33%
Grade 7	31%		33%
Grade 8	33%	32%	32%
Algebra I	32%		
Geometry	30%		
Algebra II	29%	36%	33%

Although AIR previously identified ACT college-ready cut scores on the AzMERIT ELA and math scales for the standard-setting committee's use in 2015, that study involved an indirect linkage. In that study, student performance on the grade 10 AIMS was used to predict subsequent student performance on the ACT tests, and then a linking study between the AIMS and AzMERIT allowed for the identification of the ACT cut scores on the AIMS scale to be represented onto the AzMERIT scale.

To examine directly the relationships between the AzMERIT and ACT assessments, the ADE obtained the ACT test scores for Arizona students graduating high school in spring 2016. More details of the direct linking study using AzMERIT and ACT data are presented in Section 9.5.2.

Exhibit 1.4.2 shows the location of the ACT college-ready cut scores for math and reading on the AzMERIT scale. The first column shows the location as identified via indirect linkage through AIMS and that was provided as benchmark information to AzMERIT standard-setting panelists. The second column shows the location of the ACT college-ready cut scores as identified via direct linkage between ACT and AzMERIT described here. The third column shows the location of the AzMERIT meets performance standards on the Algebra II and grade 11 ELA assessments. As indicated in the table, the location of the ACT college-ready cut scores on the AzMERIT scale were reasonably consistent across methods, especially for ELA. Importantly, the results affirm that the location of adopted AzMERIT performance standards are consistent with the ACT college-ready criteria.

Exhibit 1.4.2. Location of the ACT College-Ready Cut Scores on the AzMERIT Scales

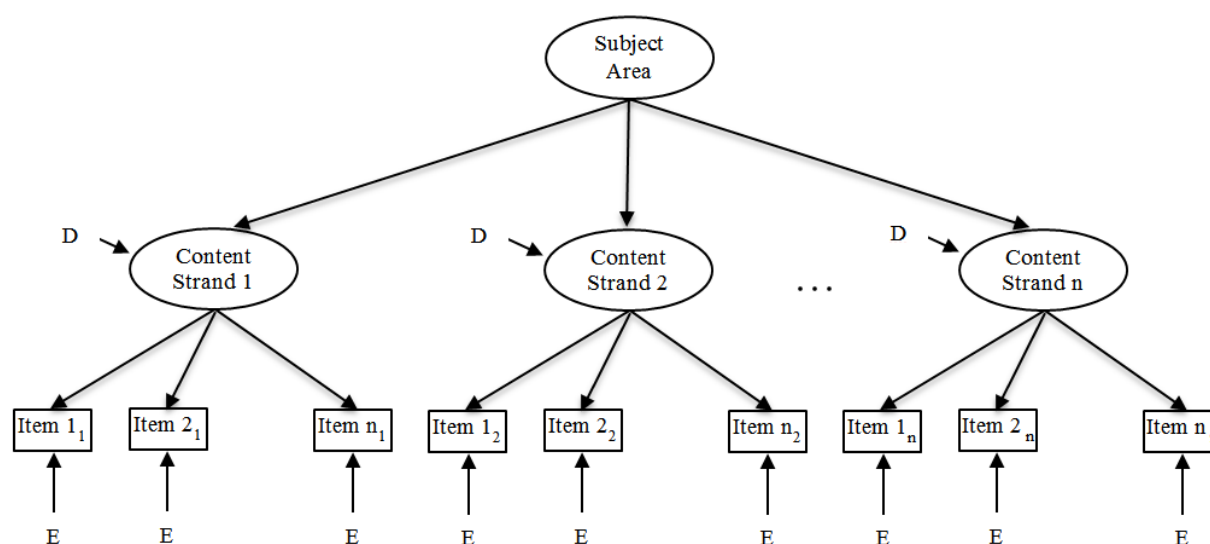
	Location of ACT College-Ready Cut on AzMERIT Scale		AzMERIT Meets Performance Standard
	Via Indirect Linkage through AIMS	Via Direct Linkage with AzMERIT	
Algebra II	3704	3727	3711
Grade 11 ELA	2579	2585	2585

The equipercentile equating method was used to verify the linkage between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT college-ready cut scores in reading was 2585 on the AzMERIT ELA scale. The location of the ACT college-ready cut score in math was 3727 for the AzMERIT math scale. Results from the equipercentile approach were thus consistent with the cut scores identified using regression models.

1.5 EVIDENCE BASED ON INTERNAL STRUCTURE

The AzMERIT assessment represents a structural model of student achievement in grade-level and course-specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., Reading Information, Reading Literature, Language, Writing). Content strands within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Exhibit 1.5.1. As the exhibit illustrates, each item is an indicator of an academic content strand. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content strand serves as an indicator of achievement in a subject area. As at the item level, the content strands include an error term indicating that the content strands are not pure indicators of overall achievement in the subject area. The paths from the content strands to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content strand construct. Similarly, the paths from subject-area achievement to the content strands represent the second-order factor loading, indicating the degree to which academic content strand constructs are correlated with the underlying construct of subject-area achievement.

Exhibit 1.5.1 Second-Order Structural Model for AzMERIT Assessments



Following the first operational test administration in spring 2015, confirmatory factor analysis was used to evaluate the fit of this structural model to student response data.¹⁰ For each of the test forms administered in spring 2015, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size, however; even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices, such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (Tucker & Lewis, 1973), and the Root Mean Square of Approximation (RMSEA) were also used to evaluate model fit.

¹⁰ Standard 1.13 – If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.

The AzMERIT assessments also claim to measure subject-area achievement using test items that probe student knowledge and skills across multiple depth of knowledge levels (DOK). As with the content standards, the classification of items by depth of knowledge also represents a structural model that can be evaluated using confirmatory factor analysis.¹¹ In this case, each item is an indicator of a DOK level first-order factor, and each depth of knowledge level is in turn an indicator of subject area achievement. Thus, confirmatory factor analysis was used to evaluate the fit of this depth of knowledge structural model to student response data from the spring 2015 AzMERIT test administration.

Exhibit 1.5.2 Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Index	Indication of Good Fit
CFI	$\geq .95$
TLI	$\geq .95$
RMSEA	$\leq .05$

In addition to testing the fit of the hypothesized AzMERIT second-order confirmatory factor analysis model, we examined the degree to which the second-order model improved fit over the more general one-factor model of academic achievement in each subject area. Because the one-factor general-achievement model was nested within the second-order model, a simple likelihood ratio test was used to determine whether the added information provided by the structure of the ACCRS frameworks improved model fit over a general-achievement model. Results indicating improved model fit for the second-order factor model provide support for the interpretation of content standard performance above that provided by the overall subject area score.¹²

1.5.1 ELA CONTENT MODEL

We began by evaluating the fit of the first-order, general-achievement model in which all items are indicators of a common subject-area factor. This model importantly evaluates the assumption of unidimensionality of the subject-area assessments, and it provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the first-order, general-achievement models in ELA are shown in Exhibit 1.5.1.1. All of the statistics indicate that the general-achievement model fits the data well. This pattern was true across all grades. The CFI and TLI values were all greater than 0.9 and generally equal to or greater than 0.95, and the RMSEA values were all below .05, indicating good fit for the base model.

Exhibit 1.5.1.1 Goodness-of-Fit for the AzMERIT ELA First-Order Model

Grade	CFI	TLI	RMSEA
3	0.93	0.93	0.05
4	0.95	0.95	0.03
5	0.97	0.96	0.04
6	0.96	0.95	0.04
7	0.97	0.97	0.04

¹¹ Standard 1.12 – If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

¹² Standard 1.14 – When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Grade	CFI	TLI	RMSEA
8	0.96	0.96	0.05
9	0.92	0.92	0.04
10	0.95	0.95	0.04
11	0.93	0.93	0.03

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.1.2. All of the statistics indicate that the second-order models posited by the AzMERIT assessments fit the data well. This pattern was true across all grades. As with the general factor model, the CFI and TLI values for the second-order models were all above or near .95, with RMSEA values well below the .05 threshold used to indicate good fit.

The results of the comparison between the hypothesized AzMERIT model and the general-achievement model are presented in Exhibit 1.5.1.3. We note that model fit for the first-order, general-achievement model was also very high and provides evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether the posited second-order reporting model adds information beyond that provided by the first-order model. The chi-square difference test shows that, across grade levels, the strand-based second-order model showed significantly better fit than the first-order, general-achievement model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.5.1.2 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.96	0.96	0.04
4	0.97	0.97	0.03
5	0.98	0.98	0.03
6	0.97	0.97	0.03
7	0.98	0.98	0.03
8	0.98	0.98	0.04
9	0.96	0.96	0.03
10	0.97	0.97	0.03
11	0.95	0.95	0.03

Exhibit 1.5.1.3 Difference in Fit Between Content Derived Second-Order and First-Order, General-Achievement Model

Grade	χ^2	<i>df</i>	<i>p</i> value
3	13560.70	3	<i>p</i> < .001
4	8460.90	3	<i>p</i> < .001
5	10944.70	3	<i>p</i> < .001
6	12019.80	3	<i>p</i> < .001
7	8848.60	3	<i>p</i> < .001
8	15590.10	3	<i>p</i> < .001
9	8896.60	3	<i>p</i> < .001
10	9084.70	3	<i>p</i> < .001
11	4412.80	3	<i>p</i> < .001

1.5.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.2.1. Across all grades, results indicate that the second-order models posited by the AzMERIT assessments fit the data well. The CFI and TLI values were all .97 to .99; RMSEA values were all .03 or lower.

Exhibit 1.5.2.1 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.99	0.99	0.02
6	0.98	0.98	0.03
7	0.99	0.99	0.02
8	0.99	0.99	0.02
9	0.98	0.98	0.02
10	0.98	0.97	0.02
11	0.98	0.98	0.02

The results of the comparison between the hypothesized AzMERIT model and the general-achievement model are presented in Exhibit 1.5.2.2. The chi-square difference test shows that, across grade levels, the DOK-based second-order model showed significantly better fit than the first-order, general-achievement model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.5.2.2 Difference in Fit Between DOK Derived Second-Order and First-Order, General-Achievement Model

Grade	χ^2	<i>df</i>	<i>p</i> value
3	21402.60	4	<i>p</i> < .001
4	12053.60	4	<i>p</i> < .001
5	17102.90	4	<i>p</i> < .001
6	18192.10	4	<i>p</i> < .001
7	16351.40	4	<i>p</i> < .001
8	25454.70	4	<i>p</i> < .001
9	14989.30	4	<i>p</i> < .001
10	14920.90	4	<i>p</i> < .001
11	8075.10	4	<i>p</i> < .001

1.5.3 MATH CONTENT MODEL

As with ELA, structural analyses of the math assessments began with an evaluation of fit for the first-order, general-achievement model in which all items are indicators of a common math subject-area factor. This model provides for an evaluation of the unidimensionality assumption of the subject-area assessments, and it provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the general-achievement models in math are shown in Exhibit 1.5.3.1. All of the statistics indicate that the general-achievement model fits the data well. This pattern was true across all grades. The CFI and TLI values were all equal to or greater than .95, and the RMSEA values are all below .05, indicating good fit for the base model.

Exhibit 1.5.3.1 Goodness-of-Fit for the AzMERIT Math First-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.97	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03

Grade	CFI	TLI	RMSEA
Algebra I	0.98	0.98	0.02
Algebra II	0.97	0.97	0.02
Geometry	0.99	0.99	0.02

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 1.5.3.2. The models show very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates well below their .05 cut-off values. All of the statistics indicate that the second-order models are a good fit for the data.

Exhibit 1.5.3.2 Goodness-of-Fit for the AzMERIT Math Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.02
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.98	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.97	0.97	0.02
Geometry	0.99	0.99	0.02

The results of the comparison between the second-order, strand-based model and the first-order, general-achievement model are presented in Exhibit 1.5.3.3. Again, model fit for the first-order, general-achievement model is very high, providing evidence for the unidimensionality of the subject-area assessments. The purpose of these analyses is to determine whether knowledge of the depth of knowledge level of items provides information beyond that provided by the more general model. The chi-square difference test shows that, across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.3.3 Difference in Fit Between Content Derived Second-Order and First-Order, General-Achievement Model

Grade	χ^2	<i>df</i>	<i>p</i> value
3	3225.00	3	<i>p</i> < .001
4	1326.30	3	<i>p</i> < .001
5	1427.00	3	<i>p</i> < .001
6	1036.20	4	<i>p</i> < .001
7	559.80	4	<i>p</i> < .001
8	1039.30	4	<i>p</i> < .001
Algebra I	750.90	3	<i>p</i> < .001
Algebra II	246.50	3	<i>p</i> < .001
Geometry	269.70	4	<i>p</i> < .001

1.5.4 MATH DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 1.5.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates well below their .05 cut-off values. All of the statistics indicate that the second-order models are a good fit for the data.

Exhibit 1.5.4.1 Goodness-of-Fit for the AzMERIT Math Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.99	0.99	0.02
Geometry	0.97	0.97	0.02

The results of the comparison between the second-order, DOK-based model and the first-order, general-achievement model are presented in Exhibit 1.5.4.2. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.4.2 Difference in Fit Between DOK Derived Second-Order and First-Order, General-Achievement Model

Grade	χ^2	df	p value
3	331.40	3	$p < .001$
4	309.50	3	$p < .001$
5	14.90	3	$p < .001$
6	14.50	3	$p < .001$
7	236.60	3	$p < .001$
8	79.20	3	$p < .001$
Algebra I	20.10	3	$p < .001$
Algebra II	26.40	3	$p < .001$
Geometry	20.90	3	$p < .001$

1.6 EVIDENCE FOR RELATIONSHIPS WITH CONCEPTUALLY RELATED CONSTRUCTS

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity. Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct-irrelevant attributes.¹³

Observed correlations between alternate indicators of student achievement of course objectives, such as locally administered assessments of student achievement and AzMERIT, should be limited only by the unreliability of the

¹³ Standard 1.16 – When validity evidence includes empirical analyses of responses to test items together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

measures. When both assessments measure student achievement in common subject areas, such as with locally administered and statewide assessments of math achievement, we expect test scores between the common subject-area assessments to be substantially correlated. In addition, we expect that the magnitude of observed correlations between test scores in different subject areas will be lower than correlations between test scores in a common subject area. Because the content domains assessed in ELA and math tests are quite different, AzMERIT ELA test scores should correlate less well with locally administered assessments of math than ELA. It is important to note, however, that test scores across subject areas and test systems are nevertheless expected to be highly correlated. This is because even though subject-area test scores measure different academic content domains, student achievement across subject areas is influenced by factors both internal (e.g., general intelligence) and external (e.g., socioeconomic status) to the student that contribute to student achievement across all academic subject areas so that student test scores across subject areas tend to be highly intercorrelated. So, while we certainly do expect correlations between test scores across subject areas to be lower than correlations between test scores within a subject area, we nevertheless expect test scores across subject areas to be quite high.

Exhibit 1.6.1 shows the correlations between student test scores on the spring 2015 statewide AzMERIT assessment with corresponding test scores on a district-wide administration of the Northwest Evaluation Association (NWEA) assessment. Sample sizes range from more than 1,400 students taking the grade 3 assessments to nearly 1,100 students taking the middle school assessments, so the observed correlations are expected to be stable. Convergent correlations are quite high, ranging from 0.82 to 0.84 between AzMERIT ELA (assessing reading, writing, and listening) and NWEA reading. Correlations between AzMERIT and NWEA math scores are even higher, ranging from 0.85 to 0.89.

Exhibit 1.6.1 Correlations Between AzMERIT and Locally Administered NWEA Test Scores

Grade	ELA Sample Size	ELA Correlation	Math Sample Size	Math Correlation
3	1426	0.82	1429	0.86
4	1214	0.84	1214	0.88
5	1303	0.84	1303	0.88
6	1119	0.82	1115	0.85
7	1081	0.82	1082	0.89
8	1090	0.82	1091	0.89

Exhibit 1.6.2 shows the discriminant correlations between AzMERIT and the locally administered NWEA assessment. As expected, correlations across subject-area assessments remain quite high, indicating considerable consistency in student achievement across subject-area assessments. Nevertheless, correlations across subject-area assessments are systematically lower than within subject correlations, indicating that the subject-area assessments are measuring domain-specific knowledge and skills in addition to common factors underlying student achievement.

Exhibit 1.6.2 Discriminant Correlations Between AzMERIT and Locally Administered NWEA Test Scores

Grade	ELA Sample Size	ELA Correlation	Math Sample Size	Math Correlation
3	1426	0.72	1428	0.70
4	1211	0.76	1217	0.72
5	1303	0.75	1303	0.72
6	1117	0.73	1117	0.71
7	1081	0.77	1080	0.74
8	1088	0.75	1093	0.71

Convergent correlations between AzMERIT and locally-administered assessments were also reported by Estrada and colleagues (Estrada, Burnham, Feld, Bergan, and Bergan, 2015). These researchers reported the mean correlations between a variety of local assessments and AzMERIT test scores for ELA and math assessments in grades 3–8. Mean correlations between AzMERIT and various local assessments of ELA ranged from .77 to .79 across the grade levels investigated. Mean correlations between AzMERIT and local assessments of mathematics ranged from .71 to .75 across grades 3 through 8. These results likewise show good convergence between AzMERIT and other locally-administered assessments purporting to measure the same constructs.

1.7 MEASUREMENT INVARIANCE ACROSS SUBGROUPS

Measurement invariance occurs when the likelihood of responding correctly conforms to the measurement model and is independent of group membership and the parameters of a measurement model are statistically equivalent across groups.¹⁴ The parameters of interest in measurement invariance testing are the factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that it is not necessary to demonstrate invariance across groups on these parameters. In general, measurement invariance testing can be conducted using a series of multiple-group confirmatory factor analysis (CFA) models, which impose identical parameters across groups. The measurement model parameters, including factor patterns (configural invariance), factor loadings (metric or weak invariance), latent intercepts/thresholds (scalar or strong invariance), and unique or residual factor variances (strict invariance), are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups.

Appendix C shows the results of measurement invariance testing by subgroups for ELA and Math. Items comprising the spring 2016 operational test administration were used to investigate measurement invariance across subgroups. The full set of tables associated with these analyses is provided for each of the grade-level and subject-area assessments. The series “a” tables (e.g., tables B.1a, B.2a, etc.) present the global model fit indices for the measurement invariance tests for each assessment. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using χ^2 difference test (at $\alpha \leq 0.05$) and the examination of significant differences of the Root Mean Square Error of Approximation (RMSEA, change in RMSEA ≤ 0.015 ; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across the following subgroups: gender (Model A), ethnicity including African American vs. White (Model B-1), Hispanic vs. White (Model B-2), Asian vs. White (Model B-3), American Indian vs. White (Model B-4), and Multi-Ethnic vs. White (Model B-5), special education program status (SPED; Model C), economic disadvantage status (Low Income; Model D), limited English proficiency status (LEP; Model E), and accommodated test forms (Accommodation, Model F). Invariance tests of subgroups were investigated separately for each grade and subject area test. Since in each ELA assessment, students were randomly assigned to one of six writing prompts for administration, the missing responses on the writing items resulted in unsuccessful model convergence. Thus, to achieve model convergence, we included the students who took a common writing prompt between online and paper-based in each ELA assessment.

¹⁴ Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

The null hypothesis of the χ^2 difference test is that the more restricted invariance model (e.g., metric) fits the data equally as well as the less restricted invariance model (e.g., configural). Given that the sensitivity of the χ^2 difference tests to sample size, we additionally examined significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retention of the more restricted invariance model (Chen, 2007).

The ethnicity B tables (e.g., tables B.1b, B.2b, etc.) show the model fit indices of scalar invariance models assuming the same factor pattern + identical factor loadings + identical latent intercept/threshold across subgroups. Global model fit indices included the Comparative Fit Index (CFI; Bentler, 1990) and Root Mean Square Error of Approximation (RMSEA). CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.870 to 0.990, and RMSEA ranged from 0.012 to 0.044. For Math, CFI values ranged from 0.905 to 0.990, and RMSEA ranged from 0.010 to 0.058.

Although the χ^2 difference test should ideally be nonsignificant, all χ^2 difference tests were significant at $\alpha = .05$ due to large sample sizes except Model B-5, where the χ^2 difference tests for most grades was nonsignificant or marginally significant at $\alpha = .05$. In spite of significant χ^2 difference tests for most models, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0.000 to 0.002 for both ELA and MATH). Based on the similar magnitudes of the RMSEA (i.e., no material change across all tested models; Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, ELA and MATH test scores have the same measurement structure across gender, ethnicity (African American vs. White, Hispanic vs. White, Asian vs. White, American Indian vs. White, and Multi-Ethnic vs. White), special education program status, economic disadvantaged status, limited English proficiency status, and accommodation test forms.

1.8 DIFFERENTIAL MODE EFFECTS ACROSS SUBGROUPS

To explore the possibility that mode of test administration may exert differential effects across subgroups, we began by identifying matched samples of students participating online and on paper. For students administered paper-based assessments, observed test scores were regressed on prior achievement and demographic variables to obtain regression weights. The resulting prediction equation was then applied to all students to yield predicted paper-based test scores. The predicted paper-based scores were used to identify matched samples of online and paper-based test takers.

To identify possible differential effects of mode across subgroups, we used the observed test score as dependent variable and then covaried the predicted test score to isolate the effects of mode. The demographic variables of interest include gender, English language learner status, special education status (SPED), free reduced lunch status (FRL), migrant status, and six ethnicity subgroups as predictors. We created dummy coded variables to represent those non-white ethnicities with 0 as no and 1 as yes. In addition, gender was coded as 0 for male and 1 for female. ELL was coded as 1 for students as English language learner and 0 for non-ELL. SPED was coded as 1 for students in a special education program and 0 for students without attending any SPED program. FRL (or Social Economic Status; SES) was coded as 1 for students having free reduced lunch and 0 as non-FRL students. Migrant was coded as 1 for students from a migrant family and 0 for non-migrant students. Significant interactions between mode of test administration and the demographic subgroup comparisons indicate differential mode effects between the specified demographic subgroups.

While many effects achieve conventional levels of statistical significance, because of the very large sample sizes, the effect sizes were quite small. Thus, Exhibit 1.8.1 shows the regression coefficient estimates for the differential mode effects by subgroup interaction only for effects where $p < .0001$.

Results indicated that mode effects were more pronounced for special education students relative to general education population. Especially for the high school EOC tests, AzMERIT tests were more difficult for special education students when administered on paper-based than when administered online.

Mode effects were more pronounced for low income students with respect to the math assessments. Math tests were generally more difficult for low income students when administered online than on paper-based.

Mode effects were also more pronounced for LEP students than for the general education population in math but not in ELA. However, the direction of this effect was not consistent across grades. Online math tests were more difficult than paper-based for LEP students in the lower grades, while paper-based math tests were more difficult than online tests for LEP students in the higher grades.

Exhibit 1.8.1 Parameter Estimates for Differential Mode Effects by Subgroups Interactions

Test	Gender	White	Black	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	Special Education	Limited English Proficiency	Free/Reduced Lunch	Migrant
ELA											
Grade 3E	0.49									0.27	
Grade 4E											
Grade 5E											
Grade 6E								-0.61			
Grade 7E								0.5			
Grade 8E					1.66	-0.34					
Grade 9E	0.45							-0.74			
Grade 10E								-1.23		-0.41	
Grade 11E	-0.33					0.36		-0.58			
MATH											
Grade 3M								0.57			
Grade 4M									0.52	-	-4.46
Grade 5M							-0.89			0.34	
Grade 6M		1.15	0.96				0.69		0.6	-0.31	
Grade 7M	-0.26									0.25	-2.87
Grade 8M		0.89					0.86		-0.58		
Algebra I						0.73		-0.8	-0.95	0.50	
Geometry						-0.44		-1.32		1.11	
Algebra II							-1.07	-0.75		0.63	

Note: Positive coefficient means that the online test is more difficult for the focal group.

1.9 EVIDENCE FOR STUDENT GROWTH – OVERALL AND BY SUBGROUPS

The AzMERIT assessments report student test scores on a vertical scale, allowing families and teachers to make inferences about student growth across school years. The validity of test score interpretations about student growth over time depends strongly on the vertical linking design used to develop the vertical scale. But even when test score interpretations are appropriate to the scaling design, it is important to examine whether student gains may be interpreted consistently across subgroups or whether differential gain rates across subgroups limit the inferences that can be made about test score gains over time.¹⁵ To address this issue, we examined rates of student growth across student gender, race/ethnicity, students with disabilities (SPED), English language learners (LEP), and low income status (Low Income).

Exhibit 1.9.1 shows the mean test scores on the spring 2016 and spring 2017 administrations of AzMERIT for students participating in both test administrations, as well as the correlation between test scores across the two assessment occasions. Correlations between test scores are quite high and indicate substantial consistency in rank ordering of student achievement between the two test administrations. The correlation between student achievement in grade 8 math and Algebra I is attenuated somewhat, and further that the distribution of student ability is somewhat less variable for this cohort, especially with respect to the spring 2017 Algebra I performance. We note that in spring 2016, grade 8 students enrolled in Algebra I were required to participate in both assessments, but in spring 2017, those high-achieving students would likely have participated in the Geometry assessment and would not have been included in these analyses. The resulting restriction of range could be responsible for the attenuated correlation.

¹⁵ Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Standard 3.17 – When aggregate scores are publicly reported for relevant subgroups— for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults— test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Exhibit 1.9.1 Test Score Stability and Performance Gains Overall

Assessment Subject_2016_2017	N	Spring 2016 Scale Score		Spring 2017 Scale Score		Change from 2016 to 2017		Percentage Scoring Lower		Correlation
		Mean	Std Dev	Mean	Std Dev	Mean	IRT based Standard Error	Expected	Observed	
ELA										
ELA_G3E–G4E	81788	2502	31.64	2522	31.43	20	14.51	0.19	0.13	0.82
ELA_G4E–G5E	80889	2518	33.73	2537	33.67	18	15.05	0.22	0.16	0.84
ELA_G5E–G6E	79850	2537	34.21	2544	32.04	7	14.99	0.38	0.35	0.83
ELA_G6E–G7E	78709	2542	33.62	2554	32.60	12	14.84	0.30	0.24	0.84
ELA_G7E–G8E	78367	2552	30.93	2555	33.89	3	14.33	0.45	0.43	0.84
ELA_G8E–G9E	68998	2558	31.55	2568	29.15	10	13.86	0.32	0.27	0.82
ELA_G9E–G10E	64016	2568	30.47	2566	29.27	-2	14.04	0.53	0.53	0.82
ELA_G10E–G11E	57207	2568	28.47	2567	29.69	-1	14.24	0.51	0.50	0.80
MATH										
Math_G3M–G4M	82368	3525	44.49	3556	44.71	31	17.37	0.16	0.11	0.82
Math_G4M–G5M	81262	3553	40.31	3589	44.37	36	16.39	0.11	0.07	0.83
Math_G5M–G6M	80191	3589	41.29	3619	43.97	30	16.27	0.15	0.10	0.83
Math_G6M–G7M	76781	3615	41.46	3633	45.49	18	17.80	0.26	0.21	0.85
Math_G7M–G8M	65633	3627	32.37	3652	38.57	25	15.73	0.17	0.12	0.81
Math_G8M–Algebra I	49983	3652	32.45	3666	31.38	14	15.28	0.29	0.24	0.76
Algebra I–Geometry	56116	3674	34.02	3685	37.08	11	15.77	0.34	0.30	0.81
Geometry–Algebra II	48030	3690	35.28	3701	33.59	11	16.23	0.34	0.31	0.78

The exhibit also shows that rate of achievement gain is somewhat higher for math than ELA, and that while gain rates decelerate across the school years, the rate of gains diminishes more rapidly for ELA than math over time. For math, large gains, typically 3/4 standard deviation (e.g., average gain of 31 scale score points in grade 3 math is 70% of the 44-point standard deviation of student test scores), are observed through the middle school grades, dropping to about 1/3 standard deviation between administrations of the high school end-of-course assessments. For ELA, while elementary school gains are strong, by middle school, annual gains are between 1/3 to 1/2 standard deviation, and by high school, drop to about 1/4 standard deviation, with no growth observed from grade 9 to 10 and from grade 10 to 11.

To evaluate differential growth across demographic subgroups, a series of regression analyses were conducted to predict 2017 test scores from 2016 test scores, controlling for demographic subgroup membership. To compare ethnic subgroup performance, we created six dummy variables contrasting white students with each of the other ethnic groups (e.g., white/Hispanic, white/African American). Gender was coded 1 for female. SPED, LEP, and Low Income students were coded as 1 to contrast with students who were not identified with those needs who were coded as 0.

Exhibit 1.9.2 shows the standardized regression coefficient estimates of the differential effect on student's growth rate from 2016 to 2017 administration across subgroups. Although many individual effects attained conventional levels of statistical significance due to large sample sizes, we focus here only on highly significant effects ($p < 0.01$)

that are associated with more practically significant effect sizes and that may point to trends across grade-level and/or subject-area assessments. Appendix D shows the regression model parameter estimates of differential growth for the ELA and math assessments, including standardized and unstandardized coefficient, standard error of the unstandardized coefficient, and p value regardless of significance level.

The 2016 test scores were centered on the reference group mean so that the intercept values at the top of the table represent the mean performance of white males on the 2017 assessment, with group parameters reflecting differences from the reference group on the spring 2017 assessment. Results indicate that females generally performed better than males for both ELA and math across grades. With respect to ethnicity, Asian students generally performed better than white students in both ELA and math. For all other ethnic group comparisons, the focal groups generally performed less well than whites. Special education students, limited English proficient students, and low-income students all performed less well than the general education population in both ELA and math.

The slope represents the association between 2016 and 2017 test scores, controlling for demographic subgroups. The overall slope parameter indicates the rate of growth in test scores between 2016 and 2017. The group-specific slope parameters indicate differential growth rate between contrasted groups.

While females tended to score higher across assessments, differential gain rates by gender were small and inconsistent. Special education students generally showed lower rates of gain than general education students, although pattern was reversed during elementary school math assessments, with special education students showing greater rates of gain. Limited English proficient students showed lower rates of gain in both ELA and math, but this effect seems to moderate in the high school grades, where differential gain rates were much less pronounced. Differential gain rates for low income students were observed for both ELA and math, generally showing lower gain rates.

With respect to ethnicity, differential gain rates were small and inconsistent in the elementary and middle school grade assessments. Compared to whites, Asian students did, however, show higher gain rates during middle school grade assessments in math and lower gain rates during elementary school grade assessments in ELA. And African American and Hispanic students showed lower gain rates than whites in math assessments.

Exhibit 1.9.2.1 Standardized Regression Coefficient of Differential Growth from 2016 to 2017 Administration Across Subgroups–ELA

2016 Administration 2017 Administration	G3E G4E	G4E G5E	G5E G6E	G6E G7E	G7E G8E	G8E G9E	G9E G10E	G10E G11E
Intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Female	0.02	0.01		0.03	0.04	0.04	-0.01	0.06
SPED	-0.07	-0.08	-0.07	-0.08	-0.08	-0.06	-0.08	-0.07
LEP	-0.08	-0.08	-0.06	-0.05	-0.05	-0.05	-0.05	-0.04
Low Income	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01
Asian	0.02	0.01	0.02	0.01	0.03	0.04	0.02	0.01
Hispanic	-0.06	-0.04	-0.05	-0.02	-0.02	-0.04	-0.05	-0.03
African American	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Hawaiian/Pacific								
American Indian	-0.05	-0.04	-0.03	-0.03	-0.02	-0.02	-0.04	-0.01

Multiple Ethnicities	-0.01	-0.01						
Slope		0.80	0.79	0.79	0.82	0.80	0.78	0.80
Female			-0.03	-0.02			-0.01	-0.03
SPED		-0.01		0.01	-0.04	-0.02	-0.02	
LEP	-0.05	-0.05	-0.03		-0.03	-0.03	-0.03	-0.02
Low Income	-0.02	-0.01	-0.01		-0.01	-0.01		-0.02
Asian	-0.01	-0.01		-0.01				
Hispanic		-0.01		0.01		-0.02		-0.01
African American		-0.01		0.01				
Hawaiian/Pacific								
American Indian		-0.01			-0.01		-0.01	
Multiple Ethnicities								

Note: Only the significant effects from the multiple regression models are presented in the table. Intercept (β_0): Standardized average test score in 2017 administration. Slope (β_{10}): Rate of gain from 2016 to 2017. For the effect of special groups, the coefficient represents the difference compared to their contrast group; SPED=Special Education Status vs. Non-SPED. LEP=Limited English Proficiency vs. Non-LEP, Low Income=Low income vs. Non-Low Income. For the effect of ethnic groups, the coefficient represents differential growth rate compared to White.

**Exhibit 1.9.2.2 Standardized Regression Coefficient of Differential Growth from 2016 to 2017 Administration
Across Subgroups-Math**

2016 Administration 2017 Administration	G3M G4M	G4M G5M	G5M G6M	G6M G7M	G7M G8M	G8M AlgI	AlgI Geo	Geo AlgII
Intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Female	-0.02		-0.01			0.03	-0.02	0.00
SPED	-0.07	-0.06	-0.07	-0.08	-0.07	-0.07	-0.05	-0.05
LEP	-0.06	-0.04	-0.05	-0.05	-0.05	-0.06	-0.04	-0.03
Low Income	-0.03	-0.03	-0.03	-0.02	-0.01	-0.02	-0.02	-0.05
Asian	0.02	0.03	0.01	0.02	0.02	0.02		0.05
Hispanic	-0.05	-0.02	-0.06	-0.04	-0.03	-0.04	-0.06	-0.04
African American	-0.03	-0.02	-0.04	-0.03	-0.01	-0.01	-0.04	-0.01
Hawaiian/Pacific	-0.01							
American Indian	-0.04	-0.03	-0.05	-0.04	-0.02	-0.03	-0.03	-0.03
Multiple Ethnicities			-0.01		-0.01		-0.01	
Slope	0.76	0.79	0.78	0.86	0.83	0.78	0.80	0.81
Female		-0.01	0.01		-0.02		-0.01	-0.03
SPED	0.03	0.03	0.02	-0.07	-0.06	-0.04		-0.03
LEP	-0.01	-0.02	-0.02	-0.05	-0.04	-0.05	-0.03	-0.03
Low Income	-0.02		-0.02	-0.02	-0.01	-0.01	-0.01	-0.03
Asian			0.01		0.01	0.01	0.01	0.01
Hispanic		-0.01	-0.01	-0.02		-0.02	-0.02	-0.03
African American				-0.01	-0.01	-0.01		-0.01
Hawaiian/Pacific								
American Indian	0.01			-0.02	-0.01	-0.01	-0.01	-0.01
Multiple Ethnicities								

Note: Only the significant effects from the multiple regression models are presented in the table. Intercept (β_{00}): Standardized average test score in 2017 administration. Slope (β_{10}): Rate of gain from 2016 to 2017. For the effect of special groups, the coefficient represents the difference compared to their contrast group; SPED=Special Education Status vs. Non-SPED. LEP=Limited English Proficiency vs. Non-LEP, Low Income=Low income vs. Non-Low Income. For the effect of ethnic groups, the coefficient represents differential growth rate compared to White.

1.10 DAY, WEEK, AND TIME OF DAY EFFECTS ON PERFORMANCE

Administration of the new AzMERIT online tests is untimed, and schools may flexibly schedule students to take the tests in computer labs throughout the testing window. Thus, students taking the same grade level or end-of-course (EOC) test are not required to test on the same day. Because the days and times on which tests can be administered is variable, the possibility arises that performance factors associated with time of day or day of week may influence student test scores.

A series of regression models were developed to predict student performance using the day of the week and time of the day variables, as well as the duration of the test administration from test start to test end. The dependent variable for these analyses was the spring 2016 AzMERIT scale score. To control for student achievement, we first covaried previous achievement using spring 2015 AzMERIT test scores. Because of the need to covary previous achievement, the analyses were limited to students participating in the grade 4 to 8 and high school EOC assessments in mathematics and ELA tests, and for whom 2015 test scores were available. The day of the week was coded as 1 to 5 (1 for Monday, 2 for Tuesday, and so on). For the regression analyses, the time of day and the duration were

continuous variables using the actual time. Time of day effects were further evaluated using paired comparisons between early morning, late morning, early afternoon, and late afternoon.

Exhibit 1.10.1 shows the standardized regression coefficient estimates of the time effect on student's performance only for effects where $p < .05$. Generally, results indicate that starting tests earlier in the week resulted in higher test performance. Tests started on Friday were consistently associated with impaired performance. There were some exceptions to this. For example, students beginning the grade 7 ELA tests on Monday scored lower than students beginning on any other day than Friday. But generally, the pattern was pronounced.

Conversely, assessments which were completed earlier in the week were associated with lower test scores. Tests ending on any day other than Monday were associated with higher test scores. And this effect was generally true for tests ending on Tuesday. That said, students appeared to perform better on tests ending Wednesday or Thursday than on Friday, although there were exceptions to this, as well (e.g., grade 9 and 10 ELA where Friday end dates were associated with greater performance).

Time of day effects were less consistent. For ELA, morning start times were associated with greater performance than afternoon start times for high school students. For middle school students, later morning start times were associated with poorer performance than early morning or late afternoon start times. And in grade 6, ELA tests with morning start times were associated with lower performance than tests with afternoon start times.

Exhibit 1.10.1. Standardized Regression Coefficients of Time Effect on Student's Performance

Test	Start Day	End Day	Start Time	End Time	Duration
ELA					
Grade 4 ELA		0.02	-0.01	0.03	-0.01
Grade 5 ELA	-0.01	0.01	-0.01	0.02	
Grade 6 ELA	0.02		0.01		
Grade 7 ELA	0.01	0.03	-0.01	-0.01	0.01
Grade 8 ELA		0.02	-0.01		0.02
Grade 9 ELA		0.01	-0.06	0.02	0.01
Grade 10 ELA	-0.02		-0.08	0.03	0.01
Grade 11 ELA	-0.03		-0.08	0.05	0.01
MATH					
Test	Start Day	End Day	Start Time	End Time	Duration
Grade 4 MATH	-0.01	0.02	-0.02		
Grade 5 MATH	-0.02	0.01	-0.03	0.04	0.01
Grade 6 MATH	-0.03	0.01		0.03	0.01
Grade 7 MATH	-0.01	0.01	-0.04	0.06	
Grade 8 MATH		0.01	-0.01	0.04	
Algebra I	-0.05	0.01	-0.12	0.08	0.04
Geometry		0.03	-0.11	0.10	0.03
Algebra II	-0.04	0.04	-0.13	0.12	0.05

Note: Standardized regression coefficient 0.01 is equivalent to 3 or 4 scale score difference.

For math tests, later start times were generally associated with better performance. An exception to this pattern was observed for Algebra I, where students beginning testing in the late morning performed better than students starting at any other time.

Tests ending early in the afternoon were generally associated with higher performance than tests ending earlier in the day, although grade 6 ELA proved an exception with tests ending early morning associated with the highest scores.

In addition, longer test administrations were associated with higher performance.

1.11 ARIZONA GLOSSARY STUDY

Construct-irrelevant barriers to accessing test content limit the validity of test score interpretations. Where use of vocabulary that is not relevant to the measured construct interferes with student ability to understand the test item, the item is not accurately assessing the intended construct. To evaluate the validity of testing accommodations such as glossaries, we expect that reducing a barrier to access will improve student performance for the disadvantaged group while having no effect on the general education population. If we see, however, a main effect of the accommodation for all groups, the accommodation is likely modifying the measurement construct.

In a previous study, students administered the grade 3 and grade 7 assessments were randomly assigned to either a glossary or no glossary condition. A sample of field-test items were glossed and if a student in the glossary condition was administered a glossed item, an introductory screen was displayed to alert students to the availability and use of the glossed items.

Results of this initial study were mixed. At grade 3, a main effect for the glossary condition indicated that providing a glossary generally impaired student performance on the ELA assessment. A significant interaction effect for math indicated that providing a glossary impaired performance of EL students.

At grade 7, the interaction effects were significant for both assessments, but the direction of the effects differed. Significant EL by condition interactions indicated that English language learners performed better on the ELA test when provided a glossary, but providing a glossary on the math items resulted in poorer performance for EL students on the math test.

Results from the initial study were limited both by the grade levels assessed and the relatively small number of items included in the study.

AIR and the ADE extended the glossary study for the spring 2017 administration. As with the previous study, the purpose of this investigation was to examine the effectiveness and validity of computer-based pop-up glossary accommodations for English language learners (ELs). The study consisted of two parts. The first part focused on establishing a method for identifying the words, terms, and expressions in items that should be glossed. The general criterion is that glossaries should be provided for terms that are easily understood by native speakers but not by English Learners, and that are not part of the standard being measured. When provided with this general criterion, raters show a very low level of agreement in their determination of terms that should receive a glossary entry. AIR developed detailed guidelines, which include glossing culturally-bound language, tagging only when understanding meaning is necessary to answer the question, more structured tagging process, and so on. The new guidelines resulted in higher levels of agreement among raters (the agreement for triplets of raters is 0.59; Kappa for triplets of raters is 0.73).

The second part of the study focused on the effectiveness and validity of glossaries. Glossary entries, if effective and valid, should increase the performance on items with glossaries for English Learners, but have no effect on the

performance of native speakers. In a randomized control trial, the pop-up glossaries were administered to students taking the Arizona spring 2017 English language arts and math state assessments. Approximately 60,000 students in each grade participated in the study. EL students range from about 1,000 to 8,000 per grade, with more in the lower grades. The participants were randomly assigned into three conditions: English glossary only; English glossary and Spanish translation; and no glossary. Exhibit 1.11.2 summarizes the number of students selected for the study by grade, subject, EL status, and experimental condition.

Exhibit 1.11.1 Number of Students Selected for the Glossary Study by Grade, Subject, EL Status and Experimental Condition

Grade	Glossary	ELA			Math		
		non-EL	EL	Total	non-EL	EL	Total
3	ENG Only	19385	2535	21920	19442	2569	22011
	ENG+SP	19780	2449	22229	19874	2481	22355
	No Gloss	19616	2532	22148	19678	2563	22241
	Total	58781	7516	66297	58994	7613	66607
4	ENG Only	19800	2425	22225	19897	2450	22347
	ENG+SP	20014	2520	22534	20121	2545	22666
	No Gloss	20140	2350	22490	20249	2375	22624
	Total	59954	7295	67249	60267	7370	67637
5	ENG Only	19802	1924	21726	19898	1935	21833
	ENG+SP	20182	1928	22110	20235	1941	22176
	No Gloss	20046	1906	21952	20133	1920	22053
	Total	60030	5758	65788	60266	5796	66062
6	ENG Only	19682	1380	21062	19716	1397	21113
	ENG+SP	20016	1343	21359	20083	1361	21444
	No Gloss	19906	1393	21299	19939	1410	21349
	Total	59604	4116	63720	59738	4168	63906
7	ENG Only	19841	1241	21082	19472	1251	20723
	ENG+SP	20092	1307	21399	19712	1306	21018
	No Gloss	19954	1316	21270	19635	1323	20958
	Total	59887	3864	63751	58819	3880	62699
8	ENG Only	20098	1044	21142	17018	1048	18066
	ENG+SP	20419	1118	21537	17365	1108	18473
	No Gloss	20370	1029	21399	17315	1025	18340
	Total	60887	3191	64078	51698	3181	54879
9	ENG Only	16243	548	16791	18482	561	19043
	ENG+SP	16477	589	17066	18676	595	19271
	No Gloss	16430	530	16960	18604	513	19117
	Total	49150	1667	50817	55762	1669	57431
10	ENG Only	15224	326	15550	15460	334	15794
	ENG+SP	15482	372	15854	15727	410	16137
	No Gloss	15279	323	15602	15688	357	16045
	Total	45985	1021	47006	46875	1101	47976
11	ENG Only	13897	183	14080	14124	182	14306

	ENG+SP	14029	218	14247	14163	175	14338
	No Gloss	13990	209	14199	14082	208	14290
	Total	41916	610	42526	42369	565	42934

To examine the effectiveness and validity of the pop-up glossaries, we ran a mixed logistic regression model on the students' responses to the experimental items. The probability of a student answering the item correctly is

$$Pr(Y_{ij} = 1|u_i) = \frac{\exp(1.7\eta_{ij})}{1+\exp(1.7\eta_{ij})},$$

$$\eta_{ij} = \mu_i + \beta_j + \alpha_1 ENG_{ij} + \alpha_2 ENG_SP_{ij} + \alpha_3 EL_i ENG_{ij} + \alpha_4 EL_i ENG_SP_{ij},$$

$$\mu_i \sim \begin{cases} N(0, \sigma^2_{non-EL}) \\ N(\mu_{EL}, \sigma^2_{EL}) \end{cases},$$

β_j effect of item j ,

$ENG_{ij} = 1$ if student i is in the English glossary condition and item j has glossaries, = 0 else

$ENG_SP_{ij} = 1$ if student i is in the English glossary + Spanish translation condition and item j has glossaries, = 0 else

$EL_i = 1$ if student i is an EL, = 0 else.

The term β_j is the fixed effect controlling the differences in difficulty across items. The term u_i is a random effect capturing the difference in achievement across students. The coefficients α 's indicate whether the glossaries affect the construct being measured or whether there's a differential effect on the EL students.

Exhibit 1.11.2. and Exhibit 1.11.3 show the coefficient estimates, the standard error of the estimates, and the z statistics for the mixed logistic regression performed for each of the ELA and Math tests. The statistics that are significant at $\alpha=0.05$ level are highlighted. The estimates include mean of u_i , which is the mean performance of the EL group (mean of the non-EL group is set to be zero). The negative mean for EL group in each grade indicates that the mean performance of EL students was below that of non-EL students. The estimates also include the main effect of the English glossary and main effect of the English glossary with Spanish translation and their interaction effects with the EL group. Because the EL group is defined as 1 and non-EL group is defined as 0 in the models, the effect of the glossary on the EL group is calculated as the sum of the main effect and the interaction effect. The effect of the glossary on the non-EL group is the main effect only. Positive coefficients indicate that the performance is improved while the negative coefficients indicate that the score is depressed.

As shown in Exhibit 1.11.2, for the ELA assessments, the effects of providing the English glossary and the English glossary with Spanish translation were significantly positive for EL students. The estimated effects ranged from 0.01 to 0.08 for elementary students and gradually increased for the middle school and high school EL students. This means that providing a glossary on the ELA tests significantly improved the performance of EL students across all grades. The main effects estimated from the models for English glossary were not significant except in grade 3, 4, and 9, and the main effects from the English glossary with Spanish translation were not significant except in grade 3, 4, and 6. This means providing a glossary had virtually no effect for non-EL students in middle school and high school grades, but a small negative effect at the elementary school grades, which might be caused by distractions.

With respect to the math assessments, Exhibit 1.11.3 shows that providing a glossary led to significant gains for EL students in almost all grades. Effects observed for the grade 5 and Algebra II assessments were not significant. For the native English speakers, providing a glossary had no impact on performance, with the exception of a slight performance gain for the English only glossary on the Geometry assessment. The results support that the glossary also significantly improved the performance of EL students in most of the mathematics tests, but did not impact the non-EL group except in the Geometry test.

Exhibit 1.11.2. The Coefficient Estimates for the Mixed Logistic Regression Model by Grade Level on Scores for the ELA Assessment

Effect	G3E	G4E	G5E	G6E	G7E	G8E	G9E	G10E	G11E
Coefficient Estimates									
EL mean of random intercept	-0.98	-0.59	-0.69	-0.64	-0.68	-0.67	-0.66	-0.64	-0.56
ENG main effect	-0.04	-0.02	-0.01	0.00	-0.01	0.00	-0.01	0.00	0.00
ENG SP main effect	-0.03	-0.03	-0.01	-0.01	-0.01	0.00	0.00	0.01	0.00
EL by ENG interaction	0.10	0.05	0.08	0.10	0.10	0.11	0.16	0.10	0.21
EL BY ENG SP interaction	0.04	0.08	0.09	0.08	0.08	0.11	0.10	0.11	0.19
ENG effect (main + interaction)	0.05	0.03	0.07	0.10	0.09	0.11	0.15	0.10	0.21
ENG SP effect (main + interaction)	0.01	0.05	0.08	0.06	0.07	0.12	0.10	0.11	0.20
Standard Errors									
EL mean of random intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
ENG main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ENG SP main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
EL by ENG interaction	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
EL BY ENG SP interaction	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04
ENG effect (main + interaction)	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
ENG SP effect (main + interaction)	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04
Z Statistics									
EL mean of random intercept	-179.59	-107.86	-117.29	-85.30	-85.37	-74.61	-72.90	-56.74	-33.35
ENG main effect	-6.86	-3.43	-1.26	-0.04	-1.69	-0.11	-2.06	0.32	-0.66
ENG SP main effect	-4.89	-5.30	-1.30	-2.08	-1.82	0.62	0.34	0.83	0.44
EL by ENG interaction	6.76	3.95	4.76	5.62	5.50	5.42	6.02	2.88	4.61
EL BY ENG SP interaction	2.79	5.97	5.67	4.27	4.88	5.67	3.68	3.26	4.61
ENG effect (main + interaction)	3.70	2.43	4.28	5.62	4.96	5.40	5.54	2.94	4.51
ENG SP effect (main + interaction)	0.64	3.61	5.17	3.58	4.27	5.86	3.76	3.43	4.68

Exhibit 1.11.3. The Coefficient Estimates for the Mixed Logistic Regression Model by Grade Level on Scores for the Mathematics Assessment

Effect	G3M	G4M	G5M	G6M	G7M	G8M	Alg I	Geometry	Alg II
Coefficient Estimates									
EL mean of random intercept	-0.83	-0.79	-0.86	-0.82	-0.83	-0.60	-0.70	-0.67	-0.44
ENG main effect	0.00	-0.01	0.00	0.00	0.01	0.01	0.01	0.03	-0.02
ENG SP main effect	-0.01	-0.01	-0.01	0.00	0.01	-0.01	0.01	0.02	-0.02
EL by ENG interaction	0.11	0.05	0.01	0.09	0.09	0.18	0.42	0.21	-0.04

EL BY ENG SP interaction	0.11	0.14	0.04	0.06	0.12	0.17	0.48	0.06	0.13
ENG effect (main + interaction)	0.12	0.04	0.01	0.08	0.10	0.19	0.43	0.24	-0.07
ENG SP effect (main + interaction)	0.10	0.12	0.03	0.06	0.13	0.16	0.48	0.08	0.11
Standard Errors									
EL mean of random intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
ENG main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ENG SP main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
EL by ENG interaction	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.10
EL BY ENG SP interaction	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.09
ENG effect (main + interaction)	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.10
ENG SP effect (main + interaction)	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.09
Z Statistics									
EL mean of random intercept	-85.51	-84.31	-82.73	-70.90	-70.91	-53.80	-62.32	-37.45	-21.00
ENG main effect	0.50	-1.00	0.00	-0.29	0.62	1.20	0.88	2.29	-1.56
ENG SP main effect	-0.82	-1.27	-0.77	0.30	0.63	-0.81	0.74	1.17	-1.12
EL by ENG interaction	5.58	2.31	0.31	2.66	2.87	5.28	8.25	2.93	-0.42
EL BY ENG SP interaction	5.33	5.99	1.41	1.90	3.84	5.01	9.67	0.87	1.41
ENG effect (main + interaction)	5.82	1.91	0.31	2.58	3.06	5.65	8.45	3.36	-0.64
ENG SP effect (main + interaction)	5.01	5.48	1.13	1.99	4.04	4.77	9.85	1.09	1.24

1.12 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretation is ongoing. Nevertheless, there currently exists sufficient evidence to support the principle claims for the test scores, including that AzMERIT test scores indicate the degree to which students have achieved the Arizona College and Career Ready Standards (ACCRS) at each grade level, and that students scoring at the proficient level or higher demonstrate levels of achievement consistent with national benchmarks indicating that they are on track to college readiness. These claims are supported by evidence of a test development process that ensures alignment of test content to the ACCRS, a standard-setting process that yielded performance standards consistent with those of rigorous, national benchmarks. Confirmatory factor analyses indicate that the subject-area assessments are unidimensional and therefore consistent with the measurement model, but also that the hypothesized reporting strand structure of the AzMERIT provides significant additional information about student achievement. In addition, test scores on the AzMERIT correlate strongly with other measures of subject-area achievement, and demonstrate differential relationships across subject-area assessments.

2. BACKGROUND OF ARIZONA STATEWIDE ASSESSMENTS

In November 2014, the Arizona State Board of Education adopted Arizona's Measurement of Educational Readiness to Inform Teaching, or AzMERIT, to measure student mastery of the Arizona academic standards and progress toward college and career readiness. The AzMERIT measures English language arts in grades 3–11, and math in grades 3–8 and following completion of high school coursework in Algebra I, Geometry, and Algebra II. The Arizona Department of Education (ADE) worked with the American Institutes for Research to develop and administer the AzMERIT beginning in the spring of 2015. In accordance with state requirements, the AzMERIT was designed to¹⁶:

- Align to the academic standards adopted by the Arizona State Board of Education in 2010 (Arizona College and Career Ready Standards, or ACCRS)
- Supply criterion referenced summative assessments for grades 3 through 8, and criterion referenced end of course assessments in identified high school math and English language arts courses for implementation beginning in the 2014–15 school year
- Assess, without bias, a range of basic knowledge and lower-level cognitive skills and higher order, analytical thinking skills in writing, analysis, and problem-solving across subjects, using multiple assessment methods
- Provide valid, reliable, and timely data to educators and policy makers to advance the academic success of Arizona students and inform the State's accountability measures
- Communicate results to students, parents and educators in a clear and timely manner to guide instruction
- Provide an accurate perspective of the quality of learning occurring within classrooms and schools
- Offer educators, students, and families critical tools to improve student achievement, including, but not limited to, formative and interim assessments, sample items, and practice tests
- Allow meaningful national or multistate comparisons of school and student achievement
- Use 21st century technology to deliver the assessment, as available infrastructure allows
- Ensure clarity, transparency, accuracy, and security in all aspects of assessment development, deployment, scoring, and reporting
- Provide for content and psychometric evaluation and validation
- Establish the involvement of Arizona stakeholders – educators, students, parents, institutions of higher education, and business – in the development of the test, test-related materials, and achievement levels indicative of college and career readiness
- Demonstrate accessibility for all students, with optimal access for English language learners and students with special needs
- Respect Arizona's local control of the selection of classroom instructional materials
- Satisfy assessment goals in a cost-efficient manner

The AzMERIT was first administered in spring 2015, assessing proficiency in ELA in grades 3–11, math in grades 3–8, and following completion of Algebra I, Geometry, and Algebra II (or similar) coursework. Following the

¹⁶ Standard 7.1 – The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

initial administration, the AzMERIT in grades 3–8 have been administered in the spring of each academic year; tests assessing high school end-of-course (EOC) tests are administered in the fall, spring, and summer of each academic year.

The Rasch model, and Masters' (1982) Partial Credit Model, an extension of the one parameter Rasch model that allows for graded responses, was used to estimate item parameters for the AzMERIT. Item pools for grade-level summative and EOC assessments were calibrated following the first operational administration in spring 2015 and then adjusted for parameter drift following the spring 2016 administration. A vertical linking design was also implemented to produce a common vertical scale across grade levels to monitor student growth across grades 3–8, as well as the high school EOC assessments. In subsequent years, pre-equated bank item parameter estimates have been applied directly for final scoring and reporting, a strategy that allows for more rapid reporting of tests administered online.

2.1 DEVELOPMENT OF ARIZONA COLLEGE AND CAREER READY STANDARDS

In 2010, the Arizona State Board of Education adopted new academic content standards in ELA and mathematics that reflect high expectations of all Arizona students and strive to ensure that high school graduates are college and career ready. The Arizona College and Career Ready Standards (ACCRS) in mathematics describe expectations for learning in grades K–8 and the first three high school courses (Algebra I, Geometry, Algebra II, or Mathematics 1,2,3) plus specific standards that could be included in a fourth high school credit mathematics course. The ACCRS in ELA describe the reading, writing, language, speaking, and listening skills that students should acquire from grade K–12. The standards can be found on ADE's website.

2.2 AZMERIT TEST DESIGN

The AzMERIT is a series of fixed-form assessments that are intended to be administered online, although it is offered as a dual mode, online and paper-based assessment to accommodate schools that are not yet ready to transition to the online testing environment. A common operational base form is administered to all students within a given test grade and subject. Each assessment is comprised of two to three discrete test sessions. The AzMERIT operational item pools include a variety of selected-response, machine-scored constructed responses (MSCR), and some handscored constructed-response items in the paper math forms where MSCR items could not readily be rendered for paper test administration. AzMERIT also includes essay responses. In spring 2016, a sample of online writing responses were handscored (100% double scoring with resolution of all discrepancies) for purposes of developing statistical models for machine scoring the remaining online responses.

Six types of MSCR items were included in the AzMERIT forms: graphic response, natural language, equation response, hot text, and table input items. The graphic response item types require students to place objects or move objects around in the answer space. A student can also plot points, draw lines, and draw shapes. The natural language item types require students to type an English language answer. The equation response items require students to enter a value or equation. Hot text items ask students to select or rearrange sentences or phrases in a passage. The table input item types require students to input numerical values into a table. The validity of computer-assigned scores for constructed-response items was evaluated following the spring 2015 online administration of the embedded field test items. Rubric validation for all operational test items was completed prior to test construction and was based on the previous field test administration of those items.

Each ELA assessment included one writing essay prompt that required an extended essay response. For the online test administrations, students were randomly administered one of two writing tasks. A random sample of student

responses to each writing task were selected for human scoring. These responses were scored by two human raters on three distinct scoring dimensions or rubrics: Statement of Purpose/Focus and Organization, Evidence/Elaboration, and Conventions/Editing, with any discrepancy adjudicated in a resolution score. This sample of essay responses and writing scores were used to develop the statistical models used for machine scoring the remaining online essay responses. All essays administered on paper tests were handscored. In addition, handscoring was required for a subset of math items administered on paper, generally equation items, where it was not possible to represent the item on paper in a way that allowed machine-scoring.

3. SUMMARY OF SUMMER 2016 AND FALL 2016 OPERATIONAL TEST ADMINISTRATION

The following tests were administered in summer 2016 and fall 2016:

- ELA (reading and writing) in grades 9–11
- Math in grades 9–11, following completion of Algebra I, Geometry, and Algebra II, or similar, coursework

Online administration of the AzMERIT occurred from June 13 through July 21, 2016 and from October 24 through December 2, 2016, for the summer 2016 and fall 2016 administrations, respectively.

The scoring and reporting of the summer and fall 2016 assessments used the items parameters calibrated following the spring 2016 administration and vertical scale and performance standards established in summer 2015. This section summarizes the operational test results for the fall 2016 administration of the AzMERIT.

3.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 3.1.1 and 3.1.2 for summer 2016 and Exhibits 3.1.3 and 3.1.4 for fall 2016 by assessment and subgroup.¹⁷

Exhibit 3.1.1 Number of Students Participating in ELA Assessments by Subgroups – Summer 2016

Group	ELA 9	ELA 10	ELA 11
All Students	567	335	214
Female	212	138	106
Male	355	197	108
Unknown			
African American	28	15	12
Asian	9	7	
Native Hawaiian/Pacific Islander	1		
Hispanic/Latino	418	208	125
American Indian or Alaskan	42	31	17
White	65	68	57
Multiple Ethnicities	4	6	3
Limited English Proficiency	26	6	6
Special Education	54	38	13
Free Reduced Lunch	364	186	99

¹⁷ Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio demographic and developmental characteristics.

Exhibit 3.1.2 Number of Students Participating in Math Assessments by Subgroups – Summer 2016

Group	Algebra I	Geometry	Algebra II
All Students	1045	907	778
Female	413	412	386
Male	632	495	392
Unknown			
African American	78	63	50
Asian	22	28	21
Native Hawaiian/Pacific Islander	1	2	
Hispanic/Latino	628	552	475
American Indian or Alaskan	56	23	21
White	240	229	200
Multiple Ethnicities	20	10	11
Limited English Proficiency	69	23	10
Special Education	98	55	25
Free Reduced Lunch	410	359	259

Exhibit 3.1.3 Number of Students Participating in ELA Assessments by Subgroups – Fall 2016

Group	ELA 9	ELA 10	ELA 11
All Students	3448	4354	4927
Female	1501	2037	2356
Male	1947	2317	2571
Unknown			
African American	207	264	323
Asian	75	98	92
Native Hawaiian/Pacific Islander	18	16	34
Hispanic/Latino	1669	1966	2210
American Indian or Alaskan	196	297	314
White	1228	1665	1896
Multiple Ethnicities	49	45	42
Limited English Proficiency	208	149	221
Special Education	301	335	393
Free Reduced Lunch	1432	1625	1876

Exhibit 3.1.4 Number of Students Participating in Math Assessments by Subgroups – Fall 2016

Group	Algebra I	Geometry	Algebra II
All Students	4882	5094	5063
Female	2228	2452	2501
Male	2654	2642	2562
Unknown			
African American	311	345	288
Asian	124	79	160
Native Hawaiian/Pacific Islander	33	29	28
Hispanic/Latino	2226	2496	2062
American Indian or Alaskan	270	202	261
White	1815	1854	2206
Multiple Ethnicities	66	63	42
Limited English Proficiency	312	217	145
Special Education	261	321	239
Free Reduced Lunch	1802	1887	1821

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibit 3.2.1 for summer 2016 and Exhibit 3.2.2 for fall 2016.

Exhibit 3.2.1 Test Score Summary Statistics – Summer 2016

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Max.	Min.
ELA					
Grade 9	567	2546	24.41	2636	2490
Grade 10	335	2549	23.59	2627	2479
Grade 11	214	2557	27.41	2638	2492
Math					
Algebra I	1045	3658	26.71	3787	3593
Geometry	907	3675	29.31	3786	3609
Algebra II	778	3691	31.72	3822	3629

Exhibit 3.2.2 Test Score Summary Statistics – Fall 2016

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Max.	Min.
ELA					
Grade 9	3448	2555	30.51	2664	2466
Grade 10	4354	2557	25.53	2660	2487
Grade 11	4927	2556	28.21	2670	2473
Math					
Algebra I	4882	3667	35.36	3787	3577
Geometry	5094	3669	29.06	3803	3609
Algebra II	5063	3694	33.22	3839	3629

The percentage of students in each performance level by grade and content area, as well as the percentage of students at or above Proficient are presented in Exhibit 3.2.3 for summer 2016 and Exhibit 3.2.4 for fall 2016.

Exhibit 3.2.3 Percentage of Students in Performance Levels – Summer 2016

Grade	Number Tested	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	At or Above Proficient
ELA						
9	567	65	25	9	2	11
10	335	77	12	10	2	12
11	214	67	15	14	4	18
Math						
Algebra I	1045	61	20	17	3	20
Geometry	907	48	30	19	3	22
Algebra II	778	50	30	14	6	20

Exhibit 3.2.4 Percentage of Students in Performance Levels – Fall 2016

Grade	Number Tested	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	At or Above Proficient
ELA						
9	3448	50	25	20	5	25
10	4354	64	17	15	5	20
11	4927	68	16	11	5	16
Math						
Algebra I	4882	51	16	24	9	33
Geometry	5094	57	25	16	2	18
Algebra II	5063	47	25	22	7	28

3.3 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 3.3.1 and 3.3.2 present the number and percentage of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian, Native Hispanic/Latino, American Indian, White, Multiple Ethnicities, limited English proficiency, special education, and free reduced lunch for summer 2016. Exhibits 3.3.3 and 3.3.4 present this information for fall 2016.

Exhibit 3.3.1 Number of Students at Each Performance Level by Subgroups – Summer 2016

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
ELA														
ELA 9	Minimally Proficient	366	130	236	23	*	*	279	24	33	*	24	45	248
	Partially Proficient	140	53	87	3	*	*	98	15	21	*	1	7	77
	Proficient	505	240	265	2	*	*	34	3	8	*	1**	2	32
	Highly Proficient	116	51	65	0	*	*	7	0	3	*	0	0	7
	At or Above Proficient	611	291	320	2	*	*	41	3	11	*	1	2	39
ELA 10	Minimally Proficient	257	108	149	14	*		166	24	46	*	*	35	150
	Partially Proficient	393	142	251	0	*		27	4	7	*	*	2	22
	Proficient	333	163	170	1**	*		15	3	9	*	*	1*	12
	Highly Proficient	63	0	6	0	*		0	0	6	*	*	0	2
	At or Above Proficient	399	163	236	1	*		15	3	15	*	*	1	14
ELA 11	Minimally Proficient	143	63	80	9			94	11	27	*	*	9	73
	Partially Proficient	333	193	140	2			17	4	8	*	*	2	14
	Proficient	299	195	104	1**			11	2	16	*	*	1	9
	Highly Proficient	99	55	44	0			3	0	6	*	*	1	3
	At or Above Proficient	338	243	145	1			14	2	22	*	*	2	12
Math														
Algebra I	Minimally Proficient	634	257	377	60	13	*	407	44	94	15	65	75	298
	Partially Proficient	205	85	120	12	3	*	119	6	63	2	3	15	66
	Proficient	718	648	114	6	2	*	89	6	72	3	1**	7	44
	Highly Proficient	288	78	21	0	4	*	13	0	11	0	0	1	2
	At or Above Proficient													

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
	At or Above Proficient	206	71	135	6	6	*	102	6	83	3	1	8	46
Geometry	Minimally Proficient	439	19	243	48	6	*	281	11	*	6	19	41	20
	Partially Proficient	270	12	148	12	9	*	163	6	*	2	3	8	80
	Proficient	175	83	92	3	12	*	95	5	*	2	1**	6	66
	Highly Proficient	231	11	12	0	1	*	13	1	*	0	0	0	13
	At or Above Proficient	199	89	104	3	13	*	108	6	*	2	1	6	79
		8	94	104	3	13	*	108	6	*	2	1	6	79
Algebra II	Minimally Proficient	388	18	202	30	5		255	13	76	9	*	15	16
	Partially Proficient	233	11	119	14	1		145	5	68	0	*	7	70
	Proficient	107	53	54	6	6		53	1*	40	1*	*	3	25
	Highly Proficient	50	33	17	0	9		22	2	16	1	*	0	2
	At or Above Proficient	157	86	71	6	15		75	3	56	2	*	3	27
		7	86	71	6	15		75	3	56	2	*	3	27

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch; *Indicates that less than 11 students participated in this assessment during this administration. These values are suppressed in compliance with federal FERPA requirements. **Indicates that more than zero students are proficient in order to protect that subgroup from discrimination.

Exhibit 3.3.2 Percentage of Students at Each Performance Level by Subgroups — Summer 2016

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
ELA														
ELA 9	Minimally Proficient	65	61	66	82	*	*	67	57	51	*	92	83	68
	Partially Proficient	25	25	25	11	*	*	23	36	32	*	7	13	21
	Proficient	9	11	7	7	*	*	8	7	12	*	1**	4	9

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
ELA														
	Highly Proficient	2	2	2	0	*	*	2	0	5	*	0	0	2
	At or Above Proficient	11	14	9	7	*	*	10	7	17	*	1	4	11
	Minimally Proficient	77	78	76	93	*		80	77	68	*	*	92	81
	Partially Proficient	12	10	13	0	*		13	13	10	*	*	5	12
	Proficient	10	12	9	7	*		7	10	13	*	*	3	6
ELA 10	Highly Proficient	2	0	3	0	*		0	0	9	*	*	0	1
	At or Above Proficient	12	12	12	7	*		7	10	22	*	*	3	8
	Minimally Proficient	67	59	74	75			75	65	47	*	*	69	74
	Partially Proficient	15	18	13	24			14	24	14	*	*	15	14
	Proficient	14	18	9	1**			9	12	28	*	*	8	9
ELA 11	Highly Proficient	4	5	4	0			2	0	11	*	*	8	3
	At or Above Proficient	18	23	13	1			11	12	39	*	*	15	12
Math														
Algebra I	Minimally Proficient	61	62	60	77	59	*	65	79	39	75	94	77	73
	Partially Proficient	20	21	19	15	14	*	19	11	26	10	4	15	16
	Proficient	17	15	18	8	9	*	14	11	30	15	1**	7	11
	Highly Proficient	3	2	3	0	18	*	2	0	5	0	0	1	0
	At or Above Proficient	20	17	21	8	27	*	16	11	35	15	1	8	11
Geometry	Minimally Proficient	48	48	49	76	21	*	51	48	38	*	83	75	56
	Partially Proficient	30	30	30	19	32	*	30	26	34	*	13	15	22
	Proficient	19	20	19	5	43	*	17	22	25	*	4	11	18
	Highly Proficient	3	3	2	0	4	*	2	4	3	*	0	0	4
	At or Above Proficient	22	23	21	5	46	*	20	26	29	*	4	11	22
Algebra II	Minimally Proficient	50	48	52	60	24		54	62	38	82	*	60	63
	Partially Proficient	30	30	30	28	5		31	24	34	0	*	28	27
	Proficient	14	14	14	12	29		11	5	20	9	*	12	10
	Highly Proficient	6	9	4	0	43		5	10	8	9	*	0	1
	At or Above Proficient	20	22	18	12	71		16	14	28	18	*	12	10

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch; *Indicates that less than 11 students participated in this assessment during this administration. These percentages are suppressed in compliance with federal FERPA requirements. **Indicates that more than zero students are proficient in order to protect that subgroup from discrimination.

Exhibit 3.3.3 Number of Students at Each Performance Level by Subgroups-Fall 2016

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
ELA														
ELA 9	Minimally Proficient	1714	635	1079	117	18	4	1031	116	399	25	170	251	867
	Partially Proficient	874	429	445	59	17	5	380	59	341	12	26	27	325
	Proficient	687	350	337	26	31	8	221	20	370	11	10	19	212
	Highly Proficient	173	87	86	5	9	1	37	1	118	1	2	4	28
	At or Above Proficient	860	437	423	31	40	9	258	21	488	12	12	23	240
ELA 10	Minimally Proficient	2778	1230	1548	180	49	9	1447	227	835	28	123	300	1166
	Partially Proficient	725	376	349	49	15	3	275	45	332	6	18	18	259
	Proficient	643	327	316	27	24	3	194	23	366	6	7	17	153
	Highly Proficient	208	104	104	8	10	1	50	2	132	5	1	0	47
	At or Above Proficient	851	431	420	35	34	4	244	25	498	11	8	17	200
ELA 11	Minimally Proficient	3334	1501	1833	240	39	22	1733	243	1016	29	175	373	1419
	Partially Proficient	782	406	376	48	19	8	266	35	395	8	26	14	271
	Proficient	560	307	253	25	19	1	161	26	323	4	15	6	142
	Highly Proficient	251	142	109	10	15	3	50	10	162	1	5	0	44
	At or Above Proficient	811	449	362	35	34	4	211	36	485	5	20	6	186
Math														
Algebra I	Minimally Proficient	2492	1097	1395	198	18	20	1440	157	593	41	251	222	1097
	Partially Proficient	770	349	421	45	11	2	351	42	301	12	37	28	289
	Proficient	1157	577	580	58	59	8	339	52	623	12	22	9	328
	Highly Proficient	463	205	258	10	36	3	96	19	298	1	2	2	88
	At or Above Proficient	1620	782	838	68	95	11	435	71	921	13	24	11	416
Geometry	Minimally Proficient	2927	1377	1550	220	27	14	1655	131	828	42	168	259	1195
	Partially Proficient	1264	642	622	83	22	8	556	39	537	12	44	45	437
	Proficient	799	390	409	39	21	7	259	29	430	9	5	17	235
	Highly Proficient	104	43	61	3	9	0	26	3	59	0	0	0	20
	At or Above Proficient	903	433	470	42	30	7	285	32	489	9	5	17	255
Algebra II	Minimally Proficient	2377	1122	1255	174	27	12	1204	172	755	21	110	203	992
	Partially Proficient	1254	662	592	66	34	5	504	58	570	15	29	28	470
	Proficient	1089	569	520	40	55	9	291	24	664	5	5	7	298
	Highly Proficient	343	148	195	8	44	2	63	7	217	1	1	1	61
	At or Above Proficient	1432	717	715	48	99	11	354	31	881	6	6	8	359

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch

Exhibit 3.3.2 Percentage of Students at Each Performance Level by Subgroups—Fall 2016

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
ELA 9	Minimally Proficient	50	42	55	57	24	22	62	59	32	51	82	83	61
	Partially Proficient	25	29	23	29	23	28	23	30	28	24	13	9	23
	Proficient	20	23	17	13	41	44	13	10	30	22	5	6	15
	Highly Proficient	5	6	4	2	12	6	2	1	10	2	1	1	2
	At or Above Proficient	25	29	22	15	53	50	15	11	40	24	6	8	17
ELA 10	Minimally Proficient	64	60	67	68	50	56	74	76	50	62	83	90	72
	Partially Proficient	17	18	15	19	15	19	14	15	20	13	12	5	16
	Proficient	15	16	14	10	24	19	10	8	22	13	5	5	9
	Highly Proficient	5	5	4	3	10	6	3	1	8	11	1	0	3
	At or Above Proficient	20	21	18	13	35	25	12	8	30	24	5	5	12
ELA 11	Minimally Proficient	68	64	71	74	42	65	78	77	54	69	79	95	76
	Partially Proficient	16	17	15	15	21	24	12	11	21	19	12	4	14
	Proficient	11	13	10	8	21	3	7	8	17	10	7	2	8
	Highly Proficient	5	6	4	3	16	9	2	3	9	2	2	0	2
	At or Above Proficient	16	19	14	11	37	12	10	11	26	12	9	2	10
Algebra I	Minimally Proficient	51	49	53	64	15	61	65	58	33	62	80	85	61
	Partially Proficient	16	16	16	14	9	6	16	16	17	18	12	11	16
	Proficient	24	26	22	19	48	24	15	19	34	18	7	3	18
	Highly Proficient	9	9	10	3	29	9	4	7	16	2	1	1	5
	At or Above Proficient	33	35	32	22	77	33	20	26	51	20	8	4	23
Geometry	Minimally Proficient	57	56	59	64	34	48	66	65	45	67	77	81	63
	Partially Proficient	25	26	24	24	28	28	22	19	29	19	20	14	23
	Proficient	16	16	15	11	27	24	10	14	23	14	2	5	12
	Highly Proficient	2	2	2	1	11	0	1	1	3	0	0	0	1
	At or Above Proficient	18	18	18	12	38	24	11	16	26	14	2	5	14
Algebra II	Minimally Proficient	47	45	49	60	17	43	58	66	34	50	76	85	54
	Partially Proficient	25	26	23	23	21	18	24	22	26	36	20	12	26
	Proficient	22	23	20	14	34	32	14	9	30	12	3	3	16
	Highly Proficient	7	6	8	3	28	7	3	3	10	2	1	0	3
	At or Above Proficient	28	29	28	17	62	39	17	12	40	14	4	3	20

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch

3.4 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. The reliability evidence of the AzMERIT ELA and math assessments are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance-level classifications.¹⁸

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.¹⁹

3.4.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 3.4.1.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and math assessments for summer 2016, and Exhibit 3.4.1.2 shows these estimates for fall 2016. Internal consistency estimates are uniformly above 0.8.,

Exhibit 3.4.1.1 Internal Consistency Reliabilities for AzMERIT Scores – Summer 2016

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.84	595.82	0.84	713.29
10/Geometry	0.83	556.35	0.83	858.80
11/Algebra II	0.87	751.57	0.84	1006.20

Note: Reliability ranges from 0 to 1. Variance is in scale score metric.

¹⁸ Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

¹⁹ Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

Exhibit 3.4.1.2 Internal Consistency Reliabilities for AzMERIT Scores – Fall 2016

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.89	930.9888	0.91	1250.278
10/Geometry	0.86	651.0646	0.81	843.549
11/Algebra II	0.88	795.3622	0.86	1103.563

Note: Reliability ranges from 0 to 1. Variance is in scale score metric.

3.4.2 STANDARD ERROR OF MEASUREMENT

Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low and high performing students, the precision of test scores decreases near the tails of the ability distribution.

Exhibit 3.4.2.1 presents the standard errors of measurement for the AzMERIT ELA and math assessments with respect to the four AzMERIT performance standards for summer 2016, and Exhibit 3.4.2.2 for fall 2016. As the tables indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Partially Proficient and Proficient performance standards.²⁰ Test scores near the tails of the ability distribution are somewhat less precise, as expected. While these numbers indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance-level classifications.

Exhibit 3.4.2.1 Conditional Standard Error of Measure at Performance Levels and Overall Standard Error of Measurement for ELA and Math – Summer 2016

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 9 ELA	10.16	9.00	9.97	11.40	9.89
Grade 10 ELA	9.90	9.00	9.50	11.19	9.78
Grade 11 ELA	9.90	9.00	9.50	11.04	9.77
Algebra I	11.21	9.62	9.46	12.82	10.68
Geometry	13.52	10.42	10.00	11.89	11.98
Algebra II	13.90	11.25	10.00	11.37	12.50

²⁰ Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 3.4.2.2 Conditional Standard Error of Measure at Performance Levels and Overall Standard Error of Measurement for ELA and Math – Fall 2016

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 9 ELA	10.35	9.00	9.95	12.06	10.04
Grade 10 ELA	9.73	9.00	9.53	11.38	9.67
Grade 11 ELA	10.04	9.00	9.46	11.24	9.88
Algebra I	11.44	9.60	9.53	12.88	10.89
Geometry	13.84	10.63	10.00	11.80	12.51
Algebra II	14.08	11.21	10.00	10.91	12.40

3.4.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed in terms of the probabilities of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).²¹ This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications is typically estimated on the scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for decision accuracy and decision consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Decision consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

3.4.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the probability of consistent classification directly using the likelihood function. The likelihood function of the achievement attribute, designated θ , given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

²¹ Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

If a student's estimated theta is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 3.4.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications differ from their observed score classifications (e.g., a student whose true score results in a Proficient level classification, but is classified incorrectly as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 3.4.4.1 Classification Accuracy

		Classification on a Form Actually Taken	
		At or Above the Cut Score	Below the Cut Score
Classification on True Score	At or Above the Cut Score	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

3.4.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 3.4.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 3.4.5.1 Classification Consistency

		Classification on the Second Form Taken	
		Above the Cut Score	Below the Cut Score
Classification on the First Form Taken	At or Above the Cut Score	N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

3.4.6 CLASSIFICATION RELIABILITY ESTIMATES

Exhibit 3.4.6.1 presents the classification accuracy and consistency indexes for the summer 2016 administration of AzMERIT, and Exhibit 3.4.6.2 for the fall 2016 administration. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Exhibit 3.4.6.1 Classification Accuracy and Consistency Indexes for Performance Standards – Summer 2016

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
9	0.90	0.95	0.99	0.85	0.93	0.98
10	0.92	0.95	0.99	0.89	0.94	0.98
11	0.92	0.94	0.97	0.89	0.92	0.96
MATH						
Algebra I	0.88	0.94	0.99	0.84	0.92	0.98
Geometry	0.87	0.93	0.99	0.82	0.90	0.98
Algebra II	0.86	0.93	0.98	0.81	0.90	0.98

Exhibit 3.4.6.2 Classification Accuracy and Consistency Indexes for Performance Standards – Fall 2016

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
9	0.91	0.93	0.97	0.88	0.90	0.96
10	0.90	0.93	0.98	0.86	0.91	0.97
11	0.92	0.94	0.98	0.89	0.92	0.97
MATH						
Algebra I	0.91	0.95	0.96	0.88	0.92	0.95
Geometry	0.88	0.94	0.99	0.83	0.92	0.99
Algebra II	0.89	0.93	0.97	0.84	0.90	0.96

3.4.7 RELIABILITY FOR SUBGROUPS IN THE POPULATION

Exhibits 3.4.7.1 and 3.4.7.2 show the mean reliability for each of the identified subgroups gender (females and males), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with IEPs [Special Education], and free or reduced lunch) for summer 2016, and Exhibits 3.4.7.3 and 3.4.7.4 show this data for fall 2016.²² Each racial and/or ethnic group was composed of approximately equal numbers of males and females. As the exhibits indicate, internal consistency reliabilities are consistent across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups. Where reliability estimates are attenuated, there is an associated decrease in variance within the subgroup population, indicating that the decrease in reliability is likely due to a restriction in range.

²² Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Exhibit 3.4.7.1 Internal Consistency Reliability by Subgroup – ELA – Summer 2016

	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.84	595.82	0.83	556.35	0.87	751.57
Female	0.83	563.71	0.78	413.47	0.88	764.02
Male	0.84	604.96	0.85	657.59	0.87	720.11
African American	0.74	391.78	0.72	376.81	0.80	477.73
Asian	0.75	385.19	0.89	927.24		
Native Hawaiian/Pacific Islander	.	.				
Hispanic/Latino	0.83	569.25	0.78	432.70	0.85	628.68
American Indian or Alaskan	0.79	463.77	0.79	457.93	0.84	604.86
White	0.87	743.24	0.88	827.81	0.89	825.25
Multiple Ethnicities	0.84	584.25	0.90	991.50	0.89	940.33
Limited English Proficiency	0.66	304.22	0.73	409.90	0.24	123.60
Special Education	0.76	418.26	0.77	468.98	0.91	1065.17
Free/Reduced Lunch	0.83	590.35	0.80	491.86	0.86	724.27

Exhibit 3.4.7.2 Internal Consistency Reliability by Subgroup – Math – Summer 2016

	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.84	713.29	0.83	858.80	0.84	1006.20
Female	0.81	591.90	0.84	896.20	0.87	1161.92
Male	0.85	792.73	0.83	827.59	0.81	839.04
African American	0.68	364.28	0.58	396.58	0.72	581.70
Asian	0.93	2017.02	0.83	705.86	0.93	2127.03
Native Hawaiian/Pacific Islander	.	.	-0.84	72.00		
Hispanic/Latino	0.81	621.12	0.83	869.89	0.80	761.00
American Indian or Alaskan	0.74	488.85	0.87	1107.27	0.85	1112.43
White	0.87	787.79	0.84	833.57	0.88	1304.52
Multiple Ethnicities	0.80	582.06	0.82	796.77	0.82	943.67
Limited English Proficiency	0.40	222.08	0.67	509.57	0.14	190.04
Special Education	0.76	524.18	0.69	565.43	0.73	628.33
Free/Reduced Lunch	0.77	522.19	0.86	1134.37	0.71	585.30

Exhibit 3.4.7.3 Internal Consistency Reliability by Subgroup – ELA Fall 2016

	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.89	931	0.86	651	0.88	795
Female	0.89	871	0.85	635	0.87	758
Male	0.89	942	0.86	658	0.88	805
African American	0.88	825	0.84	591	0.86	736
Asian	0.87	789	0.89	892	0.90	997
Native Hawaiian/Pacific Islander	0.88	825	0.83	515	0.89	850
Hispanic/Latino	0.87	768	0.83	540	0.85	644
American Indian or Alaskan	0.84	622	0.78	423	0.85	642
White	0.89	939	0.87	691	0.89	834
Multiple Ethnicities	0.90	994	0.88	825	0.85	648
Limited English Proficiency	0.82	586	0.80	488	0.87	781
Special Education	0.83	678	0.73	365	0.73	426
Free/Reduced Lunch	0.87	765	0.83	532	0.85	651

Exhibit 3.4.7.4 Internal Consistency Reliability by Subgroup – Math Fall 2016

	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.91	1250	0.81	844	0.86	1104
Female	0.90	1216	0.81	802	0.85	1005
Male	0.91	1276	0.82	881	0.87	1199
African American	0.87	912	0.79	786	0.81	880
Asian	0.90	1262	0.90	1371	0.91	1319
Native Hawaiian/Pacific Islander	0.92	1748	0.80	743	0.89	1313
Hispanic/Latino	0.86	881	0.76	686	0.81	859
American Indian or Alaskan	0.89	1121	0.80	796	0.76	732
White	0.91	1344	0.84	902	0.88	1131
Multiple Ethnicities	0.82	636	0.74	610	0.77	639
Limited English Proficiency	0.73	461	0.60	459	0.62	476
Special Education	0.71	463	0.64	525	0.58	484
Free/Reduced Lunch	0.87	953	0.78	726	0.82	908

3.4.8 SUBSCALE RELIABILITY

Coefficient alpha internal consistency reliability estimates associated with the subscales for the summer 2016 operational forms are presented in Exhibits 3.4.8.1–3.4.8.3, and in Exhibits 3.4.8.4–3.4.8.6 for fall 2016. As indicated in the exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT. The only exception is the Circles, Geometric Measurement, and Geometric Properties with Equations strand in the Geometry test.

Exhibit 3.4.8.1 Subscale Reliabilities – ELA Grades 9–11 – Summer 2016

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.68	0.63	0.66
Grade 10	0.70	0.50	0.70
Grade 11	0.76	0.66	0.71

Exhibit 3.4.8.2 Subscale Reliabilities – Algebra I & II – Summer 2016

	Algebra	Functions	Statistics
Algebra I	0.71	0.62	0.49
Algebra II	0.66	0.50	0.63

Exhibit 3.4.8.3 Subscale Reliabilities – Geometry – Summer 2016

	Circles, Geometric Measurement, & Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.44	0.49	0.54	0.53

Exhibit 3.4.8.4 Subscale Reliabilities – ELA Grades 9–11 – Fall 2016

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.75	0.74	0.74
Grade 10	0.71	0.6	0.73
Grade 11	0.75	0.66	0.72

Exhibit 3.4.8.5 Subscale Reliabilities – Algebra I & II – Fall 2016

	Algebra	Functions	Statistics
Algebra I	0.8	0.76	0.66
Algebra II	0.72	0.42	0.67

Exhibit 3.4.8.6 Subscale Reliabilities – Geometry – Fall 2016

	Circles, Geometric Measurement, & Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.34	0.44	0.5	0.47

3.5 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 3.5.1–3.5.3 for summer 2016, and in Exhibits 3.5.4–3.5.6 for fall 2016. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.²³ The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

Where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y . When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Exhibit 3.5.1 Subscale Observed and Disattenuated Intercorrelations – ELA Grades 9 to 11 – Summer 2016

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.59		0.90	
	Writing & Language	0.49	0.48	0.73	0.74
10	Literature	0.62		1.00	
	Writing & Language	0.44	0.44	0.63	0.74
11	Literature	0.68		0.96	
	Writing & Language	0.55	0.49	0.75	0.72

Exhibit 3.5. Subscale Observed and Disattenuated Intercorrelations – Algebra I & Algebra II – Summer 2016

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.65		0.98	1.00
	Statistics	0.64	0.60	1.00	1.00
Algebra II	Functions	0.64		1.00	
	Statistics	0.61	0.60	0.95	1.00

²³ Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that removes the effects of measurement error on the test should be clearly reported as adjusted estimates.

Exhibit 3.5.3 Subscale Observed and Disattenuated Intercorrelations – Geometry – Summer 2016

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	C	0.56			1.00		
	MG	0.56	0.48		1.00	0.93	
	SRTT	0.63	0.54	0.56	1.00	1.00	1.00

Note: C=Congruence; CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations; MG=Modeling with Geometry; SRTT=Similarity, Right Triangles and Trigonometry

Exhibit 3.5.4 Subscale Observed and Disattenuated Intercorrelations – ELA Grades 9–11 – Fall 2016

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.69		0.93	
	Writing & Language	0.64	0.62	0.86	0.84
10	Literature	0.61		0.94	
	Writing & Language	0.59	0.53	0.88	0.80
11	Literature	0.67		0.95	
	Writing & Language	0.63	0.59	0.91	0.86

Exhibit 3.5.5 Subscale Observed and Disattenuated Intercorrelations – Algebra I & Algebra II – Fall 2016

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.77		0.99	
	Statistics	0.75	0.73	1.00	1.00
Algebra II	Functions	0.67		1.00	
	Statistics	0.70	0.61	1.00	1.00

Exhibit 3.5.6 Subscale Observed and Disattenuated Intercorrelations – Geometry – Fall 2016

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	C	0.58			1.00		
	MG	0.53	0.51		1.00	1.00	
	SRTT	0.59	0.55	0.49	1.00	1.00	1.00

Note: C=Congruence; CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations; MG=Modeling with Geometry; SRTT=Similarity, Right Triangles and Trigonometry.

4. SUMMARY OF SPRING 2017 OPERATIONAL TEST ADMINISTRATION

The following AzMERIT assessments were administered in spring 2017:

- ELA (reading and writing) in grades 3–11
- Math in grades 3–8, Algebra I, Geometry, and Algebra II

Online administration of the AzMERIT occurred from March 27 through May 4, 2017. The paper version of the AzMERIT was administered between March 27 and April 7, 2017.

In the spring 2015 administration, item parameters for the math assessments were calibrated following the online administration to establish the AzMERIT bank scale. In the spring 2016 administration, all field test items were placed on the AzMERIT bank scale by concurrent calibrations of operational and field test items. In spring 2017, the Math tests were scored using pre-equated item parameter estimates following the spring 2016 test administration of AzMERIT. Thus, no post-equating activities were conducted prior to the scoring and reporting of the Math tests in spring 2017.

In the spring 2015 administration, item parameters for the ELA assessments were calibrated following the online administration to establish the AzMERIT bank scale. In spring 2016, in each ELA online assessment, students were randomly assigned one of six writing prompts for administration. Following the spring 2016 test administration, all operational items including reading and writing items were concurrently calibrated, and then linked back to the AzMERIT bank scale using the mean-mean equating method, while all field test items were concurrently calibrated with the mean-mean equated operational items. In spring 2017, students were assigned one of two associated with the two writing rubrics (Informative-Explanatory or Opinion for grades 3–5 or Informative-Explanatory or Argumentative for grades 6–11). The pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the spring 2017 final scoring and reporting. This section summarizes the operational test results for the spring 2017 administration of the AzMERIT. Detailed descriptions of procedures for item and test development, test administration, scaling, equating, and scoring are presented in subsequent sections.

4.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 4.1.1 and 4.1.2 by assessment and subgroup.²⁴ We note that some students participated in an end-of-course (EOC) assessment rather than a grade-level assessment, especially in grade 8, where a large number of more advanced students are enrolled in Algebra I courses. The tables in Appendix F show the demographic composition of test takers by mode of test administration.

²⁴ Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics.

Exhibit 4.1.1 Number of Students Participating in ELA Assessments by Subgroups – Spring 2017

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
All Students	88352	88680	87527	86113	85248	84511	80752	74155	68526
Female	43389	43533	42990	42562	41447	41597	39815	36928	34416
Male	44963	45147	44537	43552	43801	42914	40938	37227	34110
African American	4612	4605	4597	4615	4542	4625	4474	3882	3592
Asian	2475	2488	2435	2521	2575	2447	2443	2372	2169
Native Hawaiian/Pacific Islander	316	319	340	285	285	269	258	275	194
Hispanic/Latino	40896	40384	39452	38194	37721	37159	35688	32567	29596
American Indian or Alaskan	4309	4338	4284	4344	4197	3896	3981	3470	3139
White	32754	33591	33657	33750	33723	34073	32146	29979	28329
Multiple Ethnicities	2845	2832	2610	2301	2082	1917	1671	1525	1354
Limited English Proficiency	9267	8916	7014	5179	4983	4045	2664	1438	841
Special Education	9962	10604	10488	9981	9116	8652	7335	6596	5779
Free/Reduced Lunch	35656	35381	34197	33578	32496	31801	26224	23610	20656
Accommodation	51323	51195	50135	48038	47598	48120	41791	38283	35691

Exhibit 4.1.2 Number of Students Participating in Math Assessments by Subgroups – Spring 2017

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
All Students	88481	88961	87636	86335	83302	71886	82960	71594	63663
Female	43378	43637	43009	42648	40582	35075	40674	35683	32467
Male	45103	45324	44627	43687	42720	36811	42287	35911	31196
African American	4619	4636	4603	4644	4515	4305	4434	3895	3132
Asian	2479	2500	2434	2516	2333	1588	2081	2140	2115
Native Hawaiian/Pacific Islander	315	317	339	285	278	227	267	251	207
Hispanic/Latino	40910	40502	39529	38301	37341	33944	37582	32014	27431
American Indian or Alaskan	4332	4350	4291	4370	4158	3673	4134	3596	2839
White	32820	33687	33674	33798	32512	26408	32591	28208	26623
Multiple Ethnicities	2848	2838	2610	2310	2037	1581	1727	1376	1249
Limited English Proficiency	9369	9026	7060	5262	5013	4035	2764	1818	913
Special Education	10077	10687	10555	10055	9118	8496	7861	6072	3992
Free/Reduced Lunch	35622	35466	34222	33705	32144	29493	28222	23884	19267
Accommodation	51447	51383	50260	48194	46709	40854	45538	36721	32994

4.2 CLASSICAL ITEM ANALYSIS

Because AzMERIT is an online assessment system, classical item analysis statistics for selected-response (SR) and constructed-response (CR) items reported here are calculated based on all online student responses. Classical item analysis statistics are used to monitor item behavior and investigate irregularities in item scoring throughout the test window for online assessments, and following processing of answer documents for paper test administrations. Classical item analyses examine the degree to which the items function as intended with respect to the underlying scales. For online and paper test administrations, quality assurance reports provide the required item and test statistics for each SR and CR item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item during test administration. Key statistics computed and examined include biserial/polyserial correlations for item discrimination, biserial correlations for distractors for selected response items, and proportion correct for item difficulty.

The biserial/polyserial correlations indicate the extent to which each item differentiated between those examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The biserial correlation for dichotomous items is calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, the mean total number correct for student scoring within each of the possible score categories is used. Items are flagged for review by test development experts if the biserial correlation for the keyed (correct) response is less than .25, or changed from previous administration. For dichotomous items, we also compute the biserial correlation for each of the distractor response options.

The proportion correct score is the average number of available points achieved by students on the item. For dichotomous items, this is simply the proportion of students responding correctly. For polytomous items, the average score on the item is divided by the points available to produce a comparable index. The proportion correct score is commonly referred to as the *p*-value.

Exhibit 4.2.1 presents the average proportion of students responding correctly and average point biserial/polyserial correlations from the spring 2017 online administration of AzMERIT. As indicated in Exhibit 4.2.1, the ELA items were somewhat harder than the math items for students in grades 3–4, where this trend is reversed in grades 6 and above, with items on the ELA assessments, on average, being easier than items on the math assessments. While mean difficulty of ELA items is relatively consistent across grade-level assessments, the average difficulty of math items increases across grade level and course assessments. The proportion of students responding correctly to test items in the EOC assessments in math was relatively quite low. Mean biserial correlations for the grade-level and EOC assessments are reasonably high and consistent across assessments. Exhibit 4.2.2 shows the number of items flagged for proportion correct value, biserial/polyserial correlation, distractor biserial/polyserial, and DIF categories for the operational items in the spring 2017 online forms. The flagging criteria are presented in Sections 5.4.1 and 5.4.3.

Exhibit 4.2.1 Average Proportion Correct and Point Biserial Correlations for Operational Test Items Administered Online

Grade	Average <i>p</i> -Value	<i>p</i> -Value SD	Average Point-Biserial	Point-Biserial SD
ELA				
3	0.49	0.19	0.44	0.12
4	0.52	0.19	0.45	0.09
5	0.53	0.16	0.46	0.11
6	0.51	0.21	0.44	0.13
7	0.53	0.18	0.45	0.12
8	0.47	0.19	0.47	0.12
9	0.52	0.13	0.43	0.13
10	0.49	0.18	0.42	0.13
11	0.47	0.20	0.41	0.14
Math				
3	0.63	0.20	0.52	0.10
4	0.58	0.21	0.51	0.07
5	0.53	0.16	0.52	0.09
6	0.49	0.22	0.50	0.09
7	0.29	0.16	0.49	0.09
8	0.37	0.19	0.47	0.10
Algebra I	0.44	0.17	0.47	0.10
Geometry	0.34	0.18	0.46	0.08
Algebra II	0.30	0.16	0.42	0.10

Exhibit 4.2.2 Number of Items Flagged For P-value, Biserial/Polyserial or DIF for Operational Test Items Administered Online

Grade	Number of Flagged Operational Items			
	Proportion Correct	Biserial/Polyserial Correlation	Biserial Correlation for Distractor	Differential Item Functioning
ELA				
3	0	0	0	2
4	0	0	0	0
5	0	0	0	1
6	0	0	1	0
7	0	1	1	2
8	0	1	0	1
9	0	1	1	0
10	1	0	0	2
11	0	0	1	1
Math				
3	0	0	0	2
4	0	0	0	0
5	1	0	0	0
6	0	0	0	1
7	0	0	0	1
8	0	0	1	0
Algebra I	0	0	0	1
Geometry	1	0	0	0
Algebra II	1	0	0	1

4.3 ITEM RESPONSE THEORY ANALYSIS

Calibration is the process by which the statistical relationship between item responses and the underlying measurement construct is estimated. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^n P(z_j|\theta),$$

where Z represents the vector of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function $P(Z)$. The one-parameter model (also known as the Rasch model) is used to calibrate dichotomously scored AzMERIT items and takes the form

$$P(x_j = 1|\theta_k, b_j) = \frac{1}{1 + e^{(\theta_k - b_j)}} = P_{j1}(\theta_k).$$

The b parameter is often called the *location* or *difficulty* parameter—the greater the value of b , the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using the Rasch-family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

$$P(x_j = i | \theta_k, b_{j0} \dots b_{jm_j-1}) = \frac{e^{\sum_{v=0}^i (\theta_k - b_{jv})}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^g (\theta_k - b_{jv})}}.$$

The tables in Appendix G provide Rasch and Masters' partial credit model item parameter estimates for the spring 2017 operational test items. Since AzMERIT is an online assessment system, bank item parameters were estimated based only on online responses to test items. Exhibit 4.3.1 presents the mean and standard deviation of the Rasch item parameters by item type for each test for items administered online. The selected-response (SR) items include traditional four-option multiple-choice (MC) items, technology-enhanced (TE) selected response items, which may require students to select one or more options, and machine-scored constructed response (MSCR) items, for which students' constructed responses are scored electronically using explicit rubrics. In addition, the average Rasch difficulty is presented for each scoring dimension of the writing prompt administered at each grade. As illustrated in Exhibit 4.3.1, selected-response items are, on average, less difficult than the constructed-response item types. Within the constructed response items, Evidence and Elaboration within the writing prompts was on average, consistently found to be the most difficult.

Exhibit 4.3.1 Rasch Summary Statistics by Item Type for Items Administered Online

Grade/ Course	SR			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
ELA									
3	41	0.09	0.94	0	-	-	1.62	1.71	-1.17
4	41	0.20	0.75	0	-	-	3.63	4.15	0.17
5	39	0.03	0.66	2	1.43	1.36	2.30	2.70	-1.09
6	41	0.08	1.12	0	-	-	2.24	2.89	-1.23
7	41	0.09	0.74	0	-	-	2.47	2.75	-1.55
8	41	0.17	1.06	0	-	-	1.07	1.26	-1.60
9	43	0.07	0.52	0	-	-	1.36	1.61	-1.63
10	43	0.10	0.90	0	-	-	0.78	1.09	-2.11
11	41	-0.05	0.98	2	1.45	0.27	0.23	0.68	-1.95
Mathematics									
3	14	-0.08	1.65	31	0.07	1.27	-	-	-
4	19	0.24	1.47	26	-0.08	1.24	-	-	-
5	14	0.03	0.83	31	0.02	0.97	-	-	-
6	16	-0.28	1.41	31	0.24	1.44	-	-	-
7	14	0.63	1.28	33	1.58	0.88	-	-	-
8	21	-0.49	1.08	26	0.54	1.18	-	-	-
Algebra I	29	-0.33	0.86	18	0.61	1.01	-	-	-
Geometry	26	-0.38	1.09	21	0.31	1.06	-	-	-
Algebra II	28	-0.53	0.97	19	0.66	0.72	-	-	-

Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). The rule of thumb is that items with good model-data-fit have Infit and Outfit within the range of 0.7-1.3. Exhibit 4.3.2 summarizes the number of online administered operational test items with Infit and Outfit statistics below, within, and above the range of .7 to 1.3.

Exhibit 4.3.2 Summary of Item Fit Statistics for Items Administered Online

Grade/ Course	Infit			Outfit		
	Below	Between	Above	Below	Between	Above
	0.7	.7 - 1.3	1.3	0.7	.7 - 1.3	1.3
ELA						
3	0	46	1	1	43	3
4	0	46	1	3	43	1
5	0	47	0	0	45	2
6	3	42	2	6	33	8
7	0	46	1	1	46	0
8	0	46	1	0	44	3
9	0	47	2	1	45	3
10	0	49	0	2	46	1
11	0	49	0	1	47	1
Mathematics						
3	1	42	2	2	34	9
4	1	43	1	4	35	6
5	0	44	1	5	34	6
6	2	43	2	8	30	9
7	0	46	1	13	26	8
8	0	44	3	7	31	9
Algebra I	0	47	0	2	38	7
Geometry	2	44	1	6	36	5
Algebra II	0	47	0	6	37	4

4.4 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibits 4.4.1 to 4.4.3. The AzMERIT bank scale was established based on the spring 2015 assessments in which the item calibrations were centered on items rather than persons, resulting in operational test forms with mean difficulty of zero and standard deviation of one. Because calibrations were not centered on persons, the standard deviation of ability estimates is not expected to be 30, as might be implied by the scaling transformation.

Exhibit 4.4.1 Test Score Summary Statistics – Combined Online and Paper

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Max.	Min.
ELA					
Grade 3	88352	2502	32.52	2605	2395
Grade 4	88680	2522	31.62	2610	2400
Grade 5	87527	2536	33.96	2629	2419
Grade 6	86113	2544	32.29	2641	2431
Grade 7	85248	2553	32.95	2648	2438
Grade 8	84511	2554	34.16	2658	2448
Grade 9	80753	2566	29.75	2664	2454
Grade 10	74156	2564	29.76	2668	2458
Grade 11	68526	2565	30.24	2675	2465
Math					
Grade 3	88481	3526	47.54	3605	3395
Grade 4	88961	3555	45.16	3645	3435
Grade 5	87636	3588	44.81	3688	3478
Grade 6	86335	3618	44.30	3722	3512
Grade 7	83302	3632	45.86	3739	3529
Grade 8	71886	3652	39.28	3776	3566
Algebra I	82960	3672	36.50	3787	3577
Geometry	71594	3684	37.81	3819	3609
Algebra II	63663	3698	33.31	3839	3629

Exhibit 4.4.2 Test Score Summary Statistics – Online

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Max.	Min.
ELA					
Grade 3	72754	2502	32.71	2605	2395
Grade 4	73195	2521	31.62	2610	2400
Grade 5	72289	2536	34.04	2629	2419
Grade 6	69837	2543	32.26	2641	2431
Grade 7	69754	2553	32.87	2648	2439
Grade 8	69481	2554	34.04	2658	2448
Grade 9	62956	2565	29.26	2664	2454
Grade 10	58181	2563	29.38	2668	2458
Grade 11	54018	2563	29.96	2675	2465
Math					
Grade 3	72859	3526	47.32	3605	3395
Grade 4	73441	3556	44.68	3645	3435
Grade 5	72429	3588	44.79	3688	3478
Grade 6	70035	3617	44.17	3722	3512
Grade 7	68367	3631	45.90	3739	3529
Grade 8	59176	3651	38.87	3776	3566
Algebra I	66697	3673	37.46	3787	3577
Geometry	56138	3683	38.07	3819	3609
Algebra II	50066	3697	32.99	3839	3629

Exhibit 4.4.3 Test Score Summary Statistics – Paper (Paper + DEI)

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Max.	Min.
ELA					
Grade 3	15602	2502	31.65	2605	2395
Grade 4	15489	2524	31.47	2610	2421
Grade 5	15241	2536	33.56	2629	2423
Grade 6	16279	2548	32.00	2641	2431
Grade 7	15495	2553	33.34	2648	2438
Grade 8	15033	2558	34.46	2658	2448
Grade 9	17800	2569	31.18	2664	2477
Grade 10	15978	2567	30.91	2668	2458
Grade 11	14511	2570	30.63	2674	2465
Math					
Grade 3	15625	3524	48.50	3605	3395
Grade 4	15524	3554	47.34	3645	3435
Grade 5	15211	3588	44.93	3688	3478
Grade 6	16303	3620	44.80	3722	3512
Grade 7	14936	3636	45.51	3739	3529
Grade 8	12711	3656	40.86	3776	3566
Algebra I	16266	3669	32.14	3787	3577
Geometry	15460	3687	36.74	3819	3609
Algebra II	13598	3704	33.79	3839	3629

The percentage of students in each performance level by grade and content area, as well as the percentage of students at or above Proficient are presented in Exhibits 4.4.4 to 4.4.6.

Exhibit 4.4.4 Percentage of Students in Performance Levels – Combined Online and Paper

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	88352	45	12	30	13	43
4	88680	35	16	35	13	48
5	87527	33	23	32	12	44
6	86113	35	24	37	5	41
7	85248	37	19	35	9	44
8	84511	46	21	25	9	34
9	80753	37	27	26	10	36
10	74156	52	17	22	9	31
11	68526	52	23	17	9	25
Math						
3	88481	24	29	28	19	47
4	88961	25	28	34	13	47
5	87636	27	26	32	15	47

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
6	86335	36	23	25	16	41
7	83302	48	18	16	17	34
8	71886	49	23	18	10	28
Algebra I	82960	40	22	28	11	39
Geometry	71594	40	26	27	7	34
Algebra II	63663	45	21	27	8	34

Exhibit 4.4.5 Percentage of Students in Performance Levels – Online

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	72754	44	12	30	14	43
4	73195	36	16	35	13	48
5	72289	33	23	32	12	44
6	69837	36	24	36	4	40
7	69754	37	19	35	9	44
8	69481	47	21	24	9	33
9	62956	38	27	26	9	35
10	58181	53	17	21	8	30
11	54018	53	22	16	8	24
Math						
3	72859	24	29	28	19	47
4	73441	25	29	34	13	47
5	72429	27	26	32	15	47
6	70035	36	23	25	16	41
7	68367	49	18	16	17	33
8	59176	50	23	18	10	27
Algebra I	66697	40	21	27	12	40
Geometry	56138	41	26	26	7	33
Algebra II	50066	47	20	26	7	32

Exhibit 4.4.6 Percentage of Students in Performance Levels – Paper (Paper + DEI)

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	15602	45	11	32	12	44
4	15489	30	17	38	14	53
5	15241	32	24	32	11	43
6	16279	29	24	40	7	47
7	15495	35	21	34	10	44
8	15033	40	22	28	11	38
9	17800	33	27	27	13	40
10	15978	46	18	25	11	36
11	14511	45	25	18	13	31
Math						
3	15625	26	28	28	18	46
4	15524	27	27	34	13	46
5	15211	26	26	33	15	48
6	16303	34	23	25	18	43
7	14936	45	19	17	19	36
8	12711	47	21	19	14	32
Algebra I	16266	39	25	30	6	37
Geometry	15460	35	28	30	7	37
Algebra II	13598	37	21	31	10	41

4.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibits 4.5.1 and 4.5.2 present the number and percentage, respectively, of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian Native, Hispanic/Latino, American Indian, White, and Multiple Ethnicities, and other demographic information such as special education (SPED), limited English proficiency (LEP), free reduced lunch, and accommodation.

Exhibit 4.5.1 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
ELA															
3	Minimally Proficient	39324	17699	21624	2598	520	111	22209	2999	9736	1023	7674	7795	19799	24767
	Partially Proficient	10666	5405	5260	549	239	44	5246	499	3723	357	692	697	4434	6102
	Proficient	26676	13906	12770	1151	915	119	10597	690	12227	968	1191	700	9125	14290
	Highly Proficient	11697	6382	5315	314	801	42	2846	121	7072	499	406	75	2301	6169
4	Minimally Proficient	31331	13678	17653	2179	375	106	18256	2530	7049	751	7710	6877	16083	20712
	Partially Proficient	14368	7130	7237	813	288	59	7224	736	4813	415	1135	1153	6381	8010
	Proficient	31309	16244	15065	1302	1075	113	12170	943	14525	1164	1423	833	10637	16489
	Highly Proficient	11678	6484	5194	311	750	41	2736	130	7205	503	336	54	2281	5984
5	Minimally Proficient	28761	12409	16352	2107	335	94	16615	2413	6471	614	7756	5792	14659	18613
	Partially Proficient	20548	10282	10266	1129	423	78	10127	1073	7099	606	1465	871	8809	11296
	Proficient	27984	14568	13416	1128	1063	125	10384	690	13601	974	1021	312	8797	14889
	Highly Proficient	10237	5732	4505	234	614	43	2328	108	6486	416	246	39	1932	5337
6	Minimally Proficient	29980	12725	17253	2122	353	87	16840	2518	7368	630	7804	4251	15030	19148
	Partially Proficient	20488	10456	10032	1177	385	65	9888	1025	7426	500	1236	623	8639	11067
	Proficient	31507	17047	14459	1225	1347	113	10704	764	16319	1014	883	295	9249	15904
	Highly Proficient	4145	2336	1809	91	436	20	765	38	2637	157	59	10	662	1920
7	Minimally Proficient	31304	13066	18233	2186	403	103	17626	2605	7687	601	7384	4226	15585	19320
	Partially Proficient	16346	8095	8248	953	356	54	7799	796	5975	395	926	451	6673	8580
	Proficient	29718	15757	13961	1195	1205	96	10578	711	15078	836	697	274	8872	15655
	Highly Proficient	7890	4530	3359	208	611	32	1718	85	4984	250	110	32	1366	4043
8	Minimally Proficient	38619	16473	22138	2621	533	133	20667	2791	11046	765	7490	3618	18144	23454
	Partially Proficient	17550	9233	8316	945	441	62	7630	608	7432	412	639	248	6498	9583
	Proficient	20713	11337	9375	848	865	51	7161	433	10803	519	429	147	5797	10991
	Highly Proficient	7643	4555	3087	211	608	23	1702	65	4793	221	94	32	1363	4092

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
9	Minimally Proficient	29805	12228	17576	2182	414	99	16148	2388	8003	521	5708	2399	12626	16562
	Partially Proficient	22046	11279	10767	1151	435	69	10299	1004	8621	449	1101	204	7307	11295
	Proficient	21046	11598	9447	937	847	70	7600	507	10555	513	445	55	5174	10429
	Highly Proficient	7867	4713	3154	205	747	20	1646	83	4968	189	81	6	1119	3509
10	Minimally Proficient	38288	17711	20576	2436	625	135	20408	2648	11333	649	5762	1380	15385	20678
	Partially Proficient	12961	6751	6209	638	372	64	5307	455	5839	273	448	42	3666	6572
	Proficient	16252	8690	7561	633	762	52	5423	311	8642	415	327	16	3664	7895
	Highly Proficient	6661	3777	2884	175	613	24	1433	56	4165	188	59	0	896	3139
11	Minimally Proficient	35358	15844	19510	2255	606	107	18325	2307	11122	567	5044	780	13193	19522
	Partially Proficient	15721	8659	7060	703	495	43	6480	549	7088	347	482	48	4370	7923
	Proficient	11453	6516	4937	463	532	31	3590	231	6326	268	201	12	2319	5631
	Highly Proficient	6007	3398	2609	171	536	13	1207	53	3793	172	53	1	778	2615
Math															
3	Minimally Proficient	21403	10297	11105	1657	191	61	12280	1908	4684	519	5589	5101	11255	13939
	Partially Proficient	25466	13003	12462	1435	401	86	13349	1473	7870	817	2383	2850	11575	14940
	Proficient	25061	12611	12450	1084	730	107	10680	735	10842	869	1429	1175	8929	13804
	Highly Proficient	16561	7470	9091	443	1157	61	4604	216	9428	644	677	243	3866	8769
4	Minimally Proficient	22338	10782	11555	1783	215	83	12832	1944	4859	537	6345	5119	11525	14451
	Partially Proficient	25083	12846	12236	1415	414	89	12918	1298	8167	752	2458	2665	11374	14623
	Proficient	29997	14828	15169	1149	983	105	11857	955	13850	1084	1497	1126	10175	16352
	Highly Proficient	11549	5184	6365	289	888	40	2896	153	6814	465	388	116	2393	5957
5	Minimally Proficient	23506	10906	12600	1942	176	70	13254	1963	5437	556	6846	4455	12097	15236
	Partially Proficient	22795	11807	10988	1280	342	87	11519	1276	7609	660	2071	1757	9946	13087
	Proficient	27956	14056	13900	1069	885	120	11247	851	12863	900	1271	732	9275	14944
	Highly Proficient	13383	6241	7142	313	1031	62	3510	201	7766	494	368	116	2904	6993

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
6	Minimally Proficient	30917	14809	16105	2393	288	87	17245	2567	7584	677	7672	4010	15570	19076
	Partially Proficient	19713	9961	9751	1092	354	52	9458	990	7220	528	1335	806	8268	10747
	Proficient	21477	10816	10660	808	680	95	8035	613	10577	651	746	363	6924	11273
	Highly Proficient	14237	7065	7172	352	1194	51	3566	200	8417	454	303	83	2945	7099
7	Minimally Proficient	40066	18736	21322	2937	467	128	22001	2986	10613	822	7919	4351	19328	24231
	Partially Proficient	15068	7707	7361	731	313	46	6680	639	6264	386	643	390	5795	8035
	Proficient	13636	6959	6676	514	450	62	5011	344	6857	388	314	185	4132	7162
	Highly Proficient	14542	7181	7361	333	1103	42	3650	189	8778	441	242	87	2890	7281
8	Minimally Proficient	35277	16477	18792	2646	359	108	19311	2489	9536	708	7247	3377	16882	21179
	Partially Proficient	16360	8389	7970	856	316	52	7314	701	6706	397	801	416	6522	9028
	Proficient	12805	6621	6183	568	390	48	4908	353	6216	306	330	176	4103	6799
	Highly Proficient	7455	3588	3867	235	523	19	2412	130	3950	170	118	66	1986	3848
Algebra I	Minimally Proficient	32798	14570	18208	2290	337	103	18269	2516	8637	554	6219	2315	14136	18397
	Partially Proficient	17920	9436	8479	967	332	54	8440	891	6821	392	952	284	6353	9299
	Proficient	23221	12309	10908	966	769	75	8574	641	11612	548	545	148	6069	12390
	Highly Proficient	9068	4364	4702	215	643	35	2309	86	5522	234	148	21	1675	5454
Geometry	Minimally Proficient	28518	13813	14699	2138	357	104	15787	2100	7475	482	4683	1476	11963	15522
	Partially Proficient	18710	9808	8899	1041	421	62	8741	938	7120	359	889	253	6439	9205
	Proficient	19200	9785	9414	618	825	70	6491	500	10275	408	408	78	4782	9380
	Highly Proficient	5188	2280	2908	99	537	15	1006	58	3338	127	93	16	709	2614
Algebra II	Minimally Proficient	28656	14120	14530	1809	392	84	15110	1849	8884	480	3233	730	11100	15835
	Partially Proficient	13164	7269	5895	647	298	44	5761	541	5581	278	441	106	3987	6731
	Proficient	17078	9044	8034	589	824	62	5739	418	9060	377	266	72	3635	8317
	Highly Proficient	4785	2035	2748	87	601	17	832	31	3099	114	55	7	555	2111

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch.

Exhibit 4.5.2 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information.

Grade	Performance Level	ELA													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	45	41	48	56	21	35	54	70	30	36	77	84	56	48
	Partially Proficient	12	12	12	12	10	14	13	12	11	13	7	8	12	12
	Proficient	30	32	28	25	37	38	26	16	37	34	12	8	26	28
	Highly Proficient	13	15	12	7	32	13	7	3	22	18	4	1	6	12
4	Minimally Proficient	35	31	39	47	15	33	45	58	21	27	73	77	45	40
	Partially Proficient	16	16	16	18	12	18	18	17	14	15	11	13	18	16
	Proficient	35	37	33	28	43	35	30	22	43	41	13	9	30	32
	Highly Proficient	13	15	12	7	30	13	7	3	21	18	3	1	6	12
5	Minimally Proficient	33	29	37	46	14	28	42	56	19	24	74	83	43	37
	Partially Proficient	23	24	23	25	17	23	26	25	21	23	14	12	26	23
	Proficient	32	34	30	25	44	37	26	16	40	37	10	4	26	30
	Highly Proficient	12	13	10	5	25	13	6	3	19	16	2	1	6	11
6	Minimally Proficient	35	30	40	46	14	31	44	58	22	27	78	82	45	40
	Partially Proficient	24	25	23	26	15	23	26	24	22	22	12	12	26	23
	Proficient	37	40	33	27	53	40	28	18	48	44	9	6	28	33
	Highly Proficient	5	5	4	2	17	7	2	1	8	7	1	0	2	4
7	Minimally Proficient	37	32	42	48	16	36	47	62	23	29	81	85	48	41
	Partially Proficient	19	20	19	21	14	19	21	19	18	19	10	9	21	18
	Proficient	35	38	32	26	47	34	28	17	45	40	8	5	27	33
	Highly Proficient	9	11	8	5	24	11	5	2	15	12	1	1	4	8
8	Minimally Proficient	46	40	52	57	22	49	56	72	32	40	87	89	57	49
	Partially Proficient	21	22	19	20	18	23	21	16	22	21	7	6	20	20
	Proficient	25	27	22	18	35	19	19	11	32	27	5	4	18	23
	Highly Proficient	9	11	7	5	25	9	5	2	14	12	1	1	4	9
9	Minimally Proficient	37	31	43	49	17	38	45	60	25	31	78	90	48	40
	Partially Proficient	27	28	26	26	18	27	29	25	27	27	15	8	28	27
	Proficient	26	29	23	21	35	27	21	13	33	31	6	2	20	25
	Highly Proficient	10	12	8	5	31	8	5	2	15	11	1	0	4	8

Grade	Performance Level	Multiple Ethnicities													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
10	Minimally Proficient	52	48	55	63	26	49	63	76	38	43	87	96	65	54
	Partially Proficient	17	18	17	16	16	23	16	13	19	18	7	3	16	17
	Proficient	22	24	20	16	32	19	17	9	29	27	5	1	16	21
	Highly Proficient	9	10	8	5	26	9	4	2	14	12	1	0	4	8
11	Minimally Proficient	52	46	57	63	28	55	62	73	39	42	87	93	64	55
	Partially Proficient	23	25	21	20	23	22	22	17	25	26	8	6	21	22
	Proficient	17	19	14	13	25	16	12	7	22	20	3	1	11	16
	Highly Proficient	9	10	8	5	25	7	4	2	13	13	1	0	4	7
Math															
3	Minimally Proficient	24	24	25	36	8	19	30	44	14	18	55	54	32	27
	Partially Proficient	29	30	28	31	16	27	33	34	24	29	24	30	32	29
	Proficient	28	29	28	23	29	34	26	17	33	31	14	13	25	27
	Highly Proficient	19	17	20	10	47	19	11	5	29	23	7	3	11	17
4	Minimally Proficient	25	25	25	38	9	26	32	45	14	19	59	57	32	28
	Partially Proficient	28	29	27	31	17	28	32	30	24	26	23	30	32	28
	Proficient	34	34	33	25	39	33	29	22	41	38	14	12	29	32
	Highly Proficient	13	12	14	6	36	13	7	4	20	16	4	1	7	12
5	Minimally Proficient	27	25	28	42	7	21	34	46	16	21	65	63	35	30
	Partially Proficient	26	27	25	28	14	26	29	30	23	25	20	25	29	26
	Proficient	32	33	31	23	36	35	28	20	38	34	12	10	27	30
	Highly Proficient	15	15	16	7	42	18	9	5	23	19	3	2	8	14
6	Minimally Proficient	36	35	37	52	11	31	45	59	22	29	76	76	46	40
	Partially Proficient	23	23	22	24	14	18	25	23	21	23	13	15	25	22
	Proficient	25	25	24	17	27	33	21	14	31	28	7	7	21	23
	Highly Proficient	16	17	16	8	47	18	9	5	25	20	3	2	9	15
7	Minimally Proficient	48	46	50	65	20	46	59	72	33	40	87	87	60	52
	Partially Proficient	18	19	17	16	13	17	18	15	19	19	7	8	18	17
	Proficient	16	17	16	11	19	22	13	8	21	19	3	4	13	15
	Highly Proficient	17	18	17	7	47	15	10	5	27	22	3	2	9	16
8	Minimally Proficient	49	47	51	61	23	48	57	68	36	45	85	84	57	52
	Partially Proficient	23	24	22	20	20	23	22	19	25	25	9	10	22	22
	Proficient	18	19	17	13	25	21	14	10	24	19	4	4	14	17
	Highly Proficient	10	10	11	5	33	8	7	4	15	11	1	2	7	9

Grade	Performance Level														
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
Algebra I	Minimally Proficient	40	36	43	52	16	39	49	61	27	32	79	84	50	40
	Partially Proficient	22	23	20	22	16	20	22	22	21	23	12	10	23	20
	Proficient	28	30	26	22	37	28	23	16	36	32	7	5	22	27
	Highly Proficient	11	11	11	5	31	13	6	2	17	14	2	1	6	12
Geometry	Minimally Proficient	40	39	41	55	17	41	49	58	26	35	77	81	50	42
	Partially Proficient	26	27	25	27	20	25	27	26	25	26	15	14	27	25
	Proficient	27	27	26	16	39	28	20	14	36	30	7	4	20	26
	Highly Proficient	7	6	8	3	25	6	3	2	12	9	2	1	3	7
Algebra II	Minimally Proficient	45	43	47	58	19	41	55	65	33	38	81	80	58	48
	Partially Proficient	21	22	19	21	14	21	21	19	21	22	11	12	21	20
	Proficient	27	28	26	19	39	30	21	15	34	30	7	8	19	25
	Highly Proficient	8	6	9	3	28	8	3	1	12	9	1	1	3	6

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch.

4.6 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance-level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. Evidence of the reliability of AzMERIT ELA and math scores are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance-level classifications.²⁵

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability, such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.²⁶

²⁵ Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

²⁶ Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for

4.6.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently-constructed test form. Exhibit 4.6.1.1 shows the Cronbach's alpha internal consistency estimates for each of the spring 2017 AzMERIT ELA and math assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 4.6.1.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade	ELA		MATH	
	Reliability	Variance	Reliability	Variance
3	0.89	1070	0.92	2239
4	0.90	1000	0.92	1996
5	0.90	1159	0.93	2006
6	0.89	1041	0.93	1951
7	0.90	1080	0.90	2106
8	0.91	1159	0.91	1511
9 ELA / Algebra I	0.89	856	0.91	1403
10 ELA / Geometry	0.89	863	0.90	1449
11 ELA / Algebra II	0.88	898	0.87	1088

Note: Reliability ranges from 0 to 1. The variance is in scale score metric.

4.6.2 STANDARD ERROR OF MEASUREMENT

Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low- and high-performing students, the precision of test scores decreases near the tails of the ability distribution.

Exhibit 4.6.2.1 and Exhibit 4.6.2.2 present the standard errors of measurement for the AzMERIT ELA and math assessments with respect to the four AzMERIT performance standard cuts. As the tables indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Partially Proficient and Proficient performance standard cuts.²⁷ Test scores near the tails of the ability distribution are somewhat less precise, as expected. While these numbers indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance level classifications.

reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

²⁷ Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 4.6.2.1 Standard Errors of Measurement at Performance Level Cuts Spring 2017 – ELA

Grade	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
3	10.6	10	10.09	12.2	10.61
4	10.02	9	9.74	12.9	10.2
5	10.35	9.23	10.46	13.98	10.63
6	10.58	9.87	10.4	13.29	10.5
7	10.41	9.5	10.52	13.9	10.65
8	10.41	9.19	10.06	11.98	10.24
9	9.59	9	9.35	12.31	9.67
10	10	9	9.97	11.53	9.98
11	10.47	10	10.03	11.82	10.42

Exhibit 4.6.2.2 Standard Errors of Measurement at Performance Level Cuts Spring 2017 – Math

Grade	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
3	12.28	11	12.3	17.43	13.08
4	12.47	10.26	11.68	16.71	12.29
5	13.45	10	10.56	15.8	12.18
6	12.61	10.3	10.65	13.74	11.85
7	17.79	10.72	10	10.57	14.45
8	12.86	9.96	9.9	12.83	11.76
Algebra I	11.45	10	10.07	13.05	10.97
Geometry	14	10.5	10	12.21	12.02
Algebra II	13.44	10.46	10	11.04	11.83

4.6.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed to estimate the likelihood of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).²⁸ This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications is typically estimated on the scores from a single test administration using the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for classification accuracy and classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form actually taken and

²⁸ Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

the classifications that would be made on the basis of an alternate, equivalently-constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

4.6.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the probability of consistent classification directly using the likelihood function. The likelihood function of θ given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated ability (theta) is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as *below* the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as *below* the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 4.6.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications are different from students' observed scores (e.g., a student whose true score results in a classification as Proficient, but whose observed score results in an incorrect classification as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 4.6.4.1 Classification Accuracy

		Classification on the Form Actually Taken	
		Above the Cut Score	Below the Cut Score
Classification on True Score	At or Above the Cut Score	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

4.6.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 4.6.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 4.6.5.1 Classification Consistency

Classification on the First Form Taken		Classification on the Second Form Taken	
		Above the Cut Score	Below the Cut Score
		N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	At or Above the Cut Score		
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

4.6.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

Exhibit 4.6.6.1 presents the classification accuracy and consistency indexes for spring 2017 administration of the AzMERIT. Exhibit 4.6.6.2 presents the classification accuracy and consistency indexes for each of the identified subgroups: gender (female and male), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with special education, free or reduced lunch, and accommodations). Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency index assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score plus the true score, which does not include measurement error.

Exhibit 4.6.6.1 Classification Accuracy and Consistency Estimates for Performance Standards Overall

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
3	0.91	0.91	0.95	0.88	0.88	0.93
4	0.92	0.91	0.95	0.88	0.88	0.93
5	0.92	0.92	0.95	0.89	0.89	0.93
6	0.92	0.91	0.97	0.89	0.87	0.96
7	0.92	0.91	0.95	0.89	0.87	0.94
8	0.92	0.93	0.96	0.89	0.90	0.95
9	0.92	0.92	0.96	0.88	0.88	0.94
10	0.91	0.92	0.96	0.87	0.88	0.94
11	0.90	0.92	0.96	0.86	0.89	0.94
MATH						
3	0.95	0.93	0.94	0.93	0.90	0.91
4	0.94	0.92	0.95	0.91	0.89	0.93
5	0.94	0.93	0.95	0.92	0.90	0.93
6	0.93	0.93	0.95	0.90	0.90	0.93
7	0.92	0.94	0.96	0.89	0.92	0.94
8	0.91	0.94	0.97	0.88	0.92	0.96
Algebra I	0.92	0.93	0.96	0.89	0.90	0.94
Geometry	0.90	0.93	0.97	0.86	0.90	0.96
Algebra II	0.89	0.92	0.98	0.84	0.89	0.96

Exhibit 4.6.6.2 Classification Accuracy and Consistency Estimates for Performance Standards across Subgroups

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA							
G3E	Overall	0.91	0.91	0.95	0.88	0.88	0.93
	Female	0.91	0.91	0.94	0.88	0.87	0.92
	Male	0.91	0.92	0.95	0.88	0.88	0.93
	African American	0.91	0.92	0.97	0.87	0.89	0.95
	Hispanic/Latino	0.91	0.91	0.96	0.87	0.88	0.95
	Asian	0.94	0.92	0.91	0.91	0.89	0.87
	White	0.92	0.91	0.92	0.89	0.87	0.89
	Hawaiian/Pacific	0.91	0.91	0.94	0.87	0.87	0.92
	American Indian	0.90	0.93	0.98	0.86	0.90	0.97
	Multiple Ethnicities	0.91	0.91	0.94	0.88	0.87	0.91
	LEP	0.93	0.96	0.99	0.90	0.94	0.99
	SPED	0.94	0.96	0.98	0.91	0.94	0.98
	FRL	0.91	0.91	0.97	0.87	0.88	0.95
	Accommodations	0.91	0.92	0.95	0.88	0.88	0.93
G4E	Overall	0.92	0.91	0.95	0.88	0.88	0.93
	Female	0.92	0.91	0.94	0.89	0.88	0.92
	Male	0.92	0.92	0.95	0.88	0.88	0.93
	African American	0.91	0.92	0.97	0.87	0.88	0.95
	Hispanic/ Latino	0.90	0.91	0.97	0.87	0.88	0.95
	Asian	0.95	0.93	0.91	0.93	0.90	0.87
	White	0.93	0.91	0.92	0.91	0.88	0.89
	Hawaiian/Pacific	0.91	0.90	0.95	0.88	0.86	0.93
	American Indian	0.90	0.92	0.98	0.86	0.89	0.97
	Multiple Ethnicities	0.93	0.91	0.94	0.90	0.88	0.91
	LEP	0.91	0.95	0.99	0.88	0.93	0.99
	SPED	0.93	0.96	0.98	0.91	0.94	0.98
	FRL	0.91	0.91	0.97	0.87	0.88	0.95
	Accommodations	0.92	0.92	0.95	0.88	0.88	0.93
G5E	Overall	0.92	0.92	0.95	0.89	0.89	0.93
	Female	0.92	0.91	0.94	0.89	0.88	0.92
	Male	0.92	0.92	0.95	0.89	0.89	0.93
	African American	0.91	0.92	0.97	0.87	0.90	0.96
	Hispanic/ Latino	0.91	0.92	0.97	0.87	0.89	0.95
	Asian	0.95	0.91	0.92	0.93	0.88	0.88
	White	0.94	0.91	0.92	0.91	0.88	0.89
	Hawaiian/Pacific	0.92	0.91	0.94	0.89	0.88	0.92
	American Indian	0.89	0.94	0.98	0.86	0.91	0.98
	Multiple Ethnicities	0.93	0.91	0.93	0.90	0.87	0.91
	LEP	0.93	0.98	1.00	0.90	0.97	0.99
	SPED	0.93	0.97	0.99	0.91	0.95	0.98
	FRL	0.91	0.92	0.97	0.87	0.89	0.95
	Accommodations	0.92	0.92	0.95	0.89	0.89	0.93
G6E	Overall	0.92	0.91	0.97	0.89	0.87	0.96
	Female	0.92	0.90	0.97	0.89	0.86	0.95
	Male	0.92	0.92	0.97	0.89	0.88	0.96
	African American	0.91	0.92	0.98	0.87	0.88	0.98
	Hispanic/ Latino	0.91	0.91	0.98	0.88	0.88	0.98
	Asian	0.95	0.92	0.92	0.93	0.89	0.89
	White	0.94	0.90	0.95	0.91	0.86	0.93
	Hawaiian/Pacific	0.94	0.90	0.96	0.92	0.85	0.95
	American Indian	0.91	0.93	0.99	0.87	0.91	0.99
	Multiple Ethnicities	0.93	0.90	0.96	0.90	0.86	0.94

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	LEP	0.94	0.97	1.00	0.91	0.96	1.00
	SPED	0.94	0.97	0.99	0.91	0.95	0.99
	FRL	0.91	0.92	0.98	0.87	0.88	0.98
	Accommodations	0.92	0.92	0.97	0.89	0.88	0.96
G7E	Overall	0.92	0.91	0.95	0.89	0.87	0.94
	Female	0.92	0.90	0.95	0.89	0.87	0.93
	Male	0.92	0.91	0.96	0.89	0.88	0.95
	African American	0.91	0.91	0.97	0.87	0.88	0.96
	Hispanic/Latino	0.91	0.91	0.97	0.87	0.88	0.96
	Asian	0.95	0.91	0.91	0.93	0.87	0.88
	White	0.93	0.90	0.93	0.91	0.86	0.91
	Hawaiian/Pacific	0.91	0.92	0.95	0.88	0.89	0.93
	American Indian	0.90	0.93	0.99	0.86	0.91	0.98
	Multiple Ethnicities	0.92	0.90	0.94	0.89	0.87	0.92
	LEP	0.94	0.97	1.00	0.92	0.96	0.99
	SPED	0.94	0.97	0.99	0.91	0.96	0.99
	FRL	0.91	0.91	0.97	0.87	0.88	0.97
	Accommodations	0.92	0.91	0.96	0.89	0.88	0.94
G8E	Overall	0.92	0.93	0.96	0.89	0.90	0.95
	Female	0.92	0.92	0.95	0.89	0.89	0.94
	Male	0.93	0.93	0.97	0.90	0.91	0.95
	African American	0.92	0.94	0.97	0.89	0.92	0.97
	Hispanic/ Latino	0.92	0.94	0.98	0.89	0.91	0.97
	Asian	0.94	0.92	0.93	0.91	0.89	0.90
	White	0.92	0.91	0.94	0.89	0.88	0.92
	Hawaiian/Pacific	0.91	0.92	0.97	0.88	0.89	0.96
	American Indian	0.92	0.96	0.99	0.89	0.94	0.98
	Multiple Ethnicities	0.92	0.92	0.95	0.89	0.90	0.93
	LEP	0.97	0.98	1.00	0.95	0.98	0.99
	SPED	0.96	0.98	0.99	0.94	0.97	0.99
	FRL	0.92	0.94	0.98	0.88	0.91	0.97
	Accommodations	0.92	0.93	0.96	0.89	0.90	0.95
G9E	Overall	0.92	0.92	0.96	0.88	0.88	0.94
	Female	0.92	0.91	0.95	0.89	0.87	0.93
	Male	0.92	0.92	0.96	0.88	0.89	0.95
	African American	0.91	0.93	0.97	0.87	0.90	0.96
	Hispanic/Latino	0.91	0.92	0.97	0.87	0.89	0.96
	Asian	0.95	0.92	0.91	0.93	0.88	0.88
	White	0.93	0.91	0.94	0.90	0.87	0.91
	Hawaiian/Pacific	0.93	0.92	0.97	0.90	0.88	0.95
	American Indian	0.90	0.94	0.99	0.86	0.92	0.98
	Multiple Ethnicities	0.91	0.91	0.95	0.88	0.88	0.93
	LEP	0.96	0.99	1.00	0.93	0.98	1.00
	SPED	0.93	0.97	0.99	0.90	0.96	0.99
	FRL	0.91	0.92	0.97	0.87	0.89	0.96
	Accommodations	0.92	0.92	0.96	0.88	0.88	0.94
G10E	Overall	0.91	0.92	0.96	0.87	0.88	0.94
	Female	0.90	0.91	0.95	0.87	0.88	0.93
	Male	0.91	0.92	0.96	0.88	0.89	0.95
	African American	0.92	0.93	0.97	0.88	0.90	0.96
	Hispanic/ Latino	0.91	0.93	0.97	0.87	0.90	0.96
	Asian	0.92	0.91	0.91	0.89	0.87	0.87
	White	0.90	0.90	0.94	0.87	0.86	0.91
	Hawaiian/Pacific	0.87	0.92	0.96	0.83	0.88	0.94
	American Indian	0.92	0.95	0.99	0.89	0.93	0.98
	Multiple Ethnicities	0.90	0.90	0.94	0.87	0.87	0.92

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G11E	LEP	0.98	0.99	1.00	0.97	0.98	1.00
	SPED	0.96	0.98	0.99	0.94	0.97	0.99
	FRL	0.91	0.93	0.97	0.88	0.90	0.96
	Accommodations	0.91	0.92	0.96	0.87	0.89	0.94
	Overall	0.90	0.92	0.96	0.86	0.89	0.94
	Female	0.89	0.91	0.95	0.85	0.87	0.94
	Male	0.91	0.93	0.97	0.87	0.90	0.95
	African American	0.91	0.94	0.97	0.88	0.91	0.96
	Hispanic/ Latino	0.90	0.93	0.98	0.86	0.90	0.97
	Asian	0.91	0.90	0.93	0.88	0.85	0.90
	White	0.90	0.90	0.94	0.86	0.86	0.92
	Hawaiian/Pacific	0.91	0.92	0.96	0.87	0.88	0.95
	American Indian	0.91	0.95	0.99	0.87	0.93	0.98
	Multiple Ethnicities	0.90	0.91	0.94	0.86	0.87	0.92
	LEP	0.97	0.99	1.00	0.95	0.98	1.00
	SPED	0.95	0.98	0.99	0.94	0.97	0.99
	FRL	0.90	0.93	0.98	0.87	0.91	0.97
	Accommodations	0.90	0.92	0.96	0.87	0.89	0.95

MATH

G3M	Overall	0.95	0.93	0.94	0.93	0.90	0.91
	Female	0.95	0.92	0.94	0.92	0.89	0.92
	Male	0.95	0.93	0.94	0.93	0.90	0.91
	African American	0.94	0.93	0.96	0.91	0.90	0.95
	Hispanic/Latino	0.94	0.92	0.96	0.91	0.89	0.94
	Asian	0.98	0.94	0.90	0.97	0.92	0.85
	White	0.96	0.93	0.91	0.95	0.90	0.88
	Hawaiian/Pacific	0.95	0.92	0.94	0.94	0.89	0.91
	American Indian	0.92	0.93	0.97	0.89	0.91	0.96
	Multiple Ethnicities	0.95	0.93	0.93	0.93	0.89	0.90
	LEP	0.92	0.95	0.99	0.89	0.93	0.98
	SPED	0.94	0.96	0.98	0.92	0.94	0.97
	FRL	0.94	0.92	0.96	0.91	0.89	0.94
G4M	Accommodations	0.95	0.93	0.94	0.92	0.90	0.92
	Overall	0.94	0.92	0.95	0.91	0.89	0.93
	Female	0.93	0.92	0.95	0.91	0.88	0.93
	Male	0.94	0.92	0.95	0.92	0.89	0.93
	African American	0.92	0.93	0.97	0.89	0.90	0.96
	Hispanic/Latino	0.92	0.92	0.97	0.90	0.89	0.95
	Asian	0.97	0.94	0.91	0.96	0.91	0.87
	White	0.95	0.92	0.93	0.94	0.88	0.90
	Hawaiian/Pacific	0.93	0.92	0.95	0.90	0.89	0.92
	American Indian	0.91	0.93	0.98	0.88	0.91	0.97
	Multiple Ethnicities	0.94	0.92	0.94	0.92	0.89	0.91
	LEP	0.91	0.95	0.99	0.88	0.93	0.99
	SPED	0.93	0.96	0.98	0.91	0.94	0.98
	FRL	0.92	0.92	0.97	0.90	0.89	0.95
G5M	Accommodations	0.94	0.92	0.95	0.91	0.89	0.93
	Overall	0.94	0.93	0.95	0.92	0.90	0.93
	Female	0.94	0.93	0.95	0.92	0.90	0.93
	Male	0.94	0.93	0.95	0.92	0.91	0.92
	African American	0.93	0.94	0.97	0.90	0.91	0.96
	Hispanic/Latino	0.93	0.93	0.96	0.90	0.90	0.95
	Asian	0.97	0.94	0.90	0.96	0.92	0.86
	White	0.96	0.93	0.93	0.94	0.90	0.89
	Hawaiian/Pacific	0.95	0.93	0.94	0.92	0.90	0.92

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	American Indian	0.92	0.93	0.98	0.89	0.91	0.97
	Multiple Ethnicities	0.95	0.92	0.94	0.93	0.89	0.91
	LEP	0.92	0.96	0.99	0.88	0.95	0.99
	SPED	0.94	0.96	0.99	0.91	0.95	0.98
	FRL	0.93	0.93	0.96	0.90	0.90	0.95
	Accommodations	0.94	0.93	0.95	0.92	0.91	0.93
	Overall	0.93	0.93	0.95	0.90	0.90	0.93
	Female	0.93	0.93	0.95	0.90	0.90	0.93
	Male	0.93	0.93	0.95	0.90	0.90	0.93
	African American	0.92	0.94	0.97	0.89	0.91	0.96
G6M	Hispanic/Latino	0.92	0.93	0.97	0.88	0.90	0.95
	Asian	0.96	0.94	0.93	0.94	0.91	0.90
	White	0.94	0.92	0.93	0.92	0.89	0.91
	Hawaiian/Pacific	0.94	0.92	0.93	0.92	0.89	0.90
	American Indian	0.92	0.94	0.98	0.88	0.92	0.97
	Multiple Ethnicities	0.93	0.92	0.94	0.90	0.89	0.92
	LEP	0.93	0.97	0.99	0.90	0.95	0.99
	SPED	0.94	0.97	0.99	0.92	0.96	0.98
	FRL	0.92	0.93	0.97	0.88	0.90	0.96
	Accommodations	0.93	0.93	0.96	0.90	0.90	0.94
G7M	Overall	0.92	0.94	0.96	0.89	0.92	0.94
	Female	0.92	0.94	0.96	0.89	0.91	0.94
	Male	0.93	0.94	0.96	0.90	0.92	0.94
	African American	0.92	0.96	0.98	0.89	0.94	0.97
	Hispanic/Latino	0.92	0.95	0.97	0.88	0.92	0.96
	Asian	0.95	0.94	0.94	0.93	0.92	0.91
	White	0.93	0.93	0.94	0.90	0.90	0.92
	Hawaiian/Pacific	0.93	0.93	0.96	0.90	0.90	0.94
	American Indian	0.92	0.96	0.98	0.88	0.94	0.97
	Multiple Ethnicities	0.93	0.93	0.95	0.90	0.90	0.94
G8M	LEP	0.95	0.98	0.99	0.92	0.97	0.99
	SPED	0.96	0.98	0.99	0.93	0.97	0.99
	FRL	0.92	0.95	0.97	0.88	0.92	0.96
	Accommodations	0.93	0.94	0.96	0.89	0.92	0.95
	Overall	0.91	0.94	0.97	0.88	0.92	0.96
	Female	0.91	0.94	0.97	0.88	0.91	0.96
	Male	0.92	0.95	0.97	0.89	0.92	0.96
	African American	0.92	0.95	0.98	0.89	0.93	0.97
	Hispanic/Latino	0.91	0.95	0.98	0.88	0.93	0.97
	Asian	0.93	0.93	0.95	0.91	0.90	0.92
Algebra I	White	0.91	0.93	0.96	0.88	0.90	0.94
	Hawaiian/Pacific	0.92	0.94	0.98	0.89	0.91	0.97
	American Indian	0.91	0.96	0.99	0.88	0.95	0.98
	Multiple Ethnicities	0.91	0.93	0.97	0.88	0.91	0.96
	LEP	0.94	0.98	0.99	0.92	0.97	0.99
	SPED	0.95	0.98	1.00	0.93	0.97	0.99
	FRL	0.91	0.95	0.98	0.88	0.93	0.97
	Accommodations	0.92	0.95	0.97	0.89	0.92	0.96
	Overall	0.92	0.93	0.96	0.89	0.90	0.94
	Female	0.91	0.93	0.96	0.88	0.90	0.94
	Male	0.92	0.94	0.96	0.89	0.91	0.95
	African American	0.91	0.94	0.97	0.88	0.92	0.96
	Hispanic/Latino	0.91	0.94	0.97	0.88	0.91	0.96
	Asian	0.95	0.94	0.93	0.94	0.91	0.89
	White	0.93	0.93	0.94	0.90	0.89	0.92
	Hawaiian/Pacific	0.92	0.94	0.96	0.89	0.91	0.94
	Algebra I						

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	American Indian	0.90	0.94	0.99	0.87	0.92	0.98
	Multiple Ethnicities	0.91	0.93	0.95	0.88	0.90	0.93
	LEP	0.94	0.98	1.00	0.91	0.97	0.99
	SPED	0.94	0.98	0.99	0.91	0.96	0.99
	FRL	0.91	0.93	0.97	0.87	0.91	0.96
	Accommodations	0.92	0.93	0.96	0.89	0.91	0.94
	Overall	0.90	0.93	0.97	0.86	0.90	0.96
	Female	0.90	0.93	0.97	0.86	0.90	0.96
	Male	0.91	0.94	0.97	0.87	0.91	0.96
	African American	0.89	0.94	0.99	0.85	0.92	0.98
	Hispanic/Latino	0.89	0.94	0.99	0.85	0.91	0.98
Geometry	Asian	0.94	0.93	0.94	0.92	0.90	0.92
	White	0.92	0.92	0.96	0.89	0.89	0.94
	Hawaiian/Pacific	0.90	0.94	0.97	0.86	0.91	0.96
	American Indian	0.88	0.94	0.99	0.84	0.92	0.99
	Multiple Ethnicities	0.91	0.93	0.97	0.87	0.90	0.96
	LEP	0.91	0.98	1.00	0.88	0.97	0.99
	SPED	0.91	0.97	1.00	0.88	0.96	0.99
	FRL	0.89	0.94	0.99	0.85	0.91	0.98
	Accommodations	0.90	0.93	0.97	0.87	0.90	0.96
	Overall	0.89	0.92	0.98	0.84	0.89	0.96
	Female	0.88	0.92	0.98	0.84	0.89	0.97
Algebra II	Male	0.89	0.93	0.97	0.85	0.90	0.96
	African American	0.88	0.94	0.99	0.83	0.91	0.98
	Hispanic/Latino	0.88	0.93	0.99	0.82	0.90	0.98
	Asian	0.94	0.92	0.93	0.91	0.89	0.91
	White	0.90	0.91	0.96	0.86	0.88	0.95
	Hawaiian/Pacific	0.90	0.92	0.97	0.86	0.88	0.96
	American Indian	0.87	0.94	0.99	0.82	0.92	0.99
	Multiple Ethnicities	0.90	0.91	0.97	0.86	0.88	0.96
	LEP	0.89	0.96	0.99	0.84	0.95	0.99
	SPED	0.90	0.97	1.00	0.85	0.96	0.99
	FRL	0.88	0.93	0.99	0.82	0.90	0.98
	Accommodations	0.89	0.92	0.98	0.84	0.89	0.97

Note: Hawaiian/Pacific = Native Hawaiian/Pacific Islander; American Indian = American Indian or Alaskan; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free/Reduced Lunch

4.6.7 RELIABILITY FOR SUBGROUPS IN THE POPULATION

Exhibits 4.6.7.1 and 4.6.7.2 show the mean reliability for each of the identified subgroups: gender (female and male), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with IEPs [Special Education]²⁹, free or reduced lunch, and accommodations). As the exhibits indicate, internal consistency reliabilities are generally stable across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups, and that test scores are similarly precise across demographic subgroups. For subgroups where the reliability coefficients are attenuated, there is a corresponding decrease in the subgroup variance relative to the overall student population, indicating that attenuation of reliability in subgroups is due to a restriction of range.

²⁹ Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Exhibit 4.6.7.1 Internal Consistency Reliability by Subgroup – ELA

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
ELA															
3	Reliability	0.89	0.89	0.90	0.88	0.89	0.89	0.88	0.84	0.89	0.89	0.87	0.79	0.88	0.90
	Variance	1070	1041	1083	971	1152	1010	895	677	1047	1039	910	566	884	1087
4	Reliability	0.90	0.89	0.90	0.88	0.89	0.89	0.88	0.85	0.89	0.89	0.86	0.79	0.88	0.90
	Variance	1000	982	999	859	1086	956	821	660	974	1012	767	470	812	1017
5	Reliability	0.90	0.90	0.90	0.89	0.89	0.89	0.89	0.87	0.89	0.89	0.86	0.78	0.89	0.90
	Variance	1159	1146	1151	994	1188	981	967	784	1113	1107	836	501	959	1192
6	Reliability	0.89	0.89	0.90	0.88	0.89	0.89	0.88	0.86	0.88	0.89	0.85	0.81	0.88	0.90
	Variance	1041	971	1083	934	1105	1006	916	801	948	1000	780	613	910	1082
7	Reliability	0.90	0.89	0.90	0.89	0.89	0.89	0.89	0.87	0.88	0.89	0.85	0.82	0.88	0.90
	Variance	1080	1018	1106	997	1120	1035	953	808	973	1022	760	664	938	1130
8	Reliability	0.91	0.91	0.91	0.90	0.91	0.90	0.90	0.87	0.91	0.91	0.84	0.83	0.90	0.91
	Variance	1159	1093	1172	1072	1266	1072	1017	792	1101	1107	721	696	1000	1212
9	Reliability	0.89	0.89	0.89	0.89	0.90	0.90	0.88	0.85	0.89	0.89	0.83	0.74	0.88	0.89
	Variance	856	820	854	801	1018	884	726	601	836	809	551	400	727	874
10	Reliability	0.89	0.88	0.89	0.87	0.88	0.88	0.87	0.83	0.88	0.89	0.81	0.71	0.87	0.89
	Variance	863	807	898	792	854	775	759	573	812	875	585	413	754	880
11	Reliability	0.88	0.87	0.89	0.87	0.88	0.87	0.86	0.83	0.88	0.88	0.82	0.76	0.86	0.88
	Variance	898	802	959	844	937	848	778	633	886	929	678	526	768	908

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch.

Exhibit 4.6.7.2 Internal Consistency Reliability by Subgroup – Math

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
Mathematics															
3	Reliability	0.92	0.92	0.93	0.93	0.88	0.92	0.92	0.92	0.90	0.91	0.94	0.92	0.92	0.93
	Variance	2239	2092	2380	2289	1840	2316	2035	1780	1989	2065	2630	1738	2065	2340
4	Reliability	0.92	0.92	0.93	0.93	0.90	0.92	0.92	0.92	0.91	0.92	0.93	0.91	0.92	0.93
	Variance	1996	1835	2151	1954	1886	1884	1791	1640	1789	1918	2220	1496	1779	2086
5	Reliability	0.93	0.92	0.93	0.92	0.90	0.92	0.92	0.91	0.92	0.92	0.90	0.88	0.92	0.93
	Variance	2006	1853	2152	1936	1875	1869	1782	1578	1837	1875	1991	1397	1798	2097
6	Reliability	0.93	0.93	0.93	0.92	0.92	0.93	0.92	0.90	0.92	0.93	0.89	0.87	0.92	0.93
	Variance	1951	1859	2038	1763	2114	1956	1668	1504	1800	1850	1644	1303	1689	2027
7	Reliability	0.90	0.90	0.90	0.85	0.93	0.90	0.87	0.81	0.92	0.91	0.71	0.69	0.86	0.90
	Variance	2106	1984	2214	1778	2202	2161	1798	1427	2007	2138	1362	1215	1767	2145
8	Reliability	0.91	0.90	0.91	0.89	0.93	0.90	0.89	0.86	0.91	0.91	0.79	0.82	0.89	0.91
	Variance	1511	1411	1597	1309	2050	1463	1354	1096	1504	1462	873	971	1343	1552
Alg I	Reliability	0.91	0.91	0.92	0.89	0.92	0.92	0.89	0.85	0.91	0.91	0.82	0.74	0.89	0.91
	Variance	1403	1309	1486	1118	1678	1485	1147	821	1422	1408	751	536	1146	1418
Geo	Reliability	0.90	0.89	0.90	0.85	0.92	0.89	0.86	0.82	0.91	0.91	0.79	0.74	0.86	0.90
	Variance	1449	1318	1578	1100	1759	1315	1131	887	1493	1535	946	786	1107	1468
Alg II	Reliability	0.87	0.86	0.88	0.81	0.91	0.87	0.81	0.76	0.89	0.88	0.71	0.70	0.81	0.87
	Variance	1088	969	1212	837	1447	1107	808	653	1186	1094	634	596	811	1078

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch

4.6.8 SUBSCALE RELIABILITY

Coefficient alpha estimates of internal consistency reliability associated with the subscales for the 2017 operational forms are presented in Exhibits 4.6.8.1–4.6.8.6. As indicated in the exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT.

Exhibit 4.6.8.1 Subscale Reliabilities – ELA Grades 3–11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 3	0.75	0.75	0.77
Grade 4	0.76	0.76	0.70
Grade 5	0.78	0.73	0.75
Grade 6	0.78	0.71	0.75
Grade 7	0.78	0.72	0.73
Grade 8	0.81	0.74	0.77
Grade 9	0.73	0.75	0.78
Grade 10	0.79	0.64	0.74
Grade 11	0.72	0.71	0.74

Exhibit 4.6.8.2 Subscale Reliabilities – Math Grades 3–5

	Numbers & Operations– Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
Grade 3	0.68	0.76	0.84
Grade 4	0.78	0.61	0.87
Grade 5	0.79	0.75	0.80

Exhibit 4.6.8.3 Subscale Reliabilities – Math Grades 6–7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
Grade 6	0.79	0.76	0.69	0.66
Grade 7	0.67	0.53	0.58	0.69

Exhibit 4.6.8.4 Subscale Reliabilities – Math Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability & the Number System
Grade 8	0.79	0.64	0.69	0.54

Exhibit 4.6.8.5 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.85	0.77	0.60
Algebra II	0.73	0.62	0.62

Exhibit 4.6.8.6 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.54	0.69	0.63	0.68

4.7 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 4.7.1–4.7.6. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.³⁰ The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y . When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Exhibit 4.7.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 3–11

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
3	Literature	0.73		0.97	
	Writing & Language	0.64	0.64	0.85	0.84
4	Literature	0.74		0.97	
	Writing & Language	0.62	0.62	0.86	0.84
5	Literature	0.75		0.99	
	Writing & Language	0.68	0.65	0.89	0.87
6	Literature	0.72		0.97	
	Writing & Language	0.66	0.64	0.87	0.87
7	Literature	0.74		0.99	
	Writing & Language	0.67	0.64	0.88	0.88
8	Literature	0.74		0.96	
	Writing & Language	0.69	0.64	0.87	0.85
9	Literature	0.70		0.94	
	Writing & Language	0.66	0.63	0.88	0.83
10	Literature	0.67		0.95	
	Writing & Language	0.64	0.54	0.83	0.79
11	Literature	0.69		0.96	
	Writing & Language	0.67	0.64	0.91	0.88

³⁰ Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.

Exhibit 4.7.2 Subscale Intercorrelations – Math Grades 3–5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.69		0.96	
	OAT_NBT	0.73	0.75	0.91	0.94
4	MDG	0.67		0.97	
	OAT_NBT	0.79	0.72	1.00	0.99
5	MDG	0.78		1.00	
	OAT_NBT	0.81	0.77	1.00	1.00

Note: NF = Numbers and Operations-Fractions; MDG = Measurement, Data & Geometry; OAT_NBT = Operations and Algebraic Thinking, and Numbers in Base Ten

Exhibit 4.7.3 Subscale Intercorrelations– Math Grades 6–7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.77			0.99		
	RP	0.74	0.76		1.00	1.00	
	GSP	0.72	0.70	0.66	1.00	0.99	0.97
7	NS	0.77			1.00		
	RP	0.78	0.73		1.00	1.00	
	GSP	0.76	0.73	0.72	1.00	1.00	1.00

Note: EE = Expressions and Equations; NS = Number System; RP = Ratio and Proportional Relationships; GSP = Geometry, Statistics and Probability

Exhibit 4.7.4 Subscale Intercorrelations– Math Grade 8

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
8	Functions (F)	0.74			1.00		
	Geometry(G)	0.73	0.68		0.99	1.00	
	SPNS	0.69	0.66	0.64	1.00	1.00	1.00

Note: EE = Expressions and Equations; F = Functions; G = Geometry; SPNS = Statistics and Probability and the Number System

Exhibit 4.7.5 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.81		1.00	
	Statistics	0.71	0.69	1.00	1.00
Algebra II	Functions	0.70		1.00	
	Statistics	0.68	0.64	1.00	1.00

Exhibit 4.7.6 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	C	0.70			1.00		
	MG	0.66	0.67		1.00	1.00	
	SRTT	0.72	0.73	0.68	1.00	1.00	1.00

Note: C=Congruence; CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations; MG=Modeling with Geometry; SRTT=Similarity, Right Triangles and Trigonometry

4.8 HANDSCORING AGREEMENT RATE

For grades in which statistical models were constructed for machine scoring of essay responses, Measurement, Inc. (MI) handscored over 4,100 responses per prompt, with each response double scored and any discrepant scores routed for a final resolution score. At each grade, students responded to one of two randomly-selected writing tasks. Exhibit 4.8.1 shows the summary of the rater agreement for the writing prompts administered on the AzMERIT spring 2017 online tests. The rater agreement reports show percentages of exact agreement (Equal), adjacent scores (Adj. Low or Adj. High), and nonadjacent scores (Non-Adj Low or Non-Adj High). The tables also identify mismatched scores when there is a difference involving nonscorable condition codes (Mismatch NS), or a nonscorable/scorable mix (MM NS/Score). Exhibit 4.8.1 provides a summary of those results, showing the mean exact agreement rate for dimension scores across grades. Generally exact agreement rates ranged from 65%–70%, with little variability across the essay prompts.

Exhibit 4.8.1 ELA Writing Prompt Rater Agreement Report – Spring 2017 Administration

Grade	Dimension	Total Read	Second Read	NonAdj Low	Adj Low	Equal	Adj High	NonAdj High	Mismatch NS	MM NS/Score
3	Purpose/Organization	17,272	1,568	1.0	18.8	60.4	18.8	1.0	0.0	0.0
	Evidence/Elaboration	17,272	1,568	1.1	19.4	58.9	19.4	1.1	0.0	0.0
	Conventions	17,272	1,568	0.8	14.1	70.3	14.1	0.8	0.0	0.0
4	Purpose/Organization	17,119	1,556	0.6	17.6	63.6	17.6	0.6	0.0	0.0
	Evidence/Elaboration	17,119	1,556	0.8	16.8	64.8	16.8	0.8	0.0	0.0
	Conventions	17,119	1,556	0.5	16.6	65.7	16.6	0.5	0.0	0.0
5	Purpose/Organization	16,808	1,527	0.5	16.8	65.5	16.8	0.5	0.0	0.0
	Evidence/Elaboration	16,808	1,527	0.4	17.7	63.7	17.7	0.4	0.0	0.0
	Conventions	16,808	1,527	0.7	15.6	67.5	15.6	0.7	0.0	0.0
6	Purpose/Organization	18,021	1,638	0.4	15.5	68.3	15.5	0.4	0.0	0.0
	Evidence/Elaboration	18,021	1,638	0.6	16.1	66.7	16.1	0.6	0.0	0.0
	Conventions	18,021	1,638	0.7	13.8	71.1	13.8	0.7	0.0	0.0
7	Purpose/Organization	17,161	1,560	0.2	13.1	73.4	13.1	0.2	0.0	0.0
	Evidence/Elaboration	17,161	1,560	0.3	11.3	76.9	11.3	0.3	0.0	0.0
	Conventions	17,161	1,560	0.3	12.4	74.5	12.4	0.3	0.0	0.0
8	Purpose/Organization	16,658	1,514	0.4	17.1	64.9	17.1	0.4	0.0	0.1
	Evidence/Elaboration	16,658	1,514	0.5	16.9	65.1	16.9	0.5	0.0	0.1
	Conventions	16,658	1,514	0.7	13.2	72.2	13.2	0.7	0.0	0.1
9	Purpose/Organization	19,976	1,816	0.2	18.3	62.9	18.3	0.2	0.0	0.0
	Evidence/Elaboration	19,976	1,816	0.6	19.3	60.2	19.3	0.6	0.0	0.0
	Conventions	19,976	1,816	0.1	9.2	81.3	9.2	0.1	0.0	0.0
10	Purpose/Organization	18,005	1,636	0.5	16.9	65.3	16.9	0.5	0.0	0.1
	Evidence/Elaboration	18,005	1,636	1.1	17.3	63.1	17.3	1.1	0.0	0.1
	Conventions	18,005	1,636	0.2	10.7	78.2	10.7	0.2	0.0	0.1
11	Purpose/Organization	16,429	1,493	0.5	18.1	62.9	18.1	0.5	0.0	0.0
	Evidence/Elaboration	16,429	1,493	0.7	17.8	63.0	17.8	0.7	0.0	0.0
	Conventions	16,429	1,493	0.3	10.2	79.1	10.2	0.3	0.0	0.0

5. ITEM DEVELOPMENT AND TEST CONSTRUCTION

The AzMERIT assessments are rigorously examined in accordance to the guidelines provided in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence based on these standards that is necessary to validate assessment scores for their intended purposes.

The AzMERIT assessments were designed to measure student progress toward achievement of the Arizona College and Career Ready Standards (ACCRS). Although the validity of AzMERIT test score interpretations are evaluated along several dimensions, as a criterion-referenced system of tests, the meaning of test scores is critically evaluated by the degree to which test content was aligned with the ACCRS.³¹

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards and refers back to those standards in a highly iterative test development process that includes the ADE, test developers, and educator and stakeholder committees. Items used to develop the spring 2015 operational test forms were drawn mainly from the AIRCore pool of items developed to align with the Common Core State Standards. The development process for the summer 2016 and fall 2016 operational tests were the same as the spring 2016 operational test and described in the 2016 AzMERIT Technical Report. The items were all reviewed by Arizona content experts and educators prior to field testing in spring 2016 and subsequent operational test administration in spring 2017. Only items that were found to align well with the ACCRS were used. To supplement the AzMERIT pool of items, a few previously-developed Arizona items that also aligned to the ACCRS were used.

Items used to develop the spring 2017 operational test forms were drawn from custom Arizona item development and AIR's AIRCore pool of items. Both custom Arizona items and AIRCore items were developed to align with the Common Core State Standards. These items were all reviewed by the Arizona Department of Education, Arizona content experts and educators, and Arizona community members prior to field testing in spring 2016 and subsequent operational test administration in spring 2017. Only items that were found to align well with the Arizona ACCRS and to be free of bias or sensitivity concerns were used.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards that are covered in each test administration. Thus, the test specification blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprints determined how student achievement of the ACCRS was evaluated, alignment of test blueprints with the content standards was critical. The ELA and math blueprints are provided as an attachment in Appendix B.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet the blueprints. Developing test

³¹ Standard 1.11 – When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

forms is difficult because test blueprints could be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. In addition to meeting complex blueprint requirements, test developers worked to meet psychometric goals so that alternate test forms measure equivalently across the range of student ability.

5.1 ITEM DEVELOPMENT PROCESS³²

The content development process for AzMERIT is managed within AIR's Item Tracking System (ITS), which acts as a content development and management tool, item bank, and publication system supporting both paper and online publication. This item development workflow leads items from inception, through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes and rationales at each review and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence, and that the review is conducted (or recorded in the case of committee review) by an authorized person. As items travel through Arizona's extensive review process, every version of every item is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

ITS allows remote Internet access by item writers and reviewers while ensuring security with individualized passwords for all users, limited access for external users, and strong encryption of all information. Upon publication, ITS tracks the item's use on test forms. After items are used, ITS stores the resulting statistics, including exposure statistics, classical item statistics, and statistics based on item response theory (IRT).

The AzMERIT item development process is predicated on a high level of interaction between test developers at AIR and the ADE, as well as with Arizona educators and stakeholders. AIR's ITS manages item content throughout the entire life cycle of an item, from inception, through series of agreed-upon item review levels culminating in operational pool approval. It also manages item content beyond the operational life of the item, including migration of items for use in practice tests or other training materials. ITS ensures that every item follows through the entire sequence of development and provides Arizona and AIR management on-demand reports of the content and status of the inventory of items. Each item is directed through a sequence of reviews and sign-offs by AIR and ADE staff before it is locked for field test or operational administration.

The ITS is integrated with the item display engine used by the AzMERIT online test delivery system. This feature, combined with a "web approval" process, allows the display of online items to be "locked" well before test forms are constructed and ensures that only approved items are administered to Arizona students.

5.1.1 ITEM WRITING

Test development experts use item specifications to guide the item development process.³³ These item specifications, developed by content experts at AIR and the ADE, strategically guide the item development process.

³² Standard 4.7 – The procedures used to develop, review, and try out items and to select items from the item pool should be documented.

³³ Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

They are detailed documents that specify content limits, model tasks, and response types for a particular standard. Item writers use these specifications while developing items to make the best use of the available item types.

The item specifications were developed using a vertical alignment for each standard, wherein the suggested task demands and cognitive complexity of items build upon those of the previous grade level, just as the standards themselves do.

Additionally, the item specifications provide models for item writers. The models include item samples that target different Depth of Knowledge (DOK) and difficulty levels. These item models also annotate the information in order to communicate the intent of the standard and DOK and to clarify for the writer how to manipulate the item difficulty while keeping the cognitive demands the same.

Detailed item specifications include the following:

- **Content Limits:** This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. For example, in grade 3, fraction denominators are limited to 2, 3, 4, 6, and 8.
- **Acceptable Response Mechanisms:** This section identifies the various ways in which students may respond to a prompt—e.g., multiple choice, graphic response, proposition response, equation response, multi-select.
- **Depth of Knowledge:** The task demands of each standard can be classified as DOK 1, DOK 2, DOK 3 and/or DOK 4.
- **Task Demands:** In this section, the standards are broken down into specific task demands aligned to the standard. In addition, each task demand is assigned an appropriate response mechanism, DOK, and practice clusters specifically relevant to that particular task demand.
- **Examples and Sample Items:** In this section, sample items are delineated along with their corresponding expected difficulties (easy, medium, and hard). Notes for modifying the difficulty of each task demand are detailed with suggestions for the item writer. The suggestions for adapting the difficulty based on the task demands are research-based and have been reviewed by both content experts and a cognitive psychologist.

Item writers consistently followed the item specifications during the item development process. During each level of review, items were compared to the item specifications to ensure their alignment to the standard, grade-level appropriateness, and adherence to the content limits set forth in the item specifications.

Within each grade or course, all items are aligned according to DOK. Depth of Knowledge, or commonly DOK, refers to the cognitive complexity of the item and the cognitive demands on the student. Based on work performed by Webb (2002), there are four levels of DOK:

- **DOK 1—Recall.** Students recall basic mathematical ideas, perform basic arithmetic operations using established algorithms, and identify examples of general math principles.
- **DOK 2—Skill/Concept.** Students apply their basic knowledge (DOK 1) and extend their thinking to problem solve, identify relationships, and draw conclusions.
- **DOK 3—Strategic Thinking.** Students go beyond basic problem solving (e.g., word problems) to extend their thinking to nonroutine problem solving, hypothesize, and critique arguments or problem-solving strategies.

- DOK 4—Extended Thinking. At this highest level, students engage in extended problem-solving activities, which require integration of multiple standards. For example, students may engage in a performance task that includes a common stimulus and four to six associated items related to the stimulus.

Depending upon the subject area and grade or course assessment, the percentage of items and score points aligned to DOK 1, DOK 2, DOK 3, and DOK 4 vary. The percentage of test items aligned to each DOK level for each assessment is indicated in the test construction blueprint. Although the exact number of items on each form may vary, the test specifications ensure that students are administered a substantial proportion of items that assess higher-order thinking skills.

ELA

ELA item development often begins with development of reading passages. AzMERIT passages represent a variety of genres and topics. AIR's content experts develop informational texts from multiple content areas, such as history, science, and technical subjects. Literary texts represent authentic pieces from multiple genres, including stories, poetry, and drama. The ratio of informational to literary texts increases at each grade band with a greater percentage of informational texts in the upper grades. The AzMERIT utilizes both single passages as well as passage sets in which students are asked to synthesize information across texts.

To ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading, test developers adhere to detailed passage specifications. Content experts use passage complexity worksheets—based on the passage specifications—to perform an in-depth analysis of each passage. The passage specifications call for a close examination of both quantitative measures, such as word counts and Lexile readabilities, as well as qualitative measures, such as passage structure and levels of meaning, all of which are defined as important measures of text complexity.

AzMERIT's ELA assessments include extended writing tasks that provide students with meaningful contexts in which to construct their responses. Each writing prompt presents students with a variety of stimuli (at least two to three per task) that serve as a springboard for an informed piece of writing. Students are given research articles, charts and graphs, and narratives to serve as the basis for their written response. Students can then use this information, along with their own reasoning, to formulate an essay that is not only a clear and coherent expression of their own thinking, but that is also grounded in research and evidence. Each student is administered a single informative/explanatory or opinion/argumentative writing essay.

Informative/explanatory writing is focused on conveying information accurately. Informative writing seeks to enlighten the reader about processes or procedures, phenomena, states of affairs, and terminology. To produce this kind of writing, students draw from what they already know as well as from primary and secondary sources. Students develop a controlling idea and a primary focus as they relate facts, details, and examples.

Opinion (grades 3–5) and argumentative (grades 6–11) prompts ask students to analyze primary and secondary sources, make sound judgments, and present their opinions or arguments in a coherent way that weaves personal opinion with evidence from the texts. The stimuli present opposing points of view about a topic so that students have enough information to take a stand. The stimuli are followed by a prompt that asks students to write an opinion or argumentative essay. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support the ideas they present. For example, the prompt might require students to describe the steps in a process or describe problems that need to be solved.

Writing prompts present students with two or three passage stimuli on a single topic from science, technical subjects, or social studies. The reading level of the stimulus does not exceed the easy Lexile range for the grade level to enable the students to attend to the content of the passages and not struggle over unfamiliar language and non-content-related vocabulary. Moreover, this helps ensure that students are assessed on their writing skills and not their reading abilities.

MATH

Calculators are not allowed for assessments at grades 3–6, while students participating in high school assessments are allowed continual access to specific calculator functions. For the grades 7 and 8 assessments, where calculator usage is allowable for some item types, the test items are grouped into two segments, administered separately to students: calculator and no calculator. The construct of the items dictates in which section they are to be assessed.

5.1.2 MACHINE-SCORED CONSTRUCTED-RESPONSE ITEM DEVELOPMENT TOOLS

AzMERIT includes a number of machine-scored constructed-response (MSCR) items which leverage a sophisticated system that allows for a large variety of item types expecting varied student responses to be developed, and scored efficiently and economically.

Machine-scored constructed-response (MSCR) item development tools put the power of both item and rubric creation into the hands of item writers, and allow reviewers to score possible responses to ensure that the rubric is enacted correctly. For example, when administered a graphic-response item, students can respond by drawing, moving, arranging, or selecting graphic regions. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer. Test developers have flexibility in identifying features of student responses to score, which go beyond simple features (e.g., whether the correct object is put in the correct place) but can involve abstraction. For example, if a student is asked to design an experiment, the rubric can discern whether the objects representing the experimental variable actually vary across conditions or cover the range of inquiry, among other capabilities. These concepts are abstracted and many different responses may reflect those abstract features. This ability enables machine rubrics to “justify” the partial credit assigned in terms of the skills that particular response features exemplify.

In addition, throughout the item development and review process, test developers can mimic the many different possible student responses and review how the rubric is applied to those responses. Test developers can test the scoring rubric and make corrections to the scoring logic at each step.

When creating equation items, test developers have access to the Equation Editor tool. Student responses can be simple numeric responses or complex equations, or even sets of equations. This tool allows for multiple answers and the development of multistep items. Test developers can customize the equation palette to show the appropriate functions. Just as the key pad is customizable, the answer spaces are, as well. Additional answer spaces can be added as needed by the item writer. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer.

Such tools are integrated into the ITS, providing test developers with the power and flexibility to use technology to create sophisticated AzMERIT items.

5.1.3 ITEM TYPES

AzMERIT includes a wide variety of item types that are designed around a broad and growing catalog of response mechanisms. In addition to selected response items, which include traditional multiple-choice and more advanced multi-select and two-part items, AzMERIT tests utilize various item types including those with the following response mechanisms:

- Graphic Response, which includes any item to which students respond by drawing, moving, arranging, or selecting graphic regions
- Hot Text, in which students select or rearrange sentences or phrases in a passage
- Equation Response, in which students respond by entering an equation or number
- Word Builder, in which students respond by entering a single number or word
- Proposition Response, in which students respond in one or more English language sentences, which may be scored by our proposition-scoring engine, human scored, or a mixture of both
- Essay Response, in which the student response is a longer, written response

AzMERIT items use technology to measure deeper knowledge and application of knowledge in a more open-ended way and to machine score many such items. All MSCR items administered in AzMERIT are accessible. There may be occasions where it is necessary to sacrifice accessibility for some population to measure a critical standard, but test development staff would need to carefully consider the measurement benefit before developing that item.

Where possible, MSCR items were rendered for administration on paper test forms, using the gridded response field in the scannable answer documents. Where equation and graphic response items could not be rendered to accept a gridded response on paper forms, responses were handscored. For other MSCR items that could not readily be rendered for paper test administration, the item was replaced by another item measuring the same content standard(s).

The graphic-response mechanism supports most of the typical technology-enhanced item types, including sorting, matching, hot-spot, and drag-and-drop. In addition, it supports items where students actually draw a machine-scorable response and respond by constructing complex, open-ended diagrams, as well as many other possibilities. Because they are uniformly derived from a single response mechanism, the manipulations and interactions are consistent across these technology-enhanced item types, eliminating one possible source of construct-irrelevant variance.

Hot-text items are effectively selected-response items, though in some cases, the number of potential selections is quite large. These machine-scored items can have multiple correct answers and allow for very flexible student responses.

The equation response mechanism asks students to enter one or more numbers, expressions, or equations using a palette of symbols. Test developers can specify which symbols are available on an item-by-item basis, or the ADE can choose to have the palette remain consistent across all of the items within a grade level.

The availability of tools organized around response mechanisms creates a very flexible capability for test developers to create authentic, challenging tasks.

5.2 ITEM REVIEW

This section describes the multi-step item review process that items travel through—from inception, to several rounds of review by test developers, the ADE, and educators, to field testing and final review—prior to inclusion on operational test forms.³⁴ Items used to develop the spring 2017 operational test forms were drawn from custom Arizona item development and AIR’s AIRCore pool of items. Both custom Arizona items and AIR Core items were developed to align with the Common Core State Standards. These items were all reviewed by the Arizona Department of Education, Arizona content experts and educators, and Arizona community members, prior to field testing in spring 2016 and subsequent operational test administration in spring 2017. Only items that were found to align well with the ACCRS and to be free of bias or sensitivity concerns were used.

The item review procedures used to develop and review AzMERIT test items are designed to ensure item accuracy and alignment with the intended ACCRS. Following a standard item review process, item reviews proceed initially through a series of internal reviews before items are eligible for review by the ADE’s content experts. Most of AIR’s content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through four internal review steps before it is eligible for review by the ADE. Those steps include:

- Preliminary review, conducted by a group of AIR content-area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert

At every stage of the item review process, beginning with preliminary review, AIR’s test developers analyze each item to ensure that:

- The item is well-aligned with the intended content standard.
- The item conforms to the item specifications for the target being assessed.
- The item is based on a quality idea (i.e., it assesses something worthwhile in a reasonable way).
- The item is properly aligned to a depth of knowledge (DOK) level.
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward.
- Any accompanying graphic and stimulus materials are actually necessary to answer the question.
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question.
- For selected-response items, the set of response options is succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; and all plausible, but with only one correct option.
- There is no obvious or subtle cluing within the item.
- The score points for constructed-response items are clearly defined.

³⁴ Standard 4.8 – The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

- For machine-scored constructed-response (MSCR) items, the items score as intended at each score point in the rubric.

Based on their review of each item, the test developer can accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to the ADE for review. At this stage, items may be further revised based on any edits or changes requested by the ADE, or rejected outright. Items passing through the ADE's review then pass through a stakeholder review, in which educators review each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- Aligned to the Arizona content standards
- Appropriate for the grade level
- Accurate
- Presented online in a way that is clear and appropriate
- Free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics

Items successfully passing through this committee review process were then presented to a parent/community review committee to ensure that test content met community standards. Items successfully passing through all review levels were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is, therefore, an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass in each stage of a two-stage review before being included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct, and there are no other obvious problems with the items.

ADE content staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, the ADE determined that certain flagged items must be rejected, or deemed the item eligible for inclusion in operational test administrations.

5.3 FIELD TESTING

To establish a pool of items for constructing future AzMERIT test forms, newly developed test items were embedded in the spring 2016 and spring 2017 AzMERIT test forms for field testing. Embedding field-test items in operational assessments yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations. A number of factors that may influence item difficulty in the context of operational test administrations may be less relevant in stand-alone field-test contexts. For example, in a high-stakes test, such as high school end-of-course (EOC) exams where test performance may impact student grades, students may be motivated to expend greater effort to achieve maximum performance. Conversely, the high-stakes assessments may also be more likely to elicit anxiety in some students, thus impairing their performance on the tests. Even when assessments are low stakes for students, schools often work to convey to students the importance of statewide assessments in ways that are likely not done for independent field tests. While the impact of contextual factors may not be great, embedded field testing ensures that all aspects of the operational testing context influencing item difficulty are incorporated into the resulting item parameter estimates.

Embedded field testing is especially useful in the context of a pre-equating model for scoring and reporting test results. Because the test administration context remains the same between the embedded field test (EFT) and subsequent operational test administration, item parameter estimates are more stable over time than they may be when obtained through stand-alone field testing.

A potential drawback of the EFT approach is the increased assessment burden placed on students and schools. For this reason, AzMERIT utilizes EFT designs for purposes of item bank maintenance. Arizona uses AIR's online field-test engine for computer-administered tests, which, when combined with Arizona's large student population, serves to greatly reduce the number of EFT slots necessary to replenish and even grow the item banks for the Arizona assessments.

The field test engine randomly samples field test items for each individual test administration, essentially creating thousands of unique EFT forms. This sampling approach to embedding field-test items results in several important outcomes:³⁵

- Reduction in the number of embedded field-test items that each student must respond to and more efficient “spiraling” of items, which reduces clustering of item responses, resulting in more precise parameter estimates
- More generalizable item statistics because they are not based on items appearing in a single position
- A truly representative sample of respondents for each item

The embedded field testing algorithm actually consists of two different algorithms – one for identifying which field test items will be administered to which student (the distribution algorithm), and one for selecting the position on the test for each item administered the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field test items designated for administration to each student. This

³⁵ Standard 4.9 – When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.

randomization ensures that a) each item is seen by a representative sample of Arizona students, and b) every item is as likely as every other item to appear in a class or school, minimizing clustering effects.

In addition, a fixed block of field test items was also embedded in paper AzMERIT test forms so that the number of items responded to by students did not vary between assessment modes.

In the spring 2015 administrations, item parameters for the ELA and math assessments were calibrated following the online administration to establish the AzMERIT bank scale. Following the spring 2016 test administrations, the free calibration was performed on the operational items on each of the ELA and mathematics tests. Then, the free calibrated item parameters were linked back to the 2015 spring scale using the mean-mean equating method. The field test item calibration was conducted by anchoring on the post-equated operational item parameters for all of the ELA and mathematics tests. However, only the ELA spring 2016 operational tests were scored using the post-equated item parameters. In the spring 2017 test administration, the pre-equated parameters calibrated and equated following spring 2016 test administrations were used for final scoring and reporting for all the ELA and mathematics tests.

5.4 ITEM STATISTICS

Following the close of spring testing windows, AIR psychometrics staff worked to analyze field test data in preparation for item data review meetings and promotion of high-quality test items to operational item pools.³⁶ Analysis of field test items includes classical item statistics as well as the IRT item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by AIR and ADE psychometric and content staff for possible miskey or scoring errors. Items that pass through AIR and ADE statistical review are accepted for future operational use. Appendix H provides the slide presentation used to train reviewers for item data review. The training is designed to ensure that all reviewers understand how items are evaluated and that they are interpreting item statistics correctly.

5.4.1 CLASSICAL STATISTICS

Classical item analyses ensured that the field test items function as intended with respect to the AzMERIT's underlying scales. AIR's analysis program computed the required item and test statistics for each selected-response (SR) and constructed-response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are either extremely difficult or extremely easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For dichotomous items, the proportion of examinees in the

³⁶ Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

sample selecting the correct answer (p -value) is computed, as well as those selecting the incorrect responses. For constructed-response items, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to p -value and indicating the ratio of an item's mean score divided by the number of points possible). Items are flagged for review if the p -value was less than .25 or greater than .95.

The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low- achieving students. The discrimination index for dichotomous items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .25.

Distractor analysis for the dichotomous items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response is greater than .05. In addition, items are flagged if the proportion of students responding to a distractor exceeds the proportion selecting the keyed response. Although non-modal response keys are typically observed with difficult items, in combination with poor item discrimination it may indicate a miskeyed item.

5.4.2 IRT STATISTICS

Rasch and Masters' Partial Credit Model are used to estimate the item response theory (IRT) model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showing the item statistics resulting from the free (unanchored) estimation of parameters for items in the operational tests were reviewed, as well as the Winsteps-generated item and persons maps. Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are conservatively flagged if Infit or Outfit values are less than 0.7 or greater than 1.3.

5.4.3 ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, DIF can indicate bias, so all field-tested items were evaluated for DIF, and all items exhibiting DIF were flagged for further examination by AIR and the ADE's staff to make a final decision about whether the item should be excluded from the pool of potential items given its performance in field testing potential items.

AIR conducts DIF analysis on all field-tested items to detect potential item bias across major ethnic and gender groups. In Arizona, DIF is investigated among the following group comparisons (reference group/focal group):

- Male/Female
- White/Hispanic, Latino or Spanish origin/ Non-Hispanic
- White/Black, African American, or Negro
- White/American Indian or Alaskan Native
- White/Asian
- White/Native Hawaiian or Other Pacific Islander
- White/Multiple ethnicities selected
- Non-Special Education/ Special Education
- Non-Limited English Proficiency/Limited English Proficiency
- Non-Free or Reduced Lunch/Free or Reduced Lunch

AIR uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items, and (2) improved variance estimators to render the test statistics valid under complex sample designs. Because students within a district, school, and classroom are more similar than would be expected in a simple random sample of students statewide, the information provided by students within a school is not independent, so that standard errors based on the assumption of simple random samples are underestimated. We compute design consistent standard errors that reflect the clustered nature of educational systems. While clustering is mitigated through random administration of large numbers of embedded field test items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into a configurable number of intervals to compute the Mantel-Haenszel chi-square ($MH \chi^2$) DIF statistics. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta ($\Delta_{hat MH}$) for the dichotomous items; the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the polytomous items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed below. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e., -A, -B, or -C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Exhibit 5.4.3.1. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or reference groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992, Camilli & Shepard, 1994, Muniz, Hambleton, & Xing, 2001, Sireci & Rios, 2013).

Exhibit 5.5.3 DIF Classification Rules

Item Type	Category	Rule
Dichotomous Items	C	$MH \chi^2$ is significant and $ \Delta_{\text{hat } MH} \geq 1.5$
	B	$MH \chi^2$ is significant and $ \Delta_{\text{hat } MH} < 1.5$
	A	$MH \chi^2$ is not significant
Polytomous Items	C	$MH \chi^2$ is significant and $ SMD / SD \geq .25$
	B	$MH \chi^2$ is significant and $ SMD / SD < .25$
	A	$MH \chi^2$ is not significant

5.5 TEST CONSTRUCTION

The process for constructing fixed-form operational tests begins after field testing and review of item performance. Once an operational item pool is established, AIR content specialists begin the process of constructing test forms. Operational passages and items qualified for operational forms are those that meet all of the criteria established by the ADE in terms of content, fairness review, and data characteristics.

5.5.1 OPERATIONAL FORM CONSTRUCTION

Each AzMERIT form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it is covered, the type of items that measure the constructs, and every other content-relevant aspect of the test. The statistical targets, which are held constant across years and across modes, ensure that students receive scores of similar precision, regardless of which form of the test they receive.³⁷

AIR's test developers used FormBuilder software to help construct operational forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively create test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement curves (SEMCs) as test developers combine items to build a test form. This helps content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

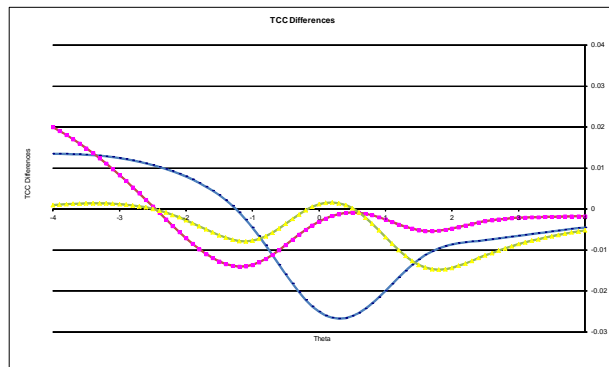
Immediately upon generation of a test form, FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allow another opportunity to ensure that poorly performing items are not included in operational test forms.

As test developers built forms, the, FormBuilder-generated TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 5.6.1.1 shows a sample graph of TCC differences. There are several important things to note when examining TCC differences. First, differences in TCCs can occur at specific

³⁷ Standard 4.12 – Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

locations in the TCCs across a range of abilities. These differences reflect different emphases in test information across forms at these ability levels. If the difficulty and error structure for the target forms is virtually identical to the reference form, as in the sample TCC and SEM curves, the item selection process concludes with multiple, parallel test forms. Once the goal of parallel forms is achieved, the information is entered into ITS, which tracks item usage and generates bookmarks (test maps) for use in scoring, forms development, and other processes.

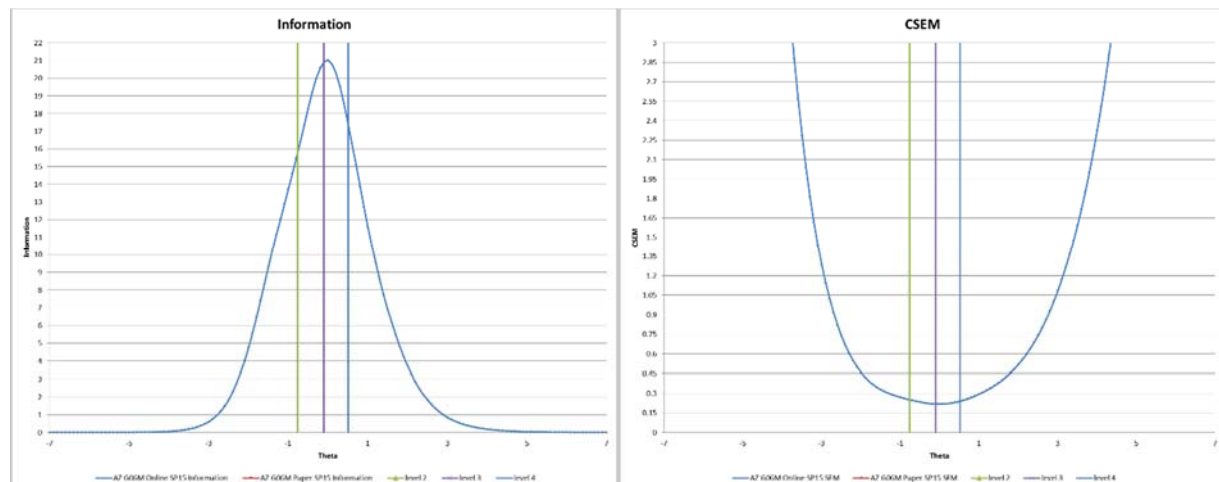
Exhibit 5.6.1.1 Test Characteristics Curve Differences



The reference form for each assessment is the operational test form administered in spring 2015. As illustrated in Exhibit 5.6.1.2, by evaluating test characteristics in reference to the base year forms, students are administered tests each year that are equivalent in difficulty across the range of ability. The Test Characteristic Curve (TCC) and SEM graphs that were used to evaluate the spring 2017 operational test forms are presented in Appendix I.

In addition, although paper test forms were developed to be as nearly identical to the online forms, there were some items that could not readily be rendered for paper test administration. In those instances, replacement items were identified and TCCs and SEMCs were evaluated to ensure equivalence between online and paper test forms.

Exhibit 5.6.1.2 Test Information and Standard Errors Relative to Performance Standards



5.5.2 ASSEMBLING TEST FORMS

The mechanical features of a test—arrangement, directions, and production—are just as important as the quality of the items. Many factors directly affect a student’s ability to demonstrate proficiency on the assessment, while others relate to the ability to score the assessment accurately and efficiently. Still others affect the inferences made from the test results.

When the test developer reviews a test form for content, in addition to making sure all the benchmark/indicator item requirements are met, he or she also makes sure that the items on the form do not cue each other—that one item does not present material that indicates the answer to another item. This is important to ensure that a student’s response on any particular test item is unaffected by, and is statistically independent of, a response to any other test item. This is called “local independence.” Independence is most commonly violated when there is a hint in one item about the answer to another item. In that case, a student’s true ability on the second item is not being assessed.

Test developers begin the form construction process by first identifying the pool of items from which forms are built. This pool of items resides at a locked operational status in the Item Tracking System. Each item contains a historical record that clearly demonstrates it has survived the full review process from internal development through client, committees, and its statistical data review.

Upon identifying and reviewing the eligible pool of items, a test developer then considers the limitations of the pool, if any. For example, there might be a shortage of high depth of knowledge (DOK 3) items at a particular benchmark. The test developer will review and select from among these items first to ensure that the constraints of the blueprint are met.

Once the items and passages for the form are selected and matched against the blueprint, the test developer reviews the form for a variety of additional content considerations, including the following:

- The items are sequentially ordered.
- Each item of the same type is presented in a consistent manner.
- The listing of the options for the multiple-choice items is consistent.
- The answer options are labeled correctly.
- All graphics are consistently presented.
- All tables and charts have titles and are consistently formatted.
- The number of the answer choice letters is approximately equal across the form.
- The answer key was checked by the initial reviewer and one additional independent reviewer.
- All stimuli have items associated with them.
- The topics of items, passages, or stimuli are not too similar to one another.
- There are no errors in spelling, grammar, or accuracy of graphics.
- The wording, layout, and appearance of the item matches how the item was field tested.
- There is gender and ethnic balance.
- The passage sets do not start with or end with a constructed-response item.
- Each item and the form are checked against the appropriate style guide.
- The directions are consistent across items and are accurate.
- All copyrighted materials have up-to-date permissions agreements.
- Word counts are within documented ranges.

After completing the initial build of the form, the test developer hands it off to another content specialist, who conducts a final review of the criteria listed above. If the test specialist reviewer finds any issues, the form is sent back for revisions. If the form meets blueprint and complies with all specified criteria, the test developer sends it to the psychometric team for review. When the psychometric team approves the form, the test developer forwards the form evaluation workbooks to the ADE's Assessment Content Experts for review, possible changes in the item selection or item position, and approval.

6. TEST ADMINISTRATION

6.1 ELIGIBILITY

Arizona public school students in grade 3 and above were required to participate in AzMERIT testing.³⁸ Additionally, any student enrolled in a private school or Bureau of Indian Education school and any home-schooled student had the option to participate, as well. Students enrolled in grades 3–8 took English language arts (ELA) and Math at the grade level in which they were enrolled. Students, in any grade, who are enrolled in high school-level English language arts courses (freshman English, sophomore English, junior English, or their equivalents) or high school-level math courses (Algebra I, Geometry, Algebra II, or their equivalents) took the respective end-of-course (EOC) test. Grade 8 students who took EOC tests in math were not required to take the grade 8 Math test.

Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them eligible for the alternate assessment for ELA and Math were excluded from AzMERIT and instead took the Multi-State Alternate Assessment (MSAA).

6.2 ADMINISTRATION PROCEDURES

Key personnel involved with AzMERIT administration include the District Test Coordinators, School Test Coordinators, and Test Administrators who proctor the test. For information about the roles and responsibilities of testing staff, see below.

A secure browser developed by AIR was required to access the computer-based AzMERIT tests. The secure browser provided a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to desktop functionalities, such as the Internet and email. Other measures that protect the integrity and security of the online test are presented in Section 6.5.

Prior to each test administration, statewide District Test Coordinator training sessions were conducted to provide information regarding both the paper and computer-based test administrations. The training also provided an overview of the Test Delivery System (TDS), Online Reporting System (ORS), and Test Information Distribution Engine (TIDE). Recorded training sessions and narrated training videos were posted online. The Test Coordinator Manual and Test Administration Directions were shipped to every testing district. Additionally, test administrators were required to complete the online TA Certification Course before administering a computer-based test.³⁹ District Test Coordinators and School Test Coordinators were responsible ensuring that all test administration personnel (paper and computer-based) were properly trained using the various resources prior to the start of testing.

³⁸ Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

³⁹ Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

Manuals and guides on test administrations are available on the AzMERIT Portal.⁴⁰ The Test Administrator User Guide was designed to familiarize Test Administrators with the Test Delivery System and contains tips and screenshots throughout the text. The guide provides enough how-to information to enable TAs to access and navigate the Test Delivery System. The user guide provides the following information:

- Steps to take prior to accessing the system and logging in
- Navigating the TA interface application
- The Student Interface, used by students for computer-based testing
- Training sites available for Test Administrators and students
- Secure browsers and keyboard shortcut keys

The *AzMERIT Test Coordinator's Manual* provides information about policies and procedures for AzMERIT Test Coordinators. This manual is updated prior to each test administration and includes test administration policies and guidance for Test Coordinators before, during, and after the testing window.

The *AzMERIT Test Administration Directions, End-of-Course* and the *AzMERIT Test Administration Directions, Grades 3–8* provide information about policies and procedures for the AzMERIT, both computer-based and paper-based versions. The *Test Administration Directions*, which is updated prior to each test administration, includes test administration information, guidance, and directions.

The *AzMERIT Test Administration Directions* provide easy-to-follow instructions for the online testing environment, such as creating online testing sessions, monitoring online sessions, verifying student information, assigning test accommodations, and starting and pausing test sessions.⁴¹ Similar guidance is provided for the paper testing environment, including instructions for the paper testing session, monitoring sessions, verifying student information, and providing test accommodations. Additional instructions for administering tests to students using Braille accommodated test booklets are provided in the *Supplemental Instructions for Braille* documents.

District and school personnel involved with AzMERIT test administration played an important role in ensuring the validity of the assessment by maintaining both standardized administration conditions and test security.

District Test Coordinators were responsible for coordinating testing at the district level. They were ultimately accountable for ensuring that testing was conducted in accordance with the test security and other policies and procedures established by ADE. They ensured that the Test Administrators in each school were appropriately trained and aware of policies and procedures, and that they were trained to use the reporting system.

Districts may also identify School Test Coordinators. School Test Coordinators may assist in the identification and training of Test Administrators. They may also create testing schedules and procedures for the school. If the school administers AzMERIT online, the School Test Coordinators may work with Technology Coordinators to ensure that the necessary secure browsers were installed and any other technical issues were resolved. During the testing window, School Test Coordinators needed to monitor testing progress, ensure that all students participate as appropriate, and handle testing incidents as necessary.

⁴⁰ Standard 7.13 – Supporting documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to the appropriate people in a timely manner.

⁴¹ Standard 4.15 – The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.

Test Administrators (TA) were responsible for reviewing necessary manuals and user guides to prepare the testing environment and ensuring that students did not have unapproved books, notes, or electronic devices available during testing. TAs were required to administer AzMERIT tests following the directions found in the *AzMERIT Test Administration Directions*.⁴² Any deviation in test administration must be reported by TAs to the School Test Coordinator, who reports it to the District Test Coordinator. The District Test Coordinator then reports it to the ADE.

Test Administrators who administered computer-based AzMERIT tests conducted a training test session using the AzMERIT Sample Tests. Test Administrators were required to pass a qualifying test before they were eligible to administer the AzMERIT online.⁴³

Test Administrators must also ensure that only resources that were allowed for specific tests were available and no additional resources were being used during the test. No calculators were permitted in AzMERIT Math tests for grades 3–6. Scientific calculators were permitted in AzMERIT Math Part 1 for grades 7 and 8. Graphing calculators were permitted in AzMERIT Math EOC Parts 1 and 2 (Algebra I, Geometry, and Algebra II). Online calculators were provided as embedded tools within the appropriate computer-based test parts. Handheld calculators could be provided to students during the appropriate test sessions. Calculator guidance was provided in both the *AzMERIT Test Coordinator's Manual* and the *AzMERIT Test Administration Directions*. The online calculators were made publicly available on the AzMERIT Portal, as well as made securely available in a secure browser for paper-based test students to access, if needed. Providing a calculator with prohibited functionality or in the incorrect test session is cause for test invalidation.

For the computer-based ELA Reading tests, headphones or earbuds were required. There were no technical specifications for headphones or earbuds. The equipment was to be checked to ensure that it worked with the computer or device the students would use for the assessment prior to the first day of testing. A sound test was also built into the computer-based assessment and students were asked to verify that headphones and earbuds were working prior to entering the test.

For the paper-based AzMERIT tests, Test Administrators needed to ensure that students used No. 2 pencils to record their responses. School Test Coordinators provided TAs with the materials needed to administer each test session. Secure materials were delivered or picked up immediately before the beginning of each test session. During math testing and when responding to the writing prompt, students were permitted to use the scratch paper as a workspace. After testing, TAs needed to return the testing materials, including all scratch paper, to the School Test Coordinator.

The School Test Coordinator and Test Administrators worked together to determine the most appropriate testing option(s), testing environment, and the average time needed to complete each test. The appropriate protocols were established to maintain a quiet testing environment throughout the testing session. TAs also needed to ensure that adequate time was available to start computers, load secure browsers, and log in students for computer-based tests or pass out and collect test materials for paper-based tests.

⁴² Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

⁴³ Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

6.2.1 MANAGING TESTING

To help schools manage their test schedule, allocate testing resources, and prioritize testing, the AzMERIT online reporting system, which is described in detail later in this chapter, offered participation reports for online testers. Within the online reporting system, educators can generate up-to-the-minute reports showing students' test status. In addition, users can set testing schedules, monitor testing progress across schools, and track students' participation based on their performance on previous tests.

ORS Online Reporting System Logged in as: Doe, Jane Contact Us Log Out

AzMERIT

Home Test Management Center This page: Help Definitions

Plan and Manage Testing

Step 1: Choose What

Test: AzMERIT
Administration: 2014-2015
Test Name: Grade 5 Mathematics
Enrolled Grade: 05

Step 2: Choose Who

District: Demo District (99)
School: Demo School (99-1234)
Personnel: Teacher: Demo

Step 3: Get Specific

☒ Students who have completed the test in the selected administration
☐ Students who have a status of any in the selected administration
☐ Students whose most recent SessionID was SessionID (optional) between 04/01/2015 and 04/15/2015

Note: If no TA or Session ID is specified, date range cannot exceed 15 days

☐ Search students by SSID : Enter up to 20 SSID(s) separated by commas

Generate Report or Export Report

AzMERIT Help Desk
1.844.560.7812
azmerithelpdesk@air.org

6.3 TESTING CONDITIONS, TOOLS, AND ACCOMMODATIONS

This section summarizes the testing conditions, tools, and accommodations that are available to AzMERIT testers, as described in the *Testing Conditions, Tools, and Accommodations Guidance* manual that is available each administration. Test tools and accommodation requirements are designed to ensure that test content is accessible for all students.

6.3.1 UNIVERSAL TEST ADMINISTRATION CONDITIONS

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment.⁴⁴ Universal test administration conditions are available for both paper-based test (PBT) and computer-based test (CBT) modes. Universal test administration conditions include:

⁴⁴ Standard 3.4 – Test takers should receive comparable treatment during the test administration and scoring process.

⁴⁴ Standard 4.5 – If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

- 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 a place holder
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT)
- 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
- Signing the scripted directions
- Having the scripted directions repeated (at student request)
- Having questions about the scripted directions or the directions that students read on their own answered
- Reading the test quietly to himself/herself as long as other students are not disrupted
- Allowing extended time (Testing session must be completed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan (IEP) as an accommodation, for AzMERIT testing purposes, these are not considered testing accommodations and are available to any student who needs them, not just to students with IEPs/Section 504 Plans.

6.3.2 UNIVERSAL TESTING TOOLS FOR COMPUTER-BASED TESTERS

The AzMERIT computer-based testing platform offers numerous testing tools. All tools are available in the AzMERIT Sample Tests, which are available to test administrators and students prior to each test administration. Test administrators are encouraged to ensure that students who will participate in the computer-based AzMERIT take the AzMERIT Sample Tests and familiarize themselves with the available tools.

Exhibit 6.3.2.1 summarizes the universal test tools that are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

⁴⁴ Standard 6.4 – The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

Exhibit 6.3.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Click anywhere on the selected-response text or button for multiple-choice options
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen <i>Test Instructions and Help</i> .
Highlighter	Highlight text in a passage or item.
Line Reader	This allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	This allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	This allows the session to be paused at any time and restarted and taken over a one-day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	This allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended-response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

6.3.3 SUBJECT AREA TOOLS FOR CBT AND PBT

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 6.3.3.1.

Exhibit 6.3.3.1 Subject Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	<p>CBT – Students have access to the dictionary/thesaurus tool. Students may opt to use a published, paper dictionary or thesaurus instead of using this tool.</p> <p>PBT – Schools must make published, paper dictionaries and thesauruses available to students.</p> <p>Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned off.</p>
Writing Guide	Writing	<p>CBT – Students have access to the writing guide tool.</p> <p>PBT – The writing guide is included within the test booklet.</p>
Scratch Paper	Writing and Math	<p>CBT – Schools must provide scratch paper (plain, lined, or graph) to students.</p> <p>PBT – Schools must provide scratch paper (plain, lined, or graph) to students.</p>
<p>Calculator</p> <p>Grades 7–8 (Part 1 only): Specific scientific calculators are acceptable.</p> <p>EOC (entire test): Specific graphing calculators are acceptable.</p>	Math	<p>CBT – Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted.</p> <p>PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.</p>

6.3.4 ACCOMMODATIONS

Accommodations are provisions made in how a student accesses or demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Learner or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. 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. Test administrators are instructed to make accommodation decisions based on individual needs, and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzMERIT test that is not already used regularly in the classroom.

Testing accommodations may not violate the construct of a test item. Testing 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, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in the *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section.⁴⁵ Arizona takes care to ensure that allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student's individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact the ADE for guidance.

Allowable accommodations are described below.⁴⁶

ACCOMMODATIONS FOR STUDENTS WITH AN INJURY

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations described in Exhibit 6.3.4.1. There are no specific CBT tools to support these accommodations.

Exhibit 6.3.4.1 Accommodations for Students with an Injury

Accommodation	Description
Adult Transcription	<p>If a student with an injury tests at a CBT school and cannot enter their own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided orally or by gestures, into the paper booklet and then into the Data Entry Interface (DEI), or directly into the DEI.</p> <p>If a student with an injury at a PBT school cannot write their own responses in a booklet, an adult must transfer the student's responses exactly as provided orally or by gestures.</p>
Assistive Technology	<p>With the use of assistive technology for the writing response and/or other open-response items, Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted.</p> <p>This accommodation also requires Adult Transcription (see above for rules on Adult Transcription).</p>
Scratch Paper	Student may take breaks during testing sessions to rest.

⁴⁵ Standard 3.10 – When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

⁴⁶ Standard 3.9 – Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

ACCOMMODATIONS FOR ENGLISH LEARNER (EL) AND FEP STUDENTS

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. Students eligible for these accommodations include English Learner (EL) students, students withdrawn from English language services at parent request, and Reclassified Fluent English Proficient (RFEP) students. Students in their monitoring period, within two school years of reclassifying as Fluent English Proficient (FEP Year 1 and FEP Year 2), may also, as appropriate, use any of the universal test administration conditions and any of the following accommodations.

The accommodations indicated as “*upon student request*” are required to be administered in a setting that does not disturb other students, such as in a one-on-one or very small group setting.

Exhibit 6.3.4.2 summarizes accommodations that may be provided for EL, RFEP, and FEP students.

Exhibit 6.3.4.2 Allowable Accommodations for EL, RFEP, and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Exact oral translation, in the student’s native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word published, paper translation dictionary. Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 6.3.4.3, as designated in their IEP or Section 504 Plan.

Exhibit 6.3.4.3 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	<p>If a student testing at a CBT school has an IEP indicating that they cannot enter their own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided orally or by gestures, into the paper booklet and then into the DEI, or directly into the DEI.</p> <p>If a student testing at a PBT school has an IEP indicating Adult Transcription, an adult must transfer the student's responses exactly as provided orally or by gestures into the paper booklet.</p>
ASL and Closed Caption	In computer-based tests, this is available for the listening items on the Reading ELA test.
Assistive Technology	<p>This is the use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted.</p> <p>This accommodation requires Adult Transcription (see above for rules on Adult Transcription).</p>
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Large Print Test Booklet	<p>CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet.</p> <p>PBT – Provide a Large Print test booklet.</p> <p>PBT Large Print Test booklet requires Adult Transcription into the DEI. See above for rules on Adult Transcription.</p>
Paper Test Booklet	CBT – Student's IEP must indicate that student cannot enter their own responses on the computer and requires a paper test or adult transcription. The school will provide a Special Paper Version booklet for the student. The student's responses must be transcribed into the paper booklet, and then entered into the DEI, or entered directly into the DEI. See above for rules on Adult Transcription.
Read Aloud Test Content	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test.</p> <p>Reading aloud the content of the Reading portion of the ELA test.</p>
Rest/Breaks	Student may take breaks during testing sessions to rest.
Sign Test Content	<p>Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the Math test.</p> <p>Signing the content of the Reading portion of the ELA test.</p>
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

6.4 SYSTEM SECURITY

6.4.1 SECURE SYSTEM DESIGN

AIR has developed a custom single sign-on application that is made available on Arizona's secure portal. This application is used to support access to AIR's system in accordance with the Arizona's user ID and password policy. Authorized users can log in to Arizona's single sign-on using their current user IDs and passwords and can be redirected to AIR's portal, where they have access to AIR's secure applications, such as the Test Information Distribution Engine (TIDE), the test delivery system (TDS), and the Online Reporting System (ORS). Nightly backups protect the data. The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or they will need to rerun the backup. The system can withstand failure of almost any component with little or no interruption of service.

AIR's hosting provider, Rackspace, has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely. Rackspace partners with nine different network providers, providing multiple, redundant data routes. Every installation is served by multiple servers, any one of which can take over for an individual test upon failure of another.

AIR's architecture ensures that data are recoverable at all times. Each disk array is internally redundant, with multiple disks containing each data element. Immediate recovery from failure of any individual disk is performed by accessing the redundant data on another disk. AIR maintains support and maintenance agreements through our hosting provider for all of the hardware used by our systems.

6.4.2 SYSTEM SECURITY COMPONENTS

AIR has built-in security controls in all of its data stores and transmissions.⁴⁷ Unique user identification is a requirement for all systems and interfaces. All of AIR's systems encrypt data at rest and in transit.

PHYSICAL SECURITY

AzMERIT data resides on servers at Rackspace, AIR's hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. All access is keycard controlled, and sensitive areas require biometric scanning.

⁴⁷ Standard 6.16 – Transmission of individually-identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information. Standard 8.6 – Test data maintained or transmitted in data files, including all personally-identifiable information (not just results), should be adequately protected from improper access, use, or disclosure, including by reasonable physical, technical, and administrative protections as appropriate to the particular data set and its risks, and in compliance with applicable legal requirements. Use of facsimile transmission, computer networks, data banks, or other electronic data-processing or transmittal systems should be restricted to situations in which confidentiality can be reasonably assured. Users should develop and/or follow policies, consistent with any legal requirements, for whether and how test takers may review and correct personal information.

Secure data are processed at AIR facilities and are accessed from AIR machines. AIR's servers are in a secure, climate-controlled location with access codes required for entry. Access to our servers is limited to our network engineers, all of whom, like all AIR employees, have undergone rigorous background checks.

Staff at both AIR and Rackspace receive formal training in security procedures to ensure that they know the procedures and implement them properly. AIR and Rackspace protect data from accidental loss through redundant storage, backup procedures, and secure off-site storage.

NETWORK SECURITY

Hardware firewalls and intrusion detection systems protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) for our secure sites.

AIR's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts.

SOFTWARE SECURITY

All of AIR's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with Arizona's privacy laws, the Family Educational Rights and Privacy Act (FERPA), and other federal laws.

AIR's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. Different states interpret the FERPA differently, and our system is designed to support these interpretations flexibly. AIR has worked with the ADE to maintain data security according to their specifications.

AIR maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results. In addition, AIR runs automated functional tests of our test delivery system every morning, and logs from these runs are available for at least one week from the time of the run.

AIR psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior and also provide a forensics analysis report. The forensics analysis report is described more completely in Section 6.6 on data forensics.

6.5 TEST SECURITY

Maintaining a secure test environment is critical to ensuring that scores represent what students know and are able to do. Because AzMERIT was administered both as a paper-based and a computer-based assessment, test security procedures must guard against item exposure, cheating on the part of test administrators or students, or other security problems for both testing modes.

The test security procedures involve the following:

- Procedures to ensure the security of test materials

- Procedures to investigate test irregularities

Test Administrators are trained on test security procedures, and both test security policies and procedures are clearly presented with the *AzMERIT Test Administration Directions*.⁴⁸

Security of Test Materials

All test items, test materials, and student-level testing information are secure documents and must be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment questions, prompts, and student results. Any deviation in test administration must be reported to ensure the validity of the assessment results. Mishandling of test administration puts student information at risk and disadvantages the student. Failure to honor security severely jeopardizes district and state accountability requirements and the accuracy of student data.

The security of all test materials must be maintained before, during, and after test administration. Under no circumstances are students permitted to assist in preparing secure materials before testing or in organizing and returning materials after testing. After any administration, initial or make-up, secure materials (e.g., test booklets, test tickets, used scratch paper) are required to be returned immediately to the School Test Coordinator and placed in locked storage. Secure materials are never to be left unsecured and are not to remain in classrooms or be taken off the school's campus overnight. Secure materials are never to be destroyed (e.g., shredded, thrown in the trash), except for soiled documents. In addition, any monitoring software that would allow test content on student workstations to be viewed or recorded on another computer or device during testing needs to be turned off.

It is unethical and viewed as a violation of test security for any person to:

- capture images of any part of the test via any electronic device;
- duplicate in any way any part of the test;
- examine, read, or review the content of any portion of the test;
- disclose or allow to be disclosed the content of any portion of the test before, during, or after test administration;
- discuss any AzMERIT test item before, during, or after test administration;
- allow students access to any test content prior to testing;
- provide any reference sheets to students during the Math test administration;
- allow students to share information during test administration;
- allow students to use scratch paper during the ELA Reading test;
- read any parts of the test to students except as indicated in the Test Administration Directions or as part of an accommodation;
- influence students' responses by making any kind of gestures (for example, pointing to items, holding up fingers to signify item numbers or answer options) while students are taking the test;
- instruct students to go back and reread/redo responses after they have finished their test since this instruction may only be given before the students take the test;
- review students' responses;
- read or review students' scratch paper; or

⁴⁸ Standard 6.7 – Test users have the responsibility of protecting the security of test materials at all times.

Standard 7.9 – If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.

- participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures.

Additional security violations for paper-based testing include:

- Reading or reviewing any test booklet during or after testing
- Changing any student response in test booklet
- Erasing any student's response in test booklet
- Erasing any stray marks in test booklet
- Failing to return all test booklets and other test materials

Test Administrators and Proctors may not assist students in answering questions. They may not translate, reword, or explain any test content. No test content may ever be discussed before, during, or after test administration.

All regular test booklets and special documents (large print and Braille) test materials are secure documents and must be protected from loss, theft, and reproduction in any medium. A unique identification number and a bar code were printed on the front cover of all test booklets. Schools were expected to maintain test security by using the security numbers to account for all secure test materials before, during, and after test administration until the time they were returned to the contractor.

To access the computer-based AzMERIT tests, a secure Internet browser is required. The secure browser provides a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to the desktop (Internet, email, and other files or programs installed on school machines). The secure browser did not display the IP address or other URL for the site. Users could not access other applications from within the secure browser, even if they knew the keystroke sequences. The "back" and "forward" browser options were not available, except as allowed in the testing environment as testing navigation tools. Students were not able to print from the secure browsers. During testing, the desktop was locked down, and students were required to "Pause" (to save the test for another session) or "Submit" a test in order to exit the secure browser. The secure browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *Test Administrator User Guide* for further details.

Throughout the testing window, test administrators were to report any test incidents (e.g., disruptive students, loss of Internet connectivity) to the School Test Coordinator immediately. A test incident could include testing that was interrupted for an extended period of time due to a local technical malfunction or severe weather. School Test Coordinators notified District Test Coordinators of any test irregularities that were reported. District Test Coordinators were responsible for submitting requests for test invalidations to the ADE via AIR's Test Information Distribution Engine, or TIDE. The ADE made the final decision on whether to approve the requested test invalidation. District Test Coordinators could track the status and final decisions of requested test invalidations in TIDE.

6.6 DATA FORENSICS PROGRAM

The validity of test score interpretation depends critically on the integrity of the test administrations on which those scores are based. Any irregularities in the administration of assessments can therefore cast doubt on the validity of the inferences based on those test scores. Multiple facets ensure that tests are administered properly, which includes clear test administration policies, effective test administrator training, and tools to identify possible irregularities in test administrations.

For online administrations, quality assurance (QA) reports are generated during and after the testing windows. These are geared toward detection of testing irregularities that may indicate possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies.

Online test administration allows Arizona's testing contractor to track information that was not possible to track in the context of the paper- pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations) as requested by the state. AIR's test delivery system (TDS) captures all of this information.

Unlike with paper assessments, where data analysis must await the close of the testing window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout each testing window, following the first operational administration. Following the base year, the analyses used to detect the testing anomalies can be run anytime within the testing window. Evidence evaluated included changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student level and summarized for each aggregate unit, including testing session, test administrator, and school.

6.6.1 CHANGES IN STUDENT PERFORMANCE

The report examines score changes between years using a regression model. The scores between the previous and current year assessments are compared, with the current-year score regressed on the test score from the previous year.

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. To detect unusual residuals, we compute the studentized t residuals. An unusual increase or decrease in student scores between opportunities is flagged when absolute studentized t residuals are greater than 3.

The number of students with a large score gain or loss is aggregated for a testing session, test administrator, and school. Unusual changes in an aggregate performance between administrations and/or years are flagged based on the average studentized t residuals in an aggregate unit g (e.g., a testing session or a test administrator). For each aggregate unit, a critical t value is computed and flagged when absolute t was greater than 3,

$$t = \frac{\text{Average residuals}}{\sqrt{\frac{s^2}{n_g} + \frac{\sum_{j=1}^{n_g} \text{var}(e_i)}{n_g^2}}}$$

where s = standard deviation of residuals in an aggregate unit; n_g is number of students in the aggregate unit g (e.g., testing session or test administrator); and $\text{var}(e_i) = \sigma^2(1 - h_{ii})$. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

The aggregate unit size for the score change is based on the number of students included in the within- or between-year regression analyses in the aggregate unit. If the aggregate unit size is 1–5 students, the aggregate unit was flagged if the percentage of flagged students was greater than 50%.

6.6.2 ITEM RESPONSE LATENCY

The online environment also allows item response latency to be captured as the item page time (the time each item page is presented) in milliseconds. Discrete items appear one item on the screen at a time. However, for stimulus-based items selected as part of an item group, all items associated with the stimulus are selected and loaded as a group. For each student, the total time taken to complete the test is computed by summing up the page time for all items and item groups.

It is expected that item response time is shorter than the average time if students have prior knowledge of test items. An example of unusual item response time would be a test record for an individual who scores very well on the test even though the average time spent for each item was far less than that required of students statewide. If students already know the answers to the questions, the response time will be much shorter than the response time for those items where the student has no prior knowledge of the item content. Conversely, if a test administrator helps students by “coaching” them to change their responses during the test, the testing time could be longer than expected.

The average and the standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units were flagged if the test-taking time was greater than |3| standard deviations of the state average. The state average and standard deviation was computed based on all students at the time the analysis was performed.

6.6.3 INCONSISTENT ITEM RESPONSE PATTERN (PERSON FIT)

In Item Response Theory (IRT) models, person-fit measurement is used to identify examinees whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test taker has prior knowledge of some test items (or is provided answers during the exam), the student will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response latency index might flag such a student.

The person-fit index is based on all item responses. An unlikely response to a single test question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session and test administrator.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), Sotaridona, Pornell, and Vallejo (2003), define aberrant response patterns as a deviation from the expected item score model. Snijders (2001) showed that the distribution of I_2 is asymptotically normal (i.e., with an

increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the “asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05” (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using I_z for systematic flagging of aberrant response patterns. Students with $|I_z|$ values greater than 3 are flagged. Aggregate units are flagged with $|t|$ greater than 3, where t is calculated by

$$t = \frac{\text{Average } I_z \text{ values}}{\sqrt{s^2/n}},$$

where s = standard deviation of I_z values in an aggregate unit; n = number of students in an aggregate unit, e.g., testing session, or test administrator. The QA report will include a list of the flagged aggregate units with the number of flagged students in the aggregate unit (school, test administrator, test session).

6.6.4 RESPONSE CHANGE AND RESPONSE SIMILARITY

Response Change in Paper-Based Tests

Erasure patterns on paper-pencil tests are also examined for unusual patterns of response changes. For paper-based assessments, we use differences in mark density to infer student erasures, which is then used to identify instances where students may have changed an initial response from incorrect to correct, from incorrect to incorrect, or from correct to incorrect. A set of flagging rules is then used to identify an unusually large number of incorrect to correct erasures at the targeted level of analysis, whether student, testing group, or school. In the online environment, students may change their responses multiple times, and each of those response changes is recorded. Unlike with the mark discrimination analyses, there is no ambiguity about which response was selected or the order in which responses were made. The ease with which response changes can be made, and the accuracy of response capture (i.e., students no longer need to worry that an “erased” response might result in the detection of multiple marks that either cannot be resolved or do not correspond to the student’s intended response) mean that students may now feel freer to change responses, even multiple times for a single item.

Response Pattern Similarity in Computer-Based Tests

In fixed-form assessment environments, students may more readily copy from one another than would be possible in a computer adaptive test environment where students are seeing different sets of items in different sequences. To detect possible copying, it can be useful to examine student response records for patterns of excessive response similarity. While similarity in student responses to test questions may be an indicator of irregularities in test administration, response similarity does not always indicate a testing irregularity. For example, in schools with high levels of academic achievement, one would expect large numbers of students to respond correctly, and therefore similarly, to most items on the test. Nevertheless, patterns of similar responding can indicate testing irregularities, especially when students respond to items incorrectly in the same way. We employ an algorithm, following the model developed by Wesolowsky (2000), for detecting overly similar student responses to multiple-choice items to evaluate patterns of student responses in schools where test irregularities are suspected. This

study uses the similarity of responses between a pair of students to estimate the probability of possible cheating. The computational steps are as follows:

1. Based on assumptions and probability theory (pp 911-912), \hat{p}_{ji} is estimated by solving the following two equations

$$\begin{cases} p_{ji} = (1 - (1 - r_i)^{a_j})^{1/a_j} \\ \frac{\sum_{i=1}^q p_{ji}}{q} = c_j \end{cases}$$

for a_j , and from \hat{a}_j and r_i to obtain $\hat{p}_{ji} = (1 - (1 - r_i)^{\hat{a}_j})^{1/\hat{a}_j}$, where r_i is the proportion of the analysis unit (e.g., school) that answered correctly on item i , c_j is the proportion of items answered correctly by student j ;

2. W_{it} is the probability that, conditional on the answer being wrong, distractor t is chosen on question i . For now, this is estimated by the proportion of students who choose option t over students who choose wrong options on this item;

3. Using estimates from steps 1 and 2 to estimate $\hat{\mu}_{jk}$ and $\hat{\sigma}_{jk}^2$, hence, Z_{jk} ;

4. Based on Z_{jk} and significant level to decide if the students j and k have significant probability to copy each other.

In order to investigate the probability of false positive of the estimating procedure, the procedure is applied to estimate the probability of cheating for each pair within each aggregate unit (school/session), and two Bonferroni adjustments are used, one of which is based on $(n-1)$, and the other of which is based on $(n(n-1)/2)$, where n is the number of students within the aggregate unit (school/session).

Aggregate units are flagged with two different methods: aggressive method and conservative method. The aggressive method uses an $\alpha=0.05$ and Bonferroni adjustment factor $(n-1)$ to flag test sessions and schools. The more conservative method uses $\alpha=0.01$ and Bonferroni adjustment factor $(n(n-1)/2)$ to flag suspect test sessions and schools.

Bonferroni adjustment with factor $(n-1)$ is used if we know the seating of the students and the possible cheating can only happen between the front and back student pair. If no seating chart is available, the factor $(n(n-1)/2)$ is usually used. Based on simulation studies, the results based on $(n(n-1)/2)$ provide a good safety buffer against the false positive, that we see only a slight chance of false positive. As for the alpha level, it seems that using $\alpha=.01$ is preferred, so only extreme pairs that are worth investigation will be flagged.

The basic unit of analysis for evaluating response similarity in fixed form assessments is the test session. For each pair of students in a session, we compute the probability of obtaining the same response for each item, including the likelihood of answering the item correctly, as well as selecting the same incorrect response option when answering an item incorrectly. The probability of two students answering an item correctly is conditioned on the average performance of other students in the school. The Bonferroni adjustment is used to correct for the large

number of pairwise comparisons, reducing the likelihood of Type I (false positive) errors. A response similarity report identifies pairs of students with overly similar patterns of responding. Exhibit 6.6.4.1 provides sample output for the response similarity analysis. Each record indicates a pair of students flagged for overly similar patterns of responding. Access to a seating chart increases the power of this approach significantly, since students with overly similar response patterns who are known to have been seated in close proximity obviously have greater opportunity to copy their responses. This method is also useful for detecting cheating rings, where the same students are identified across multiple flagged pairs. This is evident in Exhibit 6.6.4.1, where a common group of students are each flagged in multiple comparisons.

Exhibit 6.6.4.1 Sample Roster Flagging Student Pairs with Excessively Similar Responses

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 Last Name	Student1 First Name	Student2 Barcode	Student2 Last Name	Student2 First Name
SchoolA	Class1	Reading	18		Carter	Adam		Doe	Frank
SchoolA	Class1	Reading	18		Carter	Adam		Farmer	Fred
SchoolA	Class1	Reading	18		Carter	Adam		Miller	Steve
SchoolA	Class1	Reading	18		Carter	Adam		Smith	Cecil
SchoolA	Class1	Reading	18		Carter	Adam		Carter	Henry
SchoolA	Class1	Reading	18		Carter	Adam		Turner	Mark
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Carl
SchoolA	Class1	Reading	18		Carter	Adam		Hall	Robert
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Phillip
SchoolA	Class1	Reading	18		Doe	Frank		Farmer	Fred
SchoolA	Class1	Reading	18		Doe	Frank		Carter	Henry
SchoolA	Class1	Reading	18		Doe	Frank		Hall	Robert
SchoolA	Class1	Reading	18		Doe	Frank		Granger	Phillip
SchoolA	Class1	Reading	18		Farmer	Fred		Miller	Steve
SchoolA	Class1	Reading	18		Farmer	Fred		Smith	Cecil
SchoolA	Class1	Reading	18		Farmer	Fred		Carter	Henry
SchoolA	Class1	Reading	18		Farmer	Fred		Turner	Mark
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Carl
SchoolA	Class1	Reading	18		Farmer	Fred		Hall	Robert
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Phillip
SchoolA	Class1	Reading	18		Miller	Steve		Smith	Cecil
SchoolA	Class1	Reading	18		Miller	Steve		Carter	Henry
SchoolA	Class1	Reading	18		Miller	Steve		Turner	Mark
SchoolA	Class1	Reading	18		Miller	Steve		Hall	Robert
SchoolA	Class1	Reading	18		Miller	Steve		Granger	Phillip

7. REPORTING AND INTERPRETING AZMERIT SCORES

A set of score reports that summarizes student performance in each grade and content area is provided for each administration. Score reports provide data on the performance of individual students and on the aggregated performance of students at various levels — such as state, districts, schools, and teachers. The test data are based on all students who participated in the AzMERIT assessment for the 2016–2017 school year.

The score reports include reliable and valid information describing student progress toward mastery of the state content standards. Arizona provides individual student score reports that are shipped to the student’s district for delivery to families. These reports detail student performance on overall tests and subscores. In addition, Arizona offers detailed individual- and aggregate-level data to educators via AIR’s Online Reporting System (ORS), which provides score data for each AzMERIT test, both computer-based and paper-based. The ORS allows users to compare score data between individual students and the school, district, or overall state, and also provides information about performance on subscore categories.

7.1 APPROPRIATE USES FOR SCORES AND REPORTS

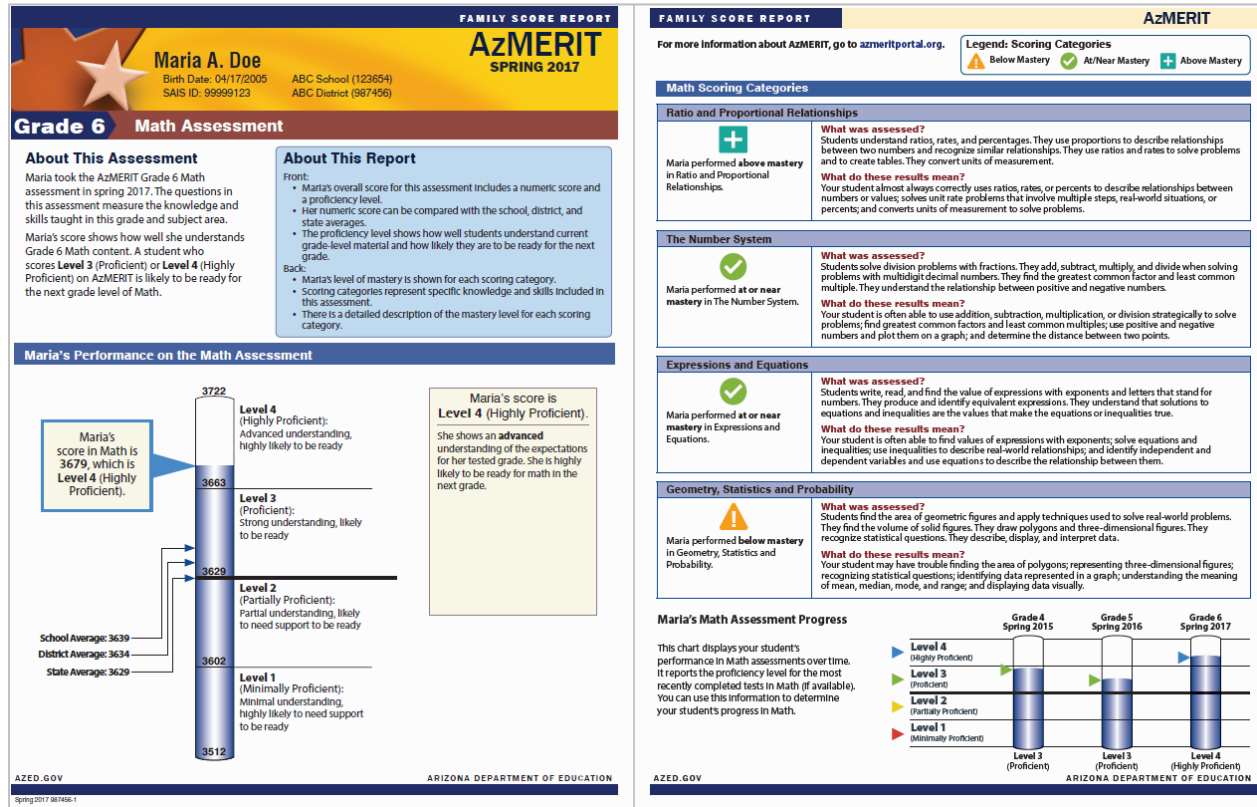
The state provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning and classroom instruction. All reporting systems for the AzMERIT, both paper and online, are designed with stakeholders in mind—such as teachers, parents and students, who are not technical measurement experts—and ensure that test results are used in ways that lead to valid inferences about student achievement and contribute to student learning.⁴⁹ For example, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy guides the reader to compare like elements and avoid comparison of dissimilar elements.

Sample reports are available at <https://azmeritportal.org>. The upcoming sections provide additional guidance for interpreting results.

⁴⁹ Standard 6.10 – When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. Standard 13.5 – Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied.

7.2 REPORTS PROVIDED

7.2.1 FAMILY REPORTS



Arizona provides full-color individual student reports to families of all AzMERIT testers. Reports are designed to be useful to families, and include:

- Full color to aid readers' interpretation of the data
- Scale scores and performance-level descriptors
- Scoring category performance, including descriptions of what was assessed and what results mean for each scoring category to guide parents and students in their understanding of student scores, including:
 - A plus (+) symbol indicates that a student is performing above mastery in a particular scoring category,
 - A checkmark indicates that a student is performing at or near mastery within the scoring category.
 - The exclamation symbol indicates a student is performing below mastery in a scoring category.
- Rubric scores for the writing portion of the ELA test, including descriptions of what those rubric scores mean
- School, district, and state average scores for comparative purposes

In addition, beginning with the spring 2016 administration, the ADE provided reports that included longitudinal data as seen at the bottom of the second page of the report. This data is designed to allow parents to track student achievement over time.

7.2.2 ONLINE REPORTING SYSTEM FOR EDUCATORS

AzMERIT results are also reported using AIR's Online Reporting System (ORS), which is designed to support educators as they evaluate the needs of their students and reflect on their own curricula and practice. Navigation in the system mirrors the instructional decision-making process, meaning the user can intuitively navigate in any of the three dimensions inherent in the data, helping the user answer three kinds of questions:

1. **Who?** The data can be displayed at levels of aggregation anywhere from the individual level for a specific student up to the entire state. Demographic breakdowns are immediately available at any level of aggregation.
2. **What?** The subject area data can be broken down into finer or coarser "chunks" of content. Navigating this dimension allows the user to travel from subject to scoring category and back.
3. **When?** When data are available over time, the system allows the user to view a data trend over time or toggle to a fixed point in time.

Each navigational step changes the reporting display, providing richer context when interpreting a class's or individual student's performance. While the system contains many reports, the interface design encourages users to think about the substantive, educational questions to which they need answers and access information from that perspective. In addition, while finding and interpreting data from multiple online assessments can easily become overwhelming, the ORS minimizes information overload for educators and administrators by organizing score information in a conceptual framework that helps users quickly locate the right level of data, evaluate its impact, and identify the concrete actions they can take to help students improve.

The AzMERIT online system produces the following online score reports: individual student reports, and aggregate reports at the teacher, school, district, and state level. The AzMERIT online score reports are structured hierarchically. Upon selecting "Home" on the Welcome page, a user is taken to the Home Page Dashboard, which displays for all grades and content areas the number of students tested and the percentage of students passing by grade and content area. Users who have access to multiple districts or schools are first required to select a single district or school. Once an aggregate unit is selected in this instance, the summary table of student performance is displayed for the selected entity. For more detailed information for a subject and a grade, the user must select that subject and grade.

On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district the school belongs to are provided above the school summary results so that the school performance can be compared with performance in the district and the state. If a teacher is selected, the summary results for state, district, and school are provided above the summary results for the teacher.

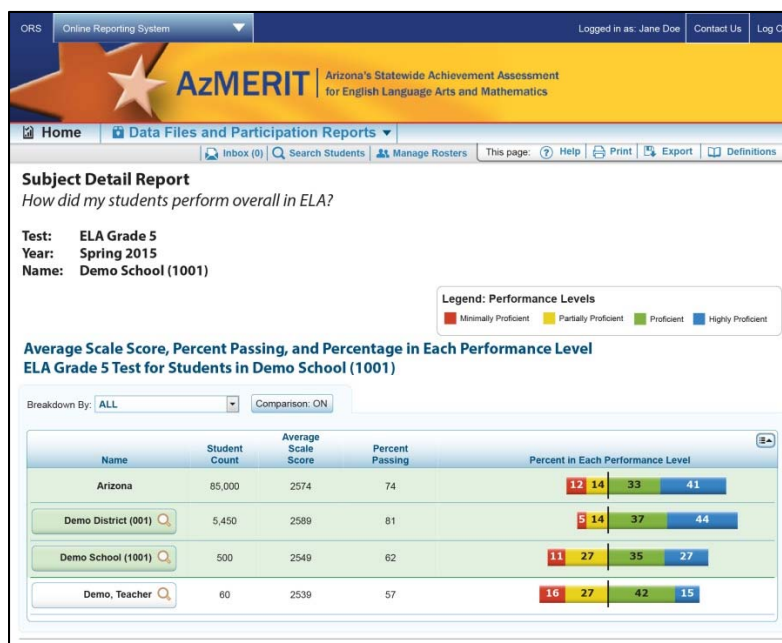
Exhibit 7.2.2.1 summarizes the types of online score reports available and the levels at which they can be viewed (e.g., student, roster, teacher, school, district).

Exhibit 7.2.2.1 AzMERIT Online Score Report Summary

Type of Report Page	Level of Aggregation	Description
Home Page Dashboard	District, school, and teacher	Summary of performance and participation (Number Tested and Percentage Passing) across grades and subjects or course

Type of Report Page	Level of Aggregation	Description
Subject Detail	District	Average scale score, percentage passing, and percentage at each performance level for a district and each school within that district; ability to disaggregate data by subgroup
	School	Average scale score, percentage passing, and percentage at each performance level for a school and each teacher within that school; ability to disaggregate data by subgroup
	Teacher	Average scale score, percentage passing, and percentage at each performance level for a teacher and each class roster associated with that teacher; ability to disaggregate data by subgroup
Scoring Category Detail	District, school, teacher, and roster	Performance on the scoring category for a subject and a grade for all students and by subgroups; relative strength and weakness indicator is also reported for each category
Student Roster	School, teacher, roster	List of students with performance on overall subject and scoring categories for a group of students associated with a school, teacher, or roster
Individual Student Report	Student	Student performance for a selected subject; report includes performance on each scoring category, and performance on the writing essay dimensions, if applicable

SUBJECT DETAIL REPORTS

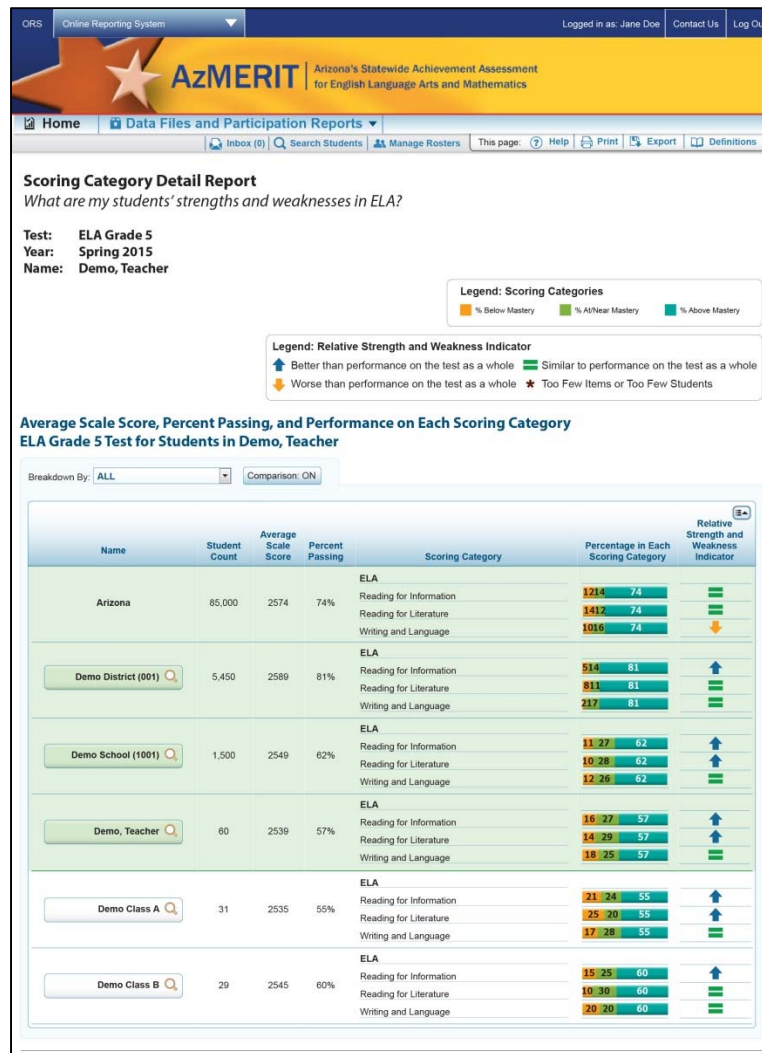


Aggregated subject reports show average performance for the state, districts, schools, teachers, and classes. Bar charts show the distribution of students' performance levels. These reports provide users with rosters of schools, teachers, and classes, allowing for simple comparisons across smaller groups.

The Subject Detail Report page shows the following data:

- **Student Count:** Number of students who have completed the selected test
- **Average Scale Score:** Average scale score of students who completed the selected test
- **Percent Passing:** The percentage of tested students reaching the proficient threshold on the selected test
- **Percent in Each Performance Level:** The distribution of students across each of the four performance levels

SCORING CATEGORY DETAIL REPORTS



Aggregated scoring category detail reports follow the layout of the subject detail reports, displaying the performance data for the state, districts, schools, teachers, and classes. In addition, these reports include a relative strength and weakness indicator for each category.

In addition to overall test scores, reporting category performance is reported as a strength and weakness indicator. The performance levels indicated on this report are relative to the test as a whole. Unlike performance levels provided at the subject level, these strengths and weaknesses do not imply proficiency. Instead, they show how

the performance of a group of students is distributed across the scoring categories relative to their overall subject performance on a test. For example, a group of students may have performed very well in a subject, but performed slightly lower in several scoring categories. Thus, the orange “down” sign for a scoring category does not imply a lack of proficiency. Instead, it simply communicates that these students’ performance on that scoring category was statistically lower than their performance across all other scoring categories put together. Although the students are doing well, an educator may want to focus instruction on these areas.

STUDENT ROSTER REPORTS

Student Roster Report – Students' Performance on Each Scoring Category
How did my students perform on the ELA test?

Test: ELA Grade 5
Year: Spring 2015
Name: Demo Class A

Legend: Performance Levels
1 Minimally Proficient 2 Partially Proficient 3 Proficient 4 Highly Proficient

Legend: Scoring Categories
Below Mastery At/Near Mastery Above Mastery

Breakdown By: ALL GO

Comparison Scores

Name	Average Scale Score
Arizona	2574
Demo District (001)	2580
Demo School (1001)	2540
Demo Teacher	2539
Demo Class A	2535

Scale Scores and Performance Levels
ELA Grade 5 Test for Students in Demo Class A

Name	SSID	Scale Score	Performance Level	Reading for Information	Reading for Literature	Writing and Language
Student A	999997890	2595	4	+	+	+
Student B	999997891	2532	2	⚠	✓	✓
Student C	999999912	2580	4	+	+	+
Student D	9999990123	INV	INV	INV	INV	INV
Student E	9999901234	2553	3	✓	✓	✓
Student F	9999912345	2527	2	✓	⚠	⚠

Student roster reports provide users with performance data for a group of students associated with a teacher or a school, as defined in TIDE. The report includes each student’s unique state ID, overall subject score, and overall subject performance level. Using the exploration menu, a user can also view each student’s scoring category performance for the selected test.

The table that appears on the Student Roster Report page shows the following data:

- **Scale score:** The score of each student who completed the test
- **Performance level:** Represents levels of overall subject mastery with respect to the Arizona College and Career Ready Standards (4, representing Highly Proficient, to 1, representing Minimally Proficient)

- **Scoring Categories:** Represents levels of scoring category mastery with respect to the Arizona College and Career Ready Standards, characterizing achievement at “above,” “at or near,” or “below” mastery on each scoring category

INDIVIDUAL STUDENT REPORTS

Individual Student Reports, which closely mirror the Family Reports, are also available through the Online Reporting System.

7.3 INTERPRETATION OF SCORES

Arizona provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning, including interpretive guides for navigating the Online Reporting System (ORS) and understanding paper family reports.⁵⁰ This section describes many of the measures presented in the paper and online score reports.

Performance levels represent levels of mastery with respect to the Arizona College and Career Ready Standards (ACCRS) for a content-area assessment. Performance levels are labeled as Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance standards are the points on the achievement scale that differentiate performance levels. Three performance standards are used to classify students into one of the four performance levels. Performance standards were recommended by panels of Arizona educators following the first administration of AzMERIT in 2015, and subsequently adopted by the Arizona State Board of Education. Panelists engaged in a rigorous, technically-sound standard-setting process that is summarized in the Performance Standards section of this technical manual, and documented in detail in the 2015 standard-setting technical report, available from the ADE.

Performance-Level Descriptors, or PLDs, define the content area knowledge, skills, and processes that examinees at a performance level are expected to possess. The descriptions of Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient performance are the public statements about what and how much Arizona educators want students to know and be able to do for each grade level and content area. The very detailed PLDs are summarized and included in score reports to provide context for the score and are designed to help parents understand what their students can and cannot do.

The student’s performance in each content area assessment is summarized in an overall test score referred to as a scale score. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale score is then used to determine how well students perform on each content area assessment. Scale scores can be used to measure how much students know and are able to do. Scale scores can also be used to compare student performance across administrations for the same grade and content area so that, for example, an average scale score of 2450 for grade 3 students in the 2015–2016 school year indicates the same level of achievement as an average scale score of 2450 for grade 3 students in the 2016–2017 school year, even though the test may include a slightly different set of items.

⁵⁰ Standard 12.18 – In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

As described in Section 9 on Scaling and Equating, for the ELA assessment, the scale score reported can range from 2395 to 2675. For the math assessment, the scale score reported can range from 3395 to 3839. Overall scale scores for ELA and math are mapped into four performance levels using three performance standards (i.e., cut scores). The AzMERIT scale score ranges can be found in Exhibit 7.3.1.

Exhibit 7.3.1 AzMERIT Scale Score Ranges

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
<i>ELA</i>				
Grade 3	2395–2496	2497–2508	2509–2540	2541–2605
Grade 4	2400–2509	2510–2522	2523–2558	2559–2610
Grade 5	2419–2519	2520–2542	2543–2577	2578–2629
Grade 6	2431–2531	2532–2552	2553–2596	2597–2641
Grade 7	2438–2542	2543–2560	2561–2599	2600–2648
Grade 8	2448–2550	2551–2571	2572–2603	2604–2658
Grade 9	2454–2554	2555–2576	2577–2605	2606–2664
Grade 10	2458–2566	2567–2580	2581–2605	2606–2668
Grade 11	2465–2568	2569–2584	2585–2607	2608–2675
<i>Math</i>				
Grade 3	3395–3494	3495–3530	3531–3572	3573–3605
Grade 4	3435–3529	3530–3561	3562–3605	3606–3645
Grade 5	3478–3562	3563–3594	3595–3634	3635–3688
Grade 6	3512–3601	3602–3628	3629–3662	3663–3722
Grade 7	3529–3628	3629–3651	3652–3679	3680–3739
Grade 8	3566–3649	3650–3672	3673–3704	3705–3776
Algebra I	3577–3660	3661–3680	3681–3719	3720–3787
Geometry	3609–3672	3673–3696	3697–3742	3743–3819
Algebra II	3629–3689	3690–3710	3711–3750	3751–3839

ELA and math assessments are reported on a vertical scale. The item response theory (IRT) vertical scale was developed in 2015 by embedding operational test items from the grade above in the embedded field test slots of each grade level assessment.

8. PERFORMANCE STANDARDS

In the summer of 2015, following the close of the first testing window, AIR convened panels of Arizona educators to recommend performance standards on each of the AzMERIT assessments. Details of the panels, procedures, and outcomes are documented in the “Recommending AzMERIT Performance Standards” technical report, which is available from ADE.⁵¹ This section briefly describes the procedures used by educators to recommend standards, and resulting performance standards.

8.1 STANDARD SETTING PROCEDURES

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the Arizona College and Career Ready Standards (ACCRS). The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade-level expectations for achievement of the ACCRS, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standards for the AzMERIT assessments are therefore central to the validity of test score interpretations.

Following the first operational administration of the AzMERIT assessments in spring 2015, a standard-setting workshop was conducted to recommend to the Arizona State Board of Education a set of performance standards for reporting student achievement of the ACCRS. The workshop consisted of a series of standardized and rigorous procedures that the Arizona educators serving as standard-setting panelists followed to recommend performance standards. The workshops employed the Bookmark procedure, a widely used method where standard-setting panelists used their expert knowledge of the ACCRS and student achievement to map the performance-level descriptors adopted by the Arizona State Board of Education onto an ordered item book (OIB) based on the first operational test form administered in spring 2015.

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college-ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standards for the grade 3–8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3–8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS

⁵¹ Standard 5.21 – When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Standard 7.4 – Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.

performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

Panelists were also provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade level assessment related to the cut-score recommendations at the other grade levels. This approach allowed panelists to view their cut-score recommendations as a coherent system of performance standards, and further reinforced the interpretation of test scores as indicating not only achievement of current grade-level standards, but also preparedness to benefit from instruction in the subsequent grade level.

8.1.1 PERFORMANCE-LEVEL DESCRIPTORS

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance-Level Descriptors (PLDs) define the content-area knowledge and skills that students at each performance level are expected to demonstrate. The standard-setting panelists based their judgments about the location of the performance standards on the PLDs as well as the Arizona College and Career Readiness Standards. The AzMERIT PLDs describe four levels of achievement:

- Minimally Proficient
- Partially Proficient
- Proficient
- Highly Proficient

Prior to convening the standard-setting workshops, AIR, in consultation with the ADE, drafted PLDs for each test that described the range of achievement encompassed by each performance level on the test. The PLDs were designed to be clear, concrete, and reflect Arizona's expectations for proficiency based on the ACCRS. Following a cycle of revisions to the draft PLDs, the ADE invited Arizona educators to review PLDs for each of the assessments. Based on feedback from 166 educators, PLDs were further revised, and the resulting drafts were used by standard-setting panelists. ADE considered any need for clarification or revision that arose throughout the standard-setting process prior to publishing the final versions of the PLDs following the standard-setting workshop. AzMERIT PLDs are available at www.azed.gov.

8.2 RECOMMENDED PERFORMANCE STANDARDS

Panelists were tasked with recommending three performance standards (Partially Proficient, Proficient, and Highly Proficient) that resulted in four performance levels (Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient). Exhibit 8.2.1 presents the performance standard associated with panelist-recommended OIB page numbers in logit value (theta), as well as the percentage of students classified as meeting or exceeding each standard. Following the standard-setting workshop, panelist recommendations were submitted to the Arizona State Board of Education; the Board formally adopted the standards in August 2015.

Exhibit 8.2.1 Final Recommended Performance Standards for AzMERIT

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% at or Above	Theta	% at or Above	Theta	% at or Above
ELA						
3	-0.09	56	0.29	41	1.36	10
4	0.14	57	0.6	39	1.8	5
5	-0.13	63	0.63	30	1.8	3
6	-0.12	61	0.58	34	2.03	4
7	-0.02	59	0.61	33	1.9	4
8	-0.06	60	0.64	33	1.72	6
9	-0.12	53	0.59	27	1.57	6
10	0.11	51	0.58	30	1.42	8
11	-0.02	46	0.52	26	1.27	8
Math						
3	-0.16	73	1.04	42	2.43	15
4	-0.31	71	0.76	42	2.2	10
5	-0.65	71	0.41	40	1.74	13
6	-0.48	62	0.41	32	1.55	11
7	-0.19	52	0.59	30	1.51	13
8	-0.69	57	0.09	32	1.15	13
Algebra I	-0.69	55	-0.03	32	1.27	9
Geometry	-1.37	53	-0.58	30	0.96	6
Algebra II	-1.49	53	-0.78	29	0.57	6

Exhibit 8.2.2 shows the percentage of students classified at each performance level in the initial year of AzMERIT administration, based on final panelist-recommended standards for the student population overall across grade levels and courses for the ELA and math assessments.

Exhibit 8.2.2 Percentage of Students at Each Performance Level based on Final Recommended Performance Standards

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
ELA				
3	44%	15%	31%	10%
4	43%	19%	33%	5%
5	37%	33%	27%	3%
6	39%	27%	30%	4%
7	41%	26%	29%	4%
8	40%	27%	26%	6%
9	47%	26%	21%	6%
10	49%	21%	22%	8%
11	54%	20%	17%	8%
Math				
3	27%	31%	27%	15%
4	29%	29%	32%	10%
5	29%	31%	27%	13%
6	38%	30%	21%	11%
7	48%	22%	18%	13%
8	43%	24%	20%	13%
Algebra I	45%	23%	23%	9%
Geometry	47%	24%	24%	6%
Algebra II	47%	24%	23%	6%

Exhibit 8.2.3 shows the percentage of students meeting the AzMERIT proficient standard for each assessment in the base year of 2015 (meaning they are categorized as Proficient or Highly Proficient), and the approximate percentage of Arizona students that would be expected to meet the ACT college-ready standard, the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8, and the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2015 field test administration. As Exhibit 8.2.3 indicates, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college-ready, and the NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 8.2.3 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Grade/ Course	Percentage of Students Meeting Standard			
	AzMERIT Proficient	Arizona ACT College-Ready	Arizona NAEP Proficient	Projected SBAC
ELA				
3	41%			38%
4	38%		28%	41%
5	30%			44%
6	34%			41%
7	33%			38%
8	32%		28%	41%
9	27%			
10	30%			
11	25%	34%		41%
Math				
3	42%			39%
4	42%		42%	38%
5	40%			33%
6	32%			33%
7	31%			33%
8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

9. SCALING AND EQUATING

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^n P(z|\theta),$$

where Z represents the pattern of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function $P(Z)$. The one-parameter model (1PL; also known as the Rasch model), is used to calibrate AzMERIT items that are scored either right or wrong, and takes the form

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i .

The b parameter is often called the *location* or *difficulty* parameter; the greater the value of b , the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i|\theta) = \frac{\exp \sum_{k=0}^{x_i} (\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l (\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^0 (\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i . Item parameters for the assessments were calibrated following the spring administration in 2015 and vertical scales were established for reporting both ELA and math. In addition, a series of linking studies were performed to allow the comparison of performance on the AzMERIT to other state and national scales. A mode comparability study was also completed to examine possible effects of test administration mode. These studies were completed prior to establishing performance standards in summer 2015 and subsequent scoring and reporting of AzMERIT results. AzMERIT ELA is reported on a scale ranging from 2395 to 2675 across the grade-level and high school End-of-Course tests. AzMERIT math is reported on a scale ranging from 3395 to 3839 across grade-level and high school End-of-Course (Algebra I, Geometry, and Algebra II) tests.

9.1.1 ITEM RESPONSE THEORY PROCEDURES

The AzMERIT assessment was administered for the first time in the spring of 2015. Following test administration, item response theory (IRT) procedures were used to calibrate item parameter estimates and create the new AzMERIT scales for scoring and reporting.⁵² This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by AIR, ADE, and HumRRO, which acts as a third-party quality assurance contractor.

Within AzMERIT, students are able to skip items in both the online and paper test platforms. While omitted items are scored as incorrect for purposes of ability estimation, all omitted responses are treated as not-administered for purposes of IRT analysis. All students who respond to at least one item within each test session are considered to have attempted a test. All attempted records are included in IRT analysis with the exclusion of students who had more than one record for the same test and records that are had been invalidated prior to scaling.

9.1.2 CALIBRATION OF AZMERIT ITEM BANKS

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for AzMERIT. Winsteps is publicly available software from Mesa Press. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters' (1982) partial credit model, an extension of the one parameter Rasch model which allows for partial credit to be given on items, estimates the responses for polytomous items.

In spring 2015, operational items for each test were freely calibrated establishing the new AzMERIT reference scales. Following the approval of final item parameter estimates for operational items, parameter estimates for the operational items were anchored to their new AzMERIT bank values and parameter estimates for field test and linking items were estimated under that constraint. This placed parameter estimates for all field test and external-linking items on the same AzMERIT scale defined by the operational item parameters.

In spring 2016, pre-equated item parameters were used to score student test records for the math assessments. For ELA, since six new writing tasks at each grade were being administered in the ELA assessments, operational ELA items were recalibrated, and the equating constant necessary to place the common items back to the reference scale was identified and applied to the recalibrated item parameters. This placed all test items on the base year AzMERIT scale. Mean equating was used to compute the linking constant, and all operational reading items were included in the linking computation.

⁵² Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

9.1.3 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

To identify the likelihood of a student's ability across the ability distribution, we begin by evaluating the likelihood of achieving a score point for an item given the underlying level of ability. Let X_i be a random variable taking a student's response on item i ($i = 1, \dots, N$) with an outcome $x_i \in \{0, 1, \dots, m_i\}$. Item i is a dichotomously scored item if $m_i = 1$, and polytomously scored item if $m_i > 1$. Based on Masters' (1982) partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i | \theta) = \frac{\exp \sum_{k=0}^{x_i} (\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l (\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^0 (\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i . Note that if item i is a dichotomously scored item, the partial credit model becomes the Rasch model and can be written as

$$P(X_i = 1 | \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i .

LIKELIHOOD FUNCTION

The likelihood function of ability θ given responses to N items, $\mathbf{x} = \{x_i\}$, can be expressed as:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^N P(x_i | \theta).$$

The maximum likelihood estimate $\hat{\theta} = \arg \max_{\theta} L(\theta | \mathbf{x})$ or equivalently, $\hat{\theta} = \arg \max_{\theta} \ln L(\theta | \mathbf{x})$.

DERIVATIVES

Finding the maximum likelihood estimate requires an iterative method, such as Newton-Raphson iterations. Since the log-likelihood is a monotonic function of the likelihood, the following derivatives based on the log-likelihood function are used:

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \theta} &= \sum_{i=1}^N \left[x_i - \sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right] \\ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} &= \sum_{i=1}^N \left[\sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right]^2 - \sum_{i=1}^N \sum_{x_i=0}^{m_i} x_i^2 P(X_i = x_i | \theta) \end{aligned}$$

The maximum likelihood estimates of θ is found via the following iterative routine:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\partial \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t} / \frac{\partial^2 \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t^2}.$$

This iterative process repeats until the difference between $\hat{\theta}_t$ and $\hat{\theta}_{t+1}$ is less than a pre-specified threshold.

ESTIMATING ZERO AND PERFECT SCORES

In the event of zero or perfect scores, a procedure recommended by Berkson (as cited in Linacre, 2004) is implemented to add (or subtract) 0.5 to (or from) the test score prior to estimating student ability. Thus, for students responding incorrectly to all items in a scale or subscale, students will be assigned a test score of 0.5. Conversely, for students responding correctly to all items in a scale or subscale, 0.5 will be subtracted from the raw score prior to calibration.

9.2 ESTABLISHING A VERTICAL SCALE IN ELA AND MATH

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade level assessment (g) were embedded in the field test slots of the assessment in the grade below (g - 1).⁵³ In this approach, the resulting linkage represents student achievement each year on the scale of the subsequent grade-level assessment for which they are preparing to receive instruction. As such, the scale scores for each assessment can be interpreted as a pre-test score for measuring student acquisition of academic content in the subsequent grade level. While this approach risks administering to students 1–2 items measuring content that they may not yet have had the opportunity to learn, it provides a more sensitive measure of student growth than could be obtained by a linking design in the linkage represents continued growth on academic content assessed in the previous year's assessment.

9.2.1 LINKING ITEMS

Since the vertical scale essentially places each AzMERIT assessment on the scale for the assessment in the grade above, we can best assure comparability of test scores between the grades by establishing the linkage using all available operational test items. Thus, to link the grade 4 assessments to the grade 5 scales, all operational items in the grade 5 assessment were made available for administration in the grade 4 embedded field test (EFT) slots. The inclusion of all operational items in the vertical linking set ensures that the item set used to link to the target adjacent grade scale fully represents the measured construct in the target grade, allowing for valid inferences to be made with respect to student baseline performance for achievement in the subsequent grade level.

Because the AzMERIT assessments of English language arts (ELA) in high school continue as end-of-course (EOC) or grade-level measures of student achievement of the Arizona College and Career Ready Standards (ACCRS), each assessment can be linked to the grade above using all available operational items.

However, AzMERIT assessments of high school math are composed of a set of EOC tests that are not as consistently associated with grade-level instruction and which measure specific subsets of the content domain. For example, while math coursework in high school follows a typical progression and it would therefore be possible to embed “grade 9” Algebra I EOC items in the grade 8 math assessment, embed the “grade 10” Geometry EOC items in the Algebra I EOC exam, and embed the “grade 11” Algebra II the Geometry exam, the constructs measured

⁵³ Standard 5.0 – Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.

Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

across the four exams vary considerably and have implications for the interpretation of growth, or lack thereof, across assessments. For example, it is not clear what the expectation for growth should be in a vertical scale established by embedding Geometry items in an Algebra I exam, since Geometry is not a focus of instruction in Algebra I courses. An alternative approach, and the one adopted by the ADE, was to link the grade 8 math scale to both the Algebra I and Geometry EOC scales, since the grade 8 assessment includes items measuring both algebra and geometry. Because Algebra II builds on the knowledge and skills assessed in Algebra I, all Algebra II items were used to link the Algebra I assessments to the Algebra II scale.

9.2.2 LINKING ANALYSIS

When feasible, it is desirable to establish linkages using both concurrent calibrations and chain-linking approaches to ensure that results are consistent across methods. An important advantage of chain linking approaches is that, because item response theory (IRT) calibrations proceed by establishing the within-grade scale, the achievement construct intended by the blueprint and enacted in the operational test form is preserved. Unfortunately, however, at each step in the linking chain, the linking error accumulates, so that linking constants for grades more distant from the reference grade are less precise than are linking constants for grades in closer proximity to the reference grade. Concurrent calibrations do not accrue linking error across grade levels, so that linking constants are similarly precise between all grade levels. However, the calibrations resulting from this approach measure the construct that is common across the linked assessments, which may be different from the intended achievement construct at each grade level, especially for subjects such as math where the assessed construct may change markedly across grade levels. Generally, both approaches tend to converge to produce vertical scales that operate similarly (Ito, Sykes, and Yao, 2008; Karkee, Lewis, Hoskens, Yao, and Haug, 2003), and we view convergence as evidence for the robustness of the vertical scale.

Final Linking Set

To facilitate the development of a vertical scale that will be sensitive to student growth over time, we first evaluated the performance of vertical linking items between the grade levels in which they were administered to identify any items that were more difficult for students in the intended grade than they were for students in the lower grade. For math, items that showed proportion correct scores lower in the intended grade than in the lower grade were dropped from the final vertical linking set. This resulted in dropping on average just over two items per linking set, with a maximum of six items dropped for the linkage between grade 6 and grade 7 math assessments.

For reading, the proportion correct values across grades were much closer, especially at the higher grade levels, so that elimination of all items where the proportion correct value in the lower grade exceeded the higher grade would result in dropping more items from the vertical linking set than would be desirable for executing a robust equating design. Thus, we modified the rule for reading to exclude from the vertical linking set those items which showed proportion correct values more than two standard errors beyond the average standard error for the total linking set (i.e., items that were reliably less difficult at the lower grade). This approach allowed us to identify a final set of linking items that would maximize detection of growth, while retaining sufficient items to establish a strong linkage between the grade-level assessments.

Exhibit 9.2.2.1 Number of Items Dropped and Remaining in the Final Vertical Linking Set

Linkage	Math Dropped Items	Math Final VL Set	ELA Dropped Items	ELA Final VL Set
G3 → G4	1	44	1	42
G4 → G5	0	45	3	46
G5 → G6	1	46	0	47
G6 → G7	6	41	5	39
G7 → G8	3	47	2	46
G8 M → Algebra I & G8 ELA → G9 ELA	3	28	11	30
G8 M → Geometry & G9 ELA → G10 ELA	2	31	7	39
Algebra I → Algebra II & G10 ELA → G11 ELA	2	32	10	35

CHAIN LINKING

The chain linking approach proceeds from the within grade item parameters identified in the initial calibrations of the operational and embedded field test items. Because operational test items at each grade were administered in the EFT slots in the grade below, each item in the vertical linking set has two sets of item parameters: on-grade (g) and below-grade (g - 1). The chain linking proceeds by identifying the linking constants necessary to place the below-grade item parameters onto the on-grade scale for the items in the final vertical linking set. The linking constant for each grade was defined as the mean difference of the item difficulty estimates for the linking items between the linked grades. The chain linking began by placing the grade 3 item parameters on the grade 4 scale for both math and ELA and proceeded upwards. For math EOC assessments, the grade 8 math scale was linked to both the Algebra I and Geometry scales, and the Algebra I scale was linked to the Algebra II scale.

CONCURRENT CALIBRATION

A vertical scale for each subject area was also established by calibrating simultaneously all items in the final vertical linking set. As with the within-grade calibrations, parameters were estimated using Winsteps. To compare results from the chain-linking and concurrent calibrations, the concurrent calibrations were placed on the grade 3 reference scale.

Exhibit 9.2.2.2 shows the vertical linking constants resulting from chain linking the within-grade scales as well as from concurrently calibrating items from across grade levels. The linking constants are applied to their respective within-grade scale to place all item parameters on the grade 3 reference scale.

Exhibit 9.2.2.2 Vertical Linking Constants Resulting from Chain Linking Within-Grade Scales and Concurrent Calibration of Items Across Grades

Linkage	Math Chain Linked	Math Concurrent	ELA Chain-Linked	ELA Concurrent
G3→G4	1.32	1.30	0.18	0.16
G4→G5	2.75	2.67	0.81	0.78
G5→G6	3.90	3.73	1.19	1.15
G6→G7	4.48	4.28	1.44	1.39
G7→G8	5.69	5.39	1.76	1.70
G8 M → Algebra I & G8 ELA → G9 ELA	6.07	5.76	1.97	1.88
G8 M → Geometry & G9 ELA → G10 ELA	7.15	6.86	2.12	1.98
Algebra I → Algebra II & G10 ELA → G11 ELA	7.81	7.45	2.32	2.16

To more directly examine the magnitude of gains across grade level assessments, Exhibit 9.2.2.3 shows the difference between linking constants between each of the grade levels assessed.

Exhibit 9.2.2.3 Linking Constant Differences Between Each of the Grade Level Scales

Linkage	Math Chain Linked	Math Concurrent	ELA Chain-Linked	ELA Concurrent
G3 → G4	1.32	1.30	0.18	0.16
G4 → G5	1.43	1.37	0.63	0.62
G5 → G6	1.15	1.06	0.38	0.37
G6 → G7	0.58	0.55	0.25	0.24
G7 → G8	1.21	1.11	0.32	0.31
G8 M → Algebra I & G8 ELA → G9 ELA	0.38	0.37	0.21	0.18
G8 M → Geometry & G9 ELA → G10 ELA	1.08	1.10	0.15	0.10
Algebra I → Algebra II & G10 ELA → G11 ELA	0.66	0.59	0.20	0.18

Relative gains are also represented graphically in Exhibit 9.2.2.4 and Exhibit 9.2.2.5 for math and ELA, respectively, which plot the linking constants across grade-level assessments. As the linking constants indicate, for math there is relatively large and steady growth across the grade level and EOC assessments. For the ELA assessments, the cross-grade gains are more modest, and tend to diminish in the higher grade-levels.

Exhibit 9.2.2.4 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: Math

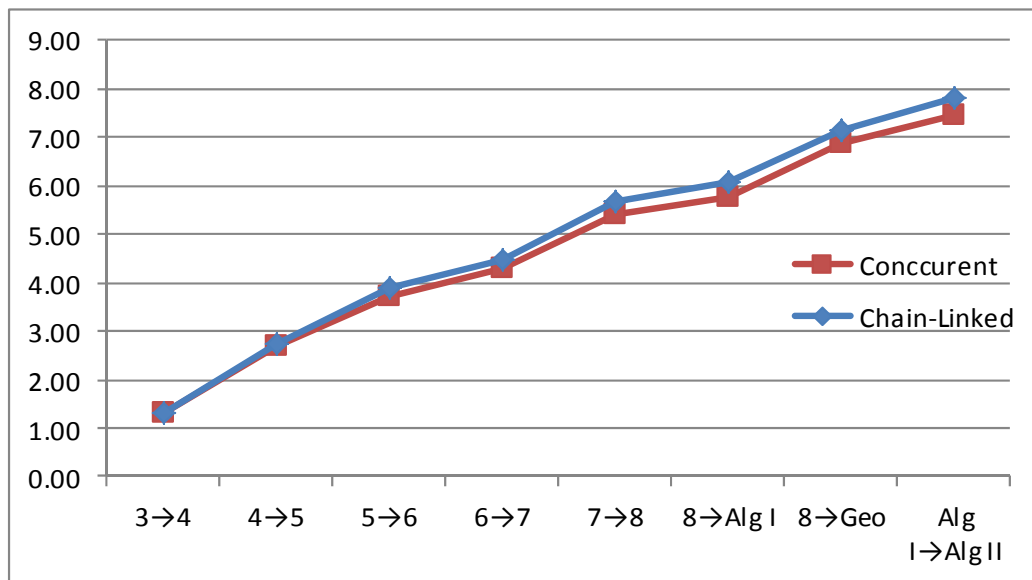
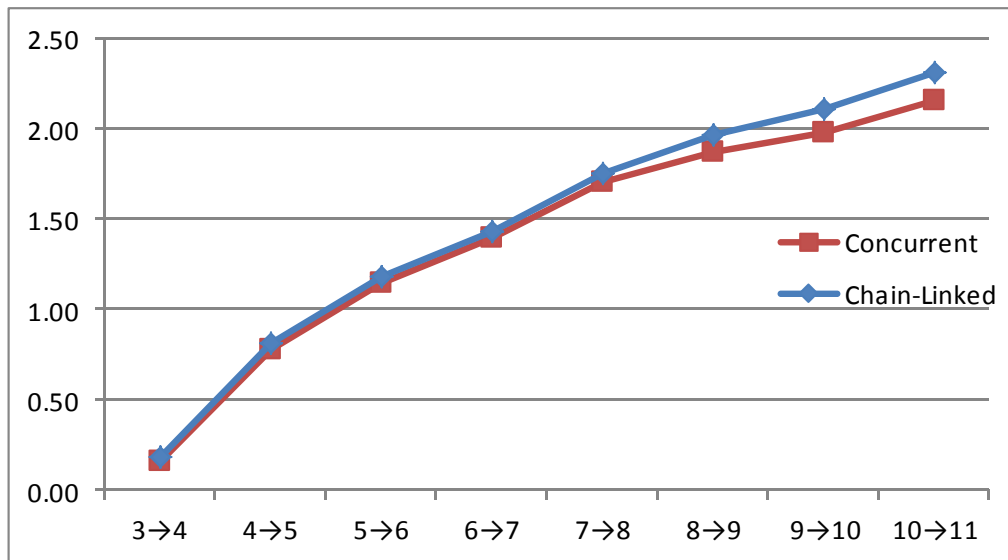


Exhibit 9.2.2.5 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: ELA



Linking constants resulting from the chain linking and concurrent calibration approach are quite consistent, indicating that both approaches converge on a common growth scale. Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain-linking method preserves the within-grade measurement construct, and was therefore selected as a preliminary vertical scale for the purpose of recommending performance standards. We note that ordered item books for the standard-setting workshop were based on the within-grade scales, so any modifications to the vertical scale would not impact the recommended performance standards.

The vertical linking constants also indicate much greater growth across grades and high school courses for math than is observed for ELA. In math, growth is on the order of about one standard deviation per year, with the exception of grade 6 to grade 7, which showed just over a half standard deviation gain. Similar one-half standard deviation gains were observed between grade 8 and Algebra I, which some students take concurrently, and between coursework in Algebra I and Algebra II. Gains in ELA are less pronounced, with somewhat larger gains in the elementary school years, with growth attenuating in the high school grades.

AZMERIT 2017 VERTICAL LINKING STUDY

It has been two years since the AzMERIT vertical scales for mathematics and English language arts (ELA) were first established in 2015. As a part of an on-going process in evaluating the stability of the vertical scales for AzMERIT, in spring 2017, the vertical linking study was repeated to evaluate results of the 2015 vertical linking study. It is noteworthy that the 2017 vertical linking study differs from the original 2015 study with respect to the linking design. In 2015, on-grade operational items were embedded in field test slots of the assessment in the grade below, whereas in 2017, on-grade field test items were embedded in field test slots of the assessment in the grade below.

Both chain linking and concurrent calibration approaches were used to produce the 2017 vertical linking constants. The robustness of the vertical linking results between the chain-linking and concurrent calibration methods was evaluated with respect to the convergence of the linking results across all grades per subject. Following the

method used in 2015 to evaluate the performance of vertical linking items between the grade levels, the items showing higher proportion correct in the lower grade than in the grade above were removed from the linking sets.

For mathematics, the linking constants produced by chain-linking and concurrent calibration didn't converge as expected. Further investigation was conducted on the behavior of the linking items. Unlike in ELA assessments, there were a large number of mathematics items for which the on-grade and linked off-grade item parameters differed substantially. The chain-linking and concurrent calibration yielded very close linking constants when these items were removed from the final linking set as well. However, this resulted in dropping on average 50% of items from a linking set. It is worth noting that the chain-linking result remained the same regardless of the number of items in the final linking set (using all available items; dropping the items with reversed proportion correct; or dropping the items showing large difference between the on-grade and linked off-grade item parameters). The vertical linking constants resulting from chain linking and concurrent calibration in math and ELA assessments are presented in 9.2.2.6 and Exhibit 9.2.2.7.

Exhibit 9.2.2.6. Vertical Linking Constants Resulting from Chain-Linking and Concurrent Calibration –Mathematics

Math	Approach 1: include all FT items used as VL items		Approach 2: remove reversed -P items		Approach 3: remove reversed-P items and 0.5-diff items	
	Chain-Linked	Concurrent	Chain-Linked	Concurrent	Chain-Linked	Concurrent
G3M	0	0	0	0	0	0
G4M	1.73	1.62	1.81	1.65	1.68	1.64
G5M	3.24	3.06	3.37	3.13	3.30	3.29
G6M	4.45	4.16	4.57	4.23	4.52	4.47
G7M	5.12	4.75	5.30	4.86	5.19	5.10
G8M	6.15	5.70	6.40	5.89	6.23	6.17
Alg I	7.04	6.26	7.32	6.48	7.17	7.04
Geometry	7.49	6.63	7.84	6.94	7.73	7.54
Alg II	8.03	6.83	8.34	7.08	8.05	7.85

Exhibit 9.2.2.7. Vertical Linking Constants Resulting from Chain-Linking and Concurrent Calibration – ELA

ELA	Approach 1: include all FT items used as VL items		Approach 2: remove reversed -P items	
	Chain-Linked	Concurrent	Chain-Linked	Concurrent
G3E	0	0	0	0
G4E	0.46	0.46	0.47	0.46
G5E	0.83	0.83	0.85	0.84
G6E	1.09	1.09	1.14	1.12
G7E	1.28	1.29	1.34	1.32
G8E	1.48	1.58	1.57	1.62
G9E	1.59	1.71	1.75	1.81
G10E	1.73	1.83	1.93	1.98
G11E	1.77	1.88	2.04	2.07

Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain-linking method preserves the within grade measurement construct. For this reason, the vertical linking constants identified via chain-linking were adopted as the AzMERIT vertical scaling constants in 2015. Comparison of the chain-linking results obtained in 2015 and 2017 is presented graphically in Exhibit 9.2.2.8 and Exhibit 9.2.2.9 for math and ELA, respectively. The vertical linking results are similar between 2015 and 2017 in terms of the overall growth patterns across grades. For each year, the vertical linking constants indicate much greater growth across grades and high school courses for mathematics than is observed for ELA. In mathematics for both years, growth is on the order of about one standard deviation per year, with the exception of grade 6 to grade 7 and grade 8 to Algebra I. Gains in ELA are less pronounced, with somewhat larger gains in the elementary school years, with growth attenuating in the high school grades for both years. Similarity between the 2015 and 2017 vertical linking results is also observed with respect to the difference between linking constants by grade. For math, although the vertical linking constants by grade in 2017 are uniformly higher than those in 2015, the difference between the 2015 and 2017 math linking constant for each grade is not larger than one standard deviation, with the exception of Algebra I, which is just over one logit at 1.10. For ELA, the vertical linking constants for grades 4 and 5 in 2017 are larger than those in 2015, while the vertical linking constants for the other grades in 2017 are smaller than those in 2015.

The relative gains from one grade to the subsequent grade are shown as the steepness of the line connecting two adjacent grades. The growth rate between adjacent grades is fairly constant between 2015 and 2017 for mathematics grade 3 to 8. Larger gain is observed in the linking between grade 8 and Algebra I in 2017 and between Algebra I to Geometry in 2015. The growth pattern is different in ELA. That is, the growth rates become similar in high school grades between two years, but quite different in the elementary and secondary schools. However, none of the difference between the 2015 and 2017 linking constant by grade is above one logit. Similar vertical linking results across years suggest that the vertical linking scale established in the first year of test administration holds for subsequent years, which supports the monitoring and evaluation of student growth over time.

Exhibit 9.2.2.8 Comparison of 2015 and 2017 Vertical Linking Constants Estimated from Chain-Linking Calibrations: Mathematics

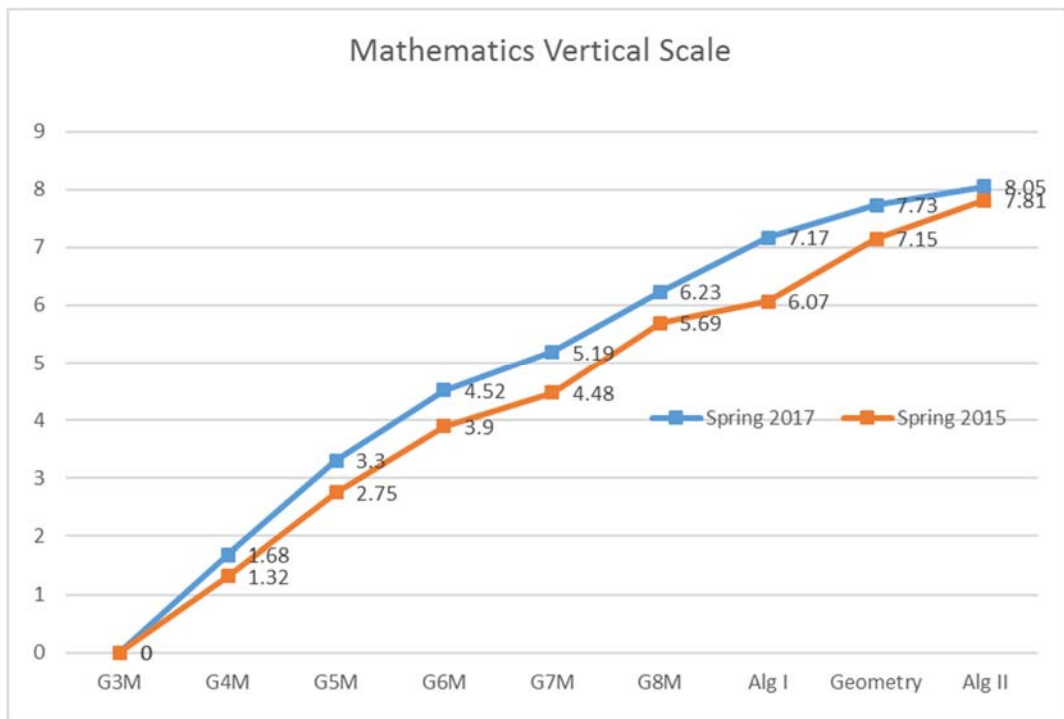
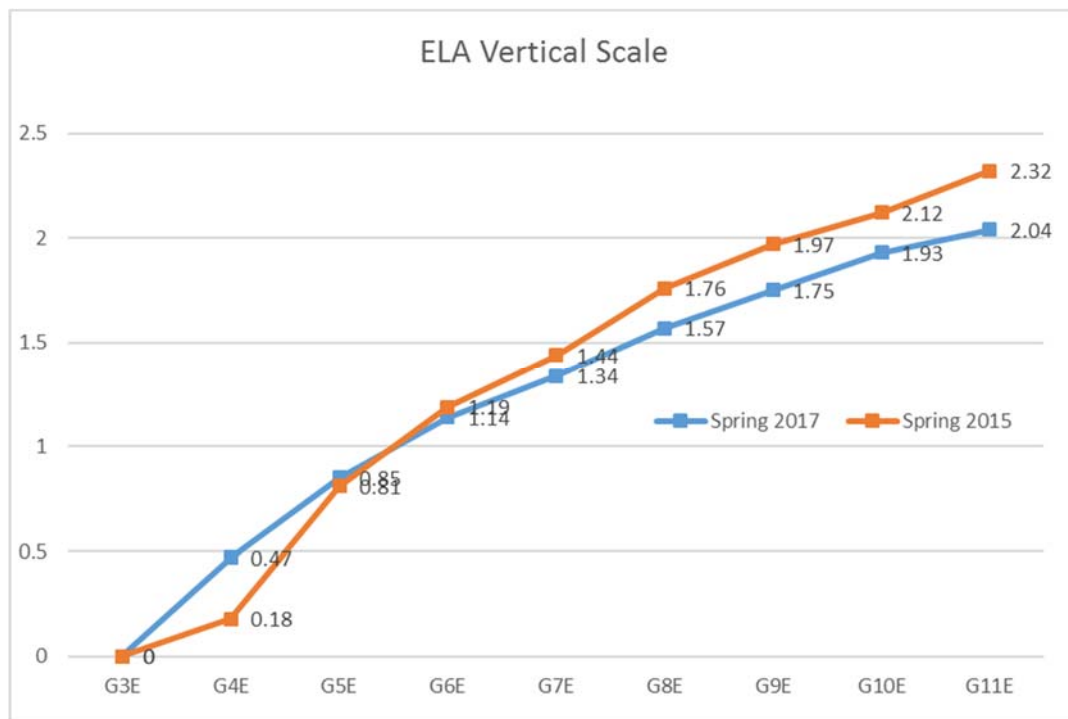


Exhibit 9.2.2.9 Comparison of 2015 and 2017 Vertical Linking Constants Estimated from Chain-Linking Calibrations: ELA



9.3 AZMERIT REPORTING SCALE (SCALE SCORES)

The AzMERIT assessments are reported on common scales within each subject (ELA and math). The IRT vertical scale scores (SS) are formed by linking each grade level assessment to the scale of the assessment in the grade level above. The vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate,⁵⁴

$$SS = a * \theta_v + d$$

where $a = 30$, $d = 2500$ for ELA tests, and $a = 30$, $d = 3500$ for Math tests. $\theta_v = \theta + c$, where θ is the on-grade ability estimate and c is a vertical linking constant listed below for each of the tests, as described in the previous section. For reporting, the on-grade ability estimate is truncated at ± 3.5 .

After transforming theta ability estimates to the vertical AzMERIT reporting scale, the observable scale scores nearest each of the performance standard cut scores are evaluated. If the observable scale score nearest the performance standard is below the cut score, the scale score is rounded up to be equal to the cut score. If the observable scale score nearest the performance standard is above the cut score, no special rounding rule is applied.

Overall scale scores for the AzMERIT are mapped into four performance levels per grade/course. The performance-level designations are: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. The performance level is evaluated using the rounded scale score.

Exhibit 9.3.1 shows the scale score ranges for the performance levels for each test.

⁵⁴ Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Exhibit 9.3.1 Scale Score Ranges for Performance Levels

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
ELA				
Grade 3	2395–2496	2497–2508	2509–2540	2541–2605
Grade 4	2400–2509	2510–2522	2523–2558	2559–2610
Grade 5	2419–2519	2520–2542	2543–2577	2578–2629
Grade 6	2431–2531	2532–2552	2553–2596	2597–2641
Grade 7	2438–2542	2543–2560	2561–2599	2600–2648
Grade 8	2448–2550	2551–2571	2572–2603	2604–2658
Grade 9	2454–2554	2555–2576	2577–2605	2606–2664
Grade 10	2458–2566	2567–2580	2581–2605	2606–2668
Grade 11	2465–2568	2569–2584	2585–2607	2608–2675
Math				
Grade 3	3395–3494	3495–3530	3531–3572	3573–3605
Grade 4	3435–3529	3530–3561	3562–3605	3606–3645
Grade 5	3478–3562	3563–3594	3595–3634	3635–3688
Grade 6	3512–3601	3602–3628	3629–3662	3663–3722
Grade 7	3529–3628	3629–3651	3652–3679	3680–3739
Grade 8	3566–3649	3650–3672	3673–3704	3705–3776
Algebra I	3577–3660	3661–3680	3681–3719	3720–3787
Geometry	3609–3672	3673–3696	3697–3742	3743–3819
Algebra II	3629–3689	3690–3710	3711–3750	3751–3839

9.4 LINKING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to reporting test scores for the spring 2015 and spring 2016 administrations of AzMERIT, AIR and ADE performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration.⁵⁵

9.4.1 MODE LINKING

A matched samples design (Way, Davis, and Fitzpatrick, 2006) was used to investigate mode comparability. A covariate regression approach was implemented to construct equivalent groups of students taking the AzMERIT assessments for both modes of test administration. For the spring 2015 mode investigation, the regression analysis identified for each student a predicted score on the paper AzMERIT assessment from previous year achievement on AIMS, covarying demographic variables that included gender, ethnicity, income level status, English language learner (EL) status, and Individualized Education Program (IEP) in the development of the prediction equation. A

⁵⁵ Standard 5.13 – When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

nearest neighbor search procedure was then applied to the predicted AzMERIT scores to select the equivalent groups of students. This procedure resulted in the identification of two matched samples for each assessment to conduct the mode comparability study.

IRT parameter estimates were then calibrated independently for the matched online and paper test administration mode samples. The linking constant necessary to bring the matched sample paper item parameters onto the matched sample online scale was then computed. Mean-mean linking was taken as the difference between the average item difficulty estimates from the matched-sample paper calibration and the average item difficulty estimates from the matched-sample online item parameter estimates.

Mode linking constants were estimated again following the spring 2016 administration of AzMERIT. Three approaches were used to identify matched samples for these analyses. In the first approach, 2014 AIMS paper test scores were used to predict student performance on the spring 2016 paper tests, with the resulting prediction model then used to identify a matched sample of online test takers. This approach allowed all available paper records to be included in the analysis, but required constructing matched samples based on achievement scores estimated two years prior. To utilize a more recent and comparable test score, a second approach was used. In this approach, we identified students who were administered AzMERIT on paper in 2015, but who participated online in spring 2016. We then identified a matched sample of students, based on AzMERIT test scores, who took the paper version of AzMERIT in both 2015 and 2016. For students at grade 3, there were no previous test scores with which to match student ability. We therefore used student performance on the multiple-choice items only on the spring 2016 AzMERIT math test to identify matched samples on the assumption that those items would be least susceptible to mode differences. To evaluate whether this approach yields results consistent with the other approaches, this approach was also applied to the grade 4 and grade 5 assessments.

Exhibit 9.4.1 presents the mode linking constants for the ELA assessments resulting from the matched sample analysis conducted on the spring 2015 administration of AzMERIT, as well as the linking constants resulting from each of the matched sample approaches used following the spring 2016 administration. In the grades 4–8 assessments, whether the matched samples are based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode-linking constants are generally small and equivalent across methods. For the high school end-of-course assessments, both approaches indicate that ELA assessments were somewhat more difficult online than on paper. The magnitude of those differences is greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. We note that the R^2 for the prediction equation used to identify matched samples for ELA based on 2014 AIMS remained quite high (R^2 around 0.65) even for the high school assessments, although matching based on spring 2015 AzMERIT achievement may nevertheless be more robust.

For grade 3 ELA, samples were matched based on student performance on the concurrently administered AzMERIT math multiple-choice items. To evaluate whether this approach yielded results consistent with the other two methods, we applied the same procedure in grades 4 and 5, where results indicated general convergence with the other methods, and indicating no effect for mode at grade 4 and a moderate mode effect at grade 5. When applied at grade 3, no mode effect was identified.

We note that any mode effect seems to interact with items, with some items easier when administered online, while others are more difficult. Thus, the mode effect is likely to be form specific and vary across test administrations. And this seems to be the case when mode linking constants are compared between the 2015 and 2016 administrations of AzMERIT. As shown in Exhibit 9.4.1, in spring 2015, mode effects were observed in grades 3, 4, and 8, but were more moderate at the other grades. In spring 2016, however, mode effects were absent or moderate in grades 3–8, but appear in the high school EOC tests.

Exhibit 9.4.1 Mode Linking Constants for AzMERIT ELA Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3E	2015	0.13	-0.01	0.13	3.90
	2016 – Math MC Match	0.17	0.16	0.01	0.30
G4E	2015	-0.09	-0.19	0.11	3.30
	2016 – 2014 AIMS Match	0.21	0.19	0.02	0.60
	2016 – 2015 AzMERIT Match	0.21	0.18	0.03	0.90
	2016 – Math MC Match	0.21	0.21	0.00	0.00
G5E	2015	0.04	-0.02	0.06	1.80
	2016 – 2014 AIMS Match	0.02	-0.02	0.04	1.20
	2016 – 2015 AzMERIT Match	0.03	-0.02	0.05	1.50
	2016 – Math MC Match	0.04	-0.04	0.08	2.40
G6E	2015	0.07	-0.02	0.09	2.70
	2016 – 2014 AIMS Match	0.18	0.21	-0.03	-0.90
	2016 – 2015 AzMERIT Match	0.20	0.16	0.04	1.20
G7E	2015	-0.08	-0.16	0.08	2.40
	2016 – 2014 AIMS Match	0.19	0.12	0.07	2.10
	2016 – 2015 AzMERIT Match	0.12	0.05	0.07	2.10
G8E	2015	-0.04	-0.22	0.18	5.40
	2016 – 2014 AIMS Match	0.01	-0.01	0.02	0.60
	2016 – 2015 AzMERIT Match	0.00	-0.05	0.05	1.50
G9E	2015	0.13	0.09	0.04	1.20
	2016 – 2014 AIMS Match	0.07	-0.12	0.20	6.00
	2016 – 2015 AzMERIT Match	0.08	-0.16	0.24	7.20
G10E	2015	-0.03	-0.10	0.07	2.10
	2016 – 2014 AIMS Match	0.10	-0.10	0.20	6.00
	2016 – 2015 AzMERIT Match	0.09	-0.04	0.13	3.90
G11E	2015	0.12	0.15	-0.03	-0.90
	2016 – 2014 AIMS Match	0.16	-0.09	0.25	7.50
	2016 – 2015 AzMERIT Match	0.14	-0.04	0.18	5.40

Exhibit 9.4.2 presents the mode linking constants computed for the spring 2015 and spring 2016 administrations of the AzMERIT math assessments. As observed for ELA, in the grade 4–8, and Algebra I math assessments, whether the spring 2016 matched samples were based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode linking constants are generally equivalent across methods. Effects of mode varied across grades, with the online form somewhat easier than paper at grade 4, somewhat more difficult at grade 7, and about the same at grades 5, 6, and 8. For the high school end-of-course assessments, both approaches indicate that math assessments were somewhat more difficult online than on paper. As with ELA, the magnitude of those differences was greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. In this case we note that the R^2 for the prediction equation used to identify matched samples for math based on 2014 AIMS remained quite a bit lower ($R^2 \approx .40$) for

the high school assessments compared to the lower grades ($R^2 \approx .65$), so that matching based on spring 2015 AzMERIT achievement are likely more robust.

Exhibit 9.4.2 Mode Linking Constants for AzMERIT Math Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3M	2015	-0.71	-0.77	0.06	1.80
	2016 – Math MC Match	-0.84	-0.57	-0.27	-8.10
G4M	2015	-0.40	-0.48	0.08	2.40
	2016 – 2014 AIMS Match	-0.43	-0.25	-0.17	-5.10
	2016 – 2015 AzMERIT Match	-0.57	-0.43	-0.14	-4.20
	2016 – Math MC Match	-0.41	-0.24	-0.17	-5.10
G5M	2015	-0.09	-0.09	-0.01	-0.30
	2016 – 2014 AIMS Match	-0.06	-0.02	-0.04	-1.20
	2016 – 2015 AzMERIT Match	-0.16	-0.12	-0.03	-0.90
	2016 – Math MC Match	-0.07	-0.06	0.00	0.00
G6M	2015	0.07	0.01	0.07	2.10
	2016 – 2014 AIMS Match	-0.01	0.04	-0.05	-1.50
	2016 – 2015 AzMERIT Match	-0.09	-0.06	-0.03	-0.90
G7M	2015	0.15	0.07	0.08	2.40
	2016 – 2014 AIMS Match	0.18	0.07	0.11	3.30
	2016 – 2015 AzMERIT Match	0.11	-0.03	0.14	4.20
G8M	2015	0.43	0.32	0.11	3.30
	2016 – 2014 AIMS Match	0.56	0.55	0.00	0.00
	2016 – 2015 AzMERIT Match	0.47	0.47	0.01	0.30
Alg I	2015	0.29	0.23	0.05	1.50
	2016 – 2014 AIMS Match	0.64	0.51	0.13	3.90
	2016 – 2015 AzMERIT Match	0.72	0.57	0.15	4.50
Geo	2015	1.12	0.99	0.13	3.90
	2016 – 2014 AIMS Match	1.34	1.15	0.20	6.00
	2016 – 2015 AzMERIT Match	1.19	1.03	0.16	4.80
Alg II	2015	1.45	1.36	0.09	2.70
	2016 – 2014 AIMS Match	1.45	1.17	0.28	8.40
	2016 – 2015 AzMERIT Match	1.06	0.91	0.15	4.50

For grade 3 math assessment, as with grade 3 ELA, samples were matched based on student performance on the math multiple-choice items. Again, this approach was applied in grades 4 and 5 to evaluate it against the other two methods, where the results indicated general convergence, indicating that items administered online were somewhat easier at grade 4 and no mode effect at grade 5. When applied at grade 3, a relatively large effect for mode was identified, indicating that items administered online were easier than on paper.

As with ELA, the identified mode effects varied across test administrations. The advantage of online over paper identified in 2016 was not observed in 2015. Likewise, observed effects of mode at grade 7 and for Algebra I and Algebra II in 2016 were not as pronounced in 2015, while effects of mode observed at grade 8 in 2015 were not observed in 2016. Thus, as with ELA, the effect of mode appears to be form specific and can be expected to vary across test administrations.

9.4.2 SCHOOL PERFORMANCE

In a separate approach to evaluating mode comparability, the ADE implemented an investigation based on the spring 2015 operational test administration statewide (Scott, 2015). In her study, Scott (2015) first identified which Arizona schools elected to administer AzMERIT online and which on paper, and then examined the two samples of schools for any differences in performance on the spring 2014 paper administration of AIMS. The rationale in selecting school-level analysis was based on schools having to choose only one of the two modes in which to assess all of their students. This increased level of matching was appropriate since the mode used by the student was, and continues to be, a school-based decision, rather than student based. Having found no difference in mean 2014 performance between the two groups, there would be no expectation for performance differences on AzMERIT except as a function of test administration mode. Following the spring 2015 administration of AzMERIT, ADE examined the performance of schools participating online and on paper, and again found performance on the AzMERIT to be comparable between the two sets of schools.

9.5 LINKING THE AZMERIT TO OTHER SCALES FOR PERFORMANCE COMPARISON

9.5.1 ESTABLISHING LINKAGES TO AIMS, SAGE, SMARTER BALANCED, PISA

To facilitate comparisons of Arizona achievement to other national and international benchmarks, a number of external linking sets were embedded in the 2015 AzMERIT field test slots. Arizona identified the locations of performance standards of other assessments systems on the AzMERIT scale; this information was used to inform panelists recommending performance standards for the AzMERIT.⁵⁶ The location of performance standards from the following assessments were identified on the AzMERIT scale:

- Smarter Balanced, by linking to AIR Core items on the Smarter Balanced scale
- PISA, by embedding PISA items in the grade 10 ELA, Algebra I, and Geometry EOC assessments
- Historical Arizona performance by embedding AIMS items to link to the AIMS scale
- Utah's SAGE via common items in the operational test form.

Subsequent to calibration of the AzMERIT operational items and establishment of the reference scale, parameter estimates for those items were anchored to their reference values and all items administered in the embedded field test (EFT) blocks were calibrated under that constraint, placing parameter estimates for all field test and external linking item sets on the same AzMERIT scale defined by the operational item parameters. All external linking items had two sets of item parameters: a) external scale, and b) AzMERIT scale. To identify the location of external scale performance standards on the AzMERIT scale, AIR identified the linking constants necessary to transform item parameters from the external reference scale to the AzMERIT scale. Where the external scale was

⁵⁶ Standard 5.23 – When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

calibrated using the Rasch model, such as with AIMS, mean-sigma equating was used to identify the location of external performance standards on the AzMERIT scale. For external scales calibrated using more general IRT models, Stocking-Lord equating was used to identify the location of external scale performance standards on the AzMERIT scale.

In the context of standard setting, this procedure enabled the ADE to identify a location in the AzMERIT ordered item book (OIB) that represented a level of difficulty similar to a particular level in the external scale. For example, after finding the linking constant necessary to put the Smarter Balanced item parameters on the AzMERIT scale, it was possible to provide standard-setting panelists with the location in the OIB that represents the level of difficulty comparable to each performance standard on the Smarter Balanced assessment.

9.5.2 IDENTIFYING THE LOCATION OF THE ACT COLLEGE-READY CUT ON AZMERIT

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the ACT college-ready cuts were identified on the AzMERIT scale and provided to panelists during performance standards workshops in 2015. In order to identify the location of the ACT college-ready cuts for the grade 11 ELA and Algebra II AzMERIT end-of-course assessments, a two-step approach was used to first identify the location of the ACT college-ready benchmark on the AIMS scale, and then use the linkages between AIMS and AzMERIT to map the ACT college-ready benchmark onto the AzMERIT scale(s). To examine directly the relationships between the AzMERIT and ACT assessments, the ADE obtained the ACT test scores for Arizona students graduating high school in spring 2016. The direct linking study using the AzMERIT and ACT data is summarized in this section.

Although AzMERIT is offered as a series of end-of-course tests in high school, most students take the Algebra II assessment at grade 11, so the focus of this investigation will be on the grade 11 ELA and Algebra II AzMERIT assessments administered in spring 2015. From among the full set of spring 2015 grade 11 ELA and Algebra II test takers, there are 58,888 (93%) and 32,945 (56%) grade 11 students, respectively. These records represent the target sample for the analyses reported in this study.

Because a large number of students did not take the ACT and the two subgroups differed systematically across demographic and achievement variables, the imputing approach is often employed to handle missing data in the analysis of the relationship between the AzMERIT scores and subsequent performance on the ACT. However, previous studies for Minnesota and Ohio showed that imputing or deleting the missing records did not impact the linkage identified between their graduation tests and the ACT test. For this study, we instead divided the complete sample of merged records into model building and cross-validation samples of equal size. The cross-validation sample allows for better estimation model fit. Because the model is built using a sample independent from that used to evaluate model fit, estimates of model fit exclude sample dependent idiosyncrasies that would be reflected as model overfit in the model development sample.

ELA: Examinees with missing ACT or AzMERIT scale scores were removed from the merged dataset. The ACT reading scale score for the remaining 25,977 students were regressed onto the applicable grade 11 ELA scale score and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2 , was identified as the best model to predict ACT reading from prior performance on the AzMERIT ELA test:

$$\hat{Y} = -290.65 + 0.12 * X_1 + 0.26 * X_2 - 2.35 * X_3 - 0.79 * X_4 + 0.57 * X_5 - 2.32 * X_6 - 1.79 * X_7 - 2.40 * X_8 - 1.82 * X_9 - 2.07 * X_{10}$$

where

\hat{Y} = ACT Reading Scale Score
 X1 = AzMERIT ELA Scale Score
 X2 = Female–Male Contrast
 X3 = American Indian–White Contrast
 X4 = Multi-ethnic Contrast
 X5 = Asian Contrast
 X6 = Hispanic–White Contrast
 X7 = African American–White Contrast
 X8 = Native Hawaiian–White Contrast
 X9 = Free and Reduced Lunch Contrast
 X10 = EL Contrast

The overall model was statistically significant ($F(10, 20388) = 1704.70$, $p < .0001$; adjusted $R^2 = 0.46$). Application of this regression model indicates that an AzMERIT ELA scale score 2585 is associated with the ACT reading college-ready cut score of 22.

Math: The records with missing ACT or AzMERIT scale scores were excluded from the analysis. Then the ACT mathematics scale scores for the remaining 13,777 students were regressed onto the applicable AzMERIT Algebra II test and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2 , was identified as the best model to predict ACT mathematics scores from prior performance on the AzMERIT Algebra II test:

$$\hat{Y} = -305.7 + 0.08 * X1 - 0.55 * X2 - 1.55 * X3 - 0.48 * X4 - 0.44 * X5 - 1.44 * X6 - 1.41 * X7 - 0.83 * X8 - 1.22 * X9 - 1.57 * X10$$

where

\hat{Y} = ACT Mathematics Scale Score
 X1 = AzMERIT Math Scale Score
 X2 = Female–Male Contrast
 X3 = American Indian–White Contrast
 X4 = Multi-ethnic Contrast
 X5 = Asian Contrast
 X6 = Hispanic–White Contrast
 X7 = African American–White Contrast
 X8 = Native Hawaiian–White Contrast
 X9 = Free and Reduced Lunch Contrast
 X10 = ELL Contrast

The overall model was statistically significant ($F(10, 13768) = 1764.13$, $p < .0001$; adjusted $R^2 = 0.51$). Application of this regression model indicates that an AzMERIT mathematics score of 3727 is associated with the ACT mathematics college-ready cut score of 22.

The validation set approach is a type of resampling method that estimates a model error rate by holding out a subset of the data from the fitting process (the testing dataset). The model is then built using the other set of observations (the training dataset). Then the model result is applied on the testing dataset in which we can then calculate the error. In summary, this general idea allows for the model to not overfit. In this study, the training dataset contained 50% randomly-selected merged records and the testing dataset had the other 50% of students.

The multiple regression built by the training set yielded the same AzMERIT cut scores (ELA 2585, Math 3727) as the ones from the full data model. Then the predictive model was applied onto the testing set. The Root Mean Squared Error (RMSE) was calculated as the square root of the average squared errors found between the actual ACT score point and the model fitted values. Furthermore, we repeated this sampling and model fitting process 100 times to see how the RMSE varied across random samples. For ELA, the average RMSE was 5.03 and the standard deviation of the RMSE was 0.02 across the 100 replications. For mathematics, the average RMSE was 2.79 and the standard deviation was 0.02. The standard deviation of the RMSE was very small indicating that the sample selected for the modeling has no significant impact on the model fitting.

In addition, the equipercentile equating method was used to verify the linking between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT cut score 22 is 2585.72 for ELA and 3727.46 for mathematics. These cut scores are consistent with those identified using regression models.

10. CONSTRUCTED-RESPONSE SCORING

The AzMERIT assessments in ELA and math utilize a variety of item types to assess students' mastery of the Arizona College and Career Ready Standards (ACCRS). ADE leverages AIR's item scoring technology to machine-score student responses to most items, including traditional selected-response (multiple-choice) item types and machine-scored constructed-response (MSCR) items types. The MSCR item types are designed to capture and score a variety of response types, such as graphing, drawing, or arranging graphic regions, selecting or rearranging sentences or phrases within passages, or entering equations or words, allowing AzMERIT items to assess a wide range of student knowledge and skills. In most cases, constructed-response machine-scored items that are developed for online administration are adapted for paper and responses are captured in a format that allows machine scoring.

In addition, some constructed-response items are scored by human raters; these items are referred to as "handscored." To support machine scoring of each essay response, in 2016, a sample of essay responses was handscored through verification, and those responses and scores were used to develop the statistical scoring models used to score the remaining responses. The statistical scoring models developed in spring 2016 will be used to score all essay responses in future test administrations. In addition, math assessments that were administered on paper included a small number of items that were scored by human raters. Generally, these were items that required students to produce an equation. The reading components of the ELA assessments, both online and paper, and the math assessments administered online are machine scored in their entirety.

AIR partners with Measurement, Inc. (MI), to fulfill all handscored requirements. AIR provides the automated electronic scoring and MI provides all handscored for the AzMERIT tests. This section describes the process for configuring and validating machine rubrics and the process for handscored, including rules, descriptions of scorer training and systems used, and mechanisms for ensuring the reliability and validity of item scores.

10.1 MACHINE SCORING

10.1.1 EXPLICIT RUBRICS

As part of the item development process for machine-scored item types which are scored with explicit rubrics, a rubric validation process was enacted to verify that rubrics are implemented as intended, and responses are scored correctly. This procedure is typically conducted following the initial administration of items, usually when the item is field-tested, and allows test developers to review the intent of the rubric versus the actual behavior. Actual student responses were reviewed by test development experts, along with resulting item scores, to ensure that the rubrics functioned as intended and awarded credit appropriately. Where necessary, test developers modified machine rubrics to address insufficiencies, automatically rescored student responses for the item, and repeating the process to finalize and approve the machine-scored rubrics. Test developers reviewed a strategic sample of responses, including responses where high achieving students scored poorly on the item, lower achieving students scored well on the item. They also reviewed randomly-selected responses from the population.

10.1.2 ESSAY AUTOSCORING

As part of the spring 2017 administration of AzMERIT, students in each grade were administered one of two writing tasks (one informational/explanatory, and the other, either opinion [grades 3–5] or argumentative [grades 6–11]) that had been calibrated during the spring 2016 administration. This section describes the processes

performed to calibrate these, and the rest of the available writing prompts completed during the spring 2016 administration. As part of the spring 2016 administration of AzMERIT, students in each grade were administered one of six writing tasks (one informational/explanatory, and the other, either opinion [grades 3–5] or argumentative [grades 6–11]) in the writing component of each of the ELA online assessments.

Two approaches were used to develop the statistical models that were used to score the essay responses. For AIRCore writing tasks that were administered online in the Florida field test (grades 8–10), ADE adopted the scoring models generated from student responses in the Florida field test administration. Because the scoring models are based on semantic and syntactic features of the text that discriminate high- versus low-scoring essays as determined by human raters, the models are highly generalizable.

For the grades where scoring models did not already exist (grades 3–7 and 11), an alternative approach was employed that allowed for autoscoreing to be implemented as part of the spring 2016 essay scoring. Because the ELA window is split into separate writing and reading assessment windows, with the online writing window closing several weeks prior to close of the reading test administration, the dual window afforded an opportunity to build and implement the statistical scoring models in time to meet spring reporting timelines.

To facilitate development of the scoring models, MI conducted rangefinding, where possible, based on student responses from the Florida assessment. The rangefinding process is designed to calibrate a sample of responses for scorer training, qualification, and monitoring. Responses exemplifying each score point are identified and annotated for scorer training. Additional responses are identified for use in qualifying readers for scoring and for establishing validity sets that are used to monitor reader performance. Thus, for grades 4–7 which were included in the Florida field test, rangefinding activities to support AzMERIT rubric scoring were completed prior to the opening of the AzMERIT assessment window.

For the grade 3 and 11 assessments, which had not been previously administered, MI pulled a sample of essay responses following the first week of the testing window with which to conduct rangefinding activities. The development of training materials and training of raters followed immediately so that handscoring could begin by the end of the fourth week of the testing window.

At the end of the second week of testing, AIR drew a random sample of 2,000 responses to each of the writing tasks administered at grades 3–7 and 11 for use in building the statistical scoring models. Those responses were routed to MI for handscoring. Each response was double scored, with any discrepancies routed for resolution scoring.

As handscoring activities were completed for each writing task, and scores were uploaded to AIR, work began to develop statistical scoring models for each rubric element, and to deploy those models to the test delivery system to score all remaining essay responses.⁵⁷

To develop the scoring models, the random sample of 2,000 responses was divided into a model building sample of 1,500 responses and a cross-validation sample of 500 responses. Model performance was evaluated on the cross-

⁵⁷ Standard 4.19 – When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

validation sample to ensure that model fit indices were not based on the model building sample, which may inflate fit indicators.

The statistical scoring models also yield an indicator of score confidence based on (1) responses with unusual features, and (2) responses scoring near rubric thresholds. For each model, a confidence threshold defined as two standard deviations below the mean confidence value for the responses in the cross-validation sample was identified. Any scored response with a confidence value below the threshold was automatically routed to MI for verification scoring.

The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and “learned” from a training set. During training, these features are related to human scores through a statistical model. The resulting estimates complete a prediction equation that predicts how a human would score a response with the measured features. Statistical rubrics are, effectively, proxy measures. Although they can directly measure some aspects of writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not make direct measures of argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Further, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may *predict* whether a response to a factual item is correct—higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the “reason” that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

As noted, the engine employs a “training set,” a set of essay responses scored with maximally valid scores, which we obtain by having all responses double-scored by expert scorers and a thorough adjudication process for adjacent or discrepant scores. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Approximately 1,500 essay responses were selected at random from the set of scored essay responses to serve as the training set.

For each dimension in the rubric, the system estimates an appropriate statistical model relating the measures to the score assigned by humans. This model, along with its final parameter estimates, is used to generate a predicted or “proxy” score.

In addition to the training set, we draw an independent random sample of responses for cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are handscored, and agreement between human- and machine-assigned scores is examined. The cross-validation process ensures that the rubric generalizes across all responses and that the statistical model identified during training does not capitalize on peculiarities in the training set.

Exhibit 10.1.2.1 presents agreement indicators for the two initial human raters, and between the resolved human and statistical rubric score, for the two writing prompts randomly assigned in each grade in the spring 2017 administration.⁵⁸ Please see the 2016 AzMERIT Technical Report, available at www.azed.gov, for the values for the complete list of prompts. Indicators include percentage exact agreement, Pearson’s correlation, a quadratic

⁵⁸ Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

weighted kappa statistic, and the standardized mean difference between the scores. Although absolute values for evaluating statistics have been advanced (Condon, 2013; Wei & Higgins, 2013), the focus of these comparisons is degradation of agreement when moving from human–human agreement to machine–human agreement. Agreement between human raters is an indicator of how reliably the responses can be scored by human raters. Since the statistical rubrics attempt to reproduce human–assigned scores, evaluation of machine–human agreement is with respect to observed human–human agreement. Responses with poor human–human agreement will not be reliably scored by either humans or machines. For the training and validation sets of the prompts administered in spring 2017, Exhibit 10.1.2.2 presents the correlations among the dimension scores.

Exhibit 10.1.2.1 Summary of Human and Machine Scores for Spring 2017 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human–Human Agreement			Human–Machine Agreement				
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted k*	SMD *	% Exact	Pearson r	Weighted k*	SMD *
3	13023	Conventions	2	2089	1.56	1.61	0.66	0.64	0.70	0.65	0.65	0.01	0.77	0.67	0.67	0.08
		Evidence	4		1.99	1.99	0.74	0.65	0.62	0.60	0.03	0.66	0.63	0.63	0.01	
		Organization	4		2.07	2.03	0.76	0.67	0.64	0.66	0.66	0.02	0.65	0.66	0.65	0.07
3	13026	Conventions	2	2088	1.48	1.59	0.67	0.65	0.71	0.66	0.66	0.00	0.75	0.68	0.67	0.17
		Evidence	4		2.01	2.03	0.78	0.66	0.66	0.68	0.04	0.66	0.66	0.65	0.02	
		Organization	4		2.04	2.04	0.80	0.68	0.64	0.68	0.68	0.05	0.65	0.67	0.66	0.01
4	13094	Conventions	2	2095	1.01	1.03	0.73	0.73	0.66	0.67	0.67	0.00	0.66	0.67	0.67	0.03
		Evidence	4		1.32	1.27	0.50	0.45	0.77	0.52	0.52	0.00	0.83	0.62	0.62	0.12
		Organization	4		1.44	1.41	0.54	0.51	0.74	0.56	0.56	0.01	0.82	0.68	0.68	0.04
4	13095	Conventions	2	2096	1.12	1.24	0.71	0.65	0.64	0.62	0.62	0.01	0.64	0.62	0.61	0.17
		Evidence	4		1.31	1.26	0.52	0.46	0.75	0.57	0.57	0.00	0.82	0.63	0.62	0.11
		Organization	4		1.51	1.58	0.60	0.56	0.71	0.59	0.59	0.03	0.74	0.61	0.60	0.12
5	13236	Conventions	2	2099	1.39	1.54	0.62	0.62	0.74	0.69	0.69	0.02	0.71	0.64	0.62	0.24
		Evidence	4		1.81	1.81	0.55	0.53	0.71	0.59	0.59	0.01	0.76	0.58	0.58	0.01
		Organization	4		1.85	1.89	0.64	0.57	0.70	0.65	0.65	0.03	0.70	0.58	0.58	0.08
5	13239	Conventions	2	2095	1.47	1.52	0.65	0.60	0.73	0.66	0.66	0.02	0.72	0.62	0.62	0.07
		Evidence	4		1.70	1.71	0.62	0.57	0.65	0.56	0.56	0.02	0.75	0.65	0.65	0.00
		Organization	4		1.88	1.95	0.62	0.55	0.71	0.65	0.65	0.02	0.73	0.60	0.59	0.12
6	13304	Conventions	2	2097	1.40	1.48	0.67	0.62	0.67	0.57	0.57	0.02	0.74	0.69	0.68	0.11
		Evidence	4		1.63	1.69	0.65	0.61	0.63	0.57	0.56	0.04	0.72	0.65	0.64	0.10
		Organization	4		1.76	1.83	0.72	0.67	0.62	0.61	0.61	0.01	0.72	0.70	0.69	0.10
6	13308	Conventions	2	2097	1.38	1.58	0.65	0.62	0.62	0.54	0.54	0.06	0.67	0.59	0.56	0.32
		Evidence	4		1.39	1.39	0.60	0.59	0.69	0.57	0.57	0.03	0.76	0.63	0.63	0.00
		Organization	4		1.62	1.60	0.68	0.64	0.63	0.60	0.60	0.03	0.73	0.67	0.66	0.02
7	13402	Conventions	2	2088	1.51	1.57	0.59	0.59	0.69	0.60	0.60	0.03	0.80	0.71	0.71	0.10
		Evidence	4		1.95	1.95	0.58	0.52	0.73	0.59	0.59	0.04	0.78	0.64	0.64	0.01
		Organization	4		1.85	1.90	0.53	0.46	0.70	0.61	0.61	0.01	0.80	0.59	0.58	0.09
7	13403	Conventions	2	2083	1.61	1.63	0.54	0.53	0.77	0.61	0.61	0.03	0.80	0.65	0.65	0.03
		Evidence	4		1.56	1.48	0.61	0.58	0.73	0.66	0.66	0.00	0.80	0.72	0.71	0.14
		Organization	4		1.82	1.71	0.61	0.58	0.68	0.61	0.61	0.02	0.76	0.66	0.65	0.19

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human–Human Agreement			Human–Machine Agreement				
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted k*	SMD *	% Exact	Pearson r	Weighted k*	SMD *
8	13437	Conventions	2	2391	1.43	1.51	0.70	0.64	0.75	0.69	0.69	0.03	0.74	0.70	0.69	0.12
		Evidence	4		1.97	1.93	0.71	0.65	0.76	0.75	0.75	0.01	0.73	0.71	0.71	0.05
		Organization	4		2.07	2.02	0.75	0.68	0.73	0.75	0.75	0.00	0.74	0.74	0.74	0.07
8	13452	Conventions	2	2631	1.58	1.70	0.59	0.54	0.79	0.67	0.67	0.02	0.80	0.70	0.68	0.20
		Evidence	4		2.11	2.05	0.74	0.63	0.77	0.77	0.77	0.01	0.72	0.72	0.71	0.09
		Organization	4		2.23	2.17	0.76	0.67	0.74	0.75	0.75	0.01	0.73	0.74	0.74	0.09
9	13557	Conventions	2	2818	1.51	1.63	0.65	0.59	0.79	0.72	0.72	0.00	0.76	0.69	0.67	0.19
		Evidence	4		1.88	1.88	0.61	0.51	0.83	0.78	0.78	0.00	0.79	0.68	0.67	0.01
		Organization	4		2.00	1.98	0.68	0.61	0.79	0.78	0.78	0.01	0.78	0.74	0.74	0.03
9	13566	Conventions	2	2852	1.53	1.62	0.64	0.59	0.82	0.76	0.76	0.04	0.78	0.71	0.70	0.14
		Evidence	4		1.96	1.99	0.63	0.58	0.84	0.80	0.80	0.00	0.81	0.75	0.74	0.05
		Organization	4		2.08	2.11	0.69	0.66	0.80	0.79	0.79	0.01	0.79	0.77	0.77	0.04
10	13639	Conventions	2	2306	1.58	1.74	0.58	0.48	0.70	0.56	0.56	0.02	0.75	0.55	0.51	0.29
		Evidence	4		1.94	1.78	0.72	0.53	0.72	0.69	0.69	0.04	0.69	0.66	0.61	0.25
		Organization	4		2.08	1.95	0.75	0.60	0.70	0.70	0.70	0.03	0.70	0.69	0.66	0.18
10	13640	Conventions	2	2399	1.67	1.73	0.53	0.47	0.73	0.60	0.60	0.03	0.81	0.61	0.60	0.12
		Evidence	4		2.15	2.06	0.76	0.62	0.69	0.72	0.72	0.02	0.71	0.70	0.68	0.12
		Organization	4		2.31	2.32	0.75	0.71	0.67	0.70	0.70	0.03	0.70	0.72	0.72	0.00
11	13722	Conventions	2	2091	1.57	1.63	0.60	0.56	0.78	0.68	0.68	0.01	0.81	0.71	0.70	0.09
		Evidence	4		2.14	2.19	0.82	0.73	0.62	0.67	0.67	0.02	0.67	0.72	0.71	0.06
		Organization	4		2.38	2.40	0.77	0.70	0.66	0.70	0.70	0.04	0.73	0.74	0.74	0.02
11	13724	Conventions	2	2090	1.53	1.56	0.62	0.59	0.73	0.61	0.61	0.03	0.79	0.71	0.70	0.06
		Evidence	4		2.14	2.16	0.76	0.74	0.63	0.68	0.68	0.00	0.76	0.78	0.78	0.03
		Organization	4		2.24	2.28	0.73	0.66	0.64	0.66	0.66	0.01	0.73	0.73	0.72	0.05

Note: Weighted K = Quadratic weighted kappa; SMD = Standardized Mean Difference

*For asterisked items, no 4-point responses were identified in the training set, so at present, statistical models for these items can only assign up to 3 points.

Exhibit 10.1.2.2. Summary of Dimension Intercorrelations for Spring 2017 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
3	13023	Conventions	2	2090	0.67	0.86
		Evidence	4			
		Organization	4			
3	13026	Conventions	2	2090	0.66	0.88
		Evidence	4			
		Organization	4			
4	13094	Conventions	2	2095	0.63	0.72
		Evidence	4			
		Organization	4			
4	13095	Conventions	2	2096	0.55	0.64
		Evidence	4			
		Organization	4			
5	13236	Conventions	2	2099	0.52	0.81
		Evidence	4			
		Organization	4			
5	13239	Conventions	2	2095	0.63	0.77
		Evidence	4			
		Organization	4			
6	13304	Conventions	2	2097	0.72	0.90
		Evidence	4			
		Organization	4			
6	13308	Conventions	2	2097	0.42	0.76
		Evidence	4			
		Organization	4			
7	13402	Conventions	2	2088	0.64	0.87
		Evidence	4			
		Organization	4			
7	13403	Conventions	2	2085	0.57	0.63
		Evidence	4			
		Organization	4			
8	13437	Conventions	2	2391	0.55	0.85
		Evidence	4			
		Organization	4			
8	13452	Conventions	2	2491	0.56	0.86
		Evidence	4			
		Organization	4			
9	13557	Conventions	2	2815	0.39	0.76
		Evidence	4			
		Organization	4			
9	13566	Conventions	2	2852	0.58	0.78
		Evidence	4			
		Organization	4			
10	13639	Conventions	2	2306	0.49	0.80
		Evidence	4			
		Organization	4			

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
10	13640	Conventions	2	2399	0.43	0.82
		Evidence	4			
		Organization	4			
11	13722	Conventions	2	2090	0.63	0.80
		Evidence	4			
		Organization	4			
11	13724	Conventions	2	2089	0.70	0.88
		Evidence	4			
		Organization	4			

10.2 HANDSCORING

Handscoring of online essay responses for statistical model building, as well as handscoring of all essay responses from paper Test Administrations, were routed to MI for scoring. As noted in Section 10.1, the sample of essay responses selected for statistical model building was independently scored by two readers. Any response assigned discrepant scores were routed for resolution scoring by a scoring trainer. In addition, all essay responses captured from paper test administrations were handscored, with 10 percent of all paper responses receiving a second reading (Reader 2) for the purpose of monitoring and maintaining sufficient inter-rater reliability, as discussed below. For ELA handscoring, where scores from Reader 1 and Reader 2 were not in adjacent agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score was based on the resolution score, when present, or else on the initial read. For math handscoring, where scores from Reader 1 and Reader 2 were not in exact agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score for math was based on the resolution score, when present, or else on the initial read.

In spring 2017, all the essays were autoscored, and the essay responses with the low confidence index were routed to MI for human verification. The final essay score was the human verification score when present.

10.2.1 HANDSCORING PROCESS

MI's handscoring efforts are managed via the Virtual Scoring Center VSC™ software, which is composed of two primary subsystems: VSC Capture™ and VSC Score™. Images of student responses to open ended items were sent to VSC Score™, which is a web-based environment for scoring constructed-response items by scorers working in an online environment. VSC Score is a secure, centrally-administered environment used by site-based scorers. The interface enabled scorers to evaluate constructed-response items and writing assessments from images. VSC Score has the following capabilities:

- Defining scorer roles and qualifications based on training, security requirements, or prior history
- Managing and randomly routing scorers' responses that require second readings in a double-blind manner
- Allowing project leaders to spot-check scorers, monitor reliability, and offer feedback
- Allowing scorers to flag responses for a variety of reasons (unusual approaches, nonscorable issues, etc.)

- Generating status reports at project milestones (such as percentage of items scored)
- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
- Accommodating paper-based scores when images are of insufficient quality
- Outputting data easily into MI's score reporting applications

Paper-pencil tests were scanned into VSC. The images were displayed to trained and qualified scorers who scored the images online. Scorers had access only to those items for which they had been qualified to score. Online assessment responses were also converted into images and displayed in an identical manner to paper-pencil student responses using the same VSC scoring application.

When logging onto VSC Score, scorers were presented with a scoring set, which is the images-scoring equivalent of a physical packet of student responses. The scoring set was generated by randomly selecting student responses from the pool of non-scored student responses. The resultant set of responses was checked out to the scorer. The images they received had no demographic information on them. The scorer did not know the name, sex, school, or location of the student whose item was being scored. The scorer evaluated the first response, entered the score by clicking the appropriate values on the scoring toolbar, and clicked the Submit button. For multi-page responses, a scorer had to view each page of the response before a score was entered. Once the Submit button was clicked, the system recorded the score and the next response in the scoring set appeared for the scorer to score and submit. This process continued until all responses in the set had been scored.

When a scorer had a question about a response, he or she transferred the image (along with a virtual note including the question and/or comments) from the current scoring set to a review set assigned to a team leader or the scoring director. The team leader or scoring director submitted the appropriate score or returned the response to the scorer with comments. This procedure was used whenever a scorer had scoring concerns or found nonscorable responses (NSR) or responses requiring condition codes. Previously, condition codes were assigned to student responses by scoring leadership per Arizona specifications, such as a code noting that the response was left blank, the response was undecipherable or illegible, the response was made in non-English, and so on. Condition codes other than blank were then recoded to the lowest score for each dimension for ability estimation. Because the statistical scoring engine cannot assign condition codes, all non-blank responses were assigned a rubric score directly, with responses that would otherwise have received a non-blank condition code being assigned the lowest score point for each dimension.

After scoring all of the responses in a set, the scorer reviewed all of the responses and modified the scores before committing them to the system. Once the scores had been committed, the set was checked in and responses were routed to other scorers as necessary. Regardless of the specific requirements, however, student responses were not marked as complete until the requisite number of independent scorers had scored the response.

VSC prioritized the available responses in the queue to make sure that the newer responses were placed toward the back of the queue.

10.2.2 HANDSCORING QUALITY CONTROL

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students. The requirements for double scorings are defined to VSC at setup time. MI assumed a double-blind scoring rate of 10% for both the essays and math constructed responses.

10.2.3 HANDSCORING RELIABILITY AND VALIDITY

MI uses a two-pronged approach to construct the scoring teams for AzMERIT. First, the scoring leadership recruits qualified, experienced scorers who have successfully scored large-scale assessments for MI, and therefore have experience understanding the approach to scoring. To ensure reliable and valid handscores, MI puts scoring directors, team leaders, and scorers through a rigorous screening and training process.⁵⁹

Scoring directors, team leaders, and scorers are hired for AzMERIT based on experience and performance. Potential new scorers are given a comprehensive content screening for reading and math. This screening is used to identify potential scorers' aptitude for content area and grade level, as well as their reading comprehension and deductive reasoning skills, which are directly related to what they may be scoring. In addition to writing an extemporaneous essay, new hires are required to read a passage and answer questions pertaining to that passage, proofread a sample essay for writing conventions, and solve a series of math problems. The results determine grade and content area placement if a scorer is to be offered a position on a project. New scorers are selected based on their scores on MI's content screening assessment given for language arts and math projects, the quality of their interview, their work history, and the references provided. The actual qualification for the scorers occurs at the end of training. In addition, the scorers are provided with ongoing validation that they are providing the state with consistency in their scoring using validation sets that are incorporated into the ongoing live scoring.

All of the Arizona training materials provided for the initial operational ELA scoring were scoring guides composed of anchor responses as well as training, qualifying, and recalibration sets approved for use by the state as a result of approval of existing documentation from AIR's Item Tracking System (ITS), which is the repository for all item attributes, including scoring rubrics. New items, approved from the previous year's field test, will be incorporated based on the materials used during the field test scoring. All materials and selected sets were submitted to Arizona for approval.⁶⁰

MI's scoring directors ensured that ELA scoring guides had detailed annotations to explain how the scoring criteria are to be applied to each response's specific features and why the response should be assigned a particular score. The approach was to focus on the precise scoring rationale, which helped scorers define the lines between score points. All scoring guides and other training materials were presented to Arizona for review and approval prior to the start of scoring.

Training sets and qualifying sets consisted of items that are most representative of the type that will be scored. MI scoring leadership selected these responses and provided them to Arizona for approval prior to their use. The training and qualifying sets contained examples of responses from all score points arranged in random score-point order. MI created an appropriate number of training sets and qualifying sets based on the complexity of the item. Essay questions were more complex than single-point math items. The sets were designed to help the scorers learn to apply the criteria illustrated in the scoring guide, ensure that the scorers become familiar with the process

⁵⁹ Standard 4.20 – The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

⁶⁰ Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

of scoring student responses, and assess the scorers' understanding of the scoring criteria before they are allowed to begin live scoring. MI worked with Arizona to finalize the number of training and qualifying sets for each item and determine the appropriate qualifying percentage. All scoring decisions and supplemental responses were submitted more than one month before the start of scoring for review and approval by the state.

MI's scoring directors trained both new and experienced scorers within the scoring rooms, giving detailed explanations of all training materials.

MI's online training interface allowed observers from ADE to witness training in real time. Through the use of TurboMeeting software, observers were able to visually see the responses being trained and discussed as each training set progressed. Observers were also allowed to hear the training through the software's audio function. In addition to observing the training of leadership virtually, representatives from Arizona also traveled to individual scoring sites to observe training in-person. This allowed Arizona to observe MI's training techniques and interact with project leadership. The State was able to provide additional guidance on scoring rationale during the training process. These observations allowed MI to further ensure reliability in the handscoring efforts.

Recruited staff followed established training methodologies to ensure the reliability and validity of scores. Scorers were trained as a group, not individually, and all scorers (whether experienced or not) were required to train on all the scoring sets and, at the end of training, pass the qualifying sets with acceptable scores to prove that they were able to understand and apply the criteria. Unless a scorer was trained and qualified for a project successfully, he or she was not permitted to score any student responses.

Each member of MI's scoring staff was required to qualify for the scoring of student responses based on standards established by Arizona. Each staff member was also expected to maintain a consistent level of scoring quality throughout the scoring effort or he or she was released from the project. MI continually monitored performance in order to guarantee scoring accuracy.

For math, MI trained scorers to handscore a limited number of math items from the paper assessment that could not be machine-scored. Scoring leadership reviewed all handscored math items prior to training. Using the scoring rubrics provided from ITS, leadership provided feedback and questions to both AIR and Arizona to ensure consistency in training methodology. Math items were trained and scored individually with the use of the provided scoring rubrics. Qualified math scorers received training that included all possible answers to each individual item.

Math handscoring was monitored in the same way as essay scoring, with consistent read-behind and validation sets incorporated into the daily scoring schedule to ensure that scorers were providing accurate scoring on a consistent basis.

10.2.4 MACHINE-SCORING VERIFICATION

In addition to the regular ELA handscoring activities, MI also provided a percentage of second readings on items that were machine-scored. These read-behind scores were used to help ensure consistency and reliability with the ELA machine-scoring. Responses requiring read-behind were generated and sent to MI, where the most experienced scorers, team leaders, and scoring directors provided a second read verification. This process utilized blind scoring, with the scorer unaware of the first score provided by machine. Where scores from Reader 1 (machine) and Reader 2 (human) were in exact agreement or adjacent, the final item score was based on the initial machine read. Where scores from Reader 1 (machine) and Reader 2 (human) were not in exact agreement or adjacent, the final item score was based on the second human read.

11. QUALITY ASSURANCE PROCEDURES

Quality assurance procedures are enforced throughout all stages of AzMERIT test development, administration, and scoring and reporting of results. This section describes quality assurance procedures associated with the following:

- Test construction
- Test production
- Answer document processing
- Data preparation
- Equating and scaling
- Scoring and reporting

Because quality assurance procedures pervade all aspects of test development, we note that discussion of quality assurance procedures is not limited to this section but is also included in sections describing all phases of test development and implementation.

11.1 QUALITY ASSURANCE IN TEST CONSTRUCTION

Section 5.5 details the form construction process. Each form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test. The statistical targets ensure that students will receive scores of similar precision, regardless of which form of the test they receive.

The form construction process is managed through AIR's FormBuilder software, which automates important form construction activities to ensure development of equated test forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement curves (SEMCs) as test developers build a test map. This helps our content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

Immediately upon generation of a test form, the FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allows another opportunity to ensure that poorly performing items are not included in operational test forms.

When submitting test forms for review by ADE, AIR produces a form evaluation workbook that includes an evaluation summary checklist, as well as summary statistics and test characteristic graphs.

All bookmaps (test maps), key files, and conversion tables were produced directly from FormBuilder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents are generated directly from information maintained in ITS. The information stored in

ITS is rigorously reviewed by multiple skilled reviewers to protect against errors. Automated production of these critical files (such as key files) virtually eliminates opportunities for errors.

11.2 QUALITY ASSURANCE IN PAPER-DELIVERED TEST PRODUCTION

Camera-ready documents are prepared after the test items have been selected, composed in forms, and reviewed per the ADE's specifications.

Paper tests go through a traditional production process. The test booklet production process starts with the creation of test maps (also referred to as bookmaps). The test map is built in the Item Tracking System (ITS) and initiates the production of printed test forms. The process includes the following five steps:

1. The 1×1s (test items printed one per page) are generated based on the test map.
2. Blackline 1 is drafted and reviewed internally.
3. Blackline 1 is delivered to the Department for review and approval.
4. Should any changes be requested in the blackline 1 review, blackline 2 forms are produced, reviewed, and delivered to the ADE.
5. The documents are taken to blueline (camera-ready copy).

Step 1 is entirely automated within ITS. ITS houses destination templates that define the format of the 1×1s and automatically generates these documents based on the test map. At this stage, items are proofread by internal editorial and test development staff and the ADE. Additionally, they are reviewed to verify that all edits from previous rounds of review have been correctly implemented. Any changes required at this stage are entered directly into ITS to ensure consistency across all item uses.

Blackline 1 is a semi-automated process. With the appropriate destination template defined and 1×1 approval, ITS generates a Quark-readable document in the specified format. Through this integration, items are automatically styled with fonts, graphics, spacing, and other formatting specifications outlined in the ADE's style guide. Our production staff may adjust page layout, including instructions, borders, and other elements, to meet the ADE's guidelines. At this stage, reviewers check the document layout and formatting. Should any egregious errors be found in the content of an item, changes must be entered into ITS and the item must be re-exported to ensure consistent item use across all test forms. Changes to blackline 1 require a second blackline proof. Changes to subsequent blackline proofs require sign-off by senior management and the ADE.

The final quality assurance step prior to printing is the blueline, or camera-ready copy, review stage. During this step, AIR and the ADE's staff review proofs from the print vendor, verifying that the file to be printed matches the previously approved blackline proof. At AIR, in addition to reviews by test development and forms production staff, two members of the technical team—who have not seen the items previously— independently take the tests. This process forces a close look at the items and gives a final opportunity to verify the keys.

During the production and review process, test book blacklines are accompanied by answer document blacklines, which are produced by MI. Answer documents reflect the demographic fields required by the ADE, as well as fields for pre-code labels and the scannable marks required for accurate data collection. The item sequence is based on test maps and corresponds directly with test books.

All blacklines in AIR's production queue are controlled by an electronic version-control server system that ensures that only the current version is immediately available to our production staff, preventing version-

control errors. Like AIR's ITS, which controls and tracks all changes to items, this production system maintains historical records (including all older versions), which senior production staff can access if necessary. Each blackline after blackline 1 and the blueline (camera-ready copy) is automatically compared with the immediately preceding version using a PDF comparison tool that highlights all changes. This step has proved useful for identifying unintended changes made during the revision process. Such changes are difficult to detect because they can appear anywhere in a document and may be subtle. The PDF comparison tool highlights these changes so differences between versions can be mapped to an intended revision. All materials delivered will go through this process, ensuring that the ADE will receive error-free materials for review and that any changes requested by the ADE are implemented promptly and accurately.

At each of the review stages, proofs will be accompanied by proof tickets that identify the document being reviewed, its review stage, the scheduled and actual delivery dates, and the return date. Sign-off by the ADE is required at each stage before proceeding with subsequent steps.

11.3 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

The production of computer-delivered assessments involves two distinct types of products, each of which follows an appropriate quality assurance process:

1. Content for online delivery shares some processes with paper versions, but also requires additional, unique steps.
2. Online test delivery software must deliver the content reliably (and, with the right tools, the accommodations, layouts, etc.).

11.3.1 PRODUCTION OF CONTENT

While the online workflow requires some additional steps, it actually removes a substantial amount of work from the time-critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. A paper form must await final forms construction before blackline proofs can show how the item will look in the booklet. Online, the appearance of the item screen can be known with certainty before the final test form is ever constructed. This characteristic of online forms enables us to lock down the final presentation of each item well before forms are constructed. In turn, this moves the final blueline review of items much earlier in the process, removing it from the critical path.

The production of computer-based tests includes five key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. Forms are finalized using the process described in Section 4.6, and final forms are approved in our FormBuilder software.
3. Complete test packages are created with our test packager, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
4. Forms are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the field.

5. The final system is deployed to a staging environment accessible to ADE for user acceptance testing and final review.

11.3.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The Item Tracking System (ITS) integrates directly with the test delivery system (TDS) display module, and displays each item exactly as it will appear to the student. This process is called web preview, and web preview is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent web preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will see it. Layout changes can be made after this process in two ways:

1. Content can be revised and re-approved for web display.
2. Online style sheets can change to revise the layout of all items on the test.

Both of these processes are subject to strict change control protocols to ensure that accidental changes are not introduced. Below, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the benefit of allowing final layout review much earlier in the process, reducing the work that must be done during the very busy period just before tests go live.

11.3.3 APPROVAL OF FINAL FORMS

Section 5.6 describes our process for constructing operational test forms, including the approval of test forms by ADE. The forms are built in FormBuilder (a component of our ITS), and upon approval, they are ready for preliminary publication.

11.3.4 PACKAGING

The test packaging system performs two simultaneous roles in the preparation of computer-based products: It compiles the form definitions and other information about how the test is to be administered (e.g., where any embedded field-test items might be inserted) and pulls together the content packaged during web approval.

The test packager assigns form identifiers to each form, evaluates the form against the blueprint, and performs a quality check against the content. The content quality check includes checks to see that every asset (e.g., graphics) referenced in the item is included in the package, confirms that the item has not changed since it was web approved, and ensures that the items have received all the approvals necessary for publication.

11.3.5 PLATFORM REVIEW

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

11.3.6 USER ACCEPTANCE TESTING AND FINAL REVIEW

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the test delivery system serves both a software evaluation and content approval role. The UAT period provides ADE with an opportunity to interact with the exact test with which the students will interact.

11.3.7 FUNCTIONALITY AND CONFIGURATION

The items, both in themselves and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of our system:

1. Testing on the system itself to ensure function, performance, and capacity
2. Capacity planning
3. Continuous monitoring

AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the testing window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

11.4 QUALITY ASSURANCE IN DOCUMENT PROCESSING

11.4.1 SCANNING ACCURACY

When test documents were returned to be scored, they must be scanned first. When they were scanned, a quality control sample of documents consisted of 10 test cases per document type (normally between 500 and 600 documents) were created so that all possible responses and all demographic grids were verified, including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured method of scan testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file

created from them to further ensure that results from the scanner, editing process (validation and data correction), data transfer to the project database, and scoring were all accurate according to the reporting rules provided by ADE.

11.4.2 QUALITY ASSURANCE IN EDITING AND DATA INPUT

At a minimum, MI implemented, maintained, and constantly updated the following quality assurance controls:

- Score key verification
- Post analysis of item keys
- Response analyses to determine score frequency distribution by item verification of bank values of item statistics
- Live data checks to verify that data/results conform to approved specifications comprehensive software test plan
- Double data entry correction process to verify student response and demographic information report data verification
- Reviewed and proofread all electronic and printed report deliverables

MI utilized a double data correction process to achieve the highest level of quality and accuracy in both Arizona CBT and PBT assessment student data. Data correction operators used their sophisticated Data Inspection, Correction and Entry (DICE) application, which retrieved flagged data records and highlighted the problem field on a computer screen for resolution. The operator compared the highlighted data on the answer document template, retrieved the original document for resolution, and made any necessary correction.

After an operator corrected a flagged record, the same flagged record was routed to a second data correction operator who repeated the data correction process. After a flagged record was edited by two independent operators, the data correction application checked to verify that both operators made identical corrections. If the two corrections differed, the record was routed to a supervisor for a third and final resolution. Agreement rate statistics were generated for the individual data correction operators, allowing the supervisor to monitor their job performance. This process continued until all flagged records were examined and resolved.

Thorough training significantly improves the accuracy of data correction. To ensure that goal, MI trained their data correction staff on the use of the data correction application and on the specific validation errors and procedures associated with the specific project. Practice sets generated by the programming staff allowed data correction staff to learn on samples of answer documents that simulated the kinds of errors they were expected to correct for the actual assessment prior to actually processing live data. Additionally, each user had an electronic copy of the data correction user's guide for reference.

MI developed verification routines as part of their standard data validation to detect duplicate student tests in the assessment, whether in a single LEA (local educational agency) or across LEAs, and student moves between schools. MI staff then worked closely with the ADE to resolve these discrepancies through processes called Barcode Processing and Tested Roster. These processes and the business rules governing them are described in a set of requirements developed in conjunction with the ADE.

11.5 QUALITY ASSURANCE IN DATA PREPARATION

AIR's test delivery system has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance-level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. For example, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the sorts of checks that are done for data review, but (a) they are done on operational data, and (b) they are conducted in real time so that our psychometricians can catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM to the Database of Record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to ADE and their quality assurance contractor. AIR psychometricians ensure that data in the extract files match the DoR prior to delivery to the ADE.

11.6 QUALITY ASSURANCE IN TEST FORM EQUATING

Item information necessary for statistical and psychometric analyses is provided to the ADE and HumRRO, ADE's independent quality assurance contractor, prior to test administration. Item information is published as part of the configuration of the online assessment system that AIR employs for administering, scoring, and reporting test scores. Information contained in these workbooks includes, but is not limited to, a unique item ID used for item tracking, test form ID, location on the test form, correct answer, item difficulty, and information about the strand, standard, and benchmark each item measures. These item files are used in quality control checks of the assessment data scoring and analysis.

To ensure security, all data is shared using ADE's SFTP site.

Prior to operational work, AIR produces simulated datasets for the purpose of testing software and analysis procedures, and shares with the ADE and the QA contractor. All parties complete a dry run of calibration and post-equating activities and compare results. The practice runs serve two functions:

1. To verify accuracy of program code and procedures
2. To evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.

Following the completion of these activities and the resolution of questions that arise, analysis specifications are finalized.

11.7 QUALITY ASSURANCE IN SCORING AND REPORTING

11.7.1 QUALITY ASSURANCE IN HANDSCORING

DOUBLE SCORING RATES, AGREEMENT RATES, VALIDITY SETS, AND ONGOING READ-BEHINDS

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students.

MI's Virtual Scoring Center software (VSC), described in Section 10.2.1, provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read-behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer's performance every day to ensure that he or she is on target and conduct one-on-one retraining sessions when necessary. MI's quality assurance procedures allow scoring staff to identify struggling scorers very early and begin retraining immediately.

We monitor their scoring intensively to ensure that all responses are scored accurately. If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, he or she is given interactive feedback and mentoring on the responses that have been scored incorrectly and is expected to change the scores. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

If a scorer's interrater agreement rate falls below the expected standard, the scorer will be re-trained. Should the scorer still be unable to score reliably, the scorer is assigned to another, non-Arizona-related project or dismissed.

In addition to using validity responses (also known as calibration or anchor responses) as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be pulled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following Arizona's review and approval. MI periodically administers validity sets to each of MI's scorers working on the scoring effort. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the State.

With the VSC program, the way in which the student responses are presented prevents scorers from having any knowledge about which responses are being single- or double-read, or which responses are validity set responses. A performance threshold of 75% is set to specify validity agreement standards as well as the frequency and total number of validity responses evaluated by each scorer based on client specifications.

HANDSCORING QA MONITORING REPORTS

MI generates detailed scorer status reports for each scoring project utilizing a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Arizona. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated

real-time reports that show both daily and cumulative (project-to-date) data are available. These reports are available to Arizona 24 hours a day via a secure website. Project leadership reviews these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

Scorers are released when they are unable to demonstrate the ability to score responses according to the criteria and standards established by MI and Arizona and perform to the level of client expectation. Should Arizona request that certain responses be rescored, we are prepared to do so, if necessary. The reporting system can produce a list of all the responses a selected scorer has scored. In these situations, all responses scored by a scorer during the time frame in question can be identified, reset, and released back into the scoring pool. The aberrant scorer's scores are deleted, and the responses are redistributed to other qualified scorers for rescoring.

MONITORING BY THE ARIZONA DEPARTMENT OF EDUCATION

ADE also directly observes MI activities, both onsite and virtually. MI provides virtual access to the training activities through the online training interface, as well as onsite training and onsite scoring. Arizona monitors the scoring process through the Client Command Center (CCC) with access to view and run specific reports during the scoring process. This ability to attend the training, qualification, and initial scoring virtually provides Arizona the most efficient use of oversight by reducing the travel requirements for onsite attendance for the ADE's staff.

IDENTIFYING, EVALUATING, AND INFORMING THE STATE ON ALERT PAPERS

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the examinee or those around him. We also flag potential security breaches identified during scoring. For possible dangerous situations, scoring project management and staff employ a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

This process is also used to notify Arizona of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow up. The ADE has processes in place to communicate the presence of and information contained within the alert paper to student's school official.

11.7.2 TEST SCORING

AIR verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the State. The ability of each of these simulated students is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they provide a check of the full range of item responses and test scores in fixed-form tests, as well. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the Online Reporting System (ORS), we merge item response data with the demographic information taken from previous year assessment data. If current year enrollment data is available by the time

simulated data files are created, we verify online reporting using current year testing information. By populating the simulated data files with real school information, it is possible to verify that specific school types and special districts are being handled properly in the reporting system.

Specifications for generating simulated data files are included in the Analysis Specifications document submitted to and approved by the ADE each year. Although the ADE does not currently provide immediate reporting, review of all simulated data is scheduled to be completed prior to the opening of the test administration, so that the integrity of item administration, data capture, item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the testing window, a series of quality assurance reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational testing window.

An additional set of forensic analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school level that may indicate cheating. The quality assurance reports can be generated on any desired schedule. Item analysis reports are evaluated frequently at the opening of the testing window to ensure that items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit 11.7.2.1 presents an overview of the quality assurance (QA) reports.

Exhibit 11.7.2.1 Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Analysis Report	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Forensic Analysis	To monitor testing irregularities	Early detection of testing irregularities

ITEM ANALYSIS REPORT

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT-based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For dichotomous items, the proportion of students selecting each response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is

also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for dichotomous items and polyserial correlations for polytomous items are computed. AIR psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between observed and expected score given the student's ability is computed for each item. The residuals for each are averaged across all students, and the average residual is used to flag an item. The item fit statistic is computed as follows:

Let X_{ij} be the variable for the response of student j to item i , and $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ be the probability that student j gets a score of x_{ij} to item i given his or her ability estimate $\hat{\theta}_j$. $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ is calculated using Rasch model

$$P(X_{ij} = x_{ij}|\hat{\theta}_j) = \frac{\exp(\hat{\theta}_j - b_i)}{1 + \exp(\hat{\theta}_j - b_i)},$$

where b_i is the difficulty parameter of item i . If item i is a polytomously scored item, $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ is calculated using the Master's Partial Credit model,

$$P(X_{ij} = x_{ij}|\hat{\theta}_j) = \frac{\exp \sum_{k=0}^{x_{ij}} (\hat{\theta}_j - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^l (\hat{\theta}_j - b_{ki})}$$

The expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as

$$E(X_{ij}|\hat{\theta}_j) = \sum_{x_{ij}=0}^{m_i} x_{ij} P(X_{ij} = x_{ij}|\hat{\theta}_j).$$

For item i , the residual between observed and expected score for student j is defined as

$$\delta_{ij} = x_{ij} - E(X_{ij}|\hat{\theta}_j).$$

The statistic δ_{ij} is aggregated across all n students for item i ,

$$\bar{\delta}_i = \frac{1}{n} \sum_{i=1}^n (\delta_{ij}).$$

The report can be configured to report all items or flag and report only those items where the fit index is above a given threshold (e.g., items could be flagged when

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > .96$$

where $se(\bar{\delta}_j) = \frac{SD(\delta_{ij})}{\sqrt{n}}$.

FORENSIC ANALYSIS

Another component in the suite of QA reports is geared toward detecting testing irregularities that may indicate possible cheating. The forensic analysis components of the QA reports are described in detail in Section 6.6. Evidence evaluated includes changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and were determined in partnership with ADE. Analyses are performed at student level and summarized for each aggregate unit, including testing session, test administrator, and school.

11.7.3 REPORTING

Scores for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Once both online and handscoring items have passed through their validity and quality checks, the handscored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the Quality Monitor (QM) system, where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies Arizona-specific scoring rules and assigns scores from the calibrated items, including calculating performance-level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively prior to deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

After passing through the series of validation checks in the QM system, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the “official” record is stored. Only after scores have passed the QM checks and are uploaded to the DoR are they passed to the Online Reporting System (ORS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QM system’s validation checks and ADE’s independent data verification checks.

12. REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*.
- AzMERIT Testing Conditions, Tools and Accommodations Guidance Manual. Arizona Department of Education (2017, February). Retrieved from: <https://cms.azed.gov/home/GetDocumentFile?id=5836103eaadebe14087eb770>
- Bentler, P.M. (1990), "Comparative Fit Indexes in Structural Models," *Psychological Bulletin*, 107(2), 238–46.
- Camilli, G., & Shepard, L. (1994). Methods for identifying biased test items. California: Sage
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. doi:10.1080/10705510701301834
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. doi:10.1080/10705510701301834
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Concon, W. (2013). Large-scale assessment, locally-developed measured, and automated scoring of essays: Fishing for the red herrings? *Assessing Writing*, 18(1), 100–108.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices, *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Estrada S., Burnham C., Feld J. K., Bergan J. R., & Bergan J. R. (2015). Can Local Assessment Data be Successfully Used as Part of an Arizona A-F Accountability System? Leawood, KS: Assessment Technology Incorporated (ATI). Retrieved from: <https://azsbe.az.gov/sites/default/files/media/ATI-Feasibility.pdf>
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253–264.
- Ito, K., Sykes, R., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling, *Applied Measurement in Education*, 21, 187–206.
- Karkee, T., Lewis, D. M., Hoskins, M., Yao, L., & Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, 5(1), 95–110.
- Livingston, S.A., & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores, *Journal of Educational Measurement*, 32(2), 179–197.
- Livingston, S. A. & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 16, 247–260.

- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52, 443–452.
- Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.) *Handbook of Structural Equation Modeling* (pp. 380–392). New York: Guilford Press.
- Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations, *International Journal of Testing*, 1(2), 115–135.
- Scott, L. (2015). Analysis of Mode Comparability of AzMERIT's Online and Paper Administrations for Spring 2015. In Arizona Department of Education, *Recommending AzMERIT Performance Standards* (pp. I-28–I-40), Retrieved from http://www.azed.gov/assessment/files/2014/11/spring-2015-azmerit-standard-setting_091415-full-report.pdf.
- Sireci, S. G. & Rios, J. A. (2013). Decisions that make difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 19(2–3), 170–187, DOI: 10.1080/13803611.2013.767621.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter, *Psychometrika*, 66, 331–342. doi:10.1007/BF02294437.
- Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing, *The Phillipine Statistician*, 52(1–4), 81–92.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test, *Journal of Educational Measurement*, 11, 265–276.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments: Synthesis Report (No. 44). Minneapolis, MN: National Center on Educational Outcomes.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. In annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wei, Y., & Higgins, J. P. (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. [Research Support, Non-U.S. Gov't]. *Stat Med*, 32(7), 1191–1205.
- Wesolowsky G.O. (2000). Detecting Excessive Similarity in Answers on Multiple Choice Exams, *Journal of Applied Statistics*, 27, 909–921.



Calculator Guidance

The AzMERIT calculator guidelines are designed to provide appropriate support for students while still measuring a student’s mastery of the standards. On tests where calculators are permitted, it is ideal for a student to use the recommended acceptable calculator. If the recommended calculator is not available, students may use a calculator with less functionality. The Desmos Scientific and Graphing calculators have been customized for AzMERIT and are embedded in online tests that allow the use of a calculator.

These guidelines are for the assessment only. They are not intended to limit instruction in the classroom. Technology is a part of the Arizona Mathematics Standards, and students should still be interacting with technology as appropriate for engaging with and learning the standards.

Grades 3-6: No calculators permitted on AzMERIT.

Grades 7-8: Scientific calculator permitted on AzMERIT Math Part 1 only.

No calculators permitted on AzMERIT Math Part 2.

Scientific calculator should include these functions: standard four functions (addition, subtraction, multiplication, and division), decimal, change sign (+/-), parentheses, square root, and π .

They may NOT include: any problem solving or programming capabilities, place values, and inequalities. *Sample acceptable calculator: TI-30X IIS or similar.*

High School End-of-Course Tests:

Graphing calculators permitted on AzMERIT Math Part 1 and Part 2.

No calculators with Computer Algebra System (CAS) features are allowed. Calculators may NOT be capable of communication with other calculators through infrared sensors. NO instruction or formula cards, or other information regarding the operation of calculators such as operating manuals are permitted. The memory of any calculator with programming capability must be cleared, reset, or disabled when students enter the testing room. Many calculators have a testing mode that will allow these features to be disabled and will meet the requirements of AzMERIT. Check the calculator documentation for instructions on enabling this mode. If the memory of any calculator is password protected, and cannot be cleared or reset, the calculator may NOT be used. Items for the EOC tests are written with these types of calculators in mind; however students may use a scientific calculator if they choose to do so. *Sample acceptable calculators: TI-84 Plus, Casio FX-9750GII, or similar.*

Additional Guidance:

- Students are not allowed to share calculators during a testing session.
- The AzMERIT online calculators available for the computer-based assessment are available for practice use on the Calculator and Tutorials site at <http://azmeritportal.org/tutorials/>.
- For EOC tests only, an online version of the scientific and graphing calculator will be available in the [Secure Browser](#) for students taking the paper-based version of the test. Students will not need to sign in to select the online calculator.
- No laptop, tablet, or phone-based calculators are allowed to be used during the AzMERIT assessment unless they are used to access the AzMERIT Secure Browser.
- The applicable portion of the computer-based assessment will include the acceptable online version of approved calculator. Providing handheld calculators is not a requirement for schools choosing the computer-based assessment. However, students may use an acceptable handheld calculator in addition to or instead of the online calculator.



AzMERIT

Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

English Language Arts Assessment Blueprint

Grade 3		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 4		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 5		
Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 6		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 7		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 8		
Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 9		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 10		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 11		
Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Listening Standards will only be assessed on the computer-based assessment.

In Grades 3-5 some items in the Reading and Language Strands will also be aligned to the standards for Reading: Foundational Skills.

Percentage of Points by Depth of Knowledge Level

Grade	DOK Level 1	DOK Level 2	DOK Level 3	DOK Level 4
3-11	10%-20%	50%-60%	15%-25%	16%-19% (Writing)

For more information go to www.azed.gov/AzMERIT

Grade 3		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	49%	53%
Number and Operations-Fractions	18%	22%
Measurement, Data, and Geometry	26%	30%

Grade 6		
Domain	Min.	Max.
Ratio and Proportional Relationships	19%	23%
The Number System	25%	29%
Expressions and Equations	29%	33%
Geometry, Statistics and Probability	17%	21%

Algebra I		
Conceptual Categories	Min.	Max.
Algebra	40%	44%
Functions	36%	40%
Statistics	17%	21%

Grade 4		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	46%	54%
Number and Operations-Fractions	29%	33%
Measurement, Data, and Geometry	15%	19%

Grade 7		
Domain	Min.	Max.
Ratio and Proportional Relationships	19%	23%
The Number System	19%	23%
Expressions and Equations	23%	27%
Geometry, Statistics and Probability	27%	35%

Geometry		
Domain	Min.	Max.
Congruence	23%	27%
Similarity, Right Triangles and Trigonometry	27%	31%
Circles, Geometric Measurement and Geometric Properties with Equations	23%	27%
Modeling with Geometry	17%	21%

Grade 5		
Domain	Min.	Max.
Operations, Algebraic Thinking, and Numbers in Base Ten	38%	42%
Number and Operations-Fractions	31%	35%
Measurement, Data, and Geometry	24%	28%

Grade 8		
Domain	Min.	Max.
Expressions and Equations	32%	36%
Functions	21%	25%
Geometry	23%	27%
Statistics and Probability and The Number System	15%	19%

Algebra II		
Conceptual Categories	Min.	Max.
Algebra	34%	38%
Functions	32%	36%
Statistics	27%	31%

Percentage of Points by Depth of Knowledge Level			
Grade	DOK Level 1	DOK Level 2	DOK Level 3
3-11	10%-20%	60%-70%	12%-30%

Within a test, approximately 70% of the assessment will be on major content within that grade or course.

Revised by ADE on 8/19/15

For more information go to www.azed.gov/AzMERIT

Appendix C.1a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	13733.772	1804				
Metric	13956.001	1847	Configural	222.230 (43)	< .01	.000
Scalar	14638.886	1890	Metric	682.885 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7117.742	1804				
Metric	7271.625	1847	Configural	153.883 (43)	< .01	.000
Scalar	7347.689	1890	Metric	76.064 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	11819.287	1804				
Metric	12378.423	1847	Configural	559.136 (43)	< .01	.000
Scalar	12648.918	1890	Metric	270.495 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6852.658	1804				
Metric	6900.130	1847	Configural	47.473 (43)	0.30	.000
Scalar	7101.063	1890	Metric	200.932 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7034.814	1804				
Metric	7278.311	1847	Configural	243.498 (43)	< .01	.000
Scalar	7436.232	1890	Metric	157.921 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7047.919	1804				
Metric	7113.779	1847	Configural	65.860 (43)	0.01	.001
Scalar	7172.388	1890	Metric	58.609 (43)	0.06	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13509.610	1804				
Metric	13867.551	1847	Configural	357.941 (43)	< .01	.000
Scalar	14304.595	1890	Metric	437.045 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	13726.756	1804				
Metric	14038.892	1847	Configural	312.136 (43)	< .01	.000
Scalar	14103.428	1890	Metric	64.536 (43)	0.02	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	13549.468	1804				
Metric	14031.524	1847	Configural	482.057 (43)	< .01	.000
Scalar	14284.069	1890	Metric	252.545 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13220.548	1804				
Metric	14268.173	1847	Configural	1047.625 (43)	< .01	.001
Scalar	14879.019	1890	Metric	610.846 (43)	< .01	.001

Appendix C.1b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	14638.886	1890	< .01	0.945	0.039
Model B-1	7347.689	1890	< .01	0.954	0.035
Model B-2	12648.918	1890	< .01	0.941	0.039
Model B-3	7101.063	1890	< .01	0.956	0.033
Model B-4	7436.232	1890	< .01	0.971	0.023
Model B-5	7172.388	1890	< .01	0.954	0.034
Model C	14304.595	1890	< .01	0.944	0.036
Model D	14103.428	1890	< .01	0.946	0.039
Model E	14284.069	1890	< .01	0.966	0.025
Model F	14879.019	1890	< .01	0.956	0.027

Appendix C.2a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	10163.750	1804				
Metric	10328.986	1847	Configural	165.236 (43)	< .01	.000
Scalar	10978.449	1890	Metric	649.463 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5563.647	1804				
Metric	5758.299	1847	Configural	194.652 (43)	< .01	.000
Scalar	5871.122	1890	Metric	112.823 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8716.399	1804				
Metric	9243.675	1847	Configural	527.276 (43)	< .01	.001
Scalar	9913.604	1890	Metric	669.929 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5405.843	1804				
Metric	5488.063	1847	Configural	82.220 (43)	< .01	.001
Scalar	5701.817	1890	Metric	213.754 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5509.142	1804				
Metric	5806.752	1847	Configural	297.610 (43)	< .01	.001
Scalar	6041.940	1890	Metric	235.189 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5468.639	1804				
Metric	5521.446	1847	Configural	52.807	0.15	.001
Scalar	5556.291	1890	Metric	34.845	0.81	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10234.397	1804				
Metric	10695.816	1847	Configural	461.419 (43)	< .01	.000
Scalar	11194.600	1890	Metric	498.784 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	10133.666	1804				
Metric	10554.761	1847	Configural	421.096 (43)	< .01	.001
Scalar	10709.722	1890	Metric	154.961 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	10034.335	1804				
Metric	10648.370	1847	Configural	614.035 (43)	< .01	.001
Scalar	10960.887	1890	Metric	312.516 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9881.793	1804				
Metric	11189.067	1847	Configural	1307.274 (43)	< .01	.002
Scalar	11780.373	1890	Metric	591.305 (43)	< .01	.000

Appendix C.2b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10978.449	1890	< .01	0.947	0.036
Model B-1	5871.122	1890	< .01	0.986	0.016
Model B-2	9913.604	1890	< .01	0.968	0.024
Model B-3	5701.817	1890	< .01	0.947	0.031
Model B-4	6041.940	1890	< .01	0.984	0.016
Model B-5	5556.291	1890	< .01	0.974	0.021
Model C	11194.600	1890	< .01	0.943	0.033
Model D	10709.722	1890	< .01	0.946	0.036
Model E	10960.887	1890	< .01	0.986	0.016
Model F	11780.373	1890	< .01	0.938	0.034

Appendix C.3a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	11011.824	1804				
Metric	11305.868	1847	Configural	294.044 (43)	< .01	.000
Scalar	11780.179	1890	Metric	474.311 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6128.336	1804				
Metric	6312.812	1847	Configural	184.476 (43)	< .01	.000
Scalar	6406.820	1890	Metric	94.008 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9646.302	1804				
Metric	10044.148	1847	Configural	397.846 (43)	< .01	.001
Scalar	10327.027	1890	Metric	282.879 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6076.608	1804				
Metric	6143.701	1847	Configural	67.094 (43)	0.01	.000
Scalar	6278.654	1890	Metric	134.953 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6160.021	1804				
Metric	6479.098	1847	Configural	319.077 (43)	< .01	.001
Scalar	6674.439	1890	Metric	195.341 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6003.154	1804				
Metric	6075.136	1847	Configural	71.982 (43)	< .01	.000
Scalar	6129.819	1890	Metric	54.682 (43)	0.11	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10971.222	1804				
Metric	11588.948	1847	Configural	617.727 (43)	< .01	.000
Scalar	12189.102	1890	Metric	600.153 (43)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11119.735	1804				
Metric	11347.250	1847	Configural	227.515 (43)	< .01	.000
Scalar	11426.483	1890	Metric	79.233 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11049.637	1804				
Metric	11494.150	1847	Configural	444.512 (43)	< .01	.000
Scalar	11680.610	1890	Metric	186.460 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	10695.809	1804				
Metric	12027.699	1847	Configural	1331.889 (43)	< .01	.002
Scalar	12617.063	1890	Metric	589.364 (43)	< .01	.001

Appendix C.3b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	11780.179	1890	< .01	0.970	0.031
Model B-1	6406.820	1890	< .01	0.986	0.016
Model B-2	10327.027	1890	< .01	0.964	0.032
Model B-3	6278.654	1890	< .01	0.969	0.027
Model B-4	6674.439	1890	< .01	0.984	0.016
Model B-5	6129.819	1890	< .01	0.970	0.027
Model C	12189.102	1890	< .01	0.966	0.029
Model D	11426.483	1890	< .01	0.970	0.031
Model E	11680.610	1890	< .01	0.987	0.015
Model F	12617.063	1890	< .01	0.981	0.017

Appendix C.4a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	9698.728	1804				
Metric	10093.239	1847	Configural	394.510 (43)	< .01	.001
Scalar	10841.059	1890	Metric	747.820 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5611.090	1804				
Metric	5743.655	1847	Configural	132.565 (43)	< .01	.000
Scalar	5860.740	1890	Metric	117.085 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8409.068	1804				
Metric	8938.388	1847	Configural	529.319 (43)	< .01	.001
Scalar	9380.459	1890	Metric	442.072 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5466.053	1804				
Metric	5516.666	1847	Configural	50.614	0.20	.000
Scalar	5655.420	1890	Metric	138.753	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5652.413	1804				
Metric	5960.932	1847	Configural	308.519 (43)	< .01	.000
Scalar	6181.560	1890	Metric	220.628 (43)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5444.331	1804				
Metric	5480.300	1847	Configural	35.969 (43)	0.77	.001
Scalar	5526.887	1890	Metric	46.587 (43)	0.33	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9547.579	1804				
Metric	10051.098	1847	Configural	503.520 (43)	< .01	.001
Scalar	10564.126	1890	Metric	513.028 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9673.726	1804				
Metric	10094.442	1847	Configural	420.715 (43)	< .01	.001
Scalar	10208.083	1890	Metric	113.641 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9627.312	1804				
Metric	9925.714	1847	Configural	298.403 (43)	< .01	.000
Scalar	10293.927	1890	Metric	368.213 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9313.958	1804				
Metric	9980.107	1847	Configural	666.149 (43)	< .01	.000
Scalar	10761.774	1890	Metric	781.667 (43)	< .01	.001

Appendix C.4b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10841.059	1890	< .01	0.880	0.032
Model B-1	5860.740	1890	< .01	0.966	0.030
Model B-2	9380.459	1890	< .01	0.961	0.033
Model B-3	5655.420	1890	< .01	0.989	0.013
Model B-4	6181.560	1890	< .01	0.986	0.016
Model B-5	5526.887	1890	< .01	0.969	0.027
Model C	10564.126	1890	< .01	0.870	0.031
Model D	10208.083	1890	< .01	0.890	0.030
Model E	10293.927	1890	< .01	0.880	0.031
Model F	10761.774	1890	< .01	0.986	0.015

Appendix C.5a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	7386.057	1804				
Metric	7573.876	1847	Configural	187.819 (43)	< .01	.000
Scalar	8502.762	1890	Metric	928.886 (43)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	4703.545	1804				
Metric	4856.004	1847	Configural	152.459 (43)	< .01	.000
Scalar	4973.729	1890	Metric	117.725 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6684.172	1804				
Metric	7103.969	1847	Configural	419.798 (43)	< .01	.000
Scalar	7451.795	1890	Metric	347.826 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4718.096	1804				
Metric	4764.669	1847	Configural	46.573 (43)	0.33	.000
Scalar	4939.561	1890	Metric	174.891 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	4623.233	1804				
Metric	4784.118	1847	Configural	160.885 (43)	< .01	.000
Scalar	4959.031	1890	Metric	174.913 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	4665.775	1804				
Metric	4707.603	1847	Configural	41.828 (43)	0.52	.001
Scalar	4769.442	1890	Metric	61.839 (43)	0.03	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7391.785	1804				
Metric	7699.688	1847	Configural	307.903 (43)	< .01	.001
Scalar	8152.133	1890	Metric	452.445 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7522.564	1804				
Metric	7856.478	1847	Configural	333.913 (43)	< .01	.000
Scalar	7923.163	1890	Metric	66.685 (43)	0.01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7541.658	1804				
Metric	7791.811	1847	Configural	250.154 (43)	< .01	.000
Scalar	8127.541	1890	Metric	335.730 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7431.922	1804				
Metric	7840.706	1847	Configural	408.784 (43)	< .01	.001
Scalar	8519.549	1890	Metric	678.843 (43)	< .01	.001

Appendix C.5b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	8502.762	1890	< .01	0.967	0.030
Model B-1	4973.729	1890	< .01	0.988	0.013
Model B-2	7451.795	1890	< .01	0.961	0.029
Model B-3	4939.561	1890	< .01	0.966	0.024
Model B-4	4959.031	1890	< .01	0.988	0.013
Model B-5	4769.442	1890	< .01	0.969	0.023
Model C	8152.133	1890	< .01	0.989	0.013
Model D	7923.163	1890	< .01	0.966	0.028
Model E	8127.541	1890	< .01	0.990	0.012
Model F	8519.549	1890	< .01	0.988	0.013

Appendix C.6a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	14027.579	1804				
Metric	14325.678	1847	Configural	298.099 (43)	< .01	.000
Scalar	15084.733	1890	Metric	759.055 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7738.802	1804				
Metric	7889.631	1847	Configural	150.829 (43)	< .01	.000
Scalar	8035.171	1890	Metric	145.540 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	12202.711	1804				
Metric	12716.273	1847	Configural	513.562 (43)	< .01	.000
Scalar	12922.089	1890	Metric	205.816 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	7613.130	1804				
Metric	7667.109	1847	Configural	53.979 (43)	0.12	.000
Scalar	7793.346	1890	Metric	126.237 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7776.336	1804				
Metric	7970.933	1847	Configural	194.597 (43)	< .01	.000
Scalar	8130.334	1890	Metric	159.402 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7529.398	1804				
Metric	7562.979	1847	Configural	33.581 (43)	0.85	.001
Scalar	7616.296	1890	Metric	53.317 (43)	0.13	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13905.703	1804				
Metric	14240.091	1847	Configural	334.388 (43)	< .01	.000
Scalar	14864.115	1890	Metric	624.024 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	14078.981	1804				
Metric	14327.866	1847	Configural	248.885 (43)	< .01	.000
Scalar	14421.238	1890	Metric	93.372 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	14095.956	1804				
Metric	14211.952	1847	Configural	115.996 (43)	< .01	.000
Scalar	14414.906	1890	Metric	202.954 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13861.740	1804				
Metric	14406.544	1847	Configural	544.805 (43)	< .01	.000
Scalar	15104.898	1890	Metric	698.353 (43)	< .01	.000

Appendix C.6b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	15084.733	1890	< .01	0.948	0.044
Model B-1	8035.171	1890	< .01	0.957	0.041
Model B-2	12922.089	1890	< .01	0.951	0.044
Model B-3	7793.346	1890	< .01	0.957	0.037
Model B-4	8130.334	1890	< .01	0.958	0.040
Model B-5	7616.296	1890	< .01	0.961	0.036
Model C	14864.115	1890	< .01	0.947	0.038
Model D	14421.238	1890	< .01	0.947	0.042
Model E	14414.906	1890	< .01	0.950	0.035
Model F	15104.898	1890	< .01	0.971	0.038

Appendix C.7a. Global Model Fit Indices of Measurement Invariance Tests for Grade 9 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	9785.018	1978				
Metric	10049.839	2023	Configural	264.821 (45)	< .01	.000
Scalar	10769.979	2068	Metric	720.140 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5881.394	1978				
Metric	5997.772	2023	Configural	116.378 (45)	< .01	.000
Scalar	6098.463	2068	Metric	100.691 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8602.375	1978				
Metric	9013.777	2023	Configural	411.402 (45)	< .01	.000
Scalar	9358.174	2068	Metric	344.397 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5811.836	1978				
Metric	5858.039	2023	Configural	46.203 (45)	0.42	.001
Scalar	5961.451	2068	Metric	103.412 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5876.617	1978				
Metric	6018.198	2023	Configural	141.581 (45)	< .01	.000
Scalar	6266.719	2068	Metric	248.521 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5702.498	1978				
Metric	5741.660	2023	Configural	39.162 (45)	0.72	.001
Scalar	5776.327	2068	Metric	34.667 (45)	0.87	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9819.145	1978				
Metric	10020.292	2023	Configural	201.147 (45)	< .01	.000
Scalar	10344.222	2068	Metric	323.930 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9860.151	1978				
Metric	10147.583	2023	Configural	287.432 (45)	< .01	.000
Scalar	10258.257	2068	Metric	110.674 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9970.516	1978				
Metric	10173.098	2023	Configural	202.582 (45)	< .01	.000
Scalar	10314.595	2068	Metric	141.497 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9912.833	1978				
Metric	10125.393	2023	Configural	212.560 (45)	< .01	.000
Scalar	10352.306	2068	Metric	226.913 (45)	< .01	.000

Appendix C.7b. Global Model Fit Indices of Scalar Invariance Model for Grade 9 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	10769.979	2068	< .01	0.965	0.023
Model B-1	6098.463	2068	< .01	0.972	0.019
Model B-2	9358.174	2068	< .01	0.963	0.023
Model B-3	5961.451	2068	< .01	0.974	0.017
Model B-4	6266.719	2068	< .01	0.973	0.019
Model B-5	5776.327	2068	< .01	0.978	0.015
Model C	10344.222	2068	< .01	0.970	0.019
Model D	10258.257	2068	< .01	0.967	0.022
Model E	10314.595	2068	< .01	0.973	0.018
Model F	10352.306	2068	< .01	0.986	0.012

Appendix C.8a. Global Model Fit Indices of Measurement Invariance Tests for Grade 10 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	11398.967	1978				
Metric	11643.610	2023	Configural	244.644 (45)	< .01	.000
Scalar	12454.369	2068	Metric	810.759 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6787.701	1978				
Metric	6887.953	2023	Configural	100.253 (45)	< .01	.000
Scalar	7000.945	2068	Metric	112.992 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9962.149	1978				
Metric	10396.724	2023	Configural	434.575 (45)	< .01	.001
Scalar	10754.618	2068	Metric	357.894 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6718.523	1978				
Metric	6794.553	2023	Configural	76.030 (45)	< .01	.000
Scalar	6955.160	2068	Metric	160.607 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6677.657	1978				
Metric	6909.453	2023	Configural	231.797 (45)	< .01	.001
Scalar	7122.427	2068	Metric	212.974 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6731.650	1978				
Metric	6784.798	2023	Configural	53.148 (45)	0.19	.000
Scalar	6837.360	2068	Metric	52.562 (45)	0.20	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	11516.948	1978				
Metric	11743.401	2023	Configural	226.452 (45)	< .01	.000
Scalar	12092.078	2068	Metric	348.678 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11482.165	1978				
Metric	11785.845	2023	Configural	303.680 (45)	< .01	.000
Scalar	11878.384	2068	Metric	92.539 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11538.971	1978				
Metric	11651.848	2023	Configural	112.877 (45)	< .01	.000
Scalar	11768.221	2068	Metric	116.373 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	11479.163	1978				
Metric	11665.006	2023	Configural	185.844 (45)	< .01	.000
Scalar	11828.482	2068	Metric	163.476 (45)	< .01	.000

Appendix C.8b. Global Model Fit Indices of Scalar Invariance Model for Grade 10 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	12454.369	2068	< .01	0.962	0.036
Model B-1	7000.945	2068	< .01	0.960	0.021
Model B-2	10754.618	2068	< .01	0.947	0.025
Model B-3	6955.160	2068	< .01	0.958	0.021
Model B-4	7122.427	2068	< .01	0.960	0.021
Model B-5	6837.360	2068	< .01	0.963	0.018
Model C	12092.078	2068	< .01	0.965	0.031
Model D	11878.384	2068	< .01	0.965	0.035
Model E	11768.221	2068	< .01	0.978	0.028
Model F	11828.482	2068	< .01	0.978	0.027

Appendix C.9a. Global Model Fit Indices of Measurement Invariance Tests for Grade 11 ELA

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	7351.566	1978				
Metric	7591.812	2023	Configural	240.246 (45)	< .01	.000
Scalar	8303.823	2068	Metric	712.011 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5120.192	1978				
Metric	5209.060	2023	Configural	88.868 (45)	< .01	.000
Scalar	5319.957	2068	Metric	110.897 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6619.298	1978				
Metric	6981.561	2023	Configural	362.263 (45)	< .01	.001
Scalar	7298.593	2068	Metric	317.032 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4973.414	1978				
Metric	5009.619	2023	Configural	36.205 (45)	0.82	.000
Scalar	5129.101	2068	Metric	119.482 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5244.742	1978				
Metric	5422.163	2023	Configural	177.422 (45)	< .01	.001
Scalar	5688.003	2068	Metric	265.839 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5065.796	1978				
Metric	5119.452	2023	Configural	53.657 (45)	0.18	.000
Scalar	5145.823	2068	Metric	26.371 (45)	0.99	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7513.769	1978				
Metric	7678.313	2023	Configural	164.544 (45)	< .01	.000
Scalar	8058.493	2068	Metric	380.180 (45)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7488.459	1978				
Metric	7754.618	2023	Configural	266.159 (45)	< .01	.001
Scalar	7914.265	2068	Metric	159.647 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7717.892	1978				
Metric	7812.260	2023	Configural	94.368 (45)	< .01	.000
Scalar	7923.082	2068	Metric	110.822 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7704.274	1978				
Metric	7847.220	2023	Configural	142.946 (45)	< .01	.000
Scalar	7993.301	2068	Metric	146.081 (45)	< .01	.000

Appendix C.9b. Global Model Fit Indices of Scalar Invariance Model for Grade 11 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	8303.823	2068	< .01	0.959	0.034
Model B-1	5319.957	2068	< .01	0.973	0.020
Model B-2	7298.593	2068	< .01	0.964	0.023
Model B-3	5129.101	2068	< .01	0.975	0.018
Model B-4	5688.003	2068	< .01	0.972	0.021
Model B-5	5145.823	2068	< .01	0.977	0.016
Model C	8058.493	2068	< .01	0.969	0.027
Model D	7914.265	2068	< .01	0.960	0.032
Model E	7923.082	2068	< .01	0.985	0.014
Model F	7993.301	2068	< .01	0.983	0.015

Appendix C.10a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	178020.127	1890				
Metric	178975.724	1934	Configural	955.597 (44)	< .01	.001
Scalar	183551.597	1978	Metric	4575.874 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	74981.430	1890				
Metric	76809.764	1934	Configural	1828.334 (44)	< .01	.000
Scalar	77843.804	1978	Metric	1034.040 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	143550.580	1890				
Metric	149168.011	1934	Configural	5617.431 (44)	< .01	.000
Scalar	154151.396	1978	Metric	4983.385 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	71303.946	1890				
Metric	71673.151	1934	Configural	369.205 (44)	< .01	.001
Scalar	72140.060	1978	Metric	466.909 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	74961.905	1890				
Metric	77173.889	1934	Configural	2211.984 (44)	< .01	.000
Scalar	78790.231	1978	Metric	1616.342 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	71635.066	1890				
Metric	71792.403	1934	Configural	157.337 (44)	< .01	.000
Scalar	71909.228	1978	Metric	116.825 (44)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	170307.577	1890				
Metric	177157.840	1934	Configural	6850.263 (44)	< .01	.001
Scalar	180389.601	1978	Metric	3231.762 (44)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	175711.661	1890				
Metric	179583.439	1934	Configural	3871.778 (44)	< .01	.000
Scalar	181201.055	1978	Metric	1617.616 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	174317.255	1890				
Metric	180824.968	1934	Configural	6507.712 (44)	< .01	.000
Scalar	182097.212	1978	Metric	1272.244 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	165211.122	1890				
Metric	181890.732	1934	Configural	16679.610 (44)	< .01	.002
Scalar	184706.069	1978	Metric	2815.336 (44)	< .01	.001

Appendix C.10b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	183551.597	1978	< .01	0.917	0.058
Model B-1	77843.804	1978	< .01	0.913	0.056
Model B-2	154151.396	1978	< .01	0.909	0.056
Model B-3	72140.060	1978	< .01	0.922	0.050
Model B-4	78790.231	1978	< .01	0.917	0.055
Model B-5	71909.228	1978	< .01	0.915	0.054
Model C	180389.601	1978	< .01	0.905	0.058
Model D	181201.055	1978	< .01	0.914	0.057
Model E	182097.212	1978	< .01	0.916	0.057
Model F	184706.069	1978	< .01	0.906	0.057

Appendix C.11a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	78513.118	1890				
Metric	79737.405	1934	Configural	1224.287 (44)	< .01	.000
Scalar	84155.308	1978	Metric	4417.903 (44)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	31709.452	1890				
Metric	33546.915	1934	Configural	1837.464 (44)	< .01	.000
Scalar	34341.647	1978	Metric	794.731 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	62665.155	1890				
Metric	66540.482	1934	Configural	3875.328 (44)	< .01	.001
Scalar	69648.954	1978	Metric	3108.472 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	29285.910	1890				
Metric	29504.006	1934	Configural	218.096 (44)	< .01	.000
Scalar	30032.709	1978	Metric	528.703 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	31111.359	1890				
Metric	32984.692	1934	Configural	1873.333 (44)	< .01	.001
Scalar	34001.235	1978	Metric	1016.544 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	29303.472	1890				
Metric	29364.693	1934	Configural	61.220 (44)	.044	.000
Scalar	29494.527	1978	Metric	129.835 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	74429.878	1890				
Metric	80370.099	1934	Configural	5940.221 (44)	< .01	.001
Scalar	84016.074	1978	Metric	3645.976 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	77747.392	1890				
Metric	80662.164	1934	Configural	2914.772 (44)	< .01	.000
Scalar	81460.601	1978	Metric	798.437 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	76537.210	1890				
Metric	82372.857	1934	Configural	5835.648 (44)	< .01	.001
Scalar	83624.630	1978	Metric	1251.772 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	71418.760	1890				
Metric	84391.883	1934	Configural	12973.123 (44)	< .01	.002
Scalar	87919.282	1978	Metric	3527.398 (44)	< .01	.001

Appendix C.11b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	84155.308	1978	< .01	0.962	0.035
Model B-1	34341.647	1978	< .01	0.961	0.032
Model B-2	69648.954	1978	< .01	0.957	0.033
Model B-3	30032.709	1978	< .01	0.960	0.029
Model B-4	34001.235	1978	< .01	0.961	0.031
Model B-5	29494.527	1978	< .01	0.964	0.029
Model C	84016.074	1978	< .01	0.959	0.033
Model D	81460.601	1978	< .01	0.962	0.034
Model E	83624.630	1978	< .01	0.964	0.033
Model F	87919.282	1978	< .01	0.957	0.033

Appendix C.12a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	72135.653	1890				
Metric	73067.921	1934	Configural	932.268 (44)	< .01	.000
Scalar	76282.010	1978	Metric	3214.089 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	29168.642	1890				
Metric	31046.373	1934	Configural	1877.730 (44)	< .01	.001
Scalar	31621.282	1978	Metric	574.909 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	57946.466	1890				
Metric	62395.316	1934	Configural	4448.850 (44)	< .01	.001
Scalar	64315.473	1978	Metric	1920.157 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	26768.137	1890				
Metric	27112.072	1934	Configural	343.935 (44)	< .01	.000
Scalar	27481.486	1978	Metric	369.413 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	29519.527	1890				
Metric	32104.916	1934	Configural	2585.389 (44)	< .01	.001
Scalar	32872.831	1978	Metric	767.915 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26763.882	1890				
Metric	26846.021	1934	Configural	82.140 (44)	< .01	.001
Scalar	26943.191	1978	Metric	97.169 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	68366.428	1890				
Metric	74117.853	1934	Configural	5751.424 (44)	< .01	.001
Scalar	79295.211	1978	Metric	5177.358 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	71225.495	1890				
Metric	74139.995	1934	Configural	2914.500 (44)	< .01	.000
Scalar	74628.714	1978	Metric	488.719 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	70880.121	1890				
Metric	74815.092	1934	Configural	3934.970 (44)	< .01	.000
Scalar	77058.496	1978	Metric	2243.404 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	66416.026	1890				
Metric	77274.731	1934	Configural	10858.705 (44)	< .01	.002
Scalar	83493.323	1978	Metric	6218.592 (44)	< .01	.001

Appendix C.12b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	76282.010	1978	< .01	0.968	0.032
Model B-1	31621.282	1978	< .01	0.968	0.029
Model B-2	64315.473	1978	< .01	0.964	0.031
Model B-3	27481.486	1978	< .01	0.968	0.027
Model B-4	32872.831	1978	< .01	0.968	0.029
Model B-5	26943.191	1978	< .01	0.970	0.027
Model C	79295.211	1978	< .01	0.963	0.031
Model D	74628.714	1978	< .01	0.967	0.031
Model E	77058.496	1978	< .01	0.968	0.030
Model F	83493.323	1978	< .01	0.962	0.031

Appendix C.13a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	90030.261	2068				
Metric	91811.680	2114	Configural	1781.419 (46)	< .01	.000
Scalar	98752.423	2160	Metric	6940.743 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	37453.505	2068				
Metric	40103.361	2114	Configural	2649.856 (46)	< .01	.001
Scalar	40679.803	2160	Metric	576.442 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	71421.466	2068				
Metric	78474.546	2114	Configural	7053.080 (46)	< .01	.002
Scalar	80280.423	2160	Metric	1805.877 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	35345.538	2068				
Metric	35560.721	2114	Configural	215.182 (46)	< .01	.001
Scalar	35866.572	2160	Metric	305.851 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	37208.036	2068				
Metric	39896.741	2114	Configural	2688.705 (46)	< .01	.001
Scalar	40737.236	2160	Metric	840.495 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	34775.418	2068				
Metric	34927.089	2114	Configural	151.671 (46)	< .01	.000
Scalar	34986.399	2160	Metric	59.310 (46)	.090	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	80371.915	2068				
Metric	87601.812	2114	Configural	7229.897 (46)	< .01	.001
Scalar	96412.066	2160	Metric	8810.254 (46)	< .01	.002
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	88516.533	2068				
Metric	93104.084	2114	Configural	4587.551 (46)	< .01	.001
Scalar	93599.712	2160	Metric	495.628 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	86690.679	2068				
Metric	91830.076	2114	Configural	5139.397 (46)	< .01	.000
Scalar	94418.244	2160	Metric	2588.169 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	77015.305	2068				
Metric	89657.183	2114	Configural	12641.879 (46)	< .01	.003
Scalar	98108.681	2160	Metric	8451.497 (46)	< .01	.001

Appendix C.13b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	98752.423	2160	< .01	0.958	0.037
Model B-1	40679.803	2160	< .01	0.960	0.033
Model B-2	80280.423	2160	< .01	0.953	0.036
Model B-3	35866.572	2160	< .01	0.957	0.031
Model B-4	40737.236	2160	< .01	0.960	0.033
Model B-5	34986.399	2160	< .01	0.962	0.031
Model C	96412.066	2160	< .01	0.952	0.034
Model D	93599.712	2160	< .01	0.957	0.036
Model E	94418.244	2160	< .01	0.964	0.033
Model F	98108.681	2160	< .01	0.956	0.033

Appendix C.14a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	37314.410	2068				
Metric	38506.401	2114	Configural	1191.991 (46)	< .01	.000
Scalar	44483.010	2160	Metric	5976.609 (46)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	18571.581	2068				
Metric	20221.959	2114	Configural	1650.378 (46)	< .01	.001
Scalar	20875.847	2160	Metric	653.888 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	30441.823	2068				
Metric	35078.471	2114	Configural	4636.648 (46)	< .01	.001
Scalar	36819.342	2160	Metric	1740.870 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	17534.361	2068				
Metric	17791.556	2114	Configural	257.195 (46)	< .01	.000
Scalar	18399.719	2160	Metric	608.162 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	18001.839	2068				
Metric	19957.793	2114	Configural	1955.954 (46)	< .01	.001
Scalar	21012.517	2160	Metric	1054.724 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	17639.958	2068				
Metric	17735.461	2114	Configural	95.503 (46)	< .01	.000
Scalar	17822.776	2160	Metric	87.315 (46)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	35767.378	2068				
Metric	39249.591	2114	Configural	3482.213 (46)	< .01	.001
Scalar	44955.521	2160	Metric	5705.930 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	36685.569	2068				
Metric	39848.742	2114	Configural	3163.173 (46)	< .01	.001
Scalar	40105.854	2160	Metric	257.113 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	36944.840	2068				
Metric	39359.374	2114	Configural	2414.535 (46)	< .01	.001
Scalar	41609.312	2160	Metric	2249.938 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	34993.541	2068				
Metric	39720.647	2114	Configural	4727.106 (46)	< .01	.001
Scalar	45818.461	2160	Metric	6097.814 (46)	< .01	.002

Appendix C.14b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	44483.010	2160	< .01	0.980	0.022
Model B-1	20875.847	2160	< .01	0.981	0.021
Model B-2	36819.342	2160	< .01	0.977	0.022
Model B-3	18399.719	2160	< .01	0.982	0.020
Model B-4	21012.517	2160	< .01	0.982	0.020
Model B-5	17822.776	2160	< .01	0.983	0.019
Model C	44955.521	2160	< .01	0.980	0.020
Model D	40105.854	2160	< .01	0.981	0.021
Model E	41609.312	2160	< .01	0.985	0.019
Model F	45818.461	2160	< .01	0.981	0.020

Appendix C.15a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 Math

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	51923.973	2068				
Metric	53374.054	2114	Configural	1450.081 (46)	< .01	.000
Scalar	57215.968	2160	Metric	3841.913 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	24553.604	2068				
Metric	25354.154	2114	Configural	800.550 (46)	< .01	.000
Scalar	26162.288	2160	Metric	808.134 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43941.130	2068				
Metric	45642.765	2114	Configural	1701.635 (46)	< .01	.000
Scalar	46780.427	2160	Metric	1137.662 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	23020.720	2068				
Metric	23325.360	2114	Configural	304.640 (46)	< .01	.000
Scalar	23646.081	2160	Metric	320.721 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	24766.121	2068				
Metric	25703.123	2114	Configural	937.002 (46)	< .01	.000
Scalar	26643.128	2160	Metric	940.004 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	23056.447	2068				
Metric	23120.850	2114	Configural	64.402 (46)	.038	.000
Scalar	23224.440	2160	Metric	103.590 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	48778.924	2068				
Metric	52531.621	2114	Configural	3752.697 (46)	< .01	.001
Scalar	57853.129	2160	Metric	5321.508 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	52081.658	2068				
Metric	53241.940	2114	Configural	1160.282 (46)	< .01	.000
Scalar	53724.454	2160	Metric	482.513 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	51790.117	2068				
Metric	52710.250	2114	Configural	920.133 (46)	< .01	.000
Scalar	54441.917	2160	Metric	1731.667 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	48522.307	2068				
Metric	51976.900	2114	Configural	3454.593 (46)	< .01	.001
Scalar	58394.447	2160	Metric	6417.548 (46)	< .01	.001

Appendix C.15b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	57215.968	2160	< .01	0.965	0.028
Model B-1	26162.288	2160	< .01	0.964	0.026
Model B-2	46780.427	2160	< .01	0.964	0.027
Model B-3	23646.081	2160	< .01	0.964	0.026
Model B-4	26643.128	2160	< .01	0.962	0.027
Model B-5	23224.440	2160	< .01	0.966	0.025
Model C	57853.129	2160	< .01	0.965	0.024
Model D	53724.454	2160	< .01	0.966	0.026
Model E	54441.917	2160	< .01	0.971	0.023
Model F	58394.447	2160	< .01	0.966	0.024

Appendix C.17a. Global Model Fit Indices of Measurement Invariance Tests for Algebra I

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	31880.312	2068				
Metric	32905.318	2114	Configural	1025.006 (46)	< .01	.000
Scalar	36406.798	2160	Metric	3501.480 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	16142.158	2068				
Metric	17121.893	2114	Configural	979.735 (46)	< .01	.001
Scalar	17604.115	2160	Metric	482.221 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	26017.305	2068				
Metric	29072.903	2114	Configural	3055.599 (46)	< .01	.001
Scalar	30186.879	2160	Metric	1113.975 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	15857.962	2068				
Metric	16107.596	2114	Configural	249.634 (46)	< .01	.000
Scalar	16536.207	2160	Metric	428.611 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	16758.206	2068				
Metric	18161.297	2114	Configural	1403.091 (46)	< .01	.001
Scalar	18954.656	2160	Metric	793.359 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	15465.159	2068				
Metric	15518.118	2114	Configural	52.959 (46)	.224	.001
Scalar	15611.302	2160	Metric	93.185 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	30806.225	2068				
Metric	32582.106	2114	Configural	1775.881 (46)	< .01	.000
Scalar	34778.963	2160	Metric	2196.858 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	31432.467	2068				
Metric	33610.533	2114	Configural	2178.066 (46)	< .01	.001
Scalar	34009.880	2160	Metric	399.346 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	31591.441	2068				
Metric	32369.885	2114	Configural	778.443 (46)	< .01	.000
Scalar	33674.958	2160	Metric	1305.073 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	31292.754	2068				
Metric	32158.284	2114	Configural	865.530 (46)	< .01	.000
Scalar	33945.363	2160	Metric	1787.079 (46)	< .01	.000

Appendix C.17b. Global Model Fit Indices of Scalar Invariance Model for Algebra I

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	36406.798	2160	< .01	0.929	0.024
Model B-1	17604.115	2160	< .01	0.981	0.021
Model B-2	30186.879	2160	< .01	0.974	0.023
Model B-3	16536.207	2160	< .01	0.981	0.021
Model B-4	18954.656	2160	< .01	0.979	0.022
Model B-5	15611.302	2160	< .01	0.983	0.020
Model C	34778.963	2160	< .01	0.927	0.024
Model D	34009.880	2160	< .01	0.930	0.023
Model E	33674.958	2160	< .01	0.936	0.023
Model F	33945.363	2160	< .01	0.982	0.020

Appendix C.16a. Global Model Fit Indices of Measurement Invariance Tests for Geometry

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	53224.910	2068				
Metric	54454.502	2114	Configural	1229.591 (46)	< .01	.000
Scalar	57940.176	2160	Metric	3485.674 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	27938.686	2068				
Metric	28828.529	2114	Configural	889.843 (46)	< .01	.000
Scalar	29338.108	2160	Metric	509.579 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43389.350	2068				
Metric	47107.528	2114	Configural	3718.178 (46)	< .01	.001
Scalar	48016.657	2160	Metric	909.129 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	27092.053	2068				
Metric	27384.954	2114	Configural	292.902 (46)	< .01	.000
Scalar	27629.871	2160	Metric	244.916 (46)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	27989.160	2068				
Metric	29872.199	2114	Configural	1883.040 (46)	< .01	.001
Scalar	30830.067	2160	Metric	957.867 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26391.366	2068				
Metric	26454.861	2114	Configural	63.495 (46)	.045	.001
Scalar	26525.935	2160	Metric	71.074 (46)	.010	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	52492.263	2068				
Metric	53783.151	2114	Configural	1290.888 (46)	< .01	.000
Scalar	55970.200	2160	Metric	2187.049 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	51894.256	2068				
Metric	54836.635	2114	Configural	2942.380 (46)	< .01	.001
Scalar	55244.890	2160	Metric	408.255 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	53475.540	2068				
Metric	54059.554	2114	Configural	584.014 (46)	< .01	.000
Scalar	54692.029	2160	Metric	632.476 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	52971.934	2068				
Metric	53842.768	2114	Configural	870.834 (46)	< .01	.000
Scalar	55431.388	2160	Metric	1588.620 (46)	< .01	.000

Appendix C.16b. Global Model Fit Indices of Scalar Invariance Model for Geometry

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	57940.176	2160	< .01	0.930	0.038
Model B-1	29338.108	2160	< .01	0.950	0.033
Model B-2	48016.657	2160	< .01	0.929	0.035
Model B-3	27629.871	2160	< .01	0.947	0.034
Model B-4	30830.067	2160	< .01	0.951	0.032
Model B-5	26525.935	2160	< .01	0.954	0.031
Model C	55970.200	2160	< .01	0.942	0.031
Model D	55244.890	2160	< .01	0.934	0.036
Model E	54692.029	2160	< .01	0.944	0.032
Model F	55431.388	2160	< .01	0.955	0.027

Appendix C.18a. Global Model Fit Indices of Measurement Invariance Tests for Algebra II

Invariance Model	χ^2	df	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
Model A: Students' Gender (Female vs. Male)						
Configural	17718.329	2068				
Metric	18874.760	2114	Configural	1156.431 (46)	< .01	.000
Scalar	20911.565	2160	Metric	2036.806 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	10607.551	2068				
Metric	11002.537	2114	Configural	394.986 (46)	< .01	.000
Scalar	11497.152	2160	Metric	494.615 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	15096.010	2068				
Metric	16321.688	2114	Configural	1225.678 (46)	< .01	.001
Scalar	17382.791	2160	Metric	1061.103 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	10235.537	2068				
Metric	10522.133	2114	Configural	286.595 (46)	< .01	.001
Scalar	10828.678	2160	Metric	306.545 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	10485.655	2068				
Metric	11186.459	2114	Configural	700.803 (46)	< .01	.001
Scalar	12270.158	2160	Metric	1083.699 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	10048.537	2068				
Metric	10133.402	2114	Configural	84.865 (46)	< .01	.001
Scalar	10216.730	2160	Metric	83.328 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	17849.740	2068				
Metric	18266.012	2114	Configural	416.272 (46)	< .01	.000
Scalar	19649.191	2160	Metric	1383.179 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	17496.271	2068				
Metric	18725.869	2114	Configural	1229.598 (46)	< .01	.001
Scalar	19370.546	2160	Metric	644.677 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	17883.903	2068				
Metric	18249.789	2114	Configural	365.886 (46)	< .01	.000
Scalar	18641.474	2160	Metric	391.685 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	18128.269	2068				
Metric	18473.548	2114	Configural	345.279 (46)	< .01	.000
Scalar	19474.733	2160	Metric	1001.185 (46)	< .01	.000

Appendix C.18b. Global Model Fit Indices of Scalar Invariance Model for Algebra II

Model	Chi-Square Test			CFI	RMSEA
	Value	<i>df</i>	P-Value		
Model A	20911.565	2160	< .01	0.979	0.017
Model B-1	11497.152	2160	< .01	0.986	0.014
Model B-2	17382.791	2160	< .01	0.978	0.017
Model B-3	10828.678	2160	< .01	0.985	0.015
Model B-4	12270.158	2160	< .01	0.986	0.015
Model B-5	10216.730	2160	< .01	0.988	0.013
Model C	19649.191	2160	< .01	0.986	0.013
Model D	19370.546	2160	< .01	0.982	0.016
Model E	18641.474	2160	< .01	0.986	0.013
Model F	19474.733	2160	< .01	0.989	0.011

Appendix D. Regression Model Parameter Estimates of Differential Growth across Subgroups-ELA

Parameter	2016_G3E to 2017_G4E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2526.01	0.13	<.0001	0.00
Female vs. Male (β_{01})	1.28	0.13	<.0001	0.02
Special Education Status vs. Non-SPED (β_{02})	-6.86	0.26	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.80	0.37	<.0001	-0.08
Low income vs. Non-Low Income (β_{04})	-1.90	0.14	<.0001	-0.03
Asian vs. White (β_{05})	3.61	0.45	<.0001	0.02
Hispanic vs. White (β_{06})	-3.86	0.15	<.0001	-0.06
African American vs. White (β_{07})	-4.50	0.32	<.0001	-0.03
Hawaiian/Pacific Islander vs. White (β_{08})	-2.40	1.07	0.0254	0.00
American Indian vs. White (β_{09})	-7.91	0.36	<.0001	-0.05
Multiple vs. White (β_{010})	-1.21	0.37	0.0012	-0.01
Slope (β_{10})	0.77	0.00	<.0001	0.77
Female vs. Male (β_{11})	0.00	0.00	0.4074	0.00
Special Education Status vs. Non-SPED (β_{12})	0.00	0.01	0.5168	0.00
Limited English Proficiency vs. Non-LEP (β_{13})	-0.15	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian vs. White (β_{15})	-0.03	0.01	0.0071	-0.01
Hispanic vs. White (β_{16})	-0.01	0.00	0.1707	0.00
African American vs. White (β_{17})	-0.01	0.01	0.6155	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.01	0.04	0.6819	0.00
American Indian vs. White (β_{19})	-0.02	0.01	0.1476	0.00
Multiple vs. White (β_{110})	0.00	0.01	0.8652	0.00

Parameter	2016_G4E to 2017_G5E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2540.03	0.14	<.0001	0.00
Female vs. Male (β_{01})	0.45	0.13	0.0004	0.01
Special Education Status vs. Non-SPED (β_{02})	-7.93	0.28	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-10.15	0.50	<.0001	-0.08
Low income vs. Non-Low Income (β_{04})	-2.06	0.14	<.0001	-0.03
Asian vs. White (β_{05})	2.78	0.47	<.0001	0.01
Hispanic vs. White (β_{06})	-3.01	0.15	<.0001	-0.04
African American vs. White (β_{07})	-4.81	0.33	<.0001	-0.03
Hawaiian/Pacific Islander vs. White (β_{08})	-0.75	1.04	0.4680	0.00
American Indian vs. White (β_{09})	-7.24	0.37	<.0001	-0.04
Multiple vs. White (β_{010})	-1.22	0.40	0.0023	-0.01
Slope (β_{10})	0.80	0.00	<.0001	0.80
Female vs. Male (β_{11})	0.01	0.00	0.1493	0.00
Special Education Status vs. Non-SPED (β_{12})	-0.03	0.01	<.0001	-0.01
Limited English Proficiency vs. Non-LEP (β_{13})	-0.16	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.01
Asian vs. White (β_{15})	-0.04	0.01	0.0017	-0.01
Hispanic vs. White (β_{16})	-0.02	0.00	<.0001	-0.01
African American vs. White (β_{17})	-0.03	0.01	0.0003	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	-0.03	0.03	0.4155	0.00
American Indian vs. White (β_{19})	-0.03	0.01	0.0050	-0.01
Multiple vs. White (β_{110})	0.00	0.01	0.8827	0.00

Parameter	2016_G5E to 2017_G6E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2547.62	0.13	<.0001	0.00
Female vs. Male (β_{01})	-0.06	0.13	0.6235	0.00
Special Education Status vs. Non-SPED (β_{02})	-6.63	0.30	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.29	0.50	<.0001	-0.06
Low income vs. Non-Low Income (β_{04})	-1.32	0.14	<.0001	-0.02
Asian vs. White (β_{05})	4.55	0.46	<.0001	0.02
Hispanic vs. White (β_{06})	-3.09	0.15	<.0001	-0.05
African American vs. White (β_{07})	-2.79	0.32	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	-0.86	1.13	0.4494	0.00
American Indian vs. White (β_{09})	-4.80	0.37	<.0001	-0.03
Multiple vs. White (β_{010})	-0.77	0.41	0.0584	0.00
Slope (β_{10})	0.74	0.00	<.0001	0.79
Female vs. Male (β_{11})	-0.03	0.00	<.0001	-0.03
Special Education Status vs. Non-SPED (β_{12})	0.00	0.01	0.9475	0.00
Limited English Proficiency vs. Non-LEP (β_{13})	-0.10	0.01	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.01	0.00	0.0011	-0.01
Asian vs. White (β_{15})	0.01	0.01	0.2202	0.00
Hispanic vs. White (β_{16})	0.01	0.00	0.2585	0.00
African American vs. White (β_{17})	-0.01	0.01	0.5104	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.05	0.03	0.1418	0.00
American Indian vs. White (β_{19})	0.02	0.01	0.0186	0.01
Multiple vs. White (β_{110})	0.03	0.01	0.0189	0.00

Parameter	2016_G6E to 2017_G7E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2555.51	0.13	<.0001	0.00
Female vs. Male (β_{01})	2.20	0.12	<.0001	0.03
Special Education Status vs. Non-SPED (β_{02})	-8.69	0.32	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-6.74	0.56	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-0.88	0.14	<.0001	-0.01
Asian vs. White (β_{05})	2.70	0.44	<.0001	0.01
Hispanic vs. White (β_{06})	-1.57	0.15	<.0001	-0.02
African American vs. White (β_{07})	-2.30	0.31	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	-1.30	1.10	0.2351	0.00
American Indian vs. White (β_{09})	-4.81	0.37	<.0001	-0.03
Multiple vs. White (β_{010})	-0.99	0.42	0.0190	0.00
Slope (β_{10})	0.76	0.00	<.0001	0.79
Female vs. Male (β_{11})	-0.02	0.00	<.0001	-0.02
Special Education Status vs. Non-SPED (β_{12})	0.02	0.01	0.0079	0.01
Limited English Proficiency vs. Non-LEP (β_{13})	-0.01	0.01	0.3098	0.00
Low income vs. Non-Low Income (β_{14})	-0.01	0.00	0.0302	-0.01
Asian vs. White (β_{15})	-0.03	0.01	0.0079	-0.01
Hispanic vs. White (β_{16})	0.01	0.00	0.0014	0.01
African American vs. White (β_{17})	0.03	0.01	0.0070	0.01
Hawaiian/Pacific Islander vs. White (β_{18})	0.02	0.03	0.5043	0.00
American Indian vs. White (β_{19})	0.02	0.01	0.0718	0.00
Multiple vs. White (β_{110})	0.00	0.01	0.9735	0.00

Parameter	2016_G7E to 2017_G8E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2555.80	0.13	<.0001	0.00
Female vs. Male (β_{01})	2.60	0.13	<.0001	0.04
Special Education Status vs. Non-SPED (β_{02})	-8.76	0.34	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-8.46	0.61	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-0.87	0.14	<.0001	-0.01
Asian vs. White (β_{05})	5.41	0.46	<.0001	0.03
Hispanic vs. White (β_{06})	-1.42	0.15	<.0001	-0.02
African American vs. White (β_{07})	-2.35	0.32	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	0.03	1.20	0.9770	0.00
American Indian vs. White (β_{09})	-4.20	0.40	<.0001	-0.02
Multiple vs. White (β_{010})	-0.57	0.46	0.2139	0.00
Slope (β_{10})	0.90	0.00	<.0001	0.82
Female vs. Male (β_{11})	0.00	0.00	0.3740	0.00
Special Education Status vs. Non-SPED (β_{12})	-0.11	0.01	<.0001	-0.04
Limited English Proficiency vs. Non-LEP (β_{13})	-0.12	0.01	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.01	0.00	0.0044	-0.01
Asian vs. White (β_{15})	-0.03	0.01	0.0217	-0.01
Hispanic vs. White (β_{16})	0.00	0.01	0.4513	0.00
African American vs. White (β_{17})	0.00	0.01	0.9459	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.03	0.04	0.4171	0.00
American Indian vs. White (β_{19})	-0.04	0.01	0.0004	-0.01
Multiple vs. White (β_{110})	0.02	0.02	0.2886	0.00

Parameter	2016_G8E to 2017_G9E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2568.44	0.12	<.0001	0.00
Female vs. Male (β_{01})	2.36	0.13	<.0001	0.04
Special Education Status vs. Non-SPED (β_{02})	-5.95	0.35	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-11.87	1.28	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-1.46	0.14	<.0001	-0.02
Asian vs. White (β_{05})	6.13	0.44	<.0001	0.04
Hispanic vs. White (β_{06})	-2.33	0.15	<.0001	-0.04
African American vs. White (β_{07})	-2.44	0.31	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	-0.49	1.13	0.6633	0.00
American Indian vs. White (β_{09})	-2.17	0.39	<.0001	-0.02
Multiple vs. White (β_{010})	-0.98	0.46	0.0315	0.00
Slope (β_{10})	0.74	0.00	<.0001	0.80
Female vs. Male (β_{11})	0.00	0.00	0.2710	0.00
Special Education Status vs. Non-SPED (β_{12})	-0.05	0.01	<.0001	-0.02
Limited English Proficiency vs. Non-LEP (β_{13})	-0.15	0.03	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.01
Asian vs. White (β_{15})	0.00	0.01	0.7223	0.00
Hispanic vs. White (β_{16})	-0.03	0.00	<.0001	-0.02
African American vs. White (β_{17})	-0.01	0.01	0.1401	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.00	0.04	0.9924	0.00
American Indian vs. White (β_{19})	-0.01	0.01	0.2950	0.00
Multiple vs. White (β_{110})	-0.01	0.01	0.7133	0.00

Parameter	2016_G9E to 2017_G10E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2568.73	0.13	<.0001	0.00
Female vs. Male (β_{01})	-0.45	0.13	0.0006	-0.01
Special Education Status vs. Non-SPED (β_{02})	-7.90	0.37	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-14.55	1.35	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-1.44	0.15	<.0001	-0.02
Asian vs. White (β_{05})	3.46	0.44	<.0001	0.02
Hispanic vs. White (β_{06})	-2.82	0.15	<.0001	-0.05
African American vs. White (β_{07})	-2.77	0.32	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	-1.55	1.11	0.1619	0.00
American Indian vs. White (β_{09})	-5.47	0.40	<.0001	-0.04
Multiple vs. White (β_{010})	-0.05	0.48	0.9127	0.00
Slope (β_{10})	0.75	0.00	<.0001	0.78
Female vs. Male (β_{11})	-0.01	0.00	0.0012	-0.01
Special Education Status vs. Non-SPED (β_{12})	-0.05	0.01	<.0001	-0.02
Limited English Proficiency vs. Non-LEP (β_{13})	-0.18	0.03	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	0.00	0.01	0.7102	0.00
Asian vs. White (β_{15})	-0.01	0.01	0.3188	0.00
Hispanic vs. White (β_{16})	0.01	0.01	0.0251	0.01
African American vs. White (β_{17})	0.00	0.01	0.6793	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.01	0.04	0.8246	0.00
American Indian vs. White (β_{19})	-0.04	0.01	0.0025	-0.01
Multiple vs. White (β_{110})	0.01	0.02	0.3800	0.00

Parameter	2016_G10E to 2017_G11E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	2566.85	0.14	<.0001	0.00
Female vs. Male (β_{01})	3.52	0.15	<.0001	0.06
Special Education Status vs. Non-SPED (β_{02})	-7.39	0.41	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-13.10	1.57	<.0001	-0.04
Low income vs. Non-Low Income (β_{04})	-0.79	0.17	<.0001	-0.01
Asian vs. White (β_{05})	2.02	0.49	<.0001	0.01
Hispanic vs. White (β_{06})	-1.83	0.17	<.0001	-0.03
African American vs. White (β_{07})	-2.29	0.36	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	-2.55	1.46	0.0816	0.00
American Indian vs. White (β_{09})	-2.19	0.45	<.0001	-0.01
Multiple vs. White (β_{010})	-0.25	0.54	0.6389	0.00
Slope (β_{10})	0.83	0.00	<.0001	0.80
Female vs. Male (β_{11})	-0.05	0.01	<.0001	-0.03
Special Education Status vs. Non-SPED (β_{12})	0.02	0.01	0.1713	0.01
Limited English Proficiency vs. Non-LEP (β_{13})	-0.13	0.04	0.0010	-0.02
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.02
Asian vs. White (β_{15})	-0.01	0.01	0.4216	0.00
Hispanic vs. White (β_{16})	-0.02	0.01	0.0002	-0.01
African American vs. White (β_{17})	0.00	0.01	0.8036	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	-0.04	0.05	0.3540	0.00
American Indian vs. White (β_{19})	-0.01	0.02	0.3580	0.00
Multiple vs. White (β_{110})	0.00	0.02	0.8985	0.00

Appendix D. Regression Model Parameter Estimates of Differential Growth across Subgroups-MATH

Parameter	2016_G3M to 2017_G4M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3562.80	0.19	<.0001	0.00
Female vs. Male (β_{01})	-1.77	0.18	<.0001	-0.02
Special Education Status vs. Non-SPED (β_{02})	-8.99	0.35	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.66	0.43	<.0001	-0.06
Low income vs. Non-Low Income (β_{04})	-2.42	0.19	<.0001	-0.03
Asian vs. White (β_{05})	5.85	0.69	<.0001	0.02
Hispanic vs. White (β_{06})	-4.16	0.21	<.0001	-0.05
African American vs. White (β_{07})	-6.39	0.46	<.0001	-0.03
Hawaiian/Pacific Islander vs. White (β_{08})	-4.81	1.52	0.0015	-0.01
American Indian vs. White (β_{09})	-8.51	0.50	<.0001	-0.04
Multiple vs. White (β_{010})	-0.54	0.52	0.3065	0.00
Slope (β_{10})	0.77	0.00	<.0001	0.76
Female vs. Male (β_{11})	-0.01	0.00	0.0400	-0.01
Special Education Status vs. Non-SPED (β_{12})	0.07	0.01	<.0001	0.03
Limited English Proficiency vs. Non-LEP (β_{13})	-0.04	0.01	<.0001	-0.01
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian vs. White (β_{15})	-0.01	0.01	0.5169	0.00
Hispanic vs. White (β_{16})	-0.01	0.00	0.2676	0.00
African American vs. White (β_{17})	0.00	0.01	0.6586	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.00	0.03	0.8876	0.00
American Indian vs. White (β_{19})	0.05	0.01	<.0001	0.01
Multiple vs. White (β_{110})	0.01	0.01	0.5810	0.00

Parameter	2016_G4M to 2017_G5M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3592.95	0.18	<.0001	0.00
Female vs. Male (β_{01})	-0.04	0.17	0.8327	0.00
Special Education Status vs. Non-SPED (β_{02})	-7.95	0.36	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-6.62	0.54	<.0001	-0.04
Low income vs. Non-Low Income (β_{04})	-2.51	0.19	<.0001	-0.03
Asian vs. White (β_{05})	7.80	0.68	<.0001	0.03
Hispanic vs. White (β_{06})	-1.96	0.20	<.0001	-0.02
African American vs. White (β_{07})	-4.23	0.45	<.0001	-0.02
Hawaiian/Pacific Islander vs. White (β_{08})	1.84	1.39	0.1851	0.00
American Indian vs. White (β_{09})	-6.89	0.48	<.0001	-0.03
Multiple vs. White (β_{010})	-0.39	0.52	0.4511	0.00
Slope (β_{10})	0.87	0.00	<.0001	0.79
Female vs. Male (β_{11})	-0.02	0.00	<.0001	-0.01
Special Education Status vs. Non-SPED (β_{12})	0.08	0.01	<.0001	0.03
Limited English Proficiency vs. Non-LEP (β_{13})	-0.06	0.01	<.0001	-0.02
Low income vs. Non-Low Income (β_{14})	0.00	0.00	0.9176	0.00
Asian vs. White (β_{15})	-0.03	0.01	0.0481	0.00
Hispanic vs. White (β_{16})	-0.01	0.01	0.0077	-0.01
African American vs. White (β_{17})	0.01	0.01	0.6221	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	-0.03	0.03	0.4044	0.00
American Indian vs. White (β_{19})	-0.01	0.01	0.3514	0.00
Multiple vs. White (β_{110})	0.01	0.01	0.5939	0.00

Parameter	2016_G5M to 2017_G6M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3625.02	0.18	<.0001	0.00
Female vs. Male (β_{01})	-1.14	0.17	<.0001	-0.01
Special Education Status vs. Non-SPED (β_{02})	-8.98	0.38	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-9.41	0.57	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-2.81	0.18	<.0001	-0.03
Asian vs. White (β_{05})	3.89	0.66	<.0001	0.01
Hispanic vs. White (β_{06})	-5.68	0.20	<.0001	-0.06
African American vs. White (β_{07})	-8.16	0.43	<.0001	-0.04
Hawaiian/Pacific Islander vs. White (β_{08})	-1.93	1.50	0.1998	0.00
American Indian vs. White (β_{09})	-10.36	0.47	<.0001	-0.05
Multiple vs. White (β_{010})	-3.01	0.54	<.0001	-0.01
Slope (β_{10})	0.83	0.00	<.0001	0.78
Female vs. Male (β_{11})	0.01	0.00	0.0005	0.01
Special Education Status vs. Non-SPED (β_{12})	0.05	0.01	<.0001	0.02
Limited English Proficiency vs. Non-LEP (β_{13})	-0.07	0.01	<.0001	-0.02
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian vs. White (β_{15})	0.05	0.01	<.0001	0.01
Hispanic vs. White (β_{16})	-0.02	0.00	0.0001	-0.01
African American vs. White (β_{17})	-0.01	0.01	0.2625	0.00
Hawaiian/Pacific Islander vs. White (β_{18})	0.07	0.04	0.0661	0.00
American Indian vs. White (β_{19})	-0.02	0.01	0.1373	0.00
Multiple vs. White (β_{110})	0.02	0.01	0.2310	0.00

Parameter	2016_G6M to 2017_G7M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3635.88	0.18	<.0001	0.00
Female vs. Male (β_{01})	0.18	0.17	0.3000	0.00
Special Education Status vs. Non-SPED (β_{02})	-11.47	0.42	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-10.86	0.61	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-1.82	0.19	<.0001	-0.02
Asian vs. White (β_{05})	6.42	0.64	<.0001	0.02
Hispanic vs. White (β_{06})	-3.90	0.20	<.0001	-0.04
African American vs. White (β_{07})	-5.20	0.44	<.0001	-0.03
Hawaiian/Pacific Islander vs. White (β_{08})	-3.70	1.53	0.0155	0.00
American Indian vs. White (β_{09})	-8.51	0.48	<.0001	-0.04
Multiple vs. White (β_{010})	-1.33	0.58	0.0215	0.00
Slope (β_{10})	0.95	0.00	<.0001	0.86
Female vs. Male (β_{11})	0.00	0.00	0.5689	0.00
Special Education Status vs. Non-SPED (β_{12})	-0.18	0.01	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{13})	-0.19	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.04	0.00	<.0001	-0.02
Asian vs. White (β_{15})	-0.02	0.01	0.0822	0.00
Hispanic vs. White (β_{16})	-0.03	0.01	<.0001	-0.02
African American vs. White (β_{17})	-0.05	0.01	<.0001	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	-0.04	0.04	0.2279	0.00
American Indian vs. White (β_{19})	-0.10	0.01	<.0001	-0.02
Multiple vs. White (β_{110})	0.01	0.01	0.4099	0.00

Parameter	2016_G7M to 2017_G8M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3654.37	0.19	<.0001	0.00
Female vs. Male (β_{01})	0.08	0.17	0.6449	0.00
Special Education Status vs. Non-SPED (β_{02})	-8.89	0.40	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.96	0.59	<.0001	-0.05
Low income vs. Non-Low Income (β_{04})	-1.03	0.18	<.0001	-0.01
Asian vs. White (β_{05})	6.19	0.72	<.0001	0.02
Hispanic vs. White (β_{06})	-1.98	0.20	<.0001	-0.03
African American vs. White (β_{07})	-2.09	0.41	<.0001	-0.01
Hawaiian/Pacific Islander vs. White (β_{08})	-1.28	1.59	0.4212	0.00
American Indian vs. White (β_{09})	-3.71	0.46	<.0001	-0.02
Multiple vs. White (β_{010})	-2.34	0.61	0.0001	-0.01
Slope (β_{10})	0.99	0.01	<.0001	0.83
Female vs. Male (β_{11})	-0.03	0.01	<.0001	-0.02
Special Education Status vs. Non-SPED (β_{12})	-0.16	0.01	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{13})	-0.19	0.02	<.0001	-0.04
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.01
Asian vs. White (β_{15})	0.06	0.02	0.0008	0.01
Hispanic vs. White (β_{16})	-0.01	0.01	0.1387	-0.01
African American vs. White (β_{17})	-0.05	0.01	0.0002	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	-0.01	0.05	0.8467	0.00
American Indian vs. White (β_{19})	-0.04	0.01	0.0020	-0.01
Multiple vs. White (β_{110})	-0.04	0.02	0.0190	-0.01

Parameter	2016_G8M to 2017_Alg I			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3667.64	0.19	<.0001	0.00
Female vs. Male (β_{01})	1.98	0.18	<.0001	0.03
Special Education Status vs. Non-SPED (β_{02})	-7.30	0.40	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-12.32	1.05	<.0001	-0.06
Low income vs. Non-Low Income (β_{04})	-1.52	0.19	<.0001	-0.02
Asian vs. White (β_{05})	4.11	0.75	<.0001	0.02
Hispanic vs. White (β_{06})	-2.72	0.21	<.0001	-0.04
African American vs. White (β_{07})	-2.01	0.42	<.0001	-0.01
Hawaiian/Pacific Islander vs. White (β_{08})	-1.85	1.65	0.2625	0.00
American Indian vs. White (β_{09})	-4.59	0.49	<.0001	-0.03
Multiple vs. White (β_{010})	-0.53	0.67	0.4280	0.00
Slope (β_{10})	0.75	0.01	<.0001	0.78
Female vs. Male (β_{11})	0.00	0.01	0.3761	0.00
Special Education Status vs. Non-SPED (β_{12})	-0.10	0.01	<.0001	-0.04
Limited English Proficiency vs. Non-LEP (β_{13})	-0.24	0.03	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.02	0.01	0.0002	-0.01
Asian vs. White (β_{15})	0.07	0.02	0.0011	0.01
Hispanic vs. White (β_{16})	-0.04	0.01	<.0001	-0.02
African American vs. White (β_{17})	-0.04	0.01	0.0030	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	-0.02	0.05	0.6855	0.00
American Indian vs. White (β_{19})	-0.06	0.01	<.0001	-0.01
Multiple vs. White (β_{110})	-0.02	0.02	0.3605	0.00

Parameter	2016_Alg I to 2017_Geo			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3689.11	0.18	<.0001	0.00
Female vs. Male (β_{01})	-1.48	0.18	<.0001	-0.02
Special Education Status vs. Non-SPED (β_{02})	-7.24	0.46	<.0001	-0.05
Limited English Proficiency vs. Non-LEP (β_{03})	-10.17	1.28	<.0001	-0.04
Low income vs. Non-Low Income (β_{04})	-1.62	0.20	<.0001	-0.02
Asian vs. White (β_{05})	1.45	0.66	0.0270	0.01
Hispanic vs. White (β_{06})	-4.81	0.21	<.0001	-0.06
African American vs. White (β_{07})	-6.42	0.45	<.0001	-0.04
Hawaiian/Pacific Islander vs. White (β_{08})	-2.31	1.58	0.1444	0.00
American Indian vs. White (β_{09})	-4.48	0.50	<.0001	-0.03
Multiple vs. White (β_{010})	-2.30	0.68	0.0007	-0.01
Slope (β_{10})	0.87	0.00	<.0001	0.80
Female vs. Male (β_{11})	-0.01	0.01	0.0063	-0.01
Special Education Status vs. Non-SPED (β_{12})	-0.01	0.01	0.5949	0.00
Limited English Proficiency vs. Non-LEP (β_{13})	-0.17	0.03	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.01
Asian vs. White (β_{15})	0.05	0.01	0.0013	0.01
Hispanic vs. White (β_{16})	-0.03	0.01	<.0001	-0.02
African American vs. White (β_{17})	-0.03	0.01	0.0318	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	0.01	0.05	0.8509	0.00
American Indian vs. White (β_{19})	-0.04	0.02	0.0094	-0.01
Multiple vs. White (β_{110})	0.01	0.02	0.5999	0.00

Parameter	2016_Geo to 2017_Alg II			
	Unstandardized Estimate	SE	p value	Standardized Estimate
Intercept (β_{00})	3702.63	0.18	<.0001	0.00
Female vs. Male (β_{01})	-0.31	0.19	0.1015	0.00
Special Education Status vs. Non-SPED (β_{02})	-6.55	0.58	<.0001	-0.05
Limited English Proficiency vs. Non-LEP (β_{03})	-9.93	1.63	<.0001	-0.03
Low income vs. Non-Low Income (β_{04})	-3.44	0.22	<.0001	-0.05
Asian vs. White (β_{05})	8.31	0.61	<.0001	0.05
Hispanic vs. White (β_{06})	-2.61	0.22	<.0001	-0.04
African American vs. White (β_{07})	-2.27	0.50	<.0001	-0.01
Hawaiian/Pacific Islander vs. White (β_{08})	2.17	1.76	0.2164	0.00
American Indian vs. White (β_{09})	-5.16	0.55	<.0001	-0.03
Multiple vs. White (β_{010})	-0.12	0.69	0.8649	0.00
Slope (β_{10})	0.77	0.00	<.0001	0.81
Female vs. Male (β_{11})	-0.04	0.01	<.0001	-0.03
Special Education Status vs. Non-SPED (β_{12})	-0.09	0.01	<.0001	-0.03
Limited English Proficiency vs. Non-LEP (β_{13})	-0.22	0.04	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.06	0.01	<.0001	-0.03
Asian vs. White (β_{15})	0.05	0.01	0.0001	0.01
Hispanic vs. White (β_{16})	-0.05	0.01	<.0001	-0.03
African American vs. White (β_{17})	-0.05	0.01	0.0004	-0.01
Hawaiian/Pacific Islander vs. White (β_{18})	0.00	0.05	0.9780	0.00
American Indian vs. White (β_{19})	-0.07	0.02	<.0001	-0.01
Multiple vs. White (β_{110})	0.00	0.02	0.9832	0.00

Appendix E. Equations and Formula for Estimating Reliability

D.1 Standard Error Formula

For the AzMERIT assessments scored using MLE, according to Masters (1982), the asymptotic estimate of the standard error for ability θ is given by

$$SE(\theta) = \left[\sum_{i=1}^N \sum_{x_i=0}^{m_i} x_i^2 P(X_i = x_i | \theta) - \sum_{i=1}^N \left[\sum_{x_i=0}^{m_i} x_i P(X_i = x_i | \theta) \right]^2 \right]^{-\frac{1}{2}},$$

which is further placed onto the reporting scale by the following transformation:

$$SE_{vs} = a \times SE(\theta),$$

where a is the slope of the scaling constants that take θ to the reporting scale. For both ELA and Mathematics tests, $a = 30$.

D.2 Student Classification Consistency Formula

For a student with estimated ability $\hat{\theta}$ and associated standard error $se(\hat{\theta})$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $se(\hat{\theta})$, that is, $\hat{\theta} \sim N(\theta, se(\hat{\theta})^2)$. The probability of the true score *at or above* the cut score θ_c is estimated as

$$P(\theta \geq \theta_c) = P\left(\frac{\theta - \hat{\theta}}{se(\hat{\theta})} \geq \frac{\theta_c - \hat{\theta}}{se(\hat{\theta})}\right) = P\left(\frac{\hat{\theta} - \theta}{se(\hat{\theta})} < \frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right),$$

where $\Phi(\cdot)$ is the cumulative function of standard normal distribution. Similarly, the probability of the true score being *below* the cut score is estimated as

$$P(\theta < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right).$$

D.2.1 Classification Accuracy Formula

The probability of a student with true ability θ being classified *at or above* the cut score θ_c , given the student's item scores $\mathbf{x} = (x_1, \dots, x_N)$, can be estimated as

$$P(\theta \geq \theta_c | \mathbf{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta},$$

where the likelihood function is

$$L(\theta | \mathbf{x}) = \prod_{i=1}^N P(x_i | \theta),$$

and $P(x_i|\theta)$ is calculated from the Rasch model or partial credit model based on the estimated item parameters.

Similarly, we can estimate the probability of *below* the cut score as:

$$P(\theta < \theta_c | \mathbf{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta}$$

Mathematically, we have

$$\begin{aligned} N_{11} &= \sum_{i \in N_1} P(\theta_i \geq \theta_c | \mathbf{x}), \\ N_{01} &= \sum_{i \in N_1} P(\theta_i < \theta_c | \mathbf{x}), \\ N_{10} &= \sum_{i \in N_0} P(\theta_i \geq \theta_c | \mathbf{x}), \text{ and} \\ N_{00} &= \sum_{i \in N_0} P(\theta_i < \theta_c | \mathbf{x}), \end{aligned}$$

where N_1 consists of the students with estimated $\hat{\theta}_i$ being *at and above* the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being *below* the cut score. The accuracy index is then computed as:

$$\frac{N_{11} + N_{00}}{N_1 + N_0}.$$

D.2.2 Classification Consistency Formula

To estimate the consistency, we assume the students are tested twice independently; hence, the probability of the student being classified as *at or above* the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

Similarly, the probability of consistency for *at or above* the cut score is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c | \mathbf{x}) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

The probability of consistency for *below* the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c | \mathbf{x}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta} \right)^2.$$

The probability of inconsistency is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 < \theta_c | \mathbf{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta \int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta \right]^2}, \text{ and}$$

$$P(\theta_1 < \theta_c, \theta_2 \geq \theta_c | \mathbf{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta | \mathbf{x}) d\theta \int_{\theta_c}^{+\infty} L(\theta | \mathbf{x}) d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta | \mathbf{x}) d\theta \right]^2}.$$

The consistent index is computed as $\frac{N_{11} + N_{00}}{N}$, where

$$N_{11} = \sum_{i \in N} P(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c | \mathbf{x}),$$

$$N_{01} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} \geq \theta_c | \mathbf{x}),$$

$$N_{10} = \sum_{i \in N} P(\theta_i \geq \theta_c, \theta_{i,2} < \theta_c | \mathbf{x}),$$

$$N_{00} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} < \theta_c | \mathbf{x}), \text{ and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

Appendix F.1 – Number of Participating Students by Demographic Subgroups – ELA Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	72754	73195	72289	69837	69754	69481	62955	58180	54018
Female	35635	35891	35418	34404	33826	34069	30924	28862	27045
Male	37119	37304	36871	35433	35928	35412	32031	29318	26973
African American	3834	3797	3819	3733	3739	3808	3411	2949	2722
Asian	1739	1665	1575	1630	1732	1736	1590	1591	1502
Native Hawaiian/Pacific	262	276	289	230	229	207	194	226	162
Hispanic/Latino	34078	33730	32937	31213	30787	30319	26596	24567	22594
American Indian or Alaskan	3565	3581	3602	3440	3546	3334	3560	3135	2819
White	26728	27666	27781	27612	27917	28375	26143	24357	23060
Multiple	2407	2371	2150	1909	1709	1585	1383	1284	1088
Special Education	8341	8913	8740	8219	7558	7166	5838	5272	4640
Limited English Proficiency	7790	7515	5980	4291	4032	3325	1952	1179	724
Free/Reduced Lunch	28652	28599	27748	26654	26520	26117	19180	17625	15352
Accommodation*	51249	51116	50056	47981	47508	48038	41694	38208	35634

Note: In 2017, Text-to-speech for instruction was counted as an accommodation but was the default option available online for all students. Starting in Spring 2018, this option is to be counted as a universal tool.

Appendix F.2 – Number of Participating Students by Demographic Subgroups – ELA Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	15602	15489	15241	16279	15495	15033	17800	15978	14511
Female	7755	7645	7573	8160	7622	7529	8892	8067	7372
Male	7847	7844	7668	8119	7873	7504	8908	7911	7139
African American	778	808	779	882	803	817	1064	933	870
Asian	736	823	860	891	843	711	853	781	667
Native Hawaiian/Pacific	54	43	51	55	56	62	64	49	32
Hispanic/Latino	6819	6656	6517	6983	6934	6841	9093	8003	7004
American Indian or Alaskan	744	757	682	905	651	563	421	335	321
White	6027	5926	5876	6138	5807	5699	6003	5622	5269
Multiple	439	462	460	392	373	332	288	241	266
Special Education	1621	1691	1748	1762	1559	1486	1497	1324	1140
Limited English Proficiency	1477	1402	1034	888	951	720	712	259	117
Free/Reduced Lunch	7005	6783	6449	6925	5976	5685	7044	5985	5304
Accommodation	74	79	79	57	90	82	97	75	57

Appendix F.3 – Number of Participating Students by Demographic Subgroups – Mathematics Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
All students	72859	73441	72429	70035	68367	59176	66697	56138	50066
Female	35639	35988	35457	34479	33227	28771	32557	27807	25476
Male	37220	37453	36972	35556	35140	30405	34140	28331	24590
African American	3844	3822	3826	3761	3724	3545	3471	2967	2369
Asian	1741	1676	1576	1628	1539	1149	1628	1549	1398
Native Hawaiian/Pacific	262	274	288	230	226	172	203	207	173
Hispanic/Latino	34100	33837	33012	31304	30546	27682	28806	23909	20785
American Indian or Alaskan	3585	3591	3608	3458	3519	3120	3718	3242	2591
White	26765	27748	27824	27661	27037	22047	27284	22996	21665
Multiple	2409	2376	2154	1915	1675	1308	1457	1162	1037
Special Education	8443	8978	8795	8275	7566	7036	6341	4913	3114
Limited English Proficiency	7883	7600	6024	4370	4079	3319	1962	1271	691
Free/Reduced Lunch	28646	28684	27794	26768	26319	24299	21091	17507	14187
Accommodation*	51370	51297	50177	48136	46614	40774	45429	36658	32951

Note: In 2017, Text-to-speech for instruction was counted as an accommodation but was the default option available online for all students. Starting in Spring 2018, this option is to be counted as a universal tool.

Appendix F.4 – Number of Participating Students by Demographic Subgroups – Mathematics Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geomet ry	Algebra II
All students	15625	15524	15211	16303	14936	12711	16266	15460	13598
Female	7740	7652	7553	8172	7356	6304	8117	7877	6991
Male	7885	7872	7658	8131	7580	6407	8149	7583	6607
African American	775	814	778	884	791	760	965	928	763
Asian	738	824	858	888	794	439	453	591	717
Native Hawaiian/Pacific	53	43	51	55	52	55	64	44	34
Hispanic/Latino	6812	6666	6518	6999	6796	6263	8777	8109	6647
American Indian or Alaskan	747	759	683	912	639	553	416	354	248
White	6056	5942	5851	6137	5475	4361	5307	5212	4958
Multiple	439	462	456	395	362	273	270	214	212
Special Education	1634	1710	1761	1780	1552	1460	1520	1160	878
Limited English Proficiency	1486	1426	1036	892	934	716	802	547	222
Free/Reduced Lunch	6977	6783	6428	6938	5826	5194	7132	6378	5080
Accommodation	77	86	83	58	95	80	109	63	43

Appendix G.1—Operational Item Parameter Used to Score Spring 17 Tests — Grade 3 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13023_e	ER	-1.15476	2.25355	4.12678	1.74186
2	13023_o	ER	-1.34689	2.14582	3.94719	1.58204
3	13023_c	ER	-1.69753	-0.82307		-1.26030
4	13026_e	ER	-1.11168	1.96713	4.20208	1.68584
5	13026_o	ER	-1.02194	1.83366	4.15705	1.65626
6	13026_c	ER	-1.53632	-0.61273		-1.07453
7	10628	MC4	-1.10023			-1.10023
8	10630	MC4	-0.90745			-0.90745
9	9418	MC4	0.70984			0.70984
10	9419	HT	-0.05709			-0.05709
11	9422	MC4	-0.87555			-0.87555
12	10632	MC4	0.48179			0.48179
13	10634	MC4	1.17725			1.17725
14	9687	HT	1.38239			1.38239
15	9690	MC4	-1.20900			-1.20900
16	9691	MC4	-0.37757			-0.37757
17	9692	MC4	-0.60073			-0.60073
18	9694	EBSR4	1.70532			1.70532
19	9697	HT	-0.26674			-0.26674
20	9698	MC4	-1.11745			-1.11745
21	9699	MC4	-0.30132			-0.30132
22	9700	MC4	1.23623			1.23623
23	11854	EBSR4	0.89082			0.89082
24	11867	MC4	-0.53376			-0.53376
25	12417	MS5	0.83368			0.83368
26	12521	MC4	0.16092			0.16092
27	12524	MC4	-1.10417			-1.10417
28	12979	ETC	0.50861			0.50861
29	12980	ETC	-1.38292	0.45809		-0.46242
30	12981	ETC	-2.23435			-2.23435
31	12983	MC4	-0.27341			-0.27341
32	12984	EBSR4	0.33388			0.33388
33	12985	MC4	-0.13365			-0.13365
34	12986	MC4	-0.03661			-0.03661
35	12987	EBSR4	1.34313			1.34313
36	12998	MC4	0.11624			0.11624
37	12133	HT	1.27299			1.27299
38	12379	EBSR4	1.68558			1.68558
39	12383	MC4	0.88868			0.88868
40	12760	MC4	0.65442			0.65442

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12758	MC4	-0.31060			-0.31060
42	12387	MC4	-0.00246			-0.00246
43	12757	EBSR4	1.66077			1.66077
44	12776	MC4	0.45152			0.45152
45	9377	ETC	0.18504			0.18504
46	9379	ETC	-1.03119			-1.03119
47	9380	ETC	-0.87865			-0.87865

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.2—Operational Item Parameter Used to Score Spring 17 Tests — Grade 4 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13095_e	ER	1.87872	4.8299	4.92767	3.87876
2	13095_o	ER	0.47386	4.35392	4.92003	3.24927
3	13095_c	ER	-1.45942	1.23827		-0.11058
4	13094_e	ER	1.87077	4.87517	6.53587	4.42727
5	13094_o	ER	1.25092	4.69117	6.1041	4.01540
6	13094_c	ER	-0.79626	1.71384		0.45879
7	11837	MC4	-0.34149			-0.34149
8	12567	MC4	-0.08581			-0.08581
9	11840	MC4	0.09196			0.09196
10	11844	EBSR4	1.01626	0.92279		0.96953
11	11841	MC4	0.87491			0.87491
12	11846	MC4	0.86215			0.86215
13	11838	MC4	1.09150			1.09150
14	11847	MC4	0.29982			0.29982
15	11967	HT	0.32005			0.32005
16	13072	HT	0.22957			0.22957
17	13073	HT	-0.20656			-0.20656
18	13042	MC4	-1.39269			-1.39269
19	13043	MC4	-0.80847			-0.80847
20	13044	MC4	-0.28400			-0.28400
21	13046	MC4	0.01302			0.01302
22	13070	MC4	0.52289			0.52289
23	13071	MC4	-0.31909			-0.31909
24	12317	MS5	0.80561			0.80561
25	12666	MC4	0.42527			0.42527
26	12647	MC4	0.22325			0.22325
27	12653	MC4	-0.53559			-0.53559
28	9428	ETC	0.79794			0.79794
29	9429	ETC	-2.21688	0.57637		-0.82026
30	9431	ETC	0.47241			0.47241
31	13040	MC4	-0.17642			-0.17642
32	13039	HT	-1.13238			-1.13238
33	13034	MC4	0.17385			0.17385
34	13037	MC4	0.22802			0.22802
35	13036	MC4	0.07255			0.07255
36	13038	MC4	0.59887			0.59887
37	13035	MC4	-0.14645			-0.14645
38	9446	MC4	-0.77651			-0.77651
39	9439	MS6	2.08491			2.08491
40	9437	MC4	-0.03208			-0.03208

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	9435	MC4	-0.45929			-0.45929
42	9451	HT	1.17836			1.17836
43	9450	MC4	-0.16915			-0.16915
44	9438	MS6	1.75326			1.75326
45	13031	ETC	-0.13928			-0.13928
46	13032	ETC	-0.51432	1.51483		0.50026
47	13033	ETC	0.62162	2.61293		1.61728

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.3—Operational Item Parameter Used to Score Spring 17 Tests — Grade 5 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13239_e	ER	0.02633	3.71815	4.71382	2.81943
2	13239_o	ER	-1.10817	3.18175	4.64235	2.23864
3	13239_c	ER	-2.15823	0.01608		-1.07108
4	13236_e	ER	-0.76876	3.70295	4.83054	2.58824
5	13236_o	ER	-0.83522	2.91473	5.00997	2.36316
6	13236_c	ER	-2.13985	-0.09297		-1.11641
7	12069	MC4	-0.68003			-0.68003
8	12068	MC4	-0.40849			-0.40849
9	12072	MC4	-0.63304			-0.63304
10	12066	GI	2.38913			2.38913
11	12067	MS5	-0.09943			-0.09943
12	12064	MC4	0.25102			0.25102
13	12065	EBSR4	0.70850			0.70850
14	9783	MC4	-0.86303			-0.86303
15	9784	MC4	0.73684			0.73684
16	9842	HT	0.63012			0.63012
17	9808	MC4	0.19637			0.19637
18	10273	HT	0.41463			0.41463
19	9782	MC4	0.54824			0.54824
20	12894	MC4	0.42103			0.42103
21	12687	MC4	-0.28139			-0.28139
22	12690	MC4	-0.12644			-0.12644
23	12706	HT	-0.46914			-0.46914
24	12671	MC4	0.00268			0.00268
25	12865	MC4	-0.57988			-0.57988
26	12663	MC4	1.09941			1.09941
27	12649	MC4	0.45115			0.45115
28	9286	ETC	-0.23947			-0.23947
29	9287	ETC	1.18837			1.18837
30	9288	ETC	-1.29281	0.66758		-0.31262
31	11799	MC4	-1.15805			-1.15805
32	11786	MC4	0.67421			0.67421
33	11802	MC4	-0.73329			-0.73329
34	11779	MC4	0.2593			0.25930
35	12439	HT	0.35913			0.35913
36	9303	MC4	0.14403			0.14403
37	9305	MC4	-1.06814			-1.06814
38	9301	MC4	0.26042			0.26042
39	9304	MS6	0.52670			0.52670
40	9302	MC4	0.39700			0.39700

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12425	MC4	1.08892			1.08892
42	12440	NL	0.46381			0.46381
43	12851	MC4	1.11609			1.11609
44	12852	MC4	-0.10604			-0.10604
45	13124	ETC	-0.37268			-0.37268
46	13129	ETC	-0.89686			-0.89686
47	13131	ETC	-1.23601			-1.23601

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.4—Operational Item Parameter Used to Score Spring 17 Tests — Grade 6 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13308_e	ER	1.23630	3.33177	5.17142	3.24650
2	13308_o	ER	0.15990	3.08300	4.37256	2.53849
3	13308_c	ER	-2.10091	-0.21027		-1.15559
4	13304_e	ER	-0.40648	2.83619	5.19328	2.540997
5	13304_o	ER	-0.76020	2.34515	4.23866	1.94120
6	13304_c	ER	-2.38004	-0.22638		-1.30321
7	9131	MC4	0.87684			0.87684
8	9134	MC4	-0.24354			-0.24354
9	9135	MS6	0.52718			0.52718
10	9137	MC4	-1.63221			-1.63221
11	9138	MC4	-1.48057			-1.48057
12	9153	MC4	-0.41536			-0.41536
13	9154	MS6	2.33626			2.33626
14	9168	EBSR4	0.35054			0.35054
15	9169	EBSR4	1.42720			1.42720
16	13259	MC4	-0.13029			-0.13029
17	13261	MC4	-0.02637			-0.02637
18	13263	MS5	1.61746			1.61746
19	13264	MC4	-1.05400			-1.05400
20	13270	MC4	-0.98317			-0.98317
21	13271	MC4	-0.32339			-0.32339
22	13272	MC4	-0.38144			-0.38144
23	13274	MC4	-0.56873			-0.56873
24	13287	EBSR4	0.02320			0.02320
25	11875	MC4	-0.04869			-0.04869
26	11876	MC4	-0.28043			-0.28043
27	11877	MS5	2.25745			2.25745
28	11879	MC4	0.50193			0.50193
29	11880	EBSR4	1.08094			1.08094
30	9724	ETC	-1.78427	0.19966		-0.792305
31	9726	ETC	-2.31186			-2.31186
32	12407	MC4	-0.06404			-0.06404
33	12408	EBSR4	2.43169			2.43169
34	12409	MC4	0.36443			0.36443
35	12411	MC4	0.62903			0.62903
36	12415	MC4	0.24204			0.24204
37	12868	MC4	0.31100			0.31100
38	12895	EBSR4	0.79934	0.88536		0.84235

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	13252	HT	-1.31937			-1.31937
40	13253	MC4	-0.13808			-0.13808
41	13255	MC4	-0.77062			-0.77062
42	13267	MC4	0.83368			0.83368
43	13268	MC4	-0.21399			-0.21399
44	13269	EBSR4	2.11059			2.11059
45	9107	ETC	-1.86994			-1.86994
46	9108	ETC	-1.32061	1.24437		-0.03812
47	9109	ETC	-1.49476	0.62001		-0.437375

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.5—Operational Item Parameter Used to Score Spring 17 Tests — Grade 7 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13403_e	ER	0.48951	3.69526	5.16521	3.11666
2	13403_o	ER	-0.54572	3.33949	5.13751	2.64376
3	13403_c	ER	-2.82848	-0.61627		-1.72238
4	13345	MC4	-0.18100			-0.18100
5	9611	HT	0.44176			0.44176
6	9787	EBSR4	0.58622			0.58622
7	13402_e	ER	-1.37884	3.89633	4.60481	2.37410
8	13402_o	ER	-1.49334	3.23935	5.1219	2.28930
9	13402_c	ER	-2.56472	-0.19658		-1.38065
10	13347	HT	-0.55621			-0.55621
11	9711	MC4	0.29644			0.29644
12	9810	MS5	-0.88849			-0.88849
13	13349	MC4	-1.14877			-1.14877
14	9713	MC4	0.2301			0.23010
15	9790	MC4	0.43432			0.43432
16	13343	MS6	0.28565			0.28565
17	9614	MC4	-1.02121			-1.02121
18	10275	MS5	1.59355			1.59355
19	13344	MC4	0.20465			0.20465
20	10613	MS5	1.17647			1.17647
21	9791	MC4	0.26709			0.26709
22	13342	MS6	0.48321			0.48321
23	10695	MC4	-0.51732			-0.51732
24	9789	MC4	-0.24857			-0.24857
25	13354	HT	1.78903			1.78903
26	9709	MC4	-0.24287			-0.24287
27	9103	ETC	0.33393			0.33393
28	13357	MC4	-0.38788			-0.38788
29	9750	MS6	-0.17847			-0.17847
30	9105	ETC	-0.14173			-0.14173
31	13356	HT	0.07253			0.07253
32	9106	ETC	0.42308			0.42308
33	13359	MC4	-0.60650			-0.60650
34	13358	MC4	0.23406			0.23406
35	13352	MS5	-0.15233			-0.15233
36	9740	MC4	-1.52845			-1.52845
37	9743	MC4	0.07478			0.07478
38	9846	MS5	1.24305			1.24305
39	9742	MC4	-0.99456			-0.99456
40	9845	MC4	1.05787			1.05787

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12553	MC4	0.68430			0.68430
42	12555	MC4	0.81559			0.81559
43	12552	MC4	0.06825			0.06825
44	12890	HT	-0.45853			-0.45853
45	9090	ETC	0.91696			0.91696
46	9091	ETC	-1.46750	-0.20095		-0.83423
47	9092	ETC	0.13818			0.13818

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.6—Operational Item Parameter Used to Score Spring 17 Tests — Grade 8 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13452_e	ER	-1.67899	1.57768	3.33549	1.07806
2	13452_o	ER	-1.89107	1.08197	3.42027	0.87039
3	13452_c	ER	-2.54121	-1.11311		-1.82716
4	13437_e	ER	-1.09959	1.99003	3.43357	1.44134
5	13437_o	ER	-1.21497	1.49481	3.52606	1.26863
6	13437_c	ER	-2.19400	-0.55737		-1.37569
7	13443	MC4	-0.27924			-0.27924
8	13442	HT	1.15462			1.15462
9	13440	MC4	-1.15597			-1.15597
10	13441	MS6	0.89435			0.89435
11	11861	EBSR4	0.69541	-0.50146		0.09698
12	11866	MC4	1.17167			1.17167
13	11906	MC4	-0.52321			-0.52321
14	11807	EBSR4	1.14334			1.14334
15	11869	MC4	-0.31254			-0.31254
16	12664	HT	-1.92911			-1.92911
17	12670	HT	1.15900			1.15900
18	12696	MC4	-0.83540			-0.83540
19	12702	EBSR4	1.64156			1.64156
20	12703	EBSR4	2.34525			2.34525
21	9029	MC4	-1.56940			-1.56940
22	9025	MC4	-0.75521			-0.75521
23	9026	MC4	0.50530			0.50530
24	10627	HT	0.47781			0.47781
25	9028	MC4	0.17704			0.17704
26	9020	MS5	1.09383			1.09383
27	9076	ETC	-0.64308			-0.64308
28	9077	ETC	-1.79630	0.32962		-0.73334
29	11815	EBSR4	0.62015			0.62015
30	11819	EBSR4	0.40419	1.84388		1.12404
31	11820	HT	-0.76310			-0.76310
32	11816	MC4	-0.07166			-0.07166
33	11811	MC4	-0.25816			-0.25816
34	11813	MC4	0.19219			0.19219
35	12429	EBSR4	-0.57203	2.62188		1.02493
36	11812	MC4	-0.25561			-0.25561
37	12427	EBSR4	1.93677			1.93677
38	11810	EBSR5	0.87267			0.87267
39	13413	EBSR4	0.35518			0.35518
40	13420	MC4	0.39856			0.39856

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	13412	HT	0.91911			0.91911
42	13416	MC4	-1.70678			-1.70678
43	13414	MS5	2.12462			2.12462
44	13419	MC4	0.60522			0.60522
45	9079	ETC	-0.73984			-0.73984
46	9080	ETC	-1.65541	-0.12677		-0.89109
47	9082	ETC	-1.43968			-1.43968

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.7—Operational Item Parameter Used to Score Spring 17 Tests — Grade 9 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13566_e	ER	-1.80017	2.23734	4.14528	1.52748
2	13566_o	ER	-1.66128	1.36788	4.09913	1.26858
3	13566_c	ER	-2.39075	-0.81852		-1.60464
4	13557_e	ER	-1.48371	2.29755	4.26558	1.69314
5	13557_o	ER	-1.33444	1.51754	4.19585	1.45965
6	13557_c	ER	-2.43540	-0.86620		-1.65080
7	13541	MC4	-0.37461			-0.37461
8	13545	MC4	-0.43877			-0.43877
9	13547	MC4	-0.39906			-0.39906
10	13543	HT	0.49797			0.49797
11	13539	MC4	-0.01696			-0.01696
12	13546	MC4	0.42348			0.42348
13	13515	MC4	-0.23416			-0.23416
14	13553	MC4	0.25005			0.25005
15	13550	EBSR4	0.08872			0.08872
16	13516	MC4	-0.71545			-0.71545
17	13518	MS5	0.23285			0.23285
18	13534	MC4	-0.67176			-0.67176
19	12118	EBSR4	0.86211			0.86211
20	12119	MC4	-0.46625			-0.46625
21	12561	MS5	0.23208			0.23208
22	12544	EBSR4	1.12006	-0.34193		0.38907
23	12191	MC4	-0.26068			-0.26068
24	12723	MC4	0.38373			0.38373
25	12190	MC4	0.02018			0.02018
26	12192	MC4	0.11611			0.11611
27	12193	HT	-1.49923			-1.49923
28	9734	ETC	-0.38567			-0.38567
29	9735	ETC	-1.01417	1.57881		0.28232
30	9736	ETC	-0.28792			-0.28792
31	12633	MC4	-0.31337			-0.31337
32	12632	MC4	0.76878			0.76878
33	12624	EBSR4	0.00841			0.00841
34	12631	MS5	0.49406			0.49406
35	12654	MC4	0.17400			0.17400
36	12629	MC4	0.29534			0.29534
37	12628	MC4	0.49349			0.49349
38	12626	EBSR4	1.07738	0.66972		0.87355
39	12621	MC4	-0.12111			-0.12111
40	11097	MC4	-0.45154			-0.45154

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	9031	MC4	-0.81045			-0.81045
42	9037	MC4	0.84443			0.84443
43	11098	MC4	0.41520			0.41520
44	9032	MC4	0.81518			0.81518
45	9038	MC4	0.55427			0.55427
46	9033	MC4	-0.15469			-0.15469
47	13455	ETC	0.34843			0.34843
48	13456	ETC	0.10449	1.39644		0.75047
49	13457	ETC	-0.84937	0.51270		-0.16834

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.8—Operational Item Parameter Used to Score Spring 17 Tests — Grade 10 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13640_e	ER	-1.90091	1.21808	2.96704	0.76140
2	13640_o	ER	-2.36052	0.51171	3.10403	0.41841
3	13640_c	ER	-3.03408	-1.28609		-2.16009
4	13639_e	ER	-1.16138	1.59043	3.81864	1.41590
5	13639_o	ER	-1.47773	1.06526	3.81875	1.13543
6	13639_c	ER	-2.96755	-1.16419		-2.06587
7	12507	MC4	-1.60151			-1.60151
8	12908	MC4	0.13614			0.13614
9	12530	MC4	0.65504			0.65504
10	12534	EBSR4	0.96610			0.96610
11	12504	HT	0.71142			0.71142
12	8812	MC4	-0.48167			-0.48167
13	8813	HT	-0.22198			-0.22198
14	8810	MC4	-0.80870			-0.80870
15	10155	HT	0.27894			0.27894
16	8852	MC4	0.40325			0.40325
17	8811	MC4	0.14036			0.14036
18	12003	MS5	0.67293			0.67293
19	12506	MC4	-0.14920			-0.14920
20	12443	MC4	-0.55645			-0.55645
21	9822	MC4	-0.54876			-0.54876
22	9813	HT	0.53987			0.53987
23	9824	MS5	0.82801			0.82801
24	9816	MC4	0.49465			0.49465
25	9814	MC4	-0.32950			-0.32950
26	9826	MC4	-0.43866			-0.43866
27	9821	MC4	0.03241			0.03241
28	9888	MC4	-0.02807			-0.02807
29	9825	MC4	-0.52709			-0.52709
30	8757	ETC	-1.36906			-1.36906
31	8758	ETC	0.32886			0.32886
32	8763	ETC	-1.27617			-1.27617
33	13586	HT	0.82135			0.82135
34	13590	HT	1.55741			1.55741
35	13588	MC4	-0.60396			-0.60396
36	13592	MC4	-2.04871			-2.04871
37	13594	MC4	-0.45131			-0.45131
38	13593	MC4	-0.44093			-0.44093
39	12446	MC4	0.04067			0.04067
40	12863	MS5	1.77151			1.77151

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12478	MC4	-0.17512			-0.17512
42	12482	MC4	0.95917			0.95917
43	12449	EBSR4	2.11892			2.11892
44	12477	EBSR4	0.69331			0.69331
45	12480	EBSR5	1.32316	2.83702		2.08009
46	12473	MC4	-0.06591			-0.06591
47	13571	ETC	0.79411			0.79411
48	13572	ETC	-0.70534			-0.70534
49	13573	ETC	-1.02335	0.82160		-0.10088

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.9—Operational Item Parameter Used to Score Spring 17 Tests — Grade 11 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13722_e	ER	-1.92023	0.91135	3.08764	0.69292
2	13722_o	ER	-2.77119	0.40624	2.72972	0.12159
3	13722_c	ER	-2.88156	-1.06637		-1.97397
4	13724_e	ER	-1.89212	0.54532	3.32819	0.66046
5	13724_o	ER	-2.63027	0.50600	3.14192	0.33922
6	13724_c	ER	-2.90567	-0.96473		-1.93520
7	9855	MC4	0.21985			0.21985
8	9853	HT	-0.75130			-0.75130
9	9851	HT	-0.72733			-0.72733
10	9858	MC4	0.99024			0.99024
11	9852	MS6	-0.84321			-0.84321
12	11926	MC4	-1.57944			-1.57944
13	11919	MC4	-1.07482			-1.07482
14	11917	MS5	0.53495			0.53495
15	11924	MC4	-0.95380			-0.95380
16	11931	MC4	-0.22058			-0.22058
17	11932	MC4	-1.45851			-1.45851
18	11922	EBSR4	1.08624			1.08624
19	11929	MC4	-0.2052			-0.20520
20	8860	HT	1.67605			1.67605
21	8864	MS5	1.95843			1.95843
22	8861	MC4	0.25879			0.25879
23	8869	MC4	-0.15979			-0.15979
24	8867	MC4	0.18188			0.18188
25	8871	MC4	0.03863			0.03863
26	12838	NL	1.25908			1.25908
27	12829	MC4	-1.00932			-1.00932
28	12837	MC4	0.28462			0.28462
29	12822	MC4	-0.14240			-0.14240
30	8778	ETC	0.49041			0.49041
31	8779	ETC	-1.89965	-0.47048		-1.18507
32	8780	ETC	-0.32980			-0.32980
33	8794	HT	0.60324			0.60324
34	8783	MC4	-0.68116			-0.68116
35	8792	MC4	-0.21395			-0.21395
36	8784	MC4	-0.25728			-0.25728
37	8781	MC4	0.34351			0.34351
38	8791	MS5	1.31404			1.31404
39	12821	MC4	0.08539			0.08539
40	12814	EBSR4	0.79201			0.79201

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12823	EBSR4	1.67864			1.67864
42	12825	MC4	-0.41243			-0.41243
43	12844	MC4	0.44267			0.44267
44	12842	MC4	0.19779			0.19779
45	12845	EBSR4	1.81600			1.81600
46	12839	NL	1.64630			1.64630
47	13644	ETC	-1.40824			-1.40824
48	13646	ETC	-1.88481			-1.88481
49	13647	ETC	-1.47527			-1.47527

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.10—Operational Item Parameter Used to Score Spring 17 Tests — Grade 3 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10460	MC4	-0.44542			-0.44542
2	12085	EQ	-0.62452			-0.62452
3	10430	EQ	2.32239			2.32239
4	9464	EQ	1.31740			1.31740
5	10408	EQ	-0.62542			-0.62542
6	9455	EQ	-0.07043			-0.07043
7	12569	MC4	0.72111			0.72111
8	10469	MC4	-1.55904			-1.55904
9	13749	MC4	-1.80241			-1.80241
10	10398	EQ	0.44980			0.44980
11	10448	EQ	1.09544			1.09544
12	10409	MC4	-1.87291			-1.87291
13	10395	EQ	-2.04152			-2.04152
14	10462	MC4	0.86520			0.86520
15	12902	EQ	-0.70461			-0.70461
16	10687	MC4	0.89578			0.89578
17	13747	EQ	-1.62651			-1.62651
18	12281	EQ	-1.24975			-1.24975
19	13751	MC4	0.05164			0.05164
20	10389	EQ	-0.05749			-0.05749
21	10683	MC4	1.42686			1.42686
22	12054	EQ	0.46601			0.46601
23	10421	EQ	-1.24741			-1.24741
24	10404	MC4	-2.42513			-2.42513
25	13767	EQ	0.78005			0.78005
26	12941	EQ	-0.57950			-0.57950
27	12296	EQ	-0.33397			-0.33397
28	13765	EQ	-0.94910			-0.94910
29	11966	EQ	0.76107			0.76107
30	12421	EQ	1.09421			1.09421
31	10679	EQ	-0.93916			-0.93916
32	10677	MS5	2.19751			2.19751
33	10434	EQ	-0.75061			-0.75061
34	11588	EQ	3.18759			3.18759
35	10396	EQ	0.38434			0.38434
36	13748	MC4	-2.44960			-2.44960
37	10685	MC4	1.69992			1.69992
38	10415	EQ	1.81916			1.81916
39	10438	EQ	-2.34273			-2.34273
40	10427	EQ	1.06820			1.06820

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10399	EQ	0.88755			0.88755
42	10477	MC4	1.51808			1.51808
43	13773	EQ	1.11913			1.11913
44	13740	EQ	-0.97445			-0.97445
45	10671	EQ	0.39429			0.39429

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.11—Operational Item Parameter Used to Score Spring 17 Tests — Grade 4 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13733	MC4	-1.76471			-1.76471
2	13738	EQ	-1.68096			-1.68096
3	10728	EQ	1.29344			1.29344
4	11608	EQ	0.79089			0.79089
5	10710	MS5	-0.44114			-0.44114
6	10709	MS5	0.78446			0.78446
7	13776	MS6	1.99349			1.99349
8	12276	EQ	-0.57774			-0.57774
9	10725	MS5	0.87782			0.87782
10	13782	MI	-1.05896			-1.05896
11	10772	MC4	-2.60398			-2.60398
12	10766	MS5	-1.46597			-1.46597
13	11675	EQ	-0.51088			-0.51088
14	13762	MI	-0.70678			-0.70678
15	12266	EQ	-1.49821			-1.49821
16	13760	EQ	-1.49980			-1.49980
17	10779	EQ	0.34347			0.34347
18	13780	MC4	-0.17136			-0.17136
19	13769	EQ	-1.25463			-1.25463
20	10705	MC4	0.15075			0.15075
21	10748	EQ	0.94535			0.94535
22	10781	MC4	0.68116			0.68116
23	13320	EQ	0.01192			0.01192
24	10713	MI	3.38519			3.38519
25	10774	EQ	1.86572			1.86572
26	13779	EQ	-0.50284			-0.50284
27	11345	EQ	2.53955			2.53955
28	10784	EQ	0.32070			0.32070
29	10731	EQ	-0.77540			-0.77540
30	9474	EQ	-0.17779			-0.17779
31	10741	MC4	0.02987			0.02987
32	12263	EQ	-2.19437			-2.19437
33	10768	MI	2.47137			2.47137
34	13772	EQ	-1.24222			-1.24222
35	10744	EQ	-0.50014			-0.50014
36	10756	EQ	1.25770			1.25770
37	11105	EQ	1.97493			1.97493
38	10735	EQ	-0.72229			-0.72229
39	13777	EQ	-1.17708			-1.17708
40	9482	GI	-1.86072	1.48656		-0.18708

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10720	MS5	1.28227			1.28227
42	10758	MS5	-0.12632			-0.12632
43	10769	MC4	0.41426			0.41426
44	13757	MS5	0.86280			0.86280
45	11331	EQ	0.99076			0.99076

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.12—Operational Item Parameter Used to Score Spring 17 Tests — Grade 5 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13067	MC4	-0.73318			-0.73318
2	13088	EQ	0.97239			0.97239
3	13089	EQ	-0.10631			-0.10631
4	13064	MI	-0.17923			-0.17923
5	10833	EQ	-0.99926			-0.99926
6	13062	EQ	0.66283			0.66283
7	10814	EQ	1.37493			1.37493
8	13047	EQ	-1.38963			-1.38963
9	12088	EQ	-0.40340			-0.40340
10	10861	EQ	1.36948			1.36948
11	10804	MC4	0.26514			0.26514
12	13329	EQ	0.31116			0.31116
13	10793	EQ	-0.99105			-0.99105
14	10844	MS5	-0.98953			-0.98953
15	10792	MS5	0.03048			0.03048
16	10863	MC4	0.62779			0.62779
17	11526	EQ	-0.17516			-0.17516
18	12223	EQ	-0.05863			-0.05863
19	11106	MS5	0.50872			0.50872
20	13324	EQ	-2.55703			-2.55703
21	10849	EQ	1.15146			1.15146
22	12221	MS5	-0.39310			-0.39310
23	10795	EQ	-1.39678			-1.39678
24	10858	EQ	0.63797			0.63797
25	10805	EQ	-0.48618			-0.48618
26	12090	EQ	0.56750			0.56750
27	13059	EQ	-0.54000			-0.54000
28	10820	MC4	1.62329			1.62329
29	13326	EQ	-0.67811			-0.67811
30	11893	EQ	1.54747			1.54747
31	10803	EQ	-0.21830			-0.21830
32	13068	EQ	-0.70605			-0.70605
33	10850	EQ	0.19694			0.19694
34	11764	EQ	1.33182			1.33182
35	10823	MS5	1.42952			1.42952
36	10796	EQ	0.28237			0.28237
37	10869	MS6	0.21386			0.21386
38	13055	EQ	-0.70247			-0.70247
39	10798	MS5	-0.17942			-0.17942
40	11710	EQ	1.23326			1.23326

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	13045	EQ	-0.26512			-0.26512
42	10811	MC4	-1.11389			-1.11389
43	13065	EQ	0.43532			0.43532
44	12198	EQ	0.15977			0.15977
45	10875	MC4	-0.69325			-0.69325

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.13—Operational Item Parameter Used to Score Spring 17 Tests — Grade 6 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	11903	EQ	0.97964			0.97964
2	10048	MC4	0.40685			0.40685
3	10093	MC4	0.03634			0.03634
4	10113	MC4	-1.43848			-1.43848
5	10082	MC4	-1.01930			-1.01930
6	10143	EQ	0.87374			0.87374
7	10095	TI	0.97298			0.97298
8	13110	EQ	-1.24624			-1.24624
9	12082	EQ	1.76972			1.76972
10	13331	EQ	-2.78919			-2.78919
11	9718	MI	2.34031			2.34031
12	10062	EQ	0.87598			0.87598
13	9492	EQ	-1.41594			-1.41594
14	11643	EQ	-0.44740			-0.44740
15	12055	EQ	1.00917			1.00917
16	10136	MC4	0.10623			0.10623
17	10047	EQ	1.59727			1.59727
18	10107	EQ	-1.82823			-1.82823
19	11774	EQ	2.22351			2.22351
20	10050	EQ	0.46610			0.46610
21	11375	EQ	0.73262			0.73262
22	11728	EQ	-1.56689			-1.56689
23	10123	EQ	2.76628			2.76628
24	10108	MC4	-0.66640			-0.66640
25	13795	MS5	0.77209			0.77209
26	10078	EQ	1.05127			1.05127
27	10064	EQ	-0.12482			-0.12482
28	13330	EQ	-1.59289			-1.59289
29	10149	GI	1.17888			1.17888
30	12345	EQ	1.08427			1.08427
31	11531	EQ	0.92185			0.92185
32	10142	MC4	-0.22168			-0.22168
33	10115	EQ	-0.21997			-0.21997
34	11904	EQ	-0.46858			-0.46858
35	10100	MC4	-2.37860			-2.37860
36	10144	EQ	-0.75359			-0.75359
37	11525	EQ	1.96060			1.96060
38	10139	MS6	2.24036			2.24036
39	13117	MC4	-0.38515			-0.38515
40	10054	MC4	-3.09329			-3.09329

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	13111	EQ	-1.25569			-1.25569
42	12050	MC4	0.07312			0.07312
43	11376	EQ	0.75812			0.75812
44	10127	MC4	-0.77211			-0.77211
45	10122	EQ	2.05238			2.05238
46	10087	EQ	-2.15564			-2.15564
47	10148	MC4	-0.41540			-0.41540

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.14—Operational Item Parameter Used to Score Spring 17 Tests — Grade 7 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10331	EQ	0.38423			0.38423
2	10289	EQ	1.13129			1.13129
3	10370	MC4	-0.35904			-0.35904
4	9504	EQ	0.16514			0.16514
5	12848	MS5	2.41103			2.41103
6	10349	MC4	-0.06195			-0.06195
7	10376	EQ	3.33419			3.33419
8	12218	EQ	1.20992			1.20992
9	10303	MC4	-1.12478			-1.12478
10	9508	EQ	1.88124			1.88124
11	13128	EQ	0.21965			0.21965
12	13334	EQ	2.91343			2.91343
13	12422	EQ	1.16001			1.16001
14	10324	EQ	1.78555			1.78555
15	10347	EQ	1.81767			1.81767
16	8698	MC4	-2.05133			-2.05133
17	10315	EQ	0.98783			0.98783
18	10701	EQ	1.03009			1.03009
19	11300	EQ	2.30616			2.30616
20	13144	MS5	1.54728			1.54728
21	12915	MI	0.41083			0.41083
22	12929	EQ	0.45298			0.45298
23	13136	EQ	2.59929			2.59929
24	10350	EQ	1.52734			1.52734
25	11972	EQ	1.86739			1.86739
26	11719	GI	0.78120			0.78120
27	10322	EQ	-0.24088			-0.24088
28	9529	GI	2.12926			2.12926
29	10355	EQ	2.32291			2.32291
30	13803	MC4	1.08843			1.08843
31	10339	EQ	2.28753			2.28753
32	13805	EQ	1.00257			1.00257
33	11332	EQ	2.28044			2.28044
34	13146	EQ	1.19622			1.19622
35	10344	MS5	1.05834			1.05834
36	12201	EQ	1.79194			1.79194
37	11348	EQ	2.42410			2.42410
38	10377	EQ	2.76641			2.76641
39	10308	MI	1.49225			1.49225
40	10310	EQ	2.12932			2.12932

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	11586	EBSR5	2.61287			2.61287
42	10309	MI	0.97888			0.97888
43	12423	EQ	0.68394			0.68394
44	9720	MI	0.53671			0.53671
45	13798	EQ	1.42370			1.42370
46	11580	MC4	0.23200			0.23200
47	9703	EQ	2.46757			2.46757

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.15—Operational Item Parameter Used to Score Spring 17 Tests — Grade 8 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10564	EQ	0.40128			0.40128
2	12460	EQ	-1.85636			-1.85636
3	13814	EQ	-0.23429			-0.23429
4	9526	EQ	2.86364	-2.36098		0.25133
5	10561	MC4	-0.22329			-0.22329
6	10541	EQ	0.75865			0.75865
7	11304	EQ	-0.65743			-0.65743
8	10582	EQ	0.88919			0.88919
9	13152	EQ	0.34924			0.34924
10	10513	MC4	-0.21755			-0.21755
11	10543	MC4	-1.05211			-1.05211
12	10532	EQ	0.16579			0.16579
13	10528	EQ	-2.45491			-2.45491
14	10517	MI	-0.60752			-0.60752
15	12476	MC4	-0.37326			-0.37326
16	10546	EQ	1.42971			1.42971
17	12462	MC4	-0.82447			-0.82447
18	10495	MS5	2.20990			2.20990
19	10565	GI	0.09384			0.09384
20	10525	MC4	-1.64422			-1.64422
21	10487	MC4	-0.12623			-0.12623
22	10531	EQ	0.03407			0.03407
23	12005	MC4	-0.97905			-0.97905
24	10552	GI	-1.27422			-1.27422
25	11690	MC4	-0.34286			-0.34286
26	10554	MC4	-1.02849			-1.02849
27	11443	MC4	-0.91880			-0.91880
28	13150	MC4	-0.22117			-0.22117
29	10512	GI	0.59204			0.59204
30	10493	MC4	0.69535			0.69535
31	12037	EQ	0.92297			0.92297
32	10482	EQ	1.38111			1.38111
33	10588	MC4	-1.06474			-1.06474
34	11686	MS5	0.81081			0.81081
35	10576	MS5	-1.36210			-1.36210
36	9525	EQ	0.25872	0.32204		0.29038
37	10589	MC4	-3.30852			-3.30852
38	8251	EQ	0.87585			0.87585
39	10503	GI	1.48134			1.48134
40	10583	MC4	-0.41648			-0.41648

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10536	EQ	2.18874			2.18874
42	10573	GI	1.31105			1.31105
43	10514	EQ	1.40567			1.40567
44	12044	EQ	0.95677			0.95677
45	10523	EQ	2.68961			2.68961
46	10526	MS5	0.62216			0.62216
47	11546	EQ	2.14921			2.14921

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.16—Operational Item Parameter Used to Score Spring 17 Tests — Algebra I

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10940	MC4	0.54039			0.54039
2	10942	MC4	0.09573			0.09573
3	10990	MC4	0.65596			0.65596
4	12074	EQ	1.78520			1.78520
5	11012	EQ	1.44670			1.44670
6	13185	MS5	1.09113			1.09113
7	12593	MC4	-0.11769			-0.11769
8	12009	MC4	0.59618			0.59618
9	10905	MC4	-0.97781			-0.97781
10	12023	EQ	-0.46606			-0.46606
11	10973	MC4	-1.36424			-1.36424
12	12020	EQ	1.19453			1.19453
13	12156	MC4	-1.83765			-1.83765
14	13168	MC4	-0.07979			-0.07979
15	12073	EQ	1.37555			1.37555
16	11052	MC4	-1.26248			-1.26248
17	13162	MC4	-0.10561			-0.10561
18	13164	MC4	-0.46630			-0.46630
19	11664	EQ	1.66230			1.66230
20	12615	MC4	-0.59744			-0.59744
21	11047	EQ	1.70037			1.70037
22	12499	MC4	-1.37038			-1.37038
23	12060	EQ	-0.03050			-0.03050
24	11537	TI	1.22253			1.22253
25	11353	GI	-1.21934			-1.21934
26	12237	MC4	-0.08308			-0.08308
27	11751	MS6	1.72472			1.72472
28	9541	GI	-0.77275	-1.14420		-0.958475
29	11574	EQ	0.52554			0.52554
30	10969	MC4	-1.75182			-1.75182
31	11338	MC4	-0.27385			-0.27385
32	10953	MC4	-0.75521			-0.75521
33	11055	MC4	0.69104			0.69104
34	10897	MC4	0.32317			0.32317
35	11548	MC4	-1.18555			-1.18555
36	13174	MC4	-0.99601			-0.99601
37	12699	EQ	1.62793			1.62793
38	9535	EQ	-0.04011			-0.04011
39	10945	MC4	-0.65134			-0.65134
40	12898	EQ	0.33053			0.33053

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10977	MC4	-0.42252			-0.42252
42	12481	EQ	0.83053			0.83053
43	10963	MC4	-0.87904			-0.87904
44	9531	EQ	0.97351			0.97351
45	12017	MC4	-0.57324			-0.57324
46	12366	MC4	0.38062			0.38062
47	10934	EQ	-0.94776			-0.94776

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.17—Operational Item Parameter Used to Score Spring 17 Tests — Geometry

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	11072	MC4	-0.56959			-0.56959
2	12369	EQ	0.04108			0.04108
3	11523	MS5	0.91410			0.91410
4	11007	EQ	-0.74951			-0.74951
5	10933	EQ	1.60831			1.60831
6	13499	MC4	-0.74574			-0.74574
7	11923	EQ	1.49293			1.49293
8	11062	MS5	0.92341			0.92341
9	11612	HT	0.95755			0.95755
10	12656	MS5	0.57618			0.57618
11	11449	EQ	1.42021			1.42021
12	13496	MC4	-1.10260			-1.10260
13	10987	EQ	1.91513			1.91513
14	11921	MS5	-0.01574			-0.01574
15	12045	EQ	0.15431			0.15431
16	10927	MC4	-0.86523			-0.86523
17	10931	EQ	0.63331			0.63331
18	11015	EQ	-1.64967			-1.64967
19	11074	MS5	0.57924			0.57924
20	9722	MI	1.01157			1.01157
21	12350	EQ	-0.75689			-0.75689
22	13521	EQ	-1.80176			-1.80176
23	10910	MC4	-0.78470			-0.78470
24	12576	MS5	0.47267			0.47267
25	11315	MS6	-1.61279			-1.61279
26	10923	MC4	-0.48643			-0.48643
27	13505	MC4	-0.83609			-0.83609
28	11448	MC4	-1.87122			-1.87122
29	11039	EQ	-0.10460			-0.10460
30	11092	MS6	0.93109			0.93109
31	12925	EQ	1.31561			1.31561
32	11547	EQ	0.05275			0.05275
33	11026	HT	-0.76141			-0.76141
34	12931	MC4	0.14744			0.14744
35	12622	MI	-3.44415			-3.44415
36	13506	MC4	-0.95238			-0.95238
37	11089	MS6	-0.76970			-0.76970
38	13497	MC4	0.56517			0.56517
39	13500	MC4	-0.35376			-0.35376
40	13532	EQ	1.10992			1.10992

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	9556	EQ	-0.21949			-0.21949
42	12091	TI	-0.35157			-0.35157
43	11792	EQ	0.43582			0.43582
44	10913	MC4	-1.79442			-1.79442
45	12152	EQ	1.23693			1.23693
46	11063	EQ	1.15008			1.15008
47	12342	EQ	-0.52835			-0.52835


Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.

Appendix G.18—Operational Item Parameter Used to Score Spring 17 Tests — Algebra II

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13204	EQ	0.47746			0.47746
2	12725	EQ	-0.35279			-0.35279
3	13206	MC4	0.27239			0.27239
4	11401	EQ	0.64249			0.64249
5	12027	MC4	-0.25980			-0.25980
6	9567	EQ	0.98222			0.98222
7	10174	MC4	-0.54727			-0.54727
8	10230	MC4	-0.36050			-0.36050
9	11725	MC4	-1.36159			-1.36159
10	12025	EQ	-0.83854			-0.83854
11	10245	EQ	0.65995			0.65995
12	12096	MS6	-0.08671			-0.08671
13	11936	MC4	-0.51723			-0.51723
14	13231	MC4	-1.90022			-1.90022
15	10193	MC4	-0.57364			-0.57364
16	11541	MC4	-1.39444			-1.39444
17	10227	MC4	-1.65582			-1.65582
18	10236	MC4	-0.89747			-0.89747
19	11121	MC4	-1.04160			-1.04160
20	9580	EQ	-0.05533			-0.05533
21	10187	MC4	-0.17866			-0.17866
22	10203	MC4	-1.02449			-1.02449
23	12565	EQ	-1.04390	2.18370		0.56990
24	13211	MC4	-0.47605			-0.47605
25	10241	MS5	1.83403			1.83403
26	10221	MC4	-0.44392			-0.44392
27	11836	EQ	1.19312			1.19312
28	12354	MC4	-0.31512			-0.31512
29	13475	MC4	-1.25771			-1.25771
30	12906	MC4	-1.84106			-1.84106
31	13212	MS5	1.61434			1.61434
32	10220	MC4	0.54450			0.54450
33	12097	EQ	-0.59842			-0.59842
34	10192	MC4	-1.21466			-1.21466
35	11380	EQ	2.01582			2.01582
36	10217	MC4	-1.30014			-1.30014
37	10256	GI	1.04876			1.04876
38	10223	EQ	0.48221			0.48221
39	10249	EQ	1.06528			1.06528
40	12076	MS5	1.05206			1.05206


Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	8253	EQ	0.56780			0.56780
42	11730	EQ	1.22715			1.22715
43	11804	EQ	1.19746			1.19746
44	11740	EQ	0.78033			0.78033
45	12611	EQ	1.48557			1.48557
46	12934	EBSR4	0.19308			0.19308
47	10210	MC4	-1.79638			-1.79638

Note: In spring 2017, the pre-equated parameters calibrated following the spring 2016 test administration of AzMERIT were used for the final scoring and reporting.




AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Statistical Review Training for ADE




AIR
American Institutes for Research



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics


Statistical Review of

- Item Quality and Performance
 - Does the item behave the way it's supposed to behave?
- Item Difficulty
 - How hard is the item?
- Differential Item Functioning
 - Does the item behave




AIR
American Institutes for Research

2


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Quality

- Do highly skilled students perform better on the item than less skilled students?
- Correlation with Test – link between selecting a response option and doing well on the rest of the test
 - For key, + is good, – is bad
 - For distractors, – is good, + is bad


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

3


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Quality Flag Criteria

- Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items; (AB)
- Adjusted biserial correlations for multiple-choice item distractors is greater than .05; (ABD)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

4


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Difficulty

- How hard is the item?
- What percent of students answer item correctly?
- MC items – % of students selecting each response option
- Non-MC items – % of students achieving each score point


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

5


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Item Difficulty Flag Criteria

- Proportion correct value is less than .25 or greater than .95 for multiple-choice items, or greater than .95 for any single score point of a constructed-response item;
- Also known as p-value (P or CR_Prop)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

6


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Non-Modal Key

- A distractor is chosen by students more often than the key is chosen


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

7


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Non-Modal Key Flag Criteria

- The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items; (NMK)


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

8


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Omit Rate

- Students do not provide a response


 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

9


 **AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Omit Rate Flag Criteria

- Omit rate is greater than .15;

 **AIR**
AMERICAN INSTITUTES FOR RESEARCH

10


AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics


Differential Item Functioning

* Fair Items behave similarly across groups


* Probability of answering correctly is the same for all students of similar ability regardless of group membership

Subgroup Comparisons:

- Female/Male
- Non-Hispanic / Hispanic, Latino or Spanish origin
- Black, African American / White
- American Indian or Alaskan Native / White
- Asian / White
- Native Hawaiian or Other Pacific Islander / White
- Multiple ethnicities selected / White



AIR
AMERICAN INSTITUTES FOR RESEARCH

11



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Differential Item Functioning (DIF)

- Direction of possible bias
 - “–” item favors reference groups
 - “+” item favors focal group
- Severity of possible bias
 - “A” No statistical evidence of DIF
 - “B” Evidence for potential mild DIF
 - “C” Evidence for potential severe DIF
- “C” indicates that the item is more difficult for one group and should be reviewed carefully for bias



AIR
AMERICAN INSTITUTES FOR RESEARCH

12



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

DIF Flag Criteria

- Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF.
- Items are categorized as **positive DIF** (i.e., +A, +B, or +C), signifying that the item **favours the focal group** (e.g., African American/Black, Hispanic, or female), or
- negative DIF** (i.e., -A, -B, or -C), signifying that the item **favours the reference group** (e.g., white or male).
- Items are flagged if their DIF statistics fall into the "C" category for any group, which indicates that the item shows **significant DIF** and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness



AIR
AMERICAN INSTITUTES FOR RESEARCH

13



AzMERIT | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Content Expert Judgments

- Statistical information is important, but not a substitute for expert judges
- Items central to a learning standard may be difficult because a concept is not currently included in curriculum
- Items may show DIF because some concepts may be less likely to be covered in all area schools



AIR
AMERICAN INSTITUTES FOR RESEARCH

14

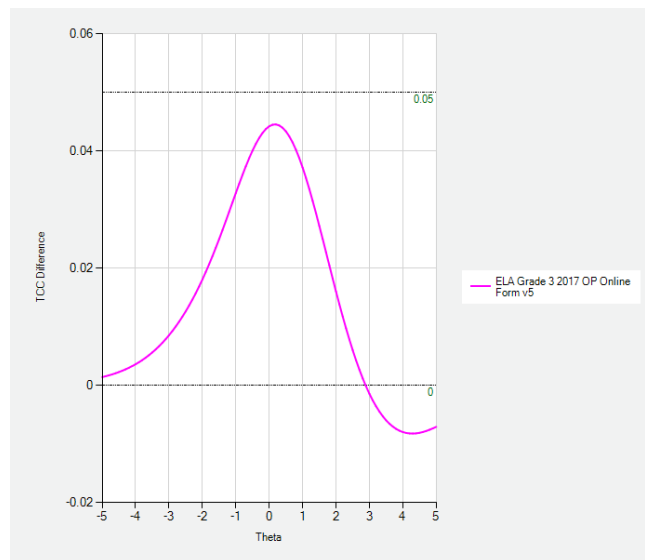
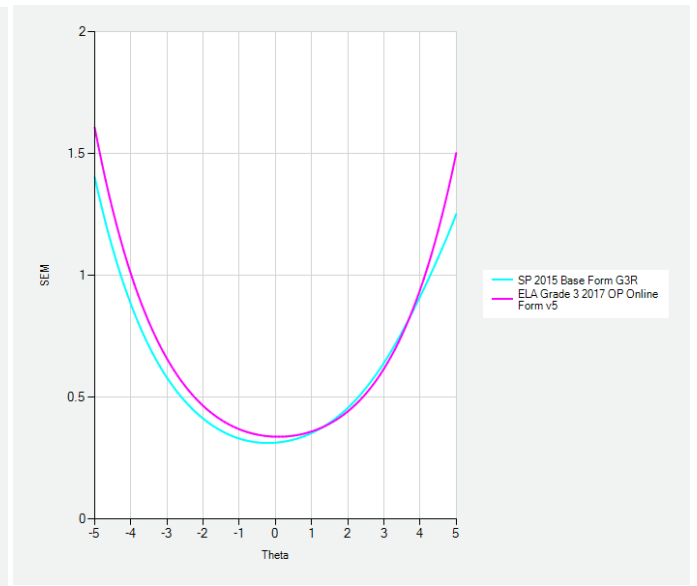
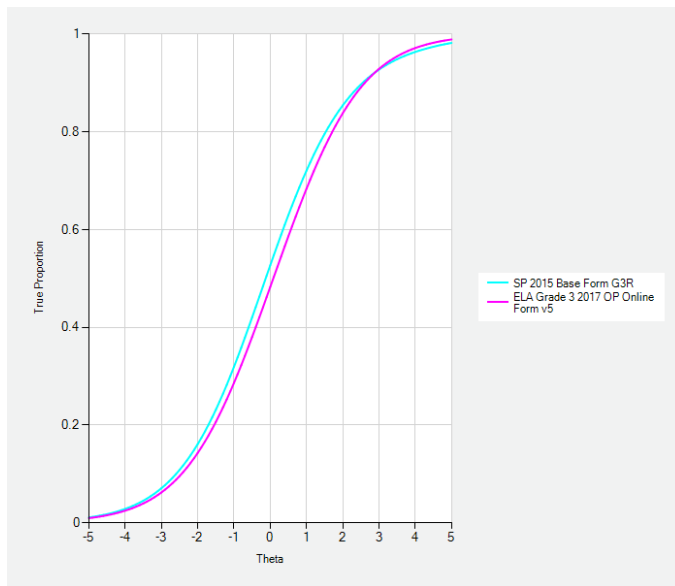
**AzMERIT** | Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics

Logistics

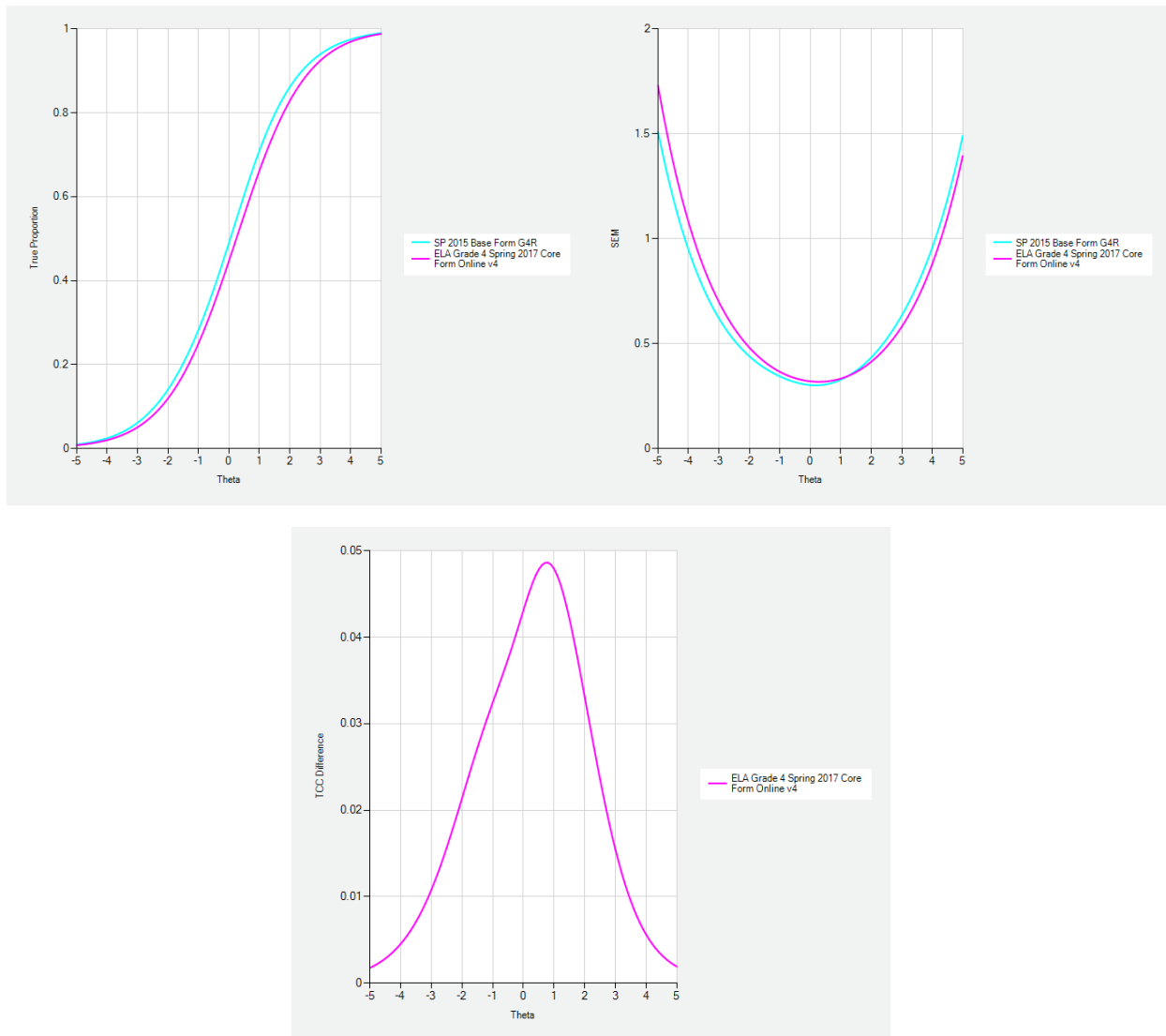
- Items can be found at the **Content and Fairness Data Review and Resolution** review level in the Arizona Assessment project in ITS
- The MDSs will be posted here on the sftp:
/files/AzMERIT/To ADE/Content Data Review/
- Please “PEND” any data comments in ITS

15

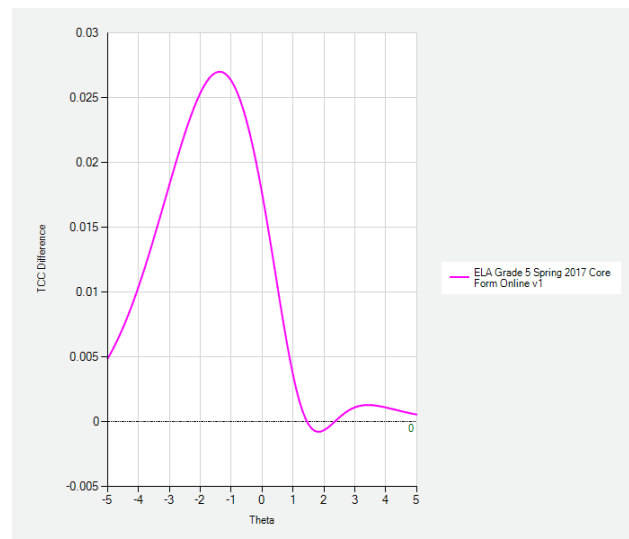
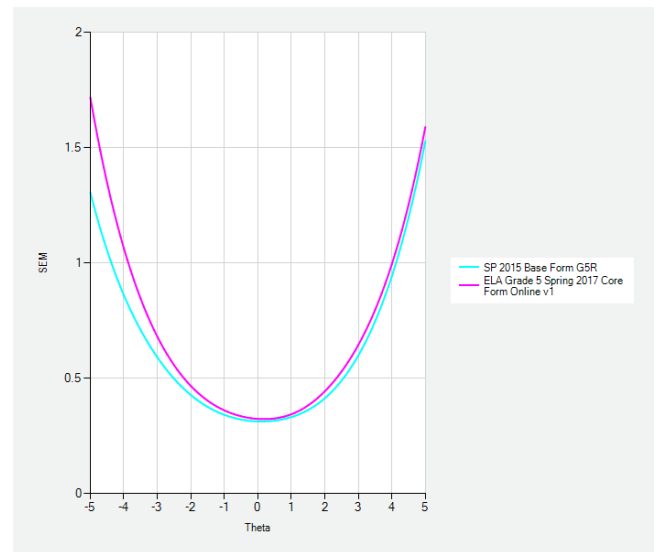
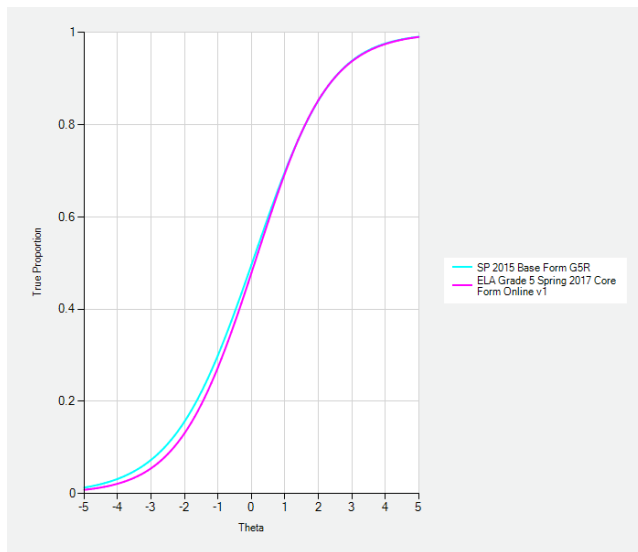
Appendix I.1 - Spring 2017 ELA Grade 3



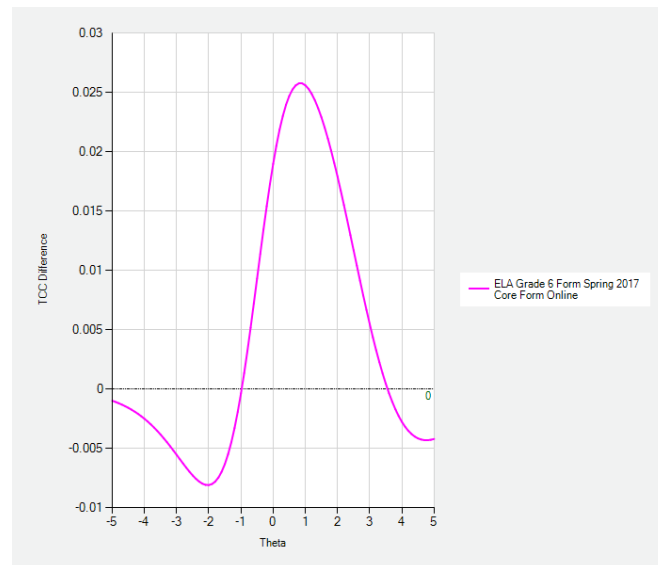
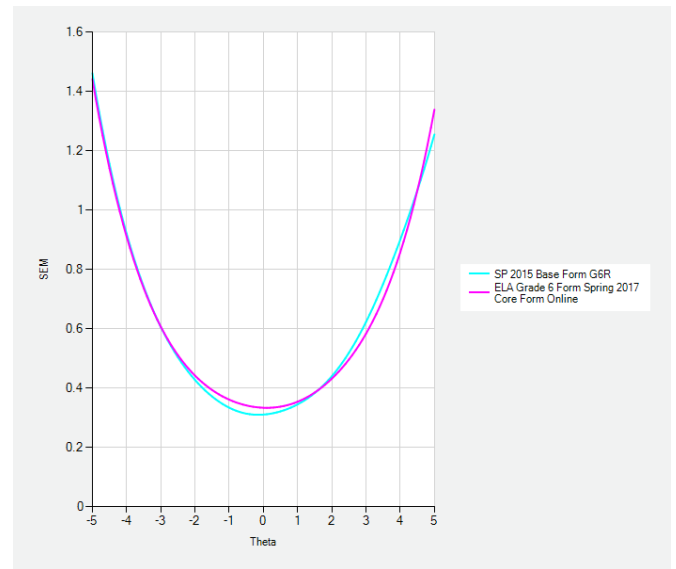
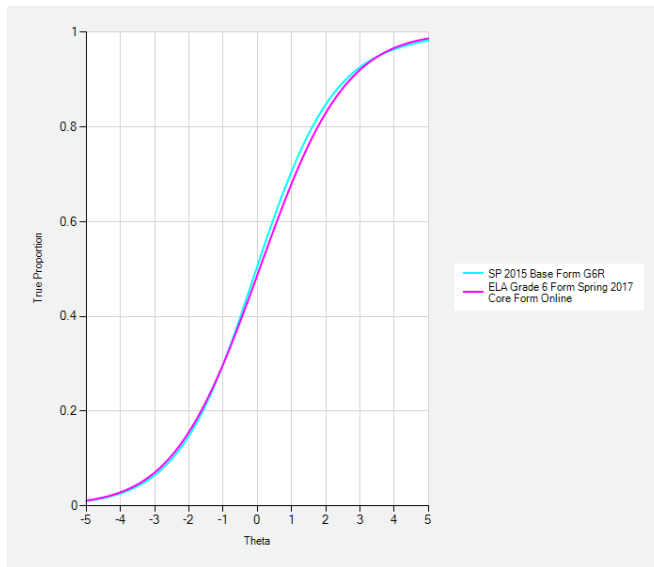
Appendix I.2 – Spring 2017 ELA Grade 4



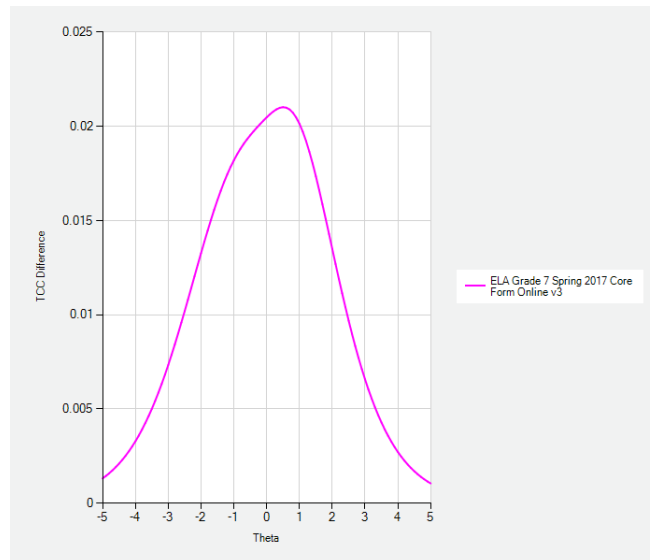
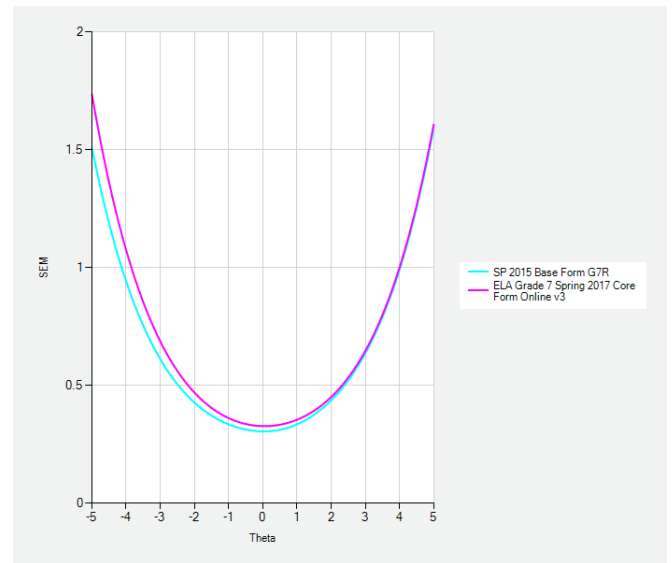
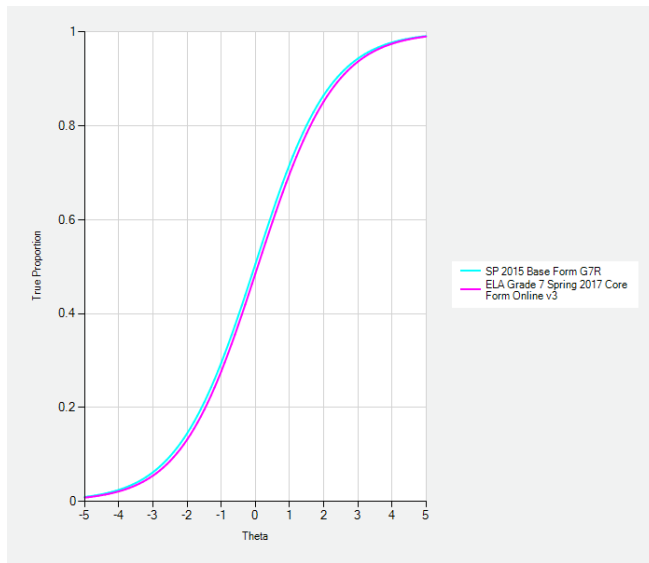
Appendix I.3 – Spring 2017 ELA Grade 5



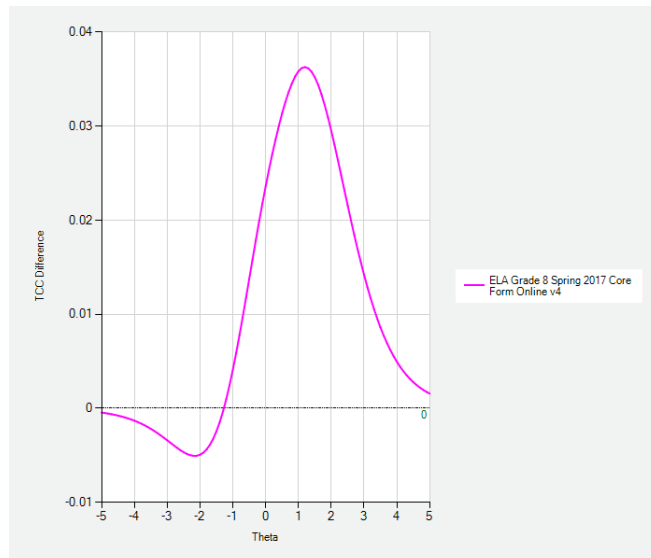
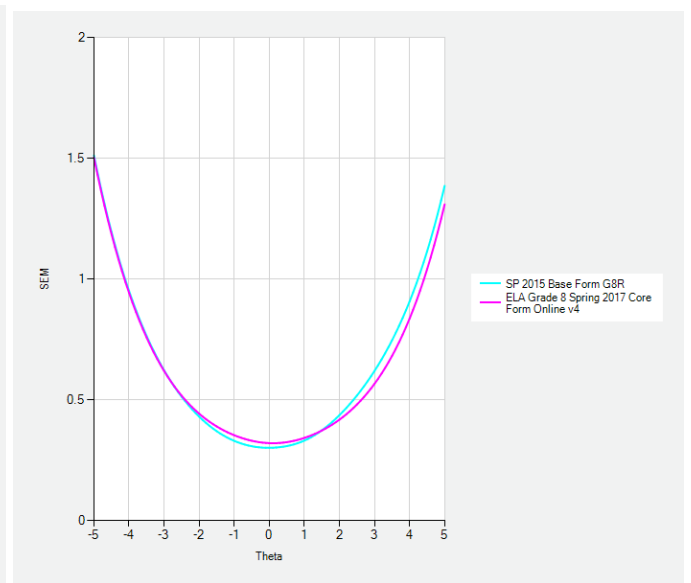
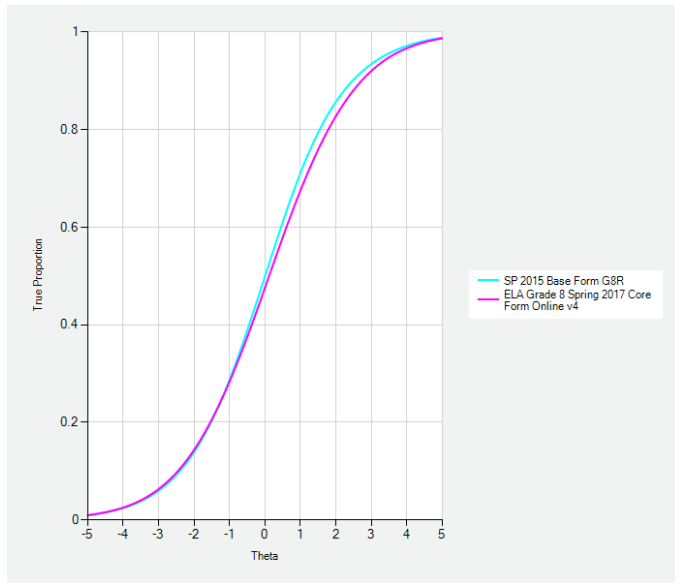
Appendix I.4 – Spring 2017 Grade 6 ELA



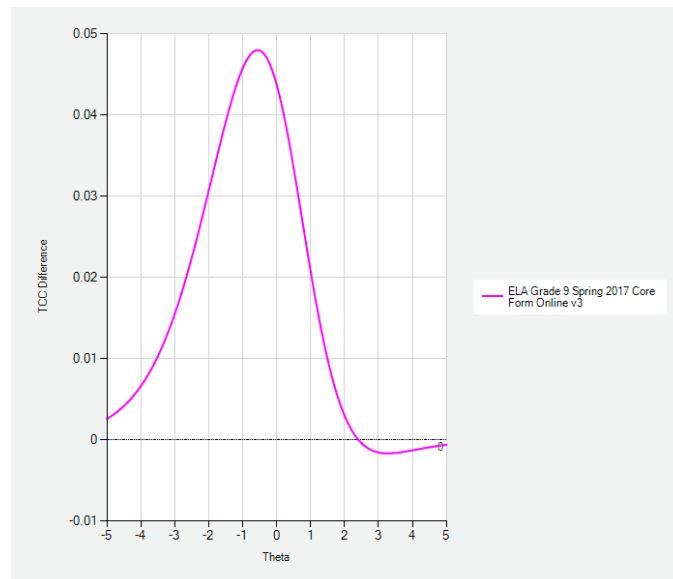
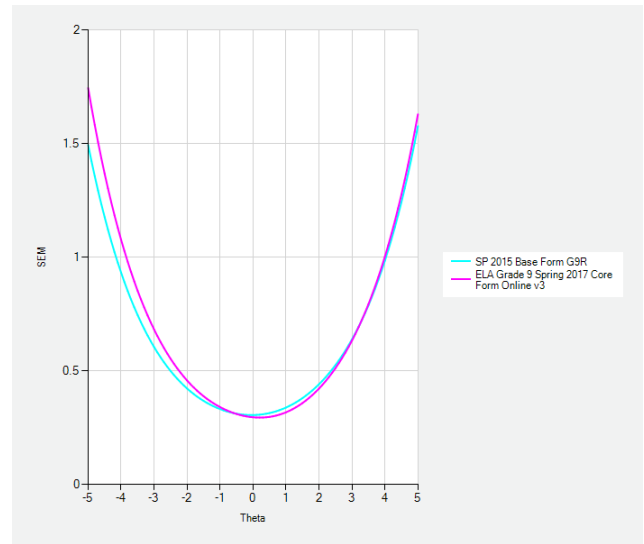
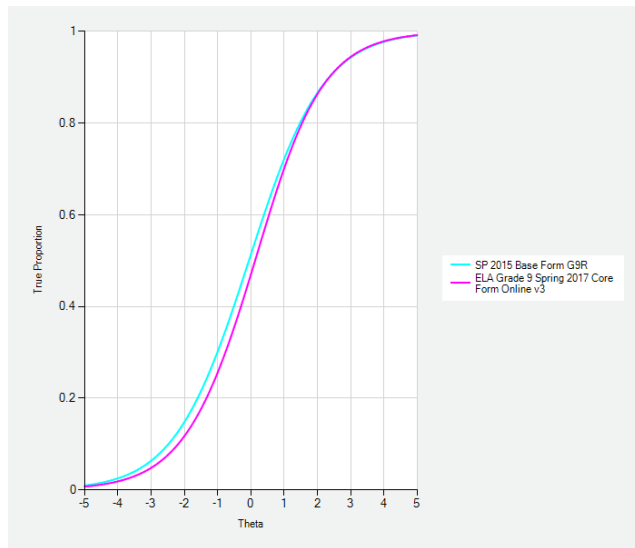
Appendix I.5 – Spring 2017 Grade 7 ELA



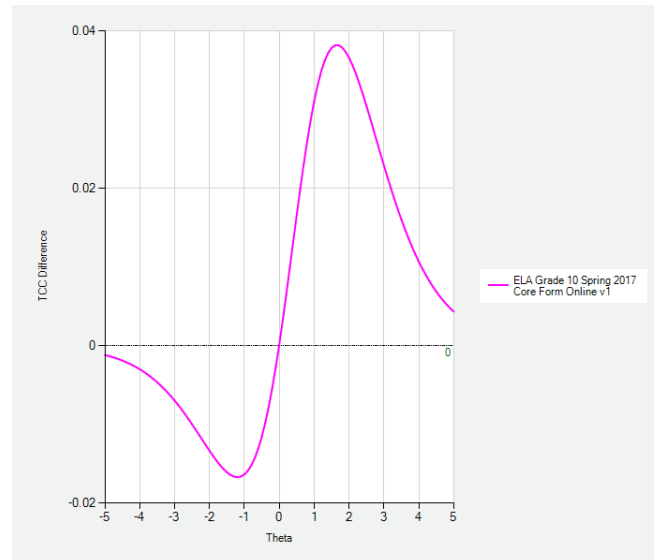
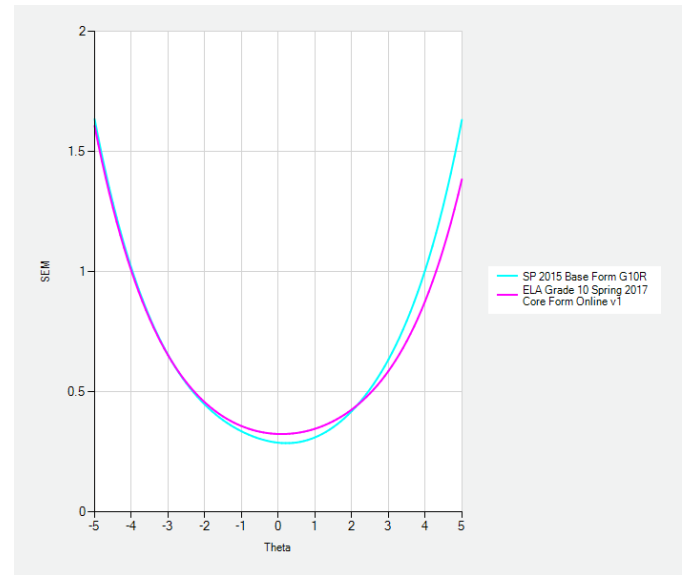
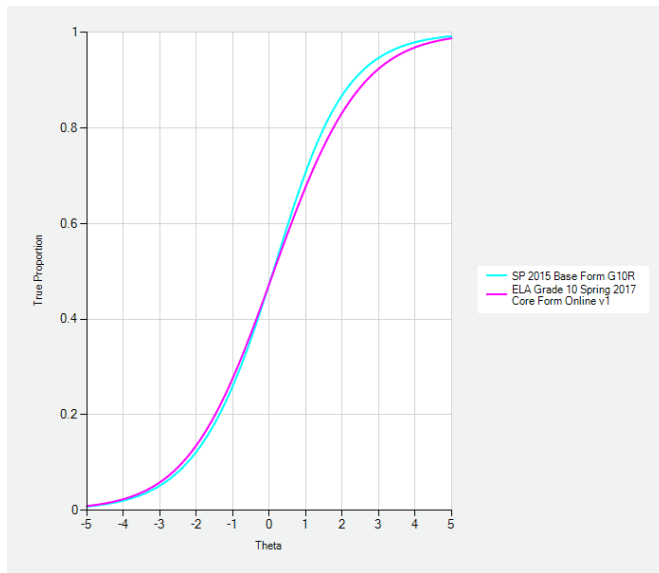
Appendix I.6 – Spring 2017 Grade 8 ELA



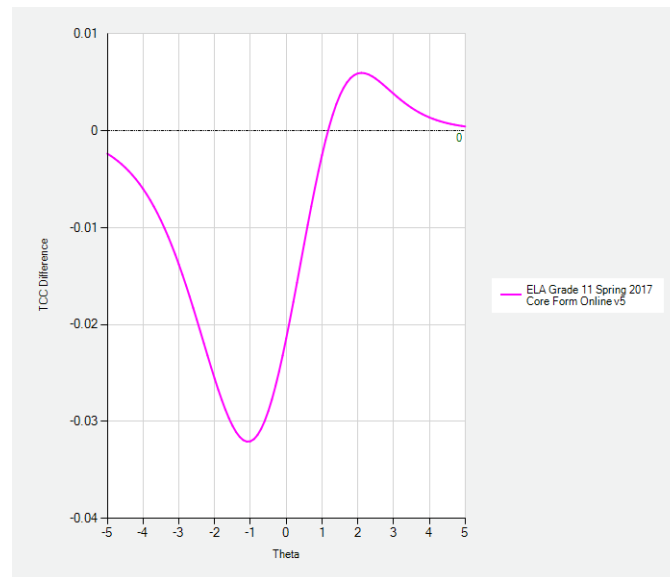
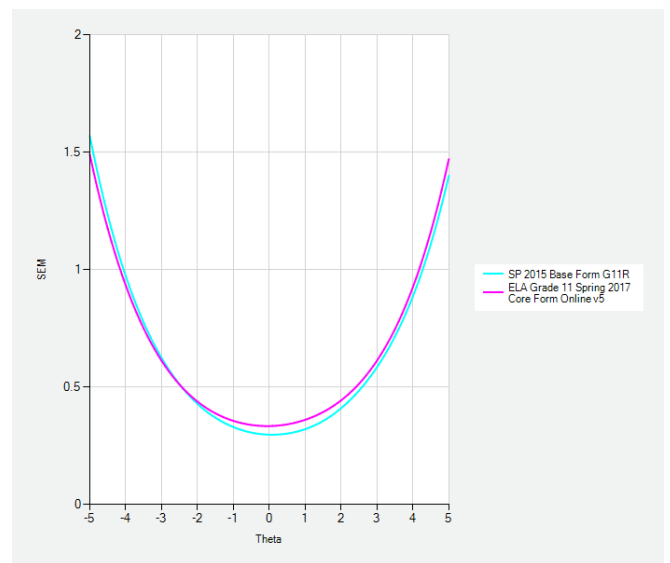
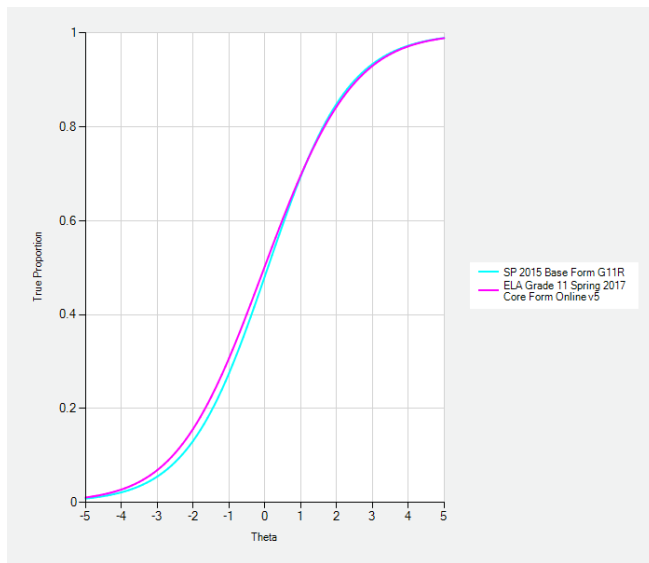
Appendix I.7 – Spring 2017 Grade 9 ELA



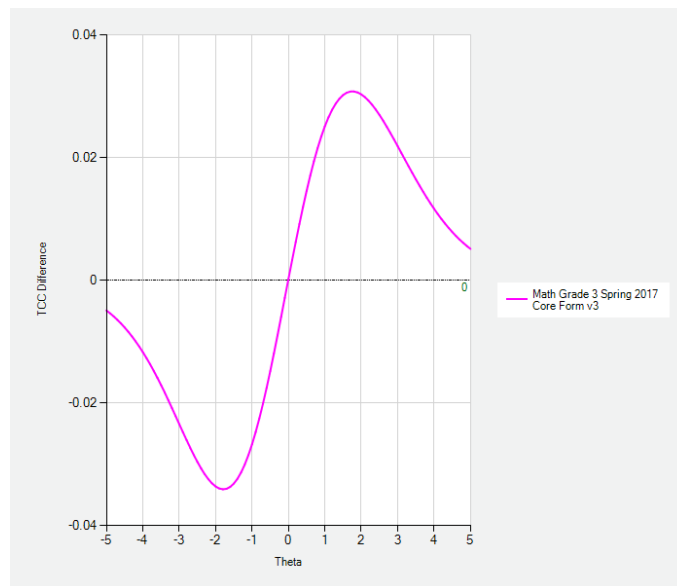
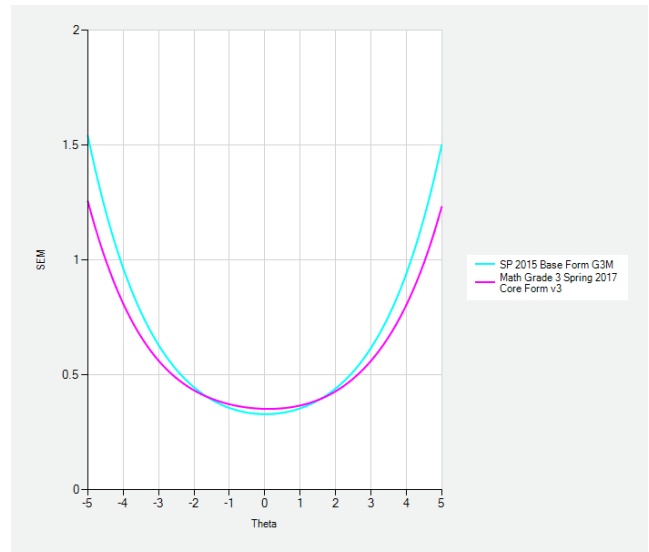
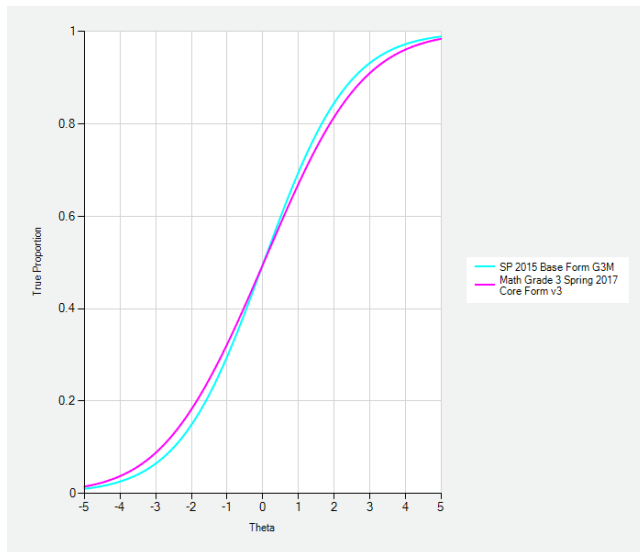
Appendix I.8 – Spring 2017 Grade 10 ELA



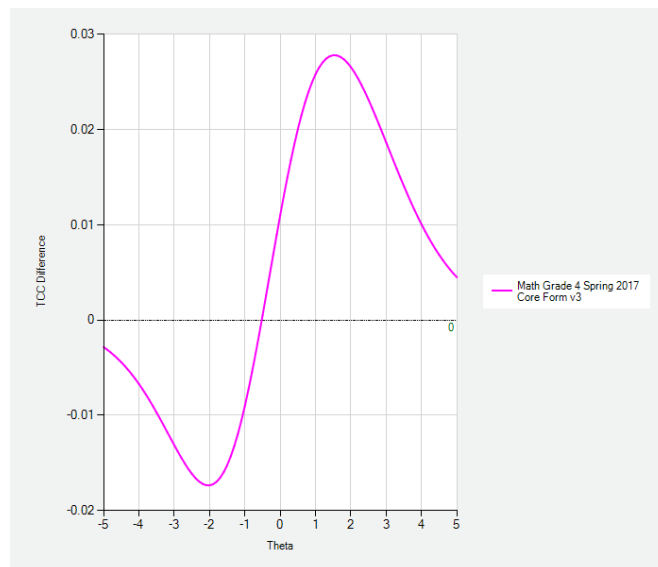
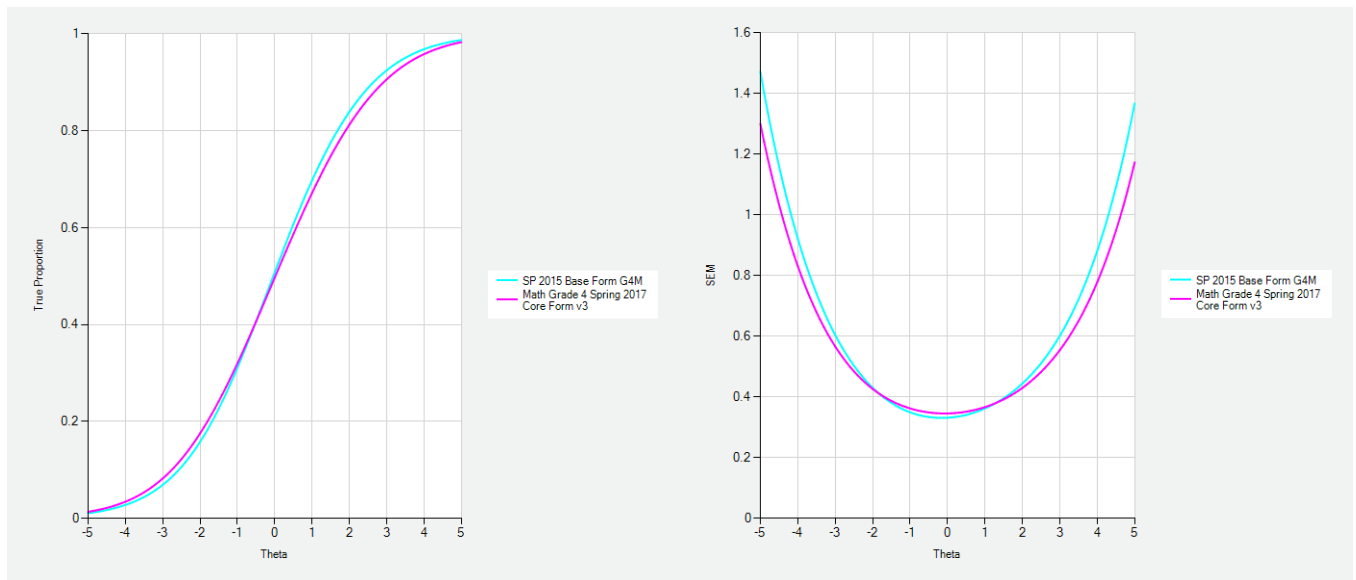
Appendix I.9 – Spring 2017 Grade 11 ELA



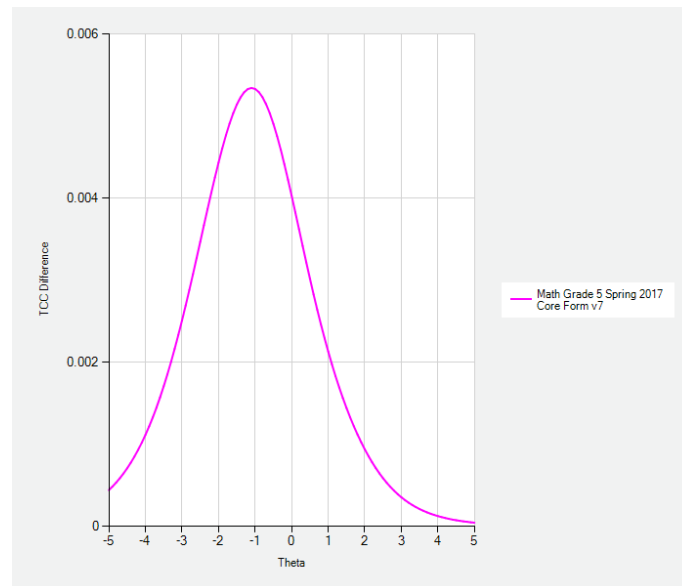
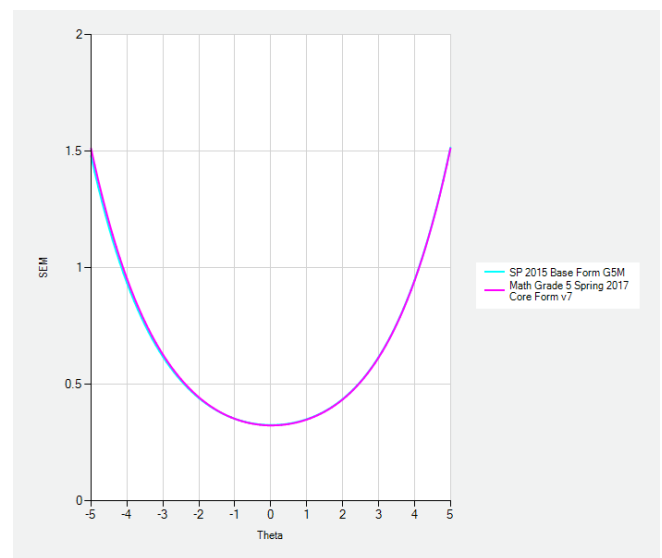
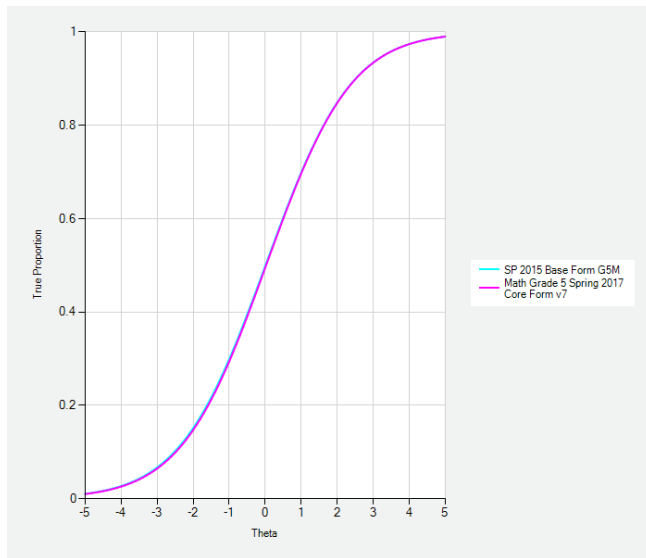
Appendix I.10 – Spring 2017 Grade 3 Math



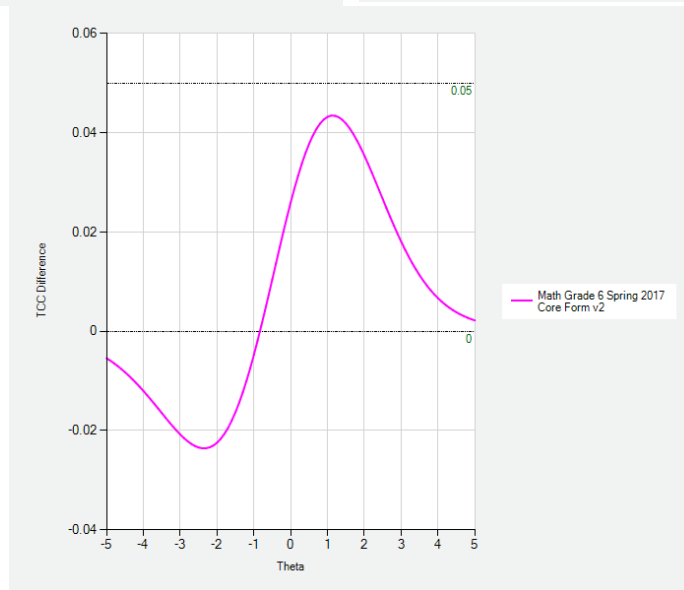
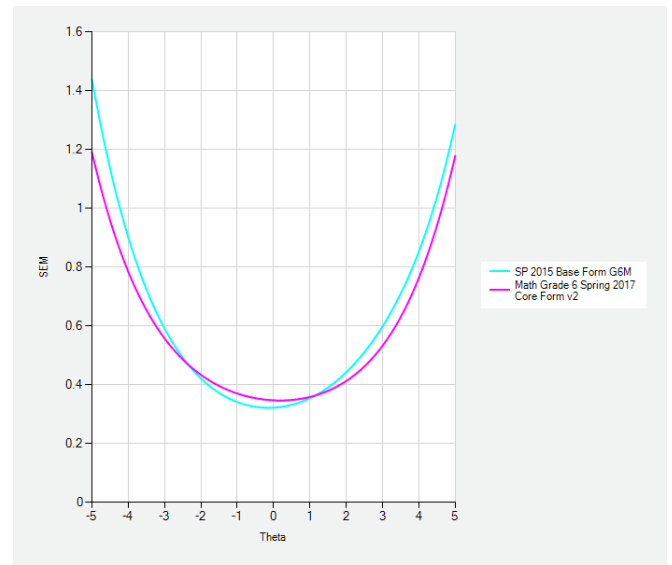
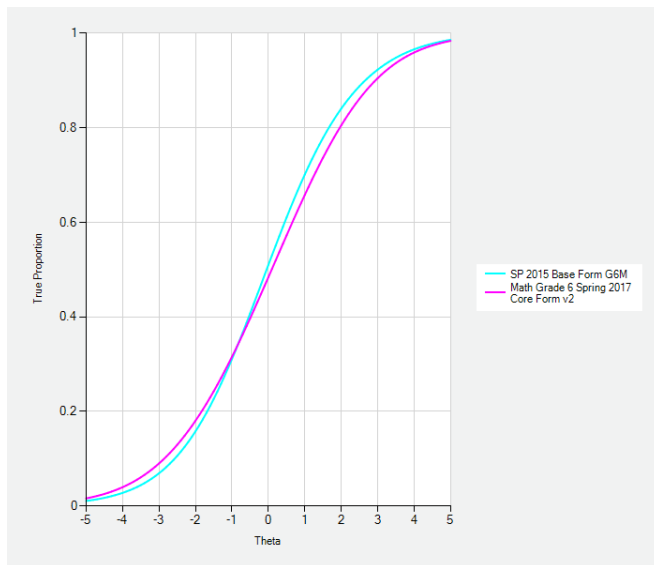
Appendix I.11 – Spring 2017 Grade 4 Math



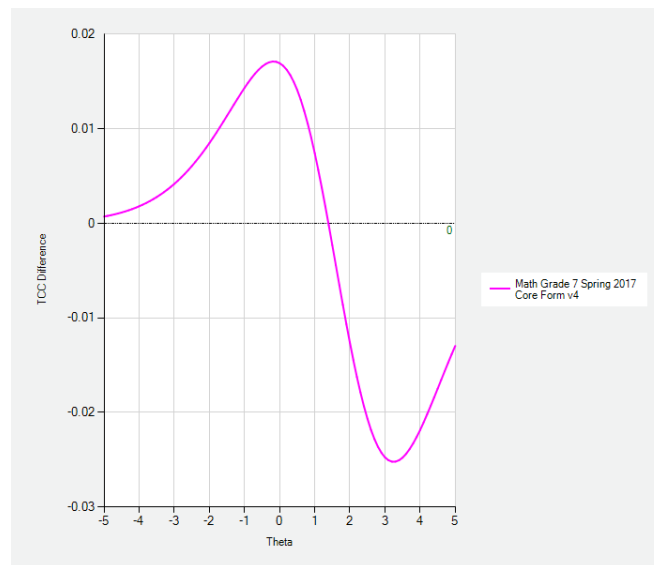
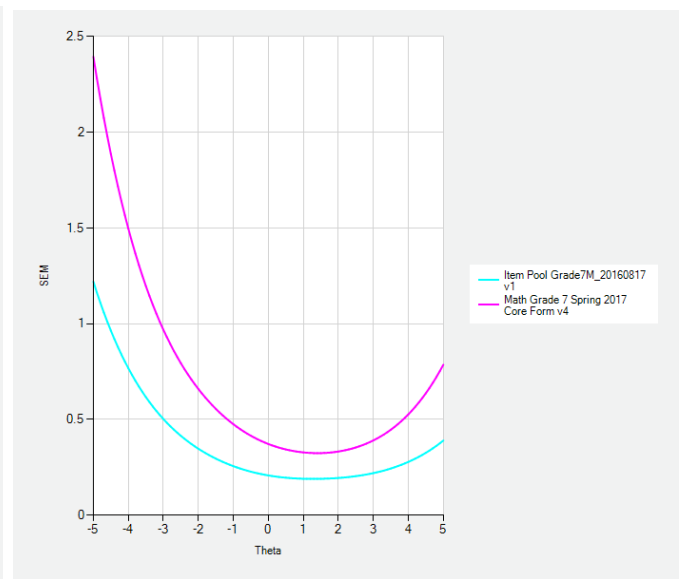
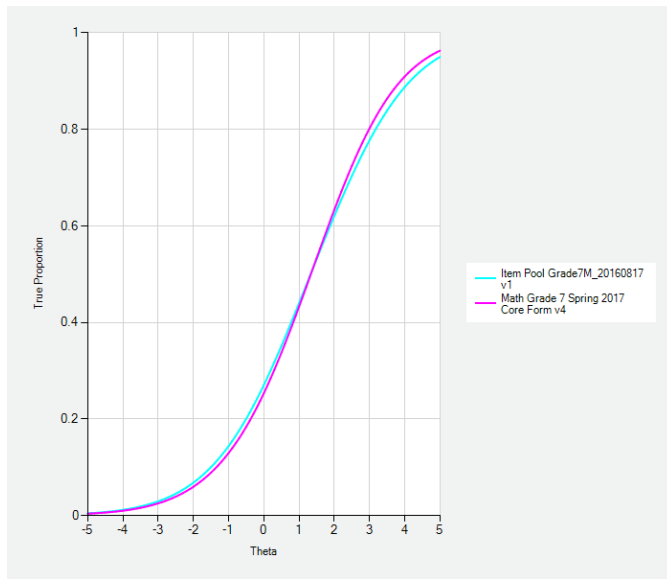
Appendix I.12 – Spring 2017 Grade 5 Math



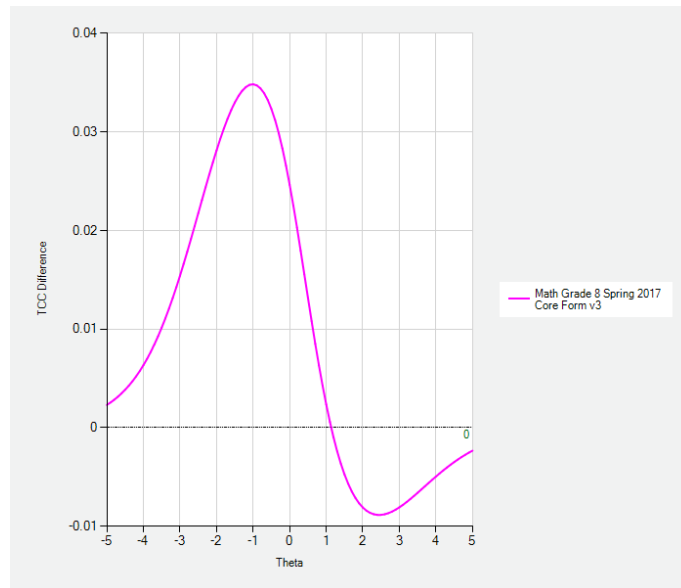
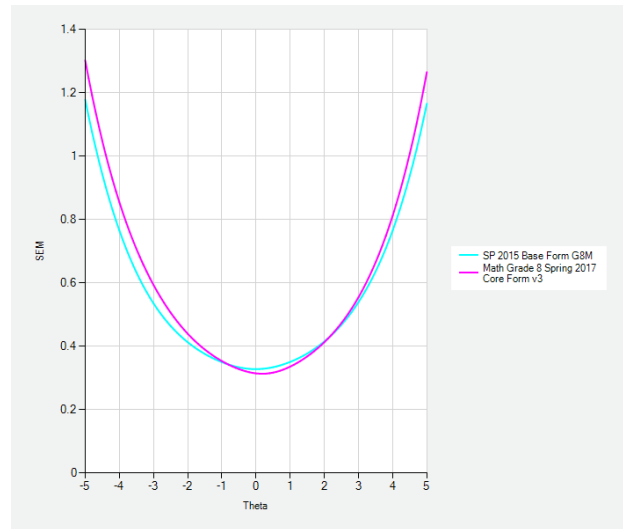
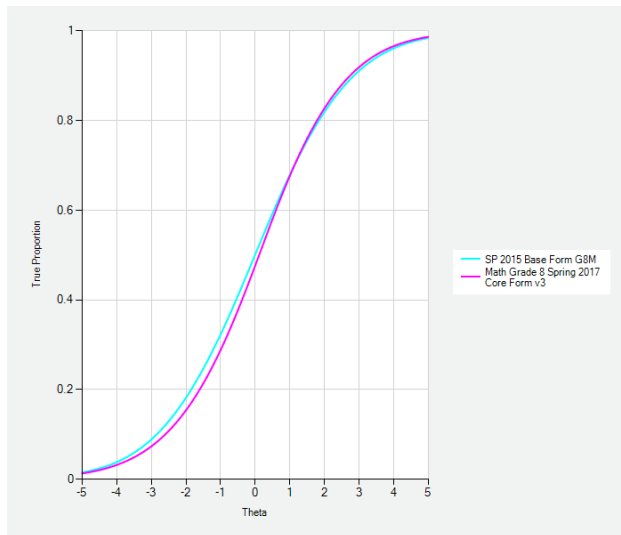
Appendix I.13 – Spring 2017 Grade 6 Math



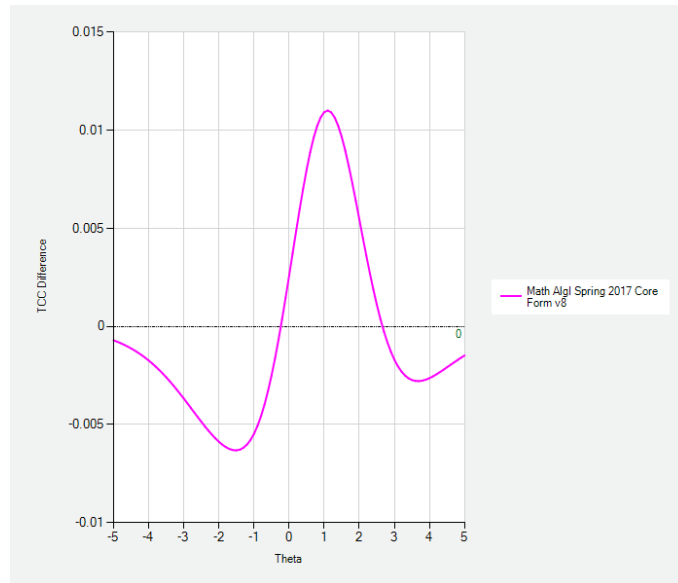
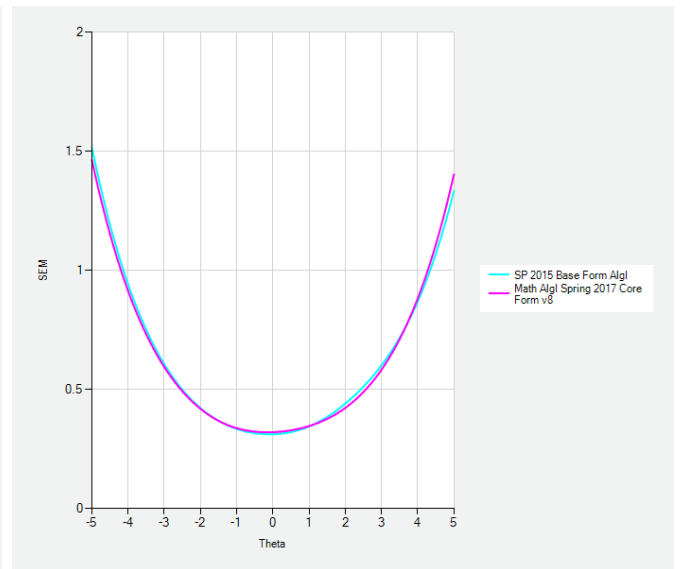
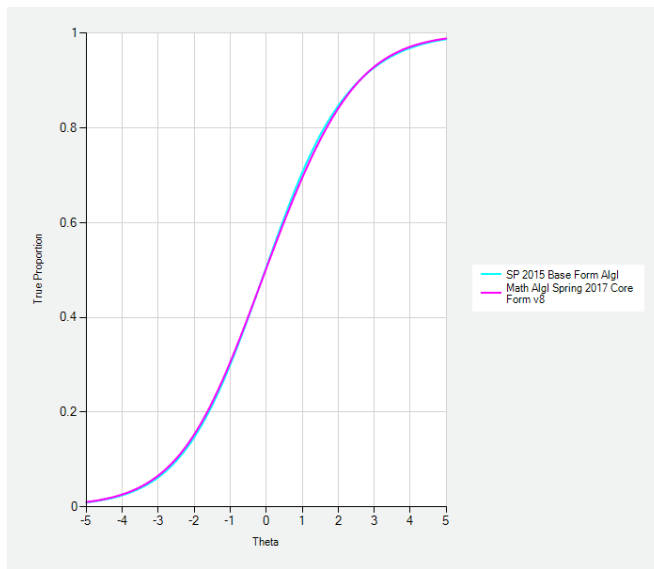
Appendix I.14 – Spring 2017 Grade 7 Math



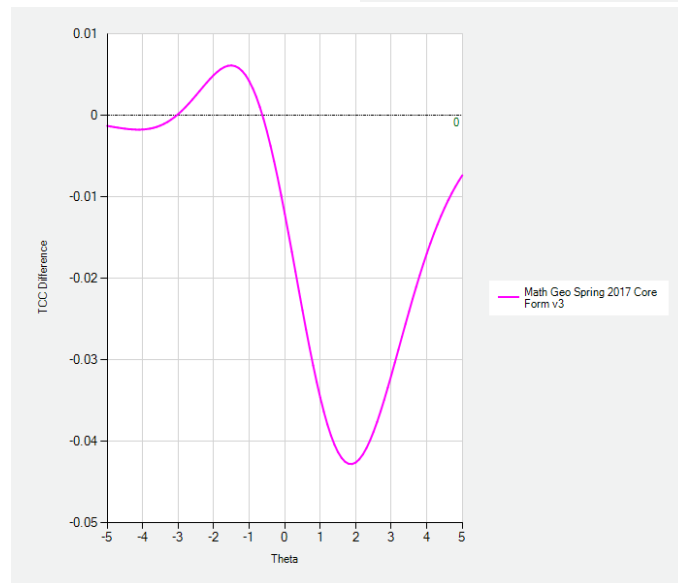
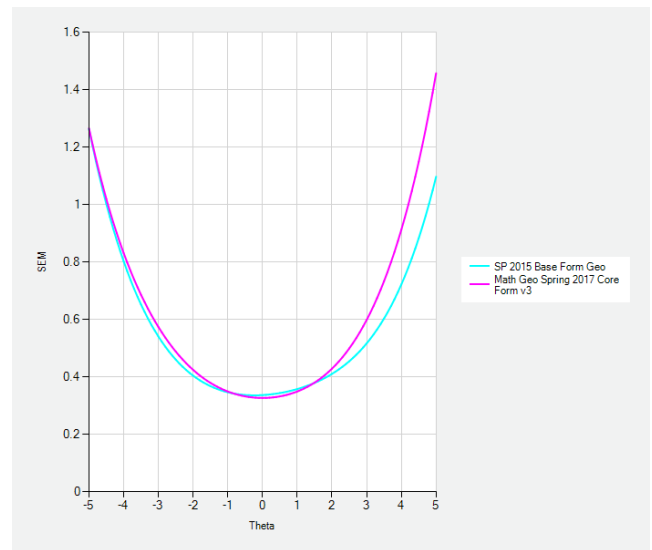
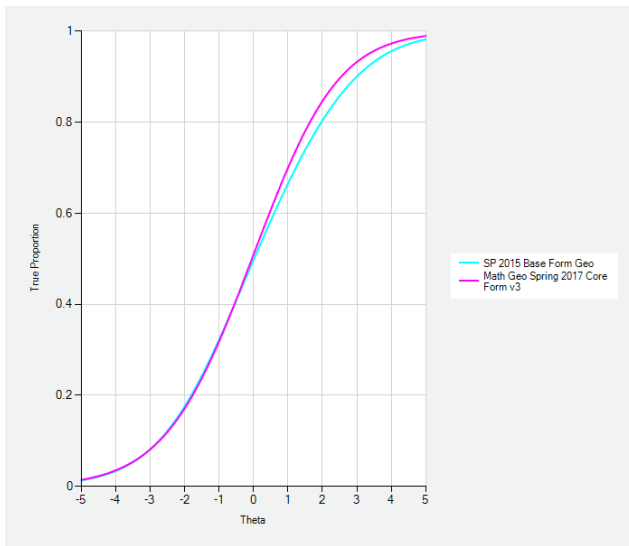
Appendix I.15 – Spring 2017 Grade 8 Math



Appendix I.16 – Spring 2017 Algebra I



Appendix I.17 – Spring 2017 Geometry



Appendix I.18 – Spring 2017 Algebra II

