



---

# Annual Technical Report

## Arizona Statewide Assessment in English Language Arts and Mathematics

**2017–2018 School Year**

September 2018

**ARIZONA STATEWIDE ASSESSMENT**

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

**ENGLISH LANGUAGE ARTS GRADES 3–11**

**MATHEMATICS GRADES 3–8, ALGEBRA I, GEOMETRY, AND ALGEBRA II**

**2017–2018 ANNUAL TECHNICAL REPORT**

**SEPTEMBER 2018**

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.....	6
1.1	Overview.....	6
1.2	Validity Evidence .....	7
1.3	Evidence Based on Test Content .....	14
1.4	Evidence for Interpretation of Performance Standards .....	17
1.5	Evidence Based on Internal Structure .....	20
1.5.1	ELA Content Model.....	21
1.5.2	ELA Depth of Knowledge .....	23
1.5.3	Mathematics Content Model .....	24
1.5.4	Mathematics Depth of Knowledge.....	25
1.6	Evidence for Relationships with Conceptually Related Constructs .....	26
1.7	Measurement Invariance Across Subgroups .....	27
1.8	Differential Mode Effects Across Subgroups .....	28
1.9	Evidence for Student Growth – Overall and by Subgroups .....	30
1.10	Day, Week, and Time-of-Day Effects on Performance .....	34
1.11	Arizona Glossary Study.....	36
1.12	Summary of Validity of Test Score Interpretations .....	40
2.	Background of Arizona Statewide Assessments.....	42
2.1	Development of Arizona State Standards .....	43
2.2	AzMERIT Test Design .....	43
3.	Summary of Summer 2017 and Fall 2017 Operational Test Administration.....	45
3.1	Student Population and Participation .....	45
3.2	Summary of Overall Student Performance .....	47
3.3	Student Performance by Subgroup .....	48
3.4	Reliability.....	53
3.4.1	Internal Consistency .....	53
3.4.2	Standard Error of Measurement .....	54
3.4.3	Student Classification Reliability .....	55
3.4.4	Classification Accuracy .....	55
3.4.5	Classification Consistency.....	56
3.4.6	Classification Reliability Estimates .....	56
3.4.7	Reliability for Subgroups in the Population.....	57
3.4.8	Subscale Reliability.....	59
3.5	Subscale Intercorrelations.....	61
4.	Summary of Spring 2018 Operational Test Administration .....	63
4.1	Student Population and Participation .....	63
4.2	Classical Item Analysis .....	65
4.3	Item Response Theory Analysis.....	67
4.4	Summary of Overall Student Performance .....	69
4.5	Student Performance by Subgroup .....	73
4.6	Reliability .....	81
4.6.1	Internal Consistency .....	81
4.6.2	Standard Error of Measurement .....	82
4.6.3	Student Classification Reliability .....	83
4.6.4	Classification Accuracy .....	84
4.6.5	Classification Consistency.....	84
4.6.6	Classification Accuracy and Consistency Estimates.....	85
4.6.7	Reliability for Subgroups in the Population.....	93
4.6.8	Subscale Reliability.....	94

4.7	Subscale Intercorrelations .....	95
4.8	Handscoring agreement rate.....	97
5.	Item Development and Test Construction .....	99
5.1	Item-Development Process .....	100
5.1.1	Item Writing.....	100
5.1.2	Machine-Scored Constructed-Response Item-Development Tools.....	103
5.1.3	Item Types .....	103
5.2	Item Review.....	105
5.3	Field Testing.....	107
5.4	Item Statistics .....	108
5.4.1	Classical Statistics .....	108
5.4.2	Item Response Theory Statistics .....	109
5.4.3	Analysis of Differential Item Functioning .....	109
5.5	Test Construction .....	111
5.5.1	Operational Form Construction .....	111
5.5.2	Assembling Test Forms.....	113
6.	Test Administration .....	115
6.1	Eligibility .....	115
6.2	Administration Procedures.....	115
6.2.1	Managing Testing.....	117
6.3	Testing Conditions, Tools, and Accommodations.....	118
6.3.1	Universal Test Administration Conditions.....	118
6.3.2	Universal Testing Tools for Computer-Based Testing .....	119
6.3.3	Subject-Area Tools for CBT and PBT.....	120
6.3.4	Accommodations.....	120
6.4	System Security .....	124
6.4.1	Secure System Design .....	124
6.4.2	System Security Components.....	124
6.5	Test Security .....	125
6.6	Data Forensics Program.....	127
6.6.1	Changes in Student Performance .....	128
6.6.2	Item Response Latency .....	128
6.6.3	Inconsistent Item Response Pattern (Person Fit) .....	129
6.6.4	Response Change and Response Similarity .....	130
7.	Reporting and Interpreting AzMERIT Scores .....	133
7.1	Appropriate Uses for Scores and Reports .....	133
7.2	Reports Provided .....	134
7.2.1	Family Reports.....	134
7.2.2	Online Reporting System for Educators .....	135
7.3	Interpretation of Scores .....	141
8.	Performance Standards.....	143
8.1	Standard-Setting Procedures.....	143
8.1.1	Performance-Level Descriptors.....	144
8.2	Recommended Performance Standards.....	144
9.	Scaling And Equating .....	148
9.1.1	Item Response Theory Procedures.....	148
9.1.2	Calibration of AzMERIT Item Banks.....	149
9.1.3	Estimating Student Ability Using Maximum Likelihood Estimation .....	149
9.2	Establishing a Vertical Scale in ELA and Mathematics.....	151
9.2.1	Linking Items .....	151

9.2.2	Linking Analysis .....	151
9.3	AzMERIT Reporting Scale (Scale Scores).....	158
9.4	Linking Paper and Online Test Scores (Mode Comparability) .....	159
9.4.1	Mode Linking .....	159
9.4.2	School Performance .....	163
9.5	Linking the AzMERIT to Other Scales for Performance Comparison .....	163
9.5.1	Establishing Linkages to AIMS, SAGE, Smarter Balanced, and PISA .....	163
9.5.2	Identifying the Location of the ACT College-Ready Cut on AzMERIT .....	164
10.	Constructed-Response Scoring.....	167
10.1	Machine Scoring .....	167
10.1.1	Explicit Rubrics .....	167
10.1.2	Essay Autoscoring.....	167
10.1.3	Machine-Identified Condition Codes.....	173
10.2	Handscoring.....	175
10.2.1	Handscoring Process .....	175
10.2.2	Handscoring Quality Control .....	176
10.2.3	Handscoring Reliability and Validity .....	176
10.2.4	Machine-Scoring Verification .....	178
11.	Quality Assurance Procedures.....	179
11.1	Quality Assurance in Test Construction.....	179
11.2	Quality Assurance in Paper-Delivered Test Production.....	180
11.3	Quality Assurance in Computer-Delivered Test Production.....	181
11.3.1	Production of Content.....	181
11.3.2	Web Approval of Content During Development .....	181
11.3.3	Approval of Final Forms .....	182
11.3.4	Packaging.....	182
11.3.5	Platform Review .....	182
11.3.6	User Acceptance Testing and Final Review .....	182
11.3.7	Functionality and Configuration .....	183
11.4	Quality Assurance in Document Processing .....	183
11.4.1	Scanning Accuracy.....	183
11.4.2	Quality Assurance in Editing and Data Input.....	183
11.5	Quality Assurance in Data Preparation .....	184
11.6	Quality Assurance in Test Form Equating.....	185
11.7	Quality Assurance in Scoring and Reporting .....	185
11.7.1	Quality Assurance in HandScoring .....	185
11.7.2	Test Scoring .....	187
11.7.3	Reporting.....	189
12.	References.....	191

## 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 2017 to Spring 2018.....	D-1
Appendix E. Equations and Formula for Estimating Reliability.....	E-1
Appendix F. Student Participation by Demographic Subgroup – Spring 2018 Administration.....	F-1
Appendix G. Operational Item Parameter Estimates – Spring 2018 Administration.....	G-1
Appendix H. Data Review Training Slides .....	H-1
Appendix I. Test Characteristic Curves – Spring 2018 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 State Standards and if individual students have achieved those standards by the end of each school year. Thus, this report begins with a review of the 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 more evidence is gained.

Chapter 2 describes the design and development of the AzMERIT assessment system, including the Arizona State Standards, 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 shows the results of the summer 2016 and fall 2017 administrations of the high school end-of-course (EOC) assessments, and Chapter 4 shows the results of the spring 2017 administration of the full AzMERIT assessment system, including end-of-course (EOC) assessments in English language arts (ELA) and mathematics for grades 3–8 and high school. These chapters provide summaries of the test-taking student population and their performance on the assessments. Additionally, 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, specifically 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 for 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 hand-scored items, and it provides summary rater agreement results. Chapter 11 provides an overview of the quality assurance (QA) 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 if 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 or invalid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Determining whether the test measures the intended construct is central to evaluating the validity of test score interpretations. Such an evaluation in turn requires a clear definition of the measurement construct. For the AzMERIT, the Arizona State Standards provides the definition of the measurement construct.

In 2010, Arizona adopted new academic content standards in ELA and mathematics. The Arizona State Standards are designed to ensure that students across grades are receiving the instruction they need to be on track for college and careers by the time they graduate.<sup>1</sup> In spring 2015, the ADE administered AzMERIT to assess proficiency on the new Arizona State Standards for the first time. The AzMERIT measures ELA and mathematics in grades 3–8 and, for high school students, follows the completion of coursework in ELA grades 9–11, as well as Algebra I, Geometry, and Algebra II.

Because measuring student achievement directly against each benchmark in the Arizona State Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the Arizona State Standards.<sup>2</sup> To ensure that each student is assessed on the intended breadth and depth of the Arizona State Standards, 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.<sup>3</sup> 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 Arizona State Standards is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzMERIT ELA and mathematics test blueprints that specify the distribution of items across reporting strands and Depth of Knowledge (DOK) levels. The ELA and mathematics blueprints are also provided in Appendix B.

---

<sup>1</sup> 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.

<sup>2</sup> 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 test-taker population.

<sup>3</sup> Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended test-taker 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).



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 mathematics 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.<sup>4</sup> 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 (TDS) provides a range of accessibility tools and accommodations to virtually all students for reducing construct-irrelevant barriers to accessing test content.<sup>5</sup> The range of accommodations, provided in the online testing environment, far exceeds the typical accommodations available in paper-based testing (PBT) administrations. Exhibits 1.2.1–1.2.5 list the accommodations and accessibility supports that are currently available for students taking the AzMERIT assessments online. Paper-pencil test forms are available as an accommodation for students testing in online schools should the accommodations provided online be insufficient to remove barriers to accessing test content. These include both large print and braille forms. Section 6.3 describes the available testing tools and accommodations for students testing online and on a paper-pencil form.

Test administrators (TAs) 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

---

<sup>4</sup> 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 test takers in the intended population.

<sup>5</sup> 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.

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 testing (PBT) and computer-based testing (CBT). 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 TA;
- 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 using American Sign Language (ASL);
- Repeating the scripted directions at student request;
- Answering questions about the scripted directions or the directions that students read on their own;
- Reading the test quietly to himself/herself, as long as other students are not disrupted; and
- Providing extended time (the 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 available to all students in all AzMERIT tests; these features cannot be disabled by TAs.

**Exhibit 1.2.1 Universal Testing Tools for CBT Available to All Students**

Universal Test Tool	Description
<b>Area Boundaries</b>	The student may click anywhere on the selected-response text or button for multiple-choice options.
<b>Expand/Collapse Passage</b>	The student may expand a passage for easier readability. Expanded passages can also be collapsed.
<b>Help</b>	The student may view the on-screen <i>Test Instructions and Help</i> .
<b>Highlighter</b>	The student may highlight text in a passage or item.
<b>Line Reader</b>	The student may track the line he or she is reading.
<b>Mark (Flag) for Review</b>	The student may mark an item for review so that it can be easily found later.
<b>Notes/Comments</b>	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 mathematics, comments are attached to a specific test item and available throughout the session.
<b>Pause and Restart</b>	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.
<b>Review Test</b>	The student may review the test before ending it.
<b>Strikethrough</b>	The student may cross out answer options for multiple-choice and multi-select items.
<b>System Settings</b>	The student may adjust the audio volume during the test.
<b>Text-to-Speech for Instructions</b>	The student may listen to test instructions.
<b>Tutorial</b>	The student may view a short video about each item type and how to respond.
<b>Writing Tools</b>	The student may use editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italics) for extended-response items.
<b>Zoom In/Zoom Out</b>	The student may zoom in to enlarge the font and images in the test and 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 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 Mathematics	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	Mathematics	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 to how a student accesses and demonstrates learning that do not substantially change the instructional level, content, or 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 (FEP) 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 is 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 (SPED) need, or language need and the accommodation(s) that are provided to the student during educational activities, including assessment. TAs 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 clues or suggestions, verbal or otherwise, 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 during AzMERIT testing 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 IEP calls for a testing accommodation that is not listed, TAs 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 Injured Students**

Accommodation	Description of Use
<b>Adult Transcription</b>	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, verbally or by gestures, directly in to the DEI or in to the paper-pencil booklet and then in to the Data Entry Interface (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 verbally or by gestures.
<b>Assistive Technology</b>	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 copies must be shredded. Any electronic copies must be deleted. This accommodation also requires adult transcription (see above for rules on adult transcription).
<b>Rest/Breaks</b>	Students may take breaks during testing sessions.

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 (RFEP) 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 accommodations.

The *upon student request* accommodations are required to be administered in a setting that does not disturb other students, such as a one-on-one setting or 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
<b>Read Aloud Test Content</b>	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and for the mathematics test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test maybe be provided upon student request.</p> <p>Reading aloud the content of the Reading portion of the ELA test is prohibited.</p>
<b>Rest/Breaks</b>	Students may take breaks during testing sessions.
<b>Simplified Directions</b>	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
<b>Translate Directions</b>	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.
<b>Translation Dictionary</b>	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
<b>Abacus</b>	Students with a visual impairment may use an abacus for any AzMERIT mathematics test without restrictions.
<b>Adult Transcription</b>	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 verbally or by gestures, directly in to the DEI or in to the paper-pencil booklet and then in to 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 verbally or by gestures in to the paper-pencil booklet.
<b>Assistive Technology</b>	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 copies must be shredded. Any electronic copies must be deleted. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
<b>Braille Test Booklet</b>	Provide a paper braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
<b>Large Print Test Booklet</b>	CBT – Either increase default zoom settings when a 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).
<b>Paper-Pencil Test Booklet</b>	CBT – Student’s IEP must indicate that student cannot enter his or her own responses on the computer and requires a paper-pencil test or Adult Transcription. The school will provide a Special Paper Version booklet for the student. The student’s responses must be entered directly into the DEI or transcribed into the paper-pencil booklet and then entered in to 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 achieving the Arizona State Standards, the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with the expectations for student learning specified in the academic standards.<sup>6</sup>

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 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 items and AIRCore items used in Arizona were

<sup>6</sup> 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.

developed to align with the Arizona State Standards. These items were all reviewed by the ADE, 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 State Standards were used. To supplement the AIRCore pool of items, a few previously developed Arizona items that also aligned to the Arizona State Standards 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.<sup>7</sup> 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 Arizona State Standards is evaluated, the alignment of test blueprints with the content standards is critical.

With the desired alignment of test blueprints to Arizona State Standards, 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, DOK, writing genre, and other criteria. 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, in which the item is reviewed by a group of American Institutes for Research (AIR) content-area experts;
- Content Review 1, in which the item is reviewed by an AIR content specialist;
- Edit, in which a copy editor checks the item for correct grammar/usage; and
- Senior Content Review, in which the item is reviewed 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 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 DOK level.
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter and considers language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward.

---

<sup>7</sup> Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended test-taker 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*, or *never*, unless absolutely necessary), and ends with a question.
- For selected-response items, the 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 clueing within the item.
- The score points for constructed-response items are clearly defined.
- For machine-scored constructed-response (MSCR) items, the 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 the 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 clearly and appropriately online; 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 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. Therefore, using the item statistics gathered in field testing to review item performance is an important step in constructing valid and equivalent operational test forms.

Additionally,, 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 rangefinding process prior to scoring in which 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 a flagged field-test item must be rejected or deemed the item eligible for inclusion in operational test administrations.

#### 1.4 EVIDENCE FOR INTERPRETATION OF PERFORMANCE STANDARDS

The alignment of test content to the Arizona State Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the Arizona State Standards. However, the interpretation of AzMERIT test scores rests fundamentally upon 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 Arizona State Standards. The cut score establishing the Proficient level of performance is the most critical because 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. Therefore, procedures used to adopt performance standards for the AzMERIT assessments are central to the validity of test score interpretations.<sup>8</sup>

Following the first operational administration of the AzMERIT in spring 2015, a standard-setting workshop was conducted to recommend a set of performance standards for reporting student achievement of the Arizona State Standards to the Arizona State Board of Education. 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 Arizona State Standards and student achievement to map the performance-level descriptors adopted by Arizona to an ordered-item booklet (OIB) 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.<sup>9</sup>

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

---

<sup>8</sup> 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.

<sup>9</sup> 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.

assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided with 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 can use benchmark information to locate performance standards that converge across assessment systems, the validity of test score interpretation is bolstered.

Additionally, 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 and 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 the 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 shows 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 for 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			Projected SBAC
	AzMERIT Proficient	Arizona ACT College-Ready	Arizona NAEP Proficient	
<b>ELA</b>				
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%
<b>Mathematics</b>				
Grade 3	42%			39%
Grade 4	42%		42%	38%
<b>Mathematics</b>				
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 mathematics 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 on the AzMERIT scale.

To directly examine 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 shown in Section 9.5.2.

Exhibit 1.4.2 shows the location of the ACT college-ready cut scores for mathematics and reading on the AzMERIT scale. The first column shows the location as identified via indirect linkage through AIMS, and this 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 was 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

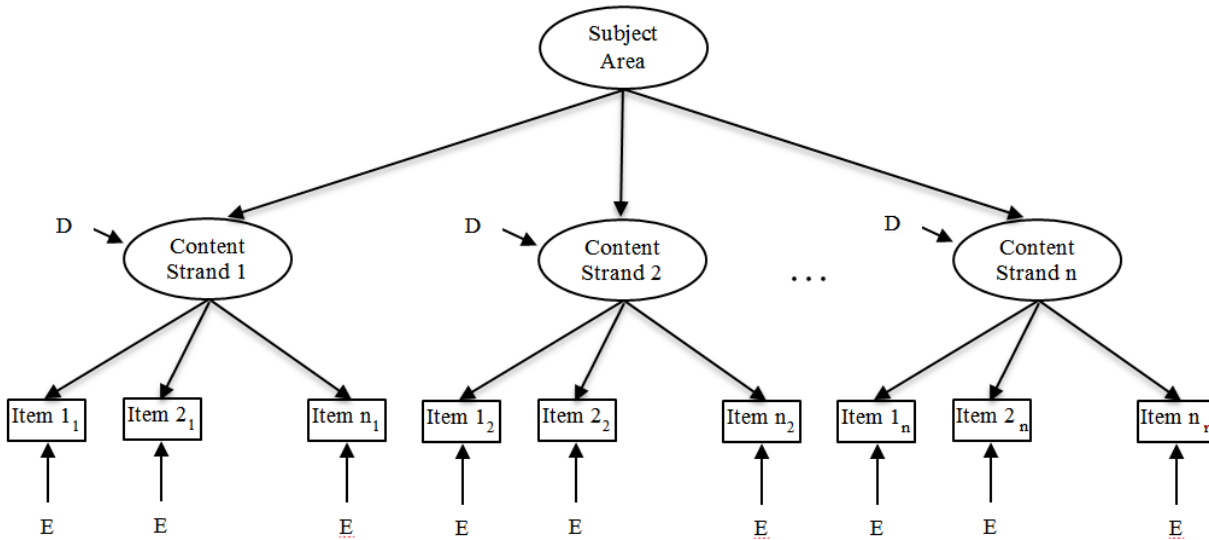
	Location of ACT College-Ready Cut on AzMERIT Scale		AzMERIT Meets Performance Standard
	Via Indirect Linkage through AIMS	Via Direct Linkage with AzMERIT	
Algebra II	3704	3727	3711
Grade 11 ELA	2579	2585	2585

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 mathematics was 3727 for the AzMERIT mathematics 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 (CFA) was used to evaluate the fit of this structural model to student response data.<sup>10</sup> 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  $\chi^2$  statistic, with good model fit indicated by a non-significant  $\chi^2$  statistic. The  $\chi^2$  statistic is sensitive to sample size, however; even well-fitting models will demonstrate highly significant  $\chi^2$  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.

The AzMERIT assessments also claim to measure subject-area achievement using test items that probe student knowledge and skills across multiple DOKs. As with the content standards, the classification of items by DOK also represents a structural model that can be evaluated using CFA.<sup>11</sup> In this case, each item is an indicator of a DOK level first-order factor, and each DOK level is in turn an indicator of subject area achievement. Thus, CFA was used to evaluate the fit of this DOK 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 CFA 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 Arizona State Standards 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.<sup>12</sup>

### 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 the statistics indicate that the general-achievement model fits the data well. This pattern was true across all grades. The CFI and TLI values

<sup>10</sup> 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.

<sup>11</sup> 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.

<sup>12</sup> 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.

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
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 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  $\chi^2_{Diff}$  *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	$\chi^2$	<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 shown 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  $\chi^2_{Diff}$  *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	$\chi^2$	<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 MATHEMATICS CONTENT MODEL

As with ELA, structural analyses of the mathematics assessments began with an evaluation of fit for the first-order, general-achievement model in which all items are indicators of a common mathematics 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 mathematics are shown in Exhibit 1.5.3.1. All 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 Mathematics 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
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 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 Mathematics 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 DOK 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  $\chi^2_{Diff}$  *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	$\chi^2$	<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 MATHEMATICS 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 RMSEA estimates well below their .05 cut-off values. All 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 Mathematics 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 shown 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  $\chi^2_{Diff}$  *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	$\chi^2$	<i>df</i>	<i>p</i> value
3	331.40	3	<i>p</i> < .001
4	309.50	3	<i>p</i> < .001
5	14.90	3	<i>p</i> < .001
6	14.50	3	<i>p</i> < .001
7	236.60	3	<i>p</i> < .001
8	79.20	3	<i>p</i> < .001
Algebra I	20.10	3	<i>p</i> < .001
Algebra II	26.40	3	<i>p</i> < .001
Geometry	20.90	3	<i>p</i> < .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.<sup>13</sup>

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 measures. When both assessments measure student achievement in common subject areas, such as with locally administered and statewide assessments of mathematics achievement, we expect test scores among the common subject-area assessments to be substantially correlated. Additionally, we expect that the magnitude of observed correlations among test scores in different subject areas will be lower than correlations among test scores in a common subject area. Because the content domains assessed in ELA and mathematics tests are quite different, AzMERIT ELA test scores should correlate less well with locally administered assessments of mathematics than ELA. It is important to note, however, that test scores across subject areas and test systems nevertheless are 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 among test scores across subject areas to be lower than correlations among test scores within a subject area, we nevertheless expect correlations among test scores across subject areas to be quite high.

Exhibit 1.6.1 shows the correlations among 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 mathematics 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	Mathematics Sample Size	Mathematics 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

<sup>13</sup> 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.

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	Mathematics Sample Size	Mathematics 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 among a variety of local assessments and AzMERIT test scores for ELA and mathematics 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 3–8. These results likewise show good convergence among 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 when the parameters of a measurement model are statistically equivalent across groups.<sup>14</sup> 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 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 mathematics. Items composing 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

<sup>14</sup> Standard 3.15 – Test developers and publishers who claim that a test can be used with test takers 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.

“a” tables (e.g., tables B.1a, B.2a, etc.) show 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  $\chi^2$  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  $\leq 0.015$ ; 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-level and subject-area test. Because 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 for online and paper-pencil tests in each ELA assessment.

The null hypothesis of the  $\chi^2$  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  $\chi^2$  difference tests to sample size, we examined additionally 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 CFI and Root Mean Square Error of Approximation (RMSEA). CFI values  $\geq 0.90$  and RMSEA values  $\leq 0.08$  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 mathematics, CFI values ranged from 0.905 to 0.990, and RMSEA ranged from 0.010 to 0.058.

Although the  $\chi^2$  difference test ideally should be nonsignificant, all  $\chi^2$  difference tests were significant at  $\alpha = .05$  due to large sample sizes except Model B-5, where the  $\chi^2$  difference tests for most grades was nonsignificant or marginally significant at  $\alpha = .05$ . Despite significant  $\chi^2$  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 mathematics). 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 mathematics 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), SPED, 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 using computer-based testing (CBT) and students participating in paper-based testing (PBT) on paper-pencil forms. For students administered paper-pencil 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 PBT scores. The predicted PBT scores were used to identify matched samples of online and paper-pencil test takers.

To identify possible differential effects of mode across subgroups, we used the observed test score as the dependent variable and then covaried the predicted test score to isolate the effects of mode. The demographic variables of interest

include gender, EL status, SPED, free reduced lunch (FRL) status, 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. Additionally, gender was coded as 0 for male and 1 for female. EL was coded as 1 for students as EL and 0 for non-ELL. SPED was coded as 1 for students in a SPED program and 0 for students not attending any SPED program. FRL (or Social Economic Status; SES) was coded as 1 for students having FRL 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 among the specified demographic subgroups.

Although 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 SPED students relative to the general education population. Especially for the high school EOC tests, AzMERIT tests were more difficult for SPED students when administered a paper-pencil test than an online test.

Mode effects were more pronounced for low income students with respect to the mathematics assessments. Mathematics tests were generally more difficult for low income students when administered an online test than a paper-pencil test.

Mode effects were also more pronounced for LEP students than for the general education population in mathematics but not in ELA. However, the direction of this effect was inconsistent across grades. Online mathematics tests were more difficult than paper-pencil tests for LEP students in the lower grades, but paper-pencil mathematics 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	Native Hawaiian/Pacific Islander	Hispanic/Latino	American Indian	Special Education	Limited English Proficiency	Free/Reduced Lunch	Migrant
<b>ELA</b>											
Grade 3E	0.49									0.27	
Grade 4E											
Grade 5E											
Grade 6E								-0.61			
Grade 7E								0.50			
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			
<b>Mathematics</b>											
Grade 3M								0.57			
Grade 4M									0.52	-	-4.46
Grade 5M							-0.89			0.34	
Grade 6M		1.15	0.96				0.69		0.60	-0.31	
Grade 7M	-0.26									0.25	-2.87
Grade 8M		0.89					0.86		-0.58		
Algebra I						0.73		-0.80	-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 differential gain rates across subgroups limit the inferences that can be made about test score gains over time.<sup>15</sup> To address this issue, we examined rates of student growth across student gender, race/ethnicity, SPED, limited English proficiency (LEP), and low income status (Low Income).

<sup>15</sup> Standard 3.15 – Test developers and publishers who claim that a test can be used with test takers from specific subgroups are responsible for providing the information necessary 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

Exhibit 1.9.1 shows the mean test scores on the spring 2017 and 2018 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 mathematics and Algebra I is attenuated somewhat, and the distribution of student ability is somewhat less variable for this cohort, especially with respect to the spring 2018 Algebra I performance. In spring 2017, grade 8 students enrolled in Algebra I were required to participate in both assessments, but in spring 2018, 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.

**Exhibit 1.9.1 Test Score Stability and Performance Gains Overall**

Assessment 2017→2018	N	Spring 2017 Scale Score		Spring 2018 Scale Score		Change from 2017 to 2018		Percentage Scoring Lower		Correlation
		Mean	Std. Dev	Mean	Std. Dev	Mean	IRT based Standard Error	Expected	Observed	
<b>ELA</b>										
<b>G3E→G4E</b>	82,784	2,503	32.28	2,520	32.70	17	14.70	0.23	0.17	0.83
<b>G4E→G5E</b>	83,485	2,522	31.49	2,539	34.76	17	14.71	0.24	0.18	0.84
<b>G5E→G6E</b>	82,093	2,536	33.79	2,543	32.27	6	14.75	0.39	0.35	0.84
<b>G6E→G7E</b>	80,319	2,544	32.09	2,554	34.17	10	14.79	0.34	0.29	0.84
<b>G7E→G8E</b>	79,810	2,554	32.78	2,560	32.65	6	14.55	0.39	0.35	0.84
<b>G8E→G9E</b>	70,499	2,557	33.29	2,570	31.71	12	14.22	0.30	0.24	0.83
<b>G9E→G10E</b>	64,539	2,569	28.80	2,566	33.13	-3	13.93	0.55	0.55	0.83
<b>G10E→G11E</b>	58,187	2,568	28.86	2,568	29.66	1	14.09	0.48	0.46	0.80
<b>Mathematics</b>										
<b>G3M→G4M</b>	83,119	3,527	47.10	3,556	44.17	29	17.49	0.18	0.13	0.82
<b>G4M→G5M</b>	83,937	3,556	44.83	3,590	46.68	34	17.04	0.14	0.09	0.84
<b>G5M→G6M</b>	82,335	3,589	44.45	3,618	46.35	29	16.69	0.16	0.11	0.85
<b>G6M→G7M</b>	80,449	3,618	43.91	3,635	42.89	17	16.33	0.26	0.21	0.86
<b>G7M→G8M</b>	67,415	3,627	42.79	3,654	38.67	28	17.92	0.17	0.11	0.83
<b>G8M→Algebra I</b>	51,030	3,654	35.59	3,668	31.92	14	15.62	0.29	0.24	0.79
<b>Algebra I→Geometry</b>	56,182	3,678	35.90	3,689	38.73	11	15.85	0.34	0.30	0.81
<b>Geometry→Algebra II</b>	48,886	3,690	36.54	3,699	34.57	9	16.40	0.37	0.32	0.81

The exhibit also shows that the rate of achievement gain is somewhat higher for mathematics than ELA, and that, although gain rates decelerate across the school years, the rate of gains diminishes more rapidly for ELA than mathematics over time. For mathematics, large gains, typically three-quarters standard deviation (e.g., average gain of 29 scale score points in grade 3 mathematics is 67% of the 44-point standard deviation of student test scores), are observed through the middle school grades, dropping to about one-third standard deviation among administrations of the high school end-of-course assessments. For ELA, although elementary school gains are strong, by middle school, annual gains are between one-third

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.



to one-half standard deviation, and drop to about one-quarter standard deviation by high school, with no observed growth from grade 9 to 10 and from grade 10 to 11.

To evaluate differential growth across demographic subgroups, a series of regression analyses was conducted to predict 2018 test scores from 2017 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 and were coded as 0.

Exhibit 1.9.2 shows the standardized regression coefficient estimates of the differential effect on student's growth rate from 2017 to 2018 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 mathematics assessments, including standardized and unstandardized coefficients, the standard error of the unstandardized coefficient, and  $p$  value regardless of significance level.

Results under intercept indicate that females generally performed better than males for both ELA and mathematics across grades in spring 2018 test scores. With respect to ethnicity, Asian students generally performed better than white students in both ELA and mathematics. For all other ethnic group comparisons, the focal groups generally performed less well than whites. Special education (SPED) students, limited English proficient (LEP) students, and low-income students all performed less well than the general education population in both ELA and mathematics.

The slope represents the association between 2017 and 2018 test scores, controlling for demographic subgroups. The overall slope parameter indicates the rate of growth in test scores between 2017 and 2018. 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. SPED students generally showed lower rates of gain than general education students, but the pattern was reversed during elementary school mathematics assessments, with SPED students showing greater rates of gain. Limited English proficient (LEP) students showed lower rates of gain in both ELA and mathematics, but this effect seems to moderate in the high school grades, in which differential gain rates were much less pronounced. Differential gain rates for low income students were observed for both ELA and mathematics, which generally showed 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 mathematics and lower gain rates during elementary school grade assessments in ELA. African American and Hispanic students showed lower gain rates than whites in mathematics assessments.

**Exhibit 1.9.2.1 Standardized Regression Coefficient of Differential Growth from 2017 to 2018 Administration Across Subgroups: ELA**

<b>2017 Administration</b>	<b>G3E</b>	<b>G4E</b>	<b>G5E</b>	<b>G6E</b>	<b>G7E</b>	<b>G8E</b>	<b>G9E</b>	<b>G10E</b>
<b>2018 Administration</b>	<b>G4E</b>	<b>G5E</b>	<b>G6E</b>	<b>G7E</b>	<b>G8E</b>	<b>G9E</b>	<b>G10E</b>	<b>G11E</b>
<b>Intercept</b>								
Female	0.02	0.02	0.04	0.06	0.03	0.03	-0.01	0.05
SPED	-0.06	-0.08	-0.06	-0.08	-0.08	-0.08	-0.05	-0.06
LEP	-0.12	-0.11	-0.10	-0.08	-0.05	-0.02	-0.02	
Low Income	-0.03	-0.04	-0.02	-0.03	-0.03	-0.02	-0.03	-0.03
Asian	0.02	0.01	0.03	0.02	0.02	0.02		0.02
Hispanic	-0.05	-0.03	-0.04	-0.03	-0.03	-0.05	-0.06	-0.04
African American	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01
Native Hawaiian/Pacific Islander		0.00			-0.01			
American Indian	-0.05	-0.04	-0.03	-0.04	-0.04	-0.02	-0.03	-0.02
Multiple Ethnicities			0.00				0.00	
<b>Slope</b>	0.79	0.78	0.80	0.80	0.81	0.80	0.82	0.83
Female		-0.01			-0.01			-0.03
SPED		0.01		-0.03	-0.02	-0.01	-0.03	-0.03
LEP	-0.09	-0.06	-0.07	-0.05	-0.03	0.01	-0.01	
Low Income	-0.02	-0.01	-0.01	-0.02		-0.01	-0.01	-0.03
Asian		-0.01				0.01	-0.01	
Hispanic			-0.01	-0.02		-0.02	-0.01	-0.03
African American							-0.01	-0.01
Native Hawaiian/Pacific Islander								
American Indian	-0.01			-0.01	-0.01		-0.01	-0.01
Multiple Ethnicities								-0.01

Note: Only significant effects from the multiple regression models are presented in the table. Intercept ( $\beta_0$ ): Standardized average test score in 2018 administration. Slope ( $\beta_{10}$ ): Rate of gain from 2017 to 2018. 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 students.

**Exhibit 1.9.2.2 Standardized Regression Coefficient of Differential Growth from 2017 to 2018 Administration Across Subgroups:  
Mathematics**

<b>2017 Administration</b>	<b>G3M</b>	<b>G4M</b>	<b>G5M</b>	<b>G6M</b>	<b>G7M</b>	<b>G8M</b>	<b>Alg I</b>	<b>Geo</b>
<b>2018 Administration</b>	<b>G4M</b>	<b>G5M</b>	<b>G6M</b>	<b>G7M</b>	<b>G8M</b>	<b>Alg I</b>	<b>Geo</b>	<b>Alg II</b>
<b>Intercept</b>								
Female			-0.01	-0.01	-0.01	0.02	-0.03	0.02
SPED	-0.06	-0.07	-0.06	-0.07	-0.09	-0.07	-0.05	-0.05
LEP	-0.08	-0.06	-0.06	-0.06	-0.05	-0.04	-0.01	0.01
Low Income	-0.03	-0.04	-0.04	-0.03	-0.02	-0.06	-0.02	-0.04
Asian	0.02	0.03	0.02	0.02	0.01	0.01	0.01	0.03
Hispanic	-0.04	-0.03	-0.05	-0.05	-0.02	-0.03	-0.06	-0.04
African American	-0.02	-0.02	-0.02	-0.01	-0.01		-0.02	
Native Hawaiian/Pacific Islander			0.00				-0.01	
American Indian	-0.04	-0.03	-0.04	-0.04	-0.03	-0.02	-0.02	-0.03
Multiple Ethnicities			-0.01				-0.01	
<b>Slope</b>	0.78	0.86	0.82	0.84	0.86	0.79	0.84	0.84
Female	-0.01	-0.03			-0.02		-0.02	-0.02
SPED	0.02	-0.02		-0.04	-0.07	-0.03	-0.03	-0.03
LEP	-0.03	-0.04	-0.04	-0.05	-0.05	-0.03	-0.02	0.01
Low Income	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02	-0.01	-0.03
Asian			0.00	-0.01	0.01	0.01	0.01	0.01
Hispanic		-0.01	-0.01		-0.02	-0.01	-0.04	-0.04
African American	0.01				-0.01		-0.02	-0.01
Native Hawaiian/Pacific Islander								
American Indian		-0.02	-0.01	0.00	-0.02	-0.01	-0.01	-0.02
Multiple Ethnicities				0.01			-0.01	

Note: Only significant effects from the multiple regression models are presented in the table. Intercept ( $\beta_{00}$ ): Standardized average test score in 2018 administration. Slope ( $\beta_{10}$ ): Rate of gain from 2017 to 2019. 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 students.

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

Administration of the new AzMERIT online tests is untimed, so schools may flexibly schedule students to take the tests in computer labs throughout the testing window. Thus, students taking the same grade-level or 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 the 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 grades 4–8 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 among 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 in which  $p < .05$ . Generally, the results indicate that starting tests earlier in the week resulted in higher test scores. Tests started on Friday were consistently associated with impaired performance, but there were some exceptions. For example, students beginning the grade 7 ELA tests on Monday scored lower than students beginning on any other day than Friday. Generally, though, the pattern was pronounced.

Conversely, assessments that 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 (e.g., grades 9 and 10 ELA, for which Friday end dates were associated with greater scores).

Time-of-day effects were less consistent. For high school students taking ELA assessments, morning start times were associated with better performance than afternoon start times. For middle school students, later morning start times were associated with poorer performance than early morning or late afternoon start times. In grade 6, ELA tests with morning start times were associated with lower scores 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
<b>ELA</b>					
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
<b>Mathematics</b>					
Grade 4 Mathematics	-0.01	0.02	-0.02		
Grade 5 Mathematics	-0.02	0.01	-0.03	0.04	0.01
Grade 6 Mathematics	-0.03	0.01		0.03	0.01
Grade 7 Mathematics	-0.01	0.01	-0.04	0.06	
Grade 8 Mathematics		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 mathematics tests, later start times were generally associated with better performance. An exception to this pattern was observed for Algebra I, in which students who began 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 scores than on tests ending earlier in the day, but grade 6 ELA proved an exception, with tests ending in the early morning associated with the highest scores.

Additionally, 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. When use of vocabulary that is not relevant to the measured construct interferes with student ability to understand the test item, the item is not assessing the intended construct accurately. 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 on 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. For 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 mathematics indicated that providing a glossary impaired performance of EL students.

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

Results from the initial study were limited both by the grade levels assessed and by 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 EL students. 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 EL students 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, implementing a 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 EL students but should 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 ELA and mathematics 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 1.11.1 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			Mathematics		
		non-EL	EL	Total	non-EL	EL	Total
3	ENG Only	19,385	2,535	21,920	19,442	2,569	22,011
	ENG+SP	19,780	2,449	22,229	19,874	2,481	22,355
	No Gloss	19,616	2,532	22,148	19,678	2,563	22,241
	Total	58,781	7,516	66,297	58,994	7,613	66,607
4	ENG Only	19,800	2,425	22,225	19,897	2,450	22,347
	ENG+SP	20,014	2,520	22,534	20,121	2,545	22,666
	No Gloss	20,140	2,350	22,490	20,249	2,375	22,624
	Total	59,954	7,295	67,249	60,267	7,370	67,637
5	ENG Only	19,802	1,924	21,726	19,898	1,935	21,833
	ENG+SP	20,182	1,928	22,110	20,235	1,941	22,176
	No Gloss	20,046	1,906	21,952	20,133	1,920	22,053
	Total	60,030	5,758	65,788	60,266	5,796	66,062
6	ENG Only	19,682	1,380	21,062	19,716	1,397	21,113
	ENG+SP	20,016	1,343	21,359	20,083	1,361	21,444
	No Gloss	19,906	1,393	21,299	19,939	1,410	21,349
	Total	59,604	4,116	63,720	59,738	4,168	63,906
7	ENG Only	19,841	1,241	21,082	19,472	1,251	20,723
	ENG+SP	20,092	1,307	21,399	19,712	1,306	21,018
	No Gloss	19,954	1,316	21,270	19,635	1,323	20,958
	Total	59,887	3,864	63,751	58,819	3,880	62,699
8	ENG Only	20,098	1,044	21,142	17,018	1,048	18,066
	ENG+SP	20,419	1,118	21,537	17,365	1,108	18,473
	No Gloss	20,370	1,029	21,399	17,315	1,025	18,340
	Total	60,887	3,191	64,078	51,698	3,181	54,879
9 / Algebra I	ENG Only	16,243	548	16,791	18,482	561	19,043
	ENG+SP	16,477	589	17,066	18,676	595	19,271
	No Gloss	16,430	530	16,960	18,604	513	19,117
	Total	49,150	1,667	50,817	55,762	1,669	57,431
10 / Geometry	ENG Only	15,224	326	15,550	15,460	334	15,794
	ENG+SP	15,482	372	15,854	15,727	410	16,137
	No Gloss	15,279	323	15,602	15,688	357	16,045
	Total	45,985	1,021	47,006	46,875	1,101	47,976
11 / Algebra II	ENG Only	13,897	183	14,080	14,124	182	14,306
	ENG+SP	14,029	218	14,247	14,163	175	14,338
	No Gloss	13,990	209	14,199	14,082	208	14,290
	Total	41,916	610	42,526	42,369	565	42,934

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

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

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

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

$\beta_j$  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  $\beta_j$  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 coefficient  $\alpha$ s indicate whether the glossaries affect the construct being measured or if there is 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 mathematics 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 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 the 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 school students and gradually increased for the middle school students and high school 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 the English glossary were not significant except in grades 3, 4, and 9, and the main effects from the English glossary with Spanish translation were not significant except in grades 3, 4, and 6. This means that providing a glossary had virtually no effect for non-EL students in middle school and high school grades, but it had a small negative effect at the elementary school grades, which might be caused by distractions.

With respect to the mathematics 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, except for a slight performance gain for the English-only glossary on the Geometry assessment. The results support that use of the glossary also significantly improved

the performance of EL students in most of the mathematics tests, but use of the glossary did not impact the non-EL group except in the Geometry test.

**Exhibit 1.11.2 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
<b>Coefficient Estimates</b>									
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
<b>Standard Errors</b>									
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
<b>Z Statistics</b>									
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 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
<b>Coefficient Estimates</b>									
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
<b>Standard Errors</b>									
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
<b>Z Statistics</b>									
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 interpretations is ongoing. Nevertheless, sufficient evidence currently exists to support the principal claims for the test scores, including that AzMERIT test scores indicate the degree to which students have achieved the Arizona State Standards 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 Arizona State Standards, 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. Additionally, 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 (AzMERIT) to measure student mastery of the Arizona academic standards and progress toward college and career readiness. The AzMERIT measures student progress in English language arts (ELA) in grades 3–11, in mathematics 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 (AIR) to develop and administer the AzMERIT beginning in the spring of 2015. In accordance with state requirements, the AzMERIT was designed to<sup>16</sup>:

- Align to the academic standards adopted by the Arizona State Board of Education in 2016 (Arizona State Standards);
- Supply criterion-referenced summative assessments for grades 3–8, and criterion-referenced end-of-course (EOC) assessments in identified high school mathematics and ELA courses for implementation beginning in the 2014–2015 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 policymakers 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 in 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, and 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 Learners (ELs) and students with special needs;
- Respect Arizona’s local control of the selection of classroom instructional materials; and
- Satisfy assessment goals in a cost-efficient manner.

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

---

<sup>16</sup> 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 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.

administration, the AzMERIT for grades 3–8 has 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, were 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 STATE STANDARDS

In 2016, 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 State Standards in mathematics describe expectations for learning in grades K–8 and the first three high school courses (Algebra I, Geometry, Algebra II; Mathematics 1, 2, 3) plus specific standards that could be included in a fourth high school credit mathematics course. The Arizona State Standards in ELA describe the reading, writing, language, speaking, and listening skills that students should acquire from grades 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, but it is offered as a dual mode, online computer-based test (CBT) and paper-based test (PBT) 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 composed of two to three discrete test sessions. The AzMERIT operational item pools include a variety of selected-response items, machine-scored constructed-response (MSCR) items, and some hand-scored, constructed-response items in the paper-pencil mathematics forms where MSCR items could not readily be rendered for paper-based testing (PBT) administration. AzMERIT also includes essay responses. In spring 2016, a sample of online writing responses was hand-scored (100% double scoring with resolution of all discrepancies) for purposes of developing statistical models to machine score 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 or move around objects 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 discrepancies adjudicated in a resolution score. This sample of essay responses and writing scores was used to develop the statistical models used for machine-scoring the remaining online essay responses. All essays administered on paper-pencil tests were hand-scored. In addition, hand-scoring was required for a subset of mathematics items administered on paper, generally equation items, for which it was not possible to represent the item on paper in a way that allowed machine-scoring.

### 3. SUMMARY OF SUMMER 2017 AND FALL 2017 OPERATIONAL TEST ADMINISTRATION

The following tests were administered in summer and fall 2017:

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

Online summer 2017 administration of Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) occurred from June 12 to July 27, 2017, and the fall 2017 administration occurred from November 6 to December 1, 2017.

The scoring and reporting of the summer and fall 2017 assessments used the item parameters calibrated following the spring 2016 administration and the vertical scale and performance standards established in summer 2015. This section summarizes the operational test results for the summer and fall 2017 administration of the AzMERIT.

#### 3.1 STUDENT POPULATION AND PARTICIPATION

The assessment data for operational analyses included Arizona students who meet minimum attempt requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in English language arts (ELA) and mathematics is shown by assessment and subgroup in Exhibits 3.1.1 and 3.1.2 for summer 2017 and Exhibits 3.1.3 and 3.1.4 for fall 2017.<sup>17</sup>

**Exhibit 3.1.1 Number of Students Participating in ELA Assessments by Subgroups: Summer 2017**

Group	ELA 9	ELA 10	ELA 11
<b>All Students</b>	554	302	241
<b>Female</b>	216	128	104
<b>Male</b>	338	174	137
<b>African American</b>	25	19	13
<b>Asian</b>	5	6	7
<b>Native Hawaiian/Pacific Islander</b>	0	0	1
<b>Hispanic/Latino</b>	366	163	129
<b>American Indian or Alaskan</b>	47	39	19
<b>White</b>	105	72	69
<b>Multiple Ethnicities</b>	6	3	3
<b>Limited English Proficiency</b>	23	6	5
<b>Special Education</b>	56	19	13
<b>Free Reduced Lunch</b>	266	122	86

<sup>17</sup> 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 Mathematics Assessments by Subgroups: Summer 2017

Group	Algebra I	Geometry	Algebra II
All Students	940	777	611
Female	424	384	260
Male	516	393	351
African American	58	55	29
Asian	18	18	17
Native Hawaiian/Pacific Islander	1	1	1
Hispanic/Latino	576	477	337
American Indian or Alaskan	29	14	11
White	239	199	199
Multiple Ethnicities	16	12	16
Limited English Proficiency	38	12	3
Special Education	51	29	16
Free Reduced Lunch	276	274	140

Exhibit 3.1.3 Number of Students Participating in ELA Assessments by Subgroups: Fall 2017

Group	ELA 9	ELA 10	ELA 11
All Students	3,895	4,563	4,937
Female	1,706	2,121	2,386
Male	2,189	2,442	2,551
African American	262	246	314
Asian	47	84	82
Native Hawaiian/Pacific Islander	10	33	33
Hispanic/Latino	1,888	2,088	2,295
American Indian or Alaskan	218	260	247
White	1,344	1,713	1,737
Multiple Ethnicities	115	130	217
Limited English Proficiency	276	229	197
Special Education	354	362	357
Free Reduced Lunch	1,204	1,014	1,216

Exhibit 3.1.4 Number of Students Participating in Mathematics Assessments by Subgroups: Fall 2017

Group	Algebra I	Geometry	Algebra II
All Students	5,436	5,893	4,571
Female	2,499	2,842	2,236
Male	2,937	3,051	2,335
African American	375	344	249
Asian	99	114	113
Native Hawaiian/Pacific Islander	25	28	16
Hispanic/Latino	2,532	2,739	1,936

Group	Algebra I	Geometry	Algebra II
American Indian or Alaskan	314	245	310
White	1,896	2,215	1,737
Multiple Ethnicities	186	195	204
Limited English Proficiency	410	384	262
Special Education	352	326	183
Free Reduced Lunch	1,630	1,221	1,048

### 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 shown in Exhibit 3.2.1 for summer 2017 and Exhibit 3.2.2 for fall 2017.

Exhibit 3.2.1 Test Score Summary Statistics: Summer 2017

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
<b>ELA</b>					
9	554	2551	23.17	2621	2477
10	302	2549	28.83	2643	2483
11	241	2546	31.04	2663	2477
<b>Mathematics</b>					
Algebra I	940	3658	28.1	3787	3577
Geometry	777	3676	34.22	3796	3609
Algebra II	611	3689	30.64	3811	3629

Exhibit 3.2.2 Test Score Summary Statistics: Fall 2017

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
<b>ELA</b>					
9	3,895	2556	27.18	2664	2454
10	4,563	2557	28.78	2650	2464
11	4,937	2554	27.34	2663	2467
<b>Mathematics</b>					
Algebra I	5,436	3668	38.3	3787	3577
Geometry	5,893	3678	34.94	3819	3609
Algebra II	4,571	3695	33.64	3839	3629



The percentage of students in each performance level by grade and content area, as well as the percentages of students at or above Proficient are shown in Exhibit 3.2.3 for summer 2017 and Exhibit 3.2.4 for fall 2017.

**Exhibit 3.2.3 Percentage of Students in Performance Levels: Summer 2017**

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
<b>ELA</b>						
9	554	60	27	12	1	13
10	302	74	10	12	4	16
11	241	71	19	8	2	10
<b>Mathematics</b>						
Algebra I	940	54	26	18	3	21
Geometry	777	48	28	20	4	24
Algebra II	611	57	23	15	5	20

**Exhibit 3.2.4 Percentage of Students in Performance Levels: Fall 2017**

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
<b>ELA</b>						
9	3,895	50	27	19	4	23
10	4,563	63	15	16	6	22
11	4,937	67	20	9	3	13
<b>Mathematics</b>						
Algebra I	5,436	48	18	22	12	33
Geometry	5,893	47	27	22	5	27
Algebra II	4,571	51	17	24	7	31

### 3.3 STUDENT PERFORMANCE BY SUBGROUP

Exhibits 3.3.1 and 3.3.2 show the number and percentage of students in each grade and subject at each performance level, by several subcategories, including female, male, African American, Asian, Native Hawaiian/Pacific Islander, Native Hispanic/Latino, American Indian, White, Multiple Ethnicities, limited English proficiency (LEP), special education (SPED), and free reduced lunch (FRL) for summer 2017. Exhibits 3.3.3 and 3.3.4 show this information for fall 2017.

Exhibit 3.3.1 Number of Students at Each Performance Level by Subgroups: Summer 2017

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	331	130	201	19	4	0	219	31	55	3	19	47	165
	Partially Proficient	150	60	90	2	0	0	106	13	26	3	4	7	73
	Proficient	67	24	43	4	1	0	37	2	23	0	0	2	25
	Highly Proficient	6	2	4	0	0	0	4	1	1	0	0	0	3
10	Minimally Proficient	224	95	129	13	3	0	128	36	42	2	6	17	100
	Partially Proficient	29	14	15	3	0	0	12	3	10	1	0	1	9
	Proficient	37	14	23	2	3	0	19	0	13	0	0	1	12
	Highly Proficient	12	5	7	1	0	0	4	0	7	0	0	0	1
11	Minimally Proficient	171	74	97	9	5	0	101	13	43	0	5	10	66
	Partially Proficient	45	17	28	3	0	1	21	5	15	0	0	2	16
	Proficient	19	11	8	1	0	0	6	1	8	3	0	1	4
	Highly Proficient	6	2	4	0	2	0	1	0	3	0	0	0	0
Algebra I	Minimally Proficient	507	208	299	31	10	0	325	18	115	8	31	43	189
	Partially Proficient	240	116	124	18	3	0	153	8	54	4	5	6	48
	Proficient	169	90	79	8	0	1	91	2	60	4	2	2	37
	Highly Proficient	24	10	14	1	5	0	7	1	10	0	0	0	2
Geometry	Minimally Proficient	373	176	197	32	6	0	243	9	77	5	10	20	157
	Partially Proficient	217	121	96	17	6	1	133	1	58	1	2	7	51
	Proficient	156	76	80	5	6	0	83	3	54	5	0	2	48
	Highly Proficient	31	11	20	1	0	0	18	1	10	1	0	0	18
Algebra II	Minimally Proficient	346	141	205	24	5	0	187	8	112	10	2	10	106
	Partially Proficient	140	62	78	5	0	0	91	1	39	3	1	4	19
	Proficient	93	38	55	0	6	1	45	2	37	2	0	2	13
	Highly Proficient	32	19	13	0	6	0	14	0	11	1	0	0	2

**Note:** Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch; \*Indicates that fewer 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 to protect that subgroup from discrimination.

Exhibit 3.3.2 Percentage of Students at Each Performance Level by Subgroups: Summer 2017

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	60	60	59	76	80	0	60	66	52	50	83	84	62
	Partially Proficient	27	28	27	8	0	0	29	28	25	50	17	13	27
	Proficient	12	11	13	16	20	0	10	4	22	0	0	4	9
	Highly Proficient	1	1	1	0	0	0	1	2	1	0	0	0	1
	At or Above Proficient	13	12	14	16	20	0	11	6	23	0	0	4	11
10	Minimally Proficient	74	74	74	68	50	0	79	92	58	67	100	89	82
	Partially Proficient	10	11	9	16	0	0	7	8	14	33	0	5	7
	Proficient	12	11	13	11	50	0	12	0	18	0	0	5	10
	Highly Proficient	4	4	4	5	0	0	2	0	10	0	0	0	1
	At or Above Proficient	16	15	17	16	50	0	14	0	28	0	0	5	11
11	Minimally Proficient	71	71	71	69	71	0	78	68	62	0	100	77	77
	Partially Proficient	19	16	20	23	0	100	16	26	22	0	0	15	19
	Proficient	8	11	6	8	0	0	5	5	12	100	0	8	5
	Highly Proficient	2	2	3	0	29	0	1	0	4	0	0	0	0
	At or Above Proficient	10	13	9	8	29	0	5	5	16	100	0	8	5
Algebra I	Minimally Proficient	54	49	58	53	56	0	56	62	48	50	82	84	68
	Partially Proficient	26	27	24	31	17	0	27	28	23	25	13	12	17
	Proficient	18	21	15	14	0	100	16	7	25	25	5	4	13
	Highly Proficient	3	2	3	2	28	0	1	3	4	0	0	0	1
	At or Above Proficient	21	24	18	16	28	100	17	10	29	25	5	4	14
Geometry	Minimally Proficient	48	46	50	58	33	0	51	64	39	42	83	69	57
	Partially Proficient	28	32	24	31	33	100	28	7	29	8	17	24	19
	Proficient	20	20	20	9	33	0	17	21	27	42	0	7	18
	Highly Proficient	4	3	5	2	0	0	4	7	5	8	0	0	7
	At or Above Proficient	24	23	25	11	33	0	21	29	32	50	0	7	24
Algebra II	Minimally Proficient	57	54	58	83	29	0	55	73	56	63	67	63	76
	Partially Proficient	23	24	22	17	0	0	27	9	20	19	33	25	14
	Proficient	15	15	16	0	35	100	13	18	19	13	0	13	9
	Highly Proficient	5	7	4	0	35	0	4	0	6	6	0	0	1
	At or Above Proficient	20	22	19	0	71	100	18	18	24	19	0	13	11

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free Reduced Lunch; \*Indicates that fewer 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 to protect that subgroup from discrimination.

Exhibit 3.3.3 Number of Students at Each Performance Level by Subgroups: Fall 2017

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	1,949	752	1,197	155	13	3	1,136	145	439	50	153	298	782
	Partially Proficient	1,059	498	561	65	8	3	469	58	422	33	74	34	273
	Proficient	743	377	366	37	21	2	256	13	390	22	46	22	136
	Highly Proficient	144	79	65	5	5	2	27	2	93	10	3	-	13
10	Minimally Proficient	2,868	1,286	1,582	170	33	22	1,508	206	849	71	163	329	767
	Partially Proficient	693	352	341	43	10	3	286	26	299	26	29	15	116
	Proficient	744	347	397	23	32	4	242	23	395	25	31	14	108
	Highly Proficient	258	136	122	10	9	4	52	5	170	8	6	4	23
11	Minimally Proficient	3,323	1,488	1,835	243	39	18	1,706	190	964	153	148	325	922
	Partially Proficient	992	565	427	56	22	7	396	39	433	37	39	25	201
	Proficient	469	254	215	13	12	5	147	16	253	23	7	5	70
	Highly Proficient	153	79	74	2	9	3	46	2	87	4	3	2	23
Algebra I	Minimally Proficient	2,632	1,141	1,491	233	16	10	1,503	191	594	80	240	275	1,019
	Partially Proficient	994	460	534	70	13	6	472	67	333	31	61	41	314
	Proficient	1,169	586	583	53	42	6	406	47	562	51	64	30	238
	Highly Proficient	641	312	329	19	28	3	151	9	407	24	45	6	59
Geometry	Minimally Proficient	2,744	1,285	1,459	209	19	15	1,584	135	692	81	213	248	735
	Partially Proficient	1,574	797	777	81	29	8	677	54	672	50	98	50	298
	Proficient	1,292	644	648	49	41	3	411	42	697	48	65	25	159
	Highly Proficient	283	116	167	5	25	2	67	14	154	16	8	3	29
Algebra II	Minimally Proficient	2,353	1,115	1,238	175	24	5	1,210	208	600	126	144	155	777
	Partially Proficient	798	416	382	36	21	3	319	66	328	25	47	10	140
	Proficient	1,106	568	538	32	43	7	342	33	608	41	53	15	108
	Highly Proficient	314	137	177	6	25	1	65	3	201	12	18	3	23

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

Exhibit 3.3.4 Percentage of Students at Each Performance Level by Subgroups: Fall 2017

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	50	44	55	59	28	30	60	67	33	43	55	84	65
	Partially Proficient	27	29	26	25	17	30	25	27	31	29	27	10	23
	Proficient	19	22	17	14	45	20	14	6	29	19	17	6	11
	Highly Proficient	4	5	3	2	11	20	1	1	7	9	1	0	1
	At or Above Proficient	23	27	20	16	55	40	15	7	36	28	18	6	12
10	Minimally Proficient	63	61	65	69	39	67	72	79	50	55	71	91	76
	Partially Proficient	15	17	14	17	12	9	14	10	17	20	13	4	11
	Proficient	16	16	16	9	38	12	12	9	23	19	14	4	11
	Highly Proficient	6	6	5	4	11	12	2	2	10	6	3	1	2
	At or Above Proficient	22	23	21	13	49	24	14	11	33	25	16	5	13
11	Minimally Proficient	67	62	72	77	48	55	74	77	55	71	75	91	76
	Partially Proficient	20	24	17	18	27	21	17	16	25	17	20	7	17
	Proficient	9	11	8	4	15	15	6	6	15	11	4	1	6
	Highly Proficient	3	3	3	1	11	9	2	1	5	2	2	1	2
	At or Above Proficient	13	14	11	5	26	24	8	7	20	12	5	2	8
Algebra I	Minimally Proficient	48	46	51	62	16	40	59	61	31	43	59	78	63
	Partially Proficient	18	18	18	19	13	24	19	21	18	17	15	12	19
	Proficient	22	23	20	14	42	24	16	15	30	27	16	9	15
	Highly Proficient	12	12	11	5	28	12	6	3	21	13	11	2	4
	At or Above Proficient	33	36	31	19	71	36	22	18	51	40	27	10	18
Geometry	Minimally Proficient	47	45	48	61	17	54	58	55	31	42	55	76	60
	Partially Proficient	27	28	25	24	25	29	25	22	30	26	26	15	24
	Proficient	22	23	21	14	36	11	15	17	31	25	17	8	13
	Highly Proficient	5	4	5	1	22	7	2	6	7	8	2	1	2
	At or Above Proficient	27	27	27	16	58	18	17	23	38	33	19	9	15
Algebra II	Minimally Proficient	51	50	53	70	21	31	63	67	35	62	55	85	74
	Partially Proficient	17	19	16	14	19	19	16	21	19	12	18	5	13
	Proficient	24	25	23	13	38	44	18	11	35	20	20	8	10
	Highly Proficient	7	6	8	2	22	6	3	1	12	6	7	2	2
	At or Above Proficient	31	32	31	15	60	50	21	12	47	26	27	10	13

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native 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 a student is to achieve the same score or to 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 mathematics assessments is provided with respect to both classical and item response theory (IRT) indices of internal consistency of test scores, and decision accuracy and consistency of performance-level classifications.<sup>18</sup>

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 could achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.<sup>19</sup>

#### 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 mathematics assessments for summer 2017, and Exhibit 3.4.1.2 shows these estimates for fall 2017. Internal consistency estimates are uniformly above 0.8.

**Exhibit 3.4.1.1 Internal Consistency Reliabilities for AzMERIT Scores: Summer 2017**

Grade/Course	ELA		Mathematics	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.84	537	0.85	789
10/Geometry	0.88	831	0.87	1171
11/Algebra II	0.88	964	0.84	939

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

<sup>18</sup> 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.

<sup>19</sup> 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 2017**

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.88	738	0.92	1467
10/Geometry	0.88	828	0.88	1221
11/Algebra II	0.86	747	0.87	1132

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 the measurement of very low- and very high -performing students, the precision of test scores decreases near the tails of the ability distribution.

Exhibit 3.4.2.1 shows the standard errors of measurement for the AzMERIT ELA and mathematics assessments, with respect to the four AzMERIT performance standards for summer 2017, and Exhibit 3.4.2.2 for fall 2017. 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.<sup>20</sup> 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 Average Standard Error of Measure at Performance Level for ELA and Mathematics: Summer 2017**

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
<b>ELA</b>					
Grade 9 ELA	9.53	9.00	9.28	11.51	9.38
Grade 10 ELA	10.27	9.00	10.00	11.47	10.17
Grade 11 ELA	10.98	10.00	10.00	12.53	10.77
<b>Mathematics</b>					
Algebra I	11.35	10.00	10.06	13.03	10.84
Geometry	14.00	10.56	10.00	11.90	12.29
Algebra II	13.60	10.45	10.00	10.94	12.30

<sup>20</sup> 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. When cut scores are specified for selection or classification, the standard errors of measurement should be reported near each cut score.

Exhibit 3.4.2.2 Average Standard Error of Measure at Performance Level for ELA and Mathematics: Fall 2017

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 9 ELA	9.57	9.00	9.30	12.63	9.50
Grade 10 ELA	10.08	9.00	10.00	11.15	9.98
Grade 11 ELA	10.52	10.00	10.00	11.47	10.40
Algebra I	11.44	10.00	10.09	13.16	11.14
Geometry	13.92	10.50	10.00	11.87	12.18
Algebra II	13.55	10.46	10.00	10.85	12.08

### 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).<sup>21</sup> This index considers the consistency of classifications for the percentage of test takers 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 taken and the classifications that would be made based on 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 taken and the classifications that would be made based on 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 in the following sections. 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 directly the probability of consistent classification using the likelihood function. The likelihood function of the achievement attribute, designated  $\theta$ , 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

<sup>21</sup> 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.



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 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 based on the true score agrees with the decision made based on the form 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 shows the classification accuracy and consistency indexes for the summer 2017 administration of AzMERIT, and Exhibit 3.4.6.2 does the same for the fall 2017 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, but 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 2017**

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
<b>ELA</b>						
<b>9</b>	0.89	0.95	0.99	0.85	0.92	0.98
<b>10</b>	0.94	0.95	0.98	0.91	0.93	0.97
<b>11</b>	0.92	0.95	0.98	0.89	0.93	0.98
<b>Mathematics</b>						
<b>Algebra I</b>	0.89	0.93	0.99	0.85	0.91	0.98
<b>Geometry</b>	0.88	0.93	0.98	0.84	0.91	0.98
<b>Algebra II</b>	0.87	0.93	0.98	0.81	0.90	0.98

**Exhibit 3.4.6.2 Classification Accuracy and Consistency Indexes for Performance Standards: Fall 2017**

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
<b>ELA</b>						
<b>9</b>	0.91	0.93	0.98	0.87	0.90	0.96
<b>10</b>	0.92	0.93	0.97	0.88	0.90	0.95
<b>11</b>	0.91	0.94	0.98	0.87	0.91	0.97
<b>Mathematics</b>						
<b>Algebra I</b>	0.91	0.94	0.96	0.88	0.92	0.95
<b>Geometry</b>	0.89	0.93	0.98	0.86	0.90	0.97
<b>Algebra II</b>	0.88	0.92	0.97	0.86	0.91	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 individualized education plans [IEPs] SPED, and free or reduced lunch [FRL] for summer 2017, and Exhibits 3.4.7.3 and 3.4.7.4 show this data for fall 2017.<sup>22</sup> 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

<sup>22</sup> 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.

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.

**Exhibit 3.4.7.1 Internal Consistency Reliability by Subgroup: ELA Summer 2017**

Subgroup	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.84	536.65	0.88	831.42	0.88	963.72
Female	0.83	516.47	0.87	817.78	0.86	806.05
Male	0.84	550.39	0.88	846.22	0.89	1070.91
African American	0.84	533.72	0.91	1177.43	0.89	1075.53
Asian	0.92	1161.00	0.81	498.57	0.96	3375.24
Native Hawaiian/Pacific Islander	0.00	0.00	0.00	0.00	0.00	0.00
Hispanic/Latino	0.81	447.35	0.85	704.82	0.84	709.07
American Indian or Alaskan	0.83	530.86	0.71	360.52	0.83	648.54
White	0.89	815.67	0.91	1178.91	0.90	1199.00
Multiple Ethnicities	0.82	504.40	0.87	723.00	-1.36	42.33
Limited English Proficiency	0.60	222.12	0.44	185.37	0.50	255.20
Special Education	0.76	388.58	0.78	527.92	0.87	941.73
Free/Reduced Lunch	0.82	470.13	0.84	635.72	0.84	711.20

**Exhibit 3.4.7.2 Internal Consistency Reliability by Subgroup: Mathematics Summer 2017**

Subgroup	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.85	789.47	0.87	1171.26	0.84	939.12
Female	0.85	775.62	0.86	1033.41	0.86	1042.96
Male	0.85	789.15	0.88	1308.90	0.82	851.39
African American	0.86	925.03	0.78	734.60	0.45	307.31
Asian	0.94	2683.08	0.90	1267.00	0.94	2205.49
Native Hawaiian/Pacific Islander	0.00	0.00	0.00	0.00	0.00	0.00
Hispanic/Latino	0.82	667.65	0.87	1185.57	0.82	847.10
American Indian or Alaskan	0.83	664.19	0.91	1603.69	0.71	524.69
White	0.87	865.25	0.88	1098.75	0.84	947.28
Multiple Ethnicities	0.88	980.53	0.92	1663.82	0.86	1036.80
Limited English Proficiency	0.74	561.56	0.63	570.36	0.31	234.33
Special Education	0.70	450.80	0.80	883.86	0.76	772.87
Free/Reduced Lunch	0.82	695.20	0.90	1699.33	0.75	701.36

Exhibit 3.4.7.3 Internal Consistency Reliability by Subgroup: ELA Fall 2017

Subgroup	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.88	738	0.88	828	0.86	747
Female	0.88	729	0.87	762	0.84	684
Male	0.88	726	0.88	873	0.86	779
African American	0.86	666	0.87	773	0.81	587
Asian	0.88	786	0.88	834	0.88	947
Native Hawaiian/Pacific Islander	0.90	1073	0.91	1212	0.91	1267
Hispanic/Latino	0.85	604	0.86	720	0.84	671
American Indian or Alaskan	0.80	448	0.82	555	0.82	618
White	0.88	754	0.88	812	0.86	766
Multiple Ethnicities	0.90	959	0.88	837	0.85	754
Limited English Proficiency	0.87	718	0.87	821	0.86	862
Special Education	0.80	493	0.81	615	0.80	607
Free/Reduced Lunch	0.85	591	0.86	724	0.84	699

Exhibit 3.4.7.4 Internal Consistency Reliability by Subgroup: Mathematics Fall 2017

Subgroup	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.92	1467	0.88	1221	0.87	1132
Female	0.92	1472	0.87	1119	0.86	1043
Male	0.91	1455	0.88	1316	0.88	1218
African American	0.88	997	0.84	1014	0.80	832
Asian	0.92	1735	0.92	1544	0.91	1320
Native Hawaiian/Pacific Islander	0.93	1765	0.89	1508	0.87	895
Hispanic/Latino	0.89	1101	0.84	1015	0.82	855
American Indian or Alaskan	0.85	782	0.88	1278	0.75	654
White	0.92	1637	0.89	1184	0.89	1221
Multiple Ethnicities	0.92	1516	0.90	1300	0.86	1145
Limited English Proficiency	0.92	1572	0.85	1113	0.86	1074
Special Education	0.82	741	0.76	836	0.75	795
Free/Reduced Lunch	0.86	889	0.84	1009	0.76	702

### 3.4.8 SUBSCALE RELIABILITY

Coefficient alpha internal consistency reliability estimates associated with the subscales for the summer 2017 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 2017. 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 2017

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.62	0.69	0.72
Grade 10	0.78	0.58	0.73
Grade 11	0.69	0.69	0.78

Exhibit 3.4.8.2 Subscale Reliabilities: Algebra I & II Summer 2017

Grade	Algebra	Functions	Statistics
Algebra I	0.76	0.64	0.45
Algebra II	0.65	0.58	0.55

Exhibit 3.4.8.3 Subscale Reliabilities: Geometry Summer 2017

Grade	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.42	0.62	0.54	0.61

Exhibit 3.4.8.4 Subscale Reliabilities: ELA Grades 9–11 Fall 2017

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.70	0.74	0.75
Grade 10	0.78	0.63	0.73
Grade 11	0.66	0.69	0.70

Exhibit 3.4.8.5 Subscale Reliabilities: Algebra I & II Fall 2017

Grade	Algebra	Functions	Statistics
Algebra I	0.85	0.77	0.61
Algebra II	0.73	0.62	0.62

Exhibit 3.4.8.6 Subscale Reliabilities: Geometry Fall 2017

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.42	0.66	0.58	0.62

### 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 2017, and in Exhibits 3.5.4–3.5.6 for fall 2017. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.<sup>23</sup> 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. Please note that disattenuated correlation equals 1 if disattenuated correlation is greater than 1.

Exhibit 3.5.1 Subscale Observed and Disattenuated Intercorrelations: ELA Grades 9–11 Summer 2017

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.58		0.80	
	Writing & Language	0.48	0.50	0.64	0.68
10	Literature	0.67		0.96	
	Writing & Language	0.60	0.46	0.88	0.68
11	Literature	0.60		0.90	
	Writing & Language	0.63	0.63	0.91	0.90

Exhibit 3.5.2 Subscale Observed and Disattenuated Intercorrelations: Algebra I & Algebra II Summer 2017

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.71		0.88	
	Statistics	0.57	0.56	0.83	0.82

<sup>23</sup> 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.

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra II	Functions	0.64		0.95	
	Statistics	0.64	0.62	1.00	1.00

Exhibit 3.5.3 Subscale Observed and Disattenuated Intercorrelations: Geometry Summer 2017

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.64			1.00		
	Modeling with Geometry (MG)	0.61	0.64		1.00	1.00	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.66	0.68	0.62	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 2017

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.66		0.92	
	Writing & Language	0.61	0.60	0.82	0.80
10	Literature	0.67		0.95	
	Writing & Language	0.61	0.52	0.90	0.77
11	Literature	0.63		0.93	
	Writing & Language	0.59	0.59	0.86	0.85

Exhibit 3.5.5 Subscale Observed and Disattenuated Intercorrelations: Algebra I & Algebra II Fall 2017

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	0.69	0.64	1.00	1.00

Exhibit 3.5.6 Subscale Observed and Disattenuated Intercorrelations: Geometry Fall 2017

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.66			1.00		
	Modeling with Geometry (MG)	0.61	0.63		1.00	1.00	

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
	Similarity, Right Triangles and Trigonometry (SRTT)	0.66	0.69	0.64	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 2018 OPERATIONAL TEST ADMINISTRATION

The following Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) assessments were administered in spring 2018:

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

Online administration of the AzMERIT occurred from April 2–27, 2018. The paper-pencil version of the AzMERIT was administered from April 2–10, 2018.

In the spring 2015 administration, item parameters for the mathematics 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 2018, the mathematics 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 mathematics tests in spring 2018.

In the spring 2015 administration, item parameters for the English language arts (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 2018, 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 2018 final scoring and reporting. This section summarizes the operational test results for the spring 2018 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 attempt requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and mathematics is presented in Exhibits 4.1.1 and 4.1.2 by assessment and subgroup.<sup>24</sup> 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

<sup>24</sup> 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.



are enrolled in Algebra I courses. The tables in Appendix F show the demographic composition of test takers by mode of test administration.

**Exhibit 4.1.1 Number of Students Participating in ELA Assessments by Subgroups: Spring 2018**

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
<b>All Students</b>	85,716	88,835	89,412	87,946	86,361	85,449	81,832	74,420	69,405
<b>Female</b>	41,673	43,783	43,943	43,196	42,703	41,528	40,420	37,034	34,846
<b>Male</b>	44,043	45,052	45,469	44,750	43,658	43,921	41,412	37,386	34,559
<b>African American</b>	2,247	2,352	2,343	2,287	2,182	2,394	2,612	2,358	2,146
<b>Asian</b>	2,856	2,905	2,882	2,851	2,821	2,843	2,823	2,498	2,452
<b>Native Hawaiian/Pacific Islander</b>	151	171	202	178	156	185	212	147	190
<b>Hispanic/Latino</b>	39,419	41,474	41,103	39,897	38,807	38,347	36,444	33,005	30,867
<b>American Indian or Alaskan</b>	4,500	4,560	4,634	4,521	4,523	4,372	4,160	3,627	3,157
<b>White</b>	34,740	35,550	36,425	36,506	36,344	35,933	34,741	31,606	29,480
<b>Multiple Ethnicities</b>	1,800	1,818	1,815	1,705	1,528	1,375	829	1,173	1,104
<b>Limited English Proficiency</b>	7,191	8,032	7,690	5,995	4,658	4,402	4,987	3,590	2,985
<b>Special Education</b>	9,883	10,759	10,812	10,051	9,394	8,824	7,012	6,043	5,325
<b>Free/Reduced Lunch</b>	39,613	41,732	41,659	40,152	38,573	37,239	26,686	23,758	20,715
<b>Accommodation</b>	5,643	6,075	5,643	4,737	4,180	3,752	1,707	1,115	890

**Exhibit 4.1.2 Number of Students Participating in Mathematics Assessments by Subgroups: Spring 2018**

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
<b>All Students</b>	86,041	89,200	89,715	88,192	86,429	73,666	87,346	71,044	64,619
<b>Female</b>	41,783	43,920	44,062	43,323	42,713	35,608	42,807	35,524	33,052
<b>Male</b>	44,258	45,280	45,653	44,869	43,716	38,058	44,539	35,520	31,567
<b>African American</b>	2,278	2,381	2,376	2,315	2,206	2,252	2,741	2,250	1,939
<b>Asian</b>	2,866	2,916	2,894	2,823	2,712	1,903	2,924	2,361	2,403
<b>Native Hawaiian/Pacific Islander</b>	153	174	203	180	155	159	228	185	190
<b>Hispanic/Latino</b>	39,559	41,640	41,260	40,027	38,946	34,336	40,146	30,604	28,560
<b>American Indian or Alaskan</b>	4,515	4,602	4,652	4,559	4,549	4,184	4,204	3,713	2,903
<b>White</b>	34,859	35,659	36,505	36,579	36,343	29,680	36,031	30,915	27,621
<b>Multiple Ethnicities</b>	1,811	1,827	1,825	1,709	1,519	1,152	1,072	1,009	999
<b>Limited English Proficiency</b>	7,243	8,085	7,736	6,034	4,706	4,017	5,333	3,272	2,710
<b>Special Education</b>	9,986	10,851	10,881	10,138	9,468	8,591	7,667	5,467	3,983
<b>Free/Reduced Lunch</b>	39,758	41,932	41,785	40,276	38,673	33,400	30,666	23,457	19,162
<b>Accommodation</b>	5,666	6,040	5,549	4,641	3,913	3,396	1,526	996	532

## 4.2 CLASSICAL ITEM ANALYSIS

Because AzMERIT is an online assessment system, classical item analysis statistics for selected-response and constructed-response 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 assessment, and following processing of answer documents for paper-based testing (PBT) administrations. Classical item analyses examine the degree to which the items function as intended with respect to the underlying scales. For online and paper-based test administrations, quality assurance (QA) reports provide the required item and test statistics for each selected-response and constructed-response 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 test takers 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 item response theory- (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 2018 online administration of AzMERIT. As indicated in Exhibit 4.2.1, the ELA items were somewhat harder than the mathematics 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 mathematics assessments. While mean difficulty of ELA items is relatively consistent across grade-level assessments, the average difficulty of mathematics items increases across grade level and course assessments. The proportion of students responding correctly to test items in the EOC assessments in mathematics was relatively 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 2018 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
<b>ELA</b>				
3	0.52	0.19	0.44	0.13
4	0.52	0.18	0.45	0.11
5	0.56	0.17	0.47	0.10

Grade	Average $p$ -Value	$p$ -Value SD	Average Point-Biserial	Point-Biserial SD
6	0.51	0.20	0.45	0.13
7	0.53	0.19	0.47	0.12
8	0.52	0.17	0.46	0.13
9	0.54	0.13	0.45	0.12
10	0.50	0.17	0.45	0.11
11	0.51	0.18	0.41	0.14
<b>Mathematics</b>				
3	0.63	0.17	0.54	0.09
4	0.58	0.17	0.52	0.08
5	0.53	0.16	0.53	0.1
6	0.49	0.21	0.52	0.1
7	0.49	0.2	0.51	0.1
8	0.42	0.18	0.47	0.11
Algebra I	0.44	0.17	0.46	0.11
Geometry	0.36	0.17	0.48	0.09
Algebra II	0.31	0.18	0.43	0.08

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

Grade	Proportion Correct	Biserial/Polyserial Correlation	Biserial Correlation for Distractor	Differential Item Functioning
<b>ELA</b>				
3	0	0	0	2
4	0	0	1	1
5	0	0	0	0
6	0	0	1	0
7	0	0	0	0
8	0	1	1	0
9	0	0	2	1
10	0	0	0	0
11	0	2	0	1
<b>Mathematics</b>				
3	0	0	0	0
4	0	0	0	1
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
Algebra I	2	0	0	0
Geometry	0	0	2	0

### 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  $\theta$  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 2018 operational test items. Because 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 items include traditional four-option multiple-choice items, technology-enhanced selected-response items, which may require students to select one or more options, and MSCR items, for which students' constructed-response items 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
<b>ELA</b>									
3	41	-0.09	1.05	-	-	-	1.62	1.73	-1.22
4	41	0.12	0.76	-	-	-	3.84	4.11	-0.10
5	41	0.00	0.72	-	-	-	2.45	2.92	-0.89
6	41	0.03	0.99	-	-	-	2.35	2.82	-1.36
7	41	0.02	0.94	-	-	-	2.24	2.49	-1.16
8	41	0.04	0.92	-	-	-	1.04	1.23	-1.55
9	43	0.03	0.53	-	-	-	1.41	1.61	-1.55
10	43	0.02	0.82	-	-	-	0.81	1.24	-2.13
11	43	-0.15	0.96	-	-	-	0.07	0.53	-1.81
<b>Mathematics</b>									
3	16	-0.05	1.28	29	0.18	1.01	-	-	-
4	14	-0.34	0.89	31	0.17	1.06	-	-	-
5	12	0.09	1.17	33	-0.02	0.96	-	-	-
6	12	-0.53	1.40	35	0.22	1.29	-	-	-
7	18	-0.70	1.10	29	0.44	0.87	-	-	-
8	23	-0.61	0.87	24	0.46	1.09	-	-	-
Algebra I	28	-0.19	0.85	19	0.42	1.01	-	-	-
Geometry	27	-0.51	0.88	20	0.56	1.05	-	-	-
Algebra II	26	-0.57	1.08	21	0.61	0.88	-	-	-

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 0.7	Between .7 - 1.3	Above 1.3	Below 0.7	Between .7 - 1.3	Above 1.3
<b>ELA</b>						
3	0	46	1	2	37	8
4	0	45	2	1	40	6
5	0	47	0	0	44	3
6	0	47	0	3	42	2
7	0	46	1	2	40	5

Grade/ Course	Infit			Outfit		
	Below 0.7	Between .7 - 1.3	Above 1.3	Below 0.7	Between .7 - 1.3	Above 1.3
8	0	47	0	2	41	4
9	0	47	2	1	44	4
10	0	49	0	0	48	1
11	0	47	2	1	45	3
<b>Mathematics</b>						
3	0	43	2	4	32	9
4	0	42	3	5	35	5
5	0	40	5	1	34	10
6	1	42	4	7	32	8
7	0	44	3	9	32	6
8	0	45	2	3	34	10
Algebra I	0	47	0	3	37	7
Geometry	0	46	1	5	36	6
Algebra II	0	44	3	7	37	3

#### 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-Based**

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
<b>ELA</b>					
3	85,716	2502	33.32	2605	2395
4	88,835	2520	32.85	2610	2400
5	89,412	2538	34.97	2629	2419
6	87,946	2542	32.41	2641	2431
7	86,361	2553	34.43	2648	2438
8	85,449	2559	32.87	2658	2448
9	81,832	2567	32.41	2664	2454
10	74,420	2564	33.52	2668	2458
11	69,405	2566	30.17	2675	2465
<b>Mathematics</b>					
3	86,041	3528	47.61	3605	3395
4	89,200	3555	44.51	3645	3435

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
5	89,715	3589	46.87	3688	3478
6	88,192	3617	46.56	3722	3512
7	86,429	3634	43.03	3739	3529
8	73,666	3654	39.34	3776	3566
Algebra I	87,346	3673	37.06	3787	3577
Geometry	71,044	3687	39.36	3819	3609
Algebra II	64,619	3698	35.12	3839	3629

Exhibit 4.4.2 Test Score Summary Statistics: Online

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
<b>ELA</b>					
3	71,562	2502	33.42	2605	2395
4	74,397	2519	32.83	2610	2400
5	74,970	2538	35.12	2629	2419
6	73,437	2542	32.45	2641	2431
7	71,436	2552	34.36	2648	2438
8	70,990	2558	32.94	2658	2448
9	69,227	2566	31.95	2664	2454
10	63,174	2564	33.25	2668	2458
11	59,553	2565	29.75	2675	2465
<b>Mathematics</b>					
3	71,852	3529	47.34	3605	3395
4	74,754	3555	44.43	3645	3435
5	75,221	3589	46.83	3688	3478
6	73,708	3617	46.58	3722	3512
7	71,605	3633	43.12	3739	3529
8	61,535	3654	39.4	3776	3566
Algebra I	74,288	3673	36.93	3787	3577
Geometry	59,877	3687	39.21	3819	3609
Algebra II	54,351	3697	34.49	3839	3629

Exhibit 4.4.3 Test Score Summary Statistics: Paper-Based (Paper-Pencil + Data Entry Interface [DEI])

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
<b>ELA</b>					
3	14,155	2504	32.73	2605	2395
4	14,439	2523	32.78	2610	2400

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
5	14,442	2539	34.16	2629	2421
6	14,509	2546	31.93	2641	2431
7	14,926	2558	34.32	2648	2438
8	14,460	2563	32.26	2658	2455
9	12,605	2572	34.45	2664	2454
10	11,246	2567	34.87	2668	2469
11	9,852	2571	32.17	2675	2465
<b>Math</b>					
3	14,190	3527	48.95	3605	3395
4	14,448	3555	44.91	3645	3435
5	14,495	3590	47.04	3688	3478
6	14,486	3620	46.37	3722	3512
7	14,827	3639	42.39	3739	3529
8	12,133	3656	39.06	3776	3566
Algebra I	13,058	3672	37.77	3787	3577
Geometry	11,167	3688	40.12	3819	3609
Algebra II	10,268	3701	38.08	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-Based**

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
<b>ELA</b>						
3	85,716	43	13	30	14	43
4	88,835	40	12	34	14	47
5	89,412	30	22	33	15	47
6	87,946	38	23	33	6	39
7	86,361	36	19	35	9	45
8	85,449	40	22	29	9	39
9	81,832	35	24	28	13	41
10	74,420	53	14	20	13	33
11	69,405	53	18	20	9	29
<b>Mathematics</b>						
3	86,041	23	24	31	22	53
4	89,200	26	27	34	13	47
5	89,715	28	25	27	19	47
6	88,192	37	20	27	16	43



Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
7	86,429	45	19	19	17	36
8	73,666	49	20	19	12	31
Algebra I	87,346	42	18	27	12	39
Geometry	71,044	39	24	27	10	37
Algebra II	64,619	44	21	25	9	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
<b>ELA</b>						
3	71,562	44	13	29	14	43
4	74,397	41	12	33	13	46
5	74,970	30	22	33	15	47
6	73,437	39	23	32	6	38
7	71,436	37	19	34	9	43
8	70,990	40	22	29	9	38
9	69,227	35	25	29	12	40
10	63,174	54	14	20	12	32
11	59,553	54	18	20	8	28
<b>Mathematics</b>						
3	71,852	23	24	31	22	53
4	74,754	26	27	34	13	47
5	75,221	28	25	27	19	46
6	73,708	37	20	27	16	43
7	71,605	46	19	18	17	35
8	61,535	49	20	19	12	31
Algebra I	74,288	42	19	27	12	39
Geometry	59,877	39	24	26	10	36
Algebra II	54,351	45	22	25	8	33

Exhibit 4.4.6 Percentage of Students in Performance Levels: Paper-Based (Paper-Pencil + DEI)

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
<b>ELA</b>						
3	14,155	40	14	32	14	46
4	14,439	35	13	37	15	52
5	14,442	30	23	35	13	48
6	14,509	33	24	37	6	43

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
7	14,926	30	19	39	12	51
8	14,460	35	22	32	10	42
9	12,605	31	22	28	18	46
10	11,246	50	14	20	17	36
11	9,852	48	18	21	13	35
<b>Mathematics</b>						
3	14,190	25	22	30	23	53
4	14,448	26	26	35	13	48
5	14,495	27	25	28	20	48
6	14,486	35	19	28	18	46
7	14,827	40	20	20	20	40
8	12,133	47	22	19	12	31
Algebra I	13,058	43	18	28	12	40
Geometry	11,167	39	22	28	10	38
Algebra II	10,268	42	19	26	13	39

#### 4.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibits 4.5.1 through 4.5.4 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, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian, White, and Multiple Ethnicities, and other demographic information such as special education (SPED), LEP, free reduced lunch (FRL), and accommodation.

Exhibit 4.5.1 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based ELA

Grade	Performance Level	Demographic Information													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	37,165	16,415	20,721	1,220	644	80	20,942	2,975	10,661	612	7,594	6,412	21,563	4,738
	Partially Proficient	11,314	5,604	5,709	326	311	15	5,378	577	4,431	275	727	432	5,516	409
	Proficient	25,300	12,981	12,315	535	1,072	40	9,875	776	12,392	606	1,158	319	9,574	430
	Highly Proficient	11,973	6,674	5,299	166	829	16	3,225	172	7,257	307	404	28	2,962	87
4	Minimally Proficient	35,867	16,103	19,744	1,188	537	54	21,039	2,935	9,546	543	8,234	7,000	21,450	4,949
	Partially Proficient	11,033	5,505	5,527	311	239	23	5,536	570	4,120	233	784	563	5,627	467
	Proficient	29,787	15,488	14,297	671	1,192	70	11,806	882	14,445	718	1,377	441	11,742	563
	Highly Proficient	12,172	6,687	5,485	182	937	24	3,093	173	7,439	324	364	28	2,914	73
5	Minimally Proficient	27,009	11,384	15,603	958	373	67	15,823	2,415	7,006	338	7,827	5,994	16,539	4,250
	Partially Proficient	20,038	10,050	9,987	529	415	49	10,450	1,082	7,107	405	1,496	1,311	10,572	847
	Proficient	29,384	15,335	14,046	667	1,181	62	11,572	977	14,218	703	1,180	367	11,499	458
	Highly Proficient	13,007	7,174	5,833	189	913	24	3,258	160	8,094	369	309	18	3,049	75
6	Minimally Proficient	33,437	14,175	19,247	1,146	408	69	19,076	2,710	9,486	526	7,963	5,353	19,594	3,874
	Partially Proficient	20,645	10,409	10,233	583	502	44	9,912	1,045	8,145	411	1,229	518	9,891	551
	Proficient	28,868	15,607	13,259	507	1,469	52	9,930	724	15,549	635	781	121	9,707	275
	Highly Proficient	5,016	3,005	2,011	51	472	13	979	42	3,326	133	78	3	960	30
7	Minimally Proficient	31,107	12,692	18,400	987	375	52	17,587	2,741	8,936	414	7,529	4,151	18,022	3,458
	Partially Proficient	16,719	8,631	8,088	452	364	28	8,122	843	6,617	293	1,016	352	8,081	407
	Proficient	30,395	16,534	13,858	638	1,281	61	11,326	835	15,626	624	752	143	10,805	273
	Highly Proficient	8,162	4,847	3,314	105	801	15	1,772	104	5,167	197	97	12	1,665	20
8	Minimally Proficient	33,807	14,055	19,735	1,189	486	80	18,876	2,830	9,877	452	7,358	3,850	18,935	3,140
	Partially Proficient	18,512	9,505	9,003	514	498	47	8,785	855	7,550	259	859	338	8,443	350
	Proficient	25,097	13,328	11,767	563	1,150	45	8,901	605	13,351	480	522	179	8,229	218
	Highly Proficient	8,060	4,641	3,417	128	709	13	1,786	82	5,156	184	85	35	1,632	38

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
9	Minimally Proficient	28,269	11,435	16,834	1,181	453	79	15,682	2,233	8,392	241	5,476	3,882	12,027	1,403
	Partially Proficient	19,994	10,156	9,838	644	464	59	9,635	1,094	7,896	201	978	675	7,006	195
	Proficient	23,306	12,696	10,610	577	968	49	8,803	696	11,947	264	454	378	6,155	71
	Highly Proficient	10,267	6,134	4,133	210	938	25	2,327	137	6,507	123	104	55	1,500	20
10	Minimally Proficient	39,520	18,339	21,181	1,495	673	76	21,263	2,731	12,708	569	5,293	3,100	15,686	1,021
	Partially Proficient	10,563	5,533	5,030	321	328	25	4,494	385	4,829	180	328	230	3,149	36
	Proficient	14,632	7,717	6,915	375	683	29	5,122	400	7,774	249	299	200	3,509	34
	Highly Proficient	9,706	5,446	4,260	167	814	17	2,127	111	6,295	175	123	60	1,414	21
11	Minimally Proficient	36,619	16,788	19,831	1,317	642	103	19,758	2,325	11,969	498	4,682	2,425	13,408	803
	Partially Proficient	12,653	7,013	5,640	367	414	30	5,432	472	5,732	205	357	289	3,642	38
	Proficient	13,928	7,659	6,269	337	758	36	4,464	303	7,759	270	227	224	2,902	35
	Highly Proficient	6,211	3,388	2,823	125	638	21	1,219	57	4,020	131	60	47	765	12

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

Exhibit 4.5.2 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based Mathematics

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	19,852	9,390	10,436	749	204	41	11,408	1,798	5,287	339	5,395	4,236	11,900	3,265
	Partially Proficient	20,420	10,495	9,918	644	331	30	10,786	1,317	6,915	390	2,176	1,820	10,870	1,320
	Proficient	26,650	13,228	13,420	604	907	55	11,610	1,041	11,841	590	1,592	975	11,422	832
	Highly Proficient	19,155	8,671	10,484	281	1,424	27	5,755	359	10,817	492	823	212	5,567	240
4	Minimally Proficient	23,580	11,271	12,308	906	250	31	13,848	2,142	6,012	389	6,689	5,246	14,384	3,876
	Partially Proficient	23,681	12,070	11,611	659	440	48	12,446	1,315	8,312	460	2,270	2,013	12,590	1,405
	Proficient	30,646	15,438	15,206	660	1,182	73	12,373	979	14,703	674	1,501	744	12,134	657
	Highly Proficient	11,298	5,141	6,157	156	1,044	22	2,974	166	6,632	304	391	82	2,826	97
5	Minimally Proficient	25,030	11,634	13,374	975	250	59	14,464	2,135	6,757	367	7,190	5,035	15,412	3,725
	Partially Proficient	22,767	11,825	10,939	669	416	53	11,650	1,323	8,208	445	2,069	1,837	11,873	1,105
	Proficient	24,651	12,867	11,780	505	844	55	10,361	910	11,434	538	1,088	731	10,071	537
	Highly Proficient	17,300	7,736	9,561	227	1,384	36	4,785	284	10,106	475	535	133	4,430	177
6	Minimally Proficient	32,355	15,622	16,719	1,242	320	76	18,447	2,657	9,079	520	7,757	4,810	19,028	3,604
	Partially Proficient	17,681	9,053	8,625	455	340	25	8,661	913	6,920	363	1,173	805	8,733	580
	Proficient	23,695	11,854	11,838	446	882	50	9,247	779	11,799	489	885	359	9,075	351
	Highly Proficient	14,483	6,795	7,688	172	1,281	29	3,672	211	8,781	337	323	60	3,442	103
7	Minimally Proficient	38,749	18,947	19,781	1,349	443	66	21,817	3,126	11,363	564	8,024	4,154	21,996	3,405
	Partially Proficient	16,451	8,419	8,032	398	374	36	7,404	717	7,218	304	798	357	7,372	307
	Proficient	16,224	8,037	8,187	261	649	25	5,952	472	8,542	323	420	147	5,737	141
	Highly Proficient	15,033	7,312	7,719	198	1,246	28	3,775	234	9,222	328	226	48	3,569	56
8	Minimally Proficient	35,746	16,751	18,976	1,371	436	80	19,596	2,833	10,899	512	7,360	3,423	19,443	2,931
	Partially Proficient	14,999	7,604	7,392	430	325	38	6,735	749	6,495	224	720	375	6,564	288
	Proficient	14,347	7,213	7,131	302	499	27	5,441	453	7,379	243	378	166	5,130	150
	Highly Proficient	8,608	4,045	4,563	149	643	14	2,569	152	4,908	173	134	54	2,264	52
Algebra I	Minimally Proficient	37,079	17,000	20,079	1,442	485	93	21,066	2,640	11,001	352	6,331	4,097	16,819	1,278
	Partially Proficient	16,107	8,424	7,683	537	358	45	7,666	757	6,570	174	721	670	5,658	155

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	Proficient	23,760	12,387	11,373	578	1,084	72	8,814	671	12,195	346	476	461	6,344	71
	Highly Proficient	10,402	4,998	5,404	184	997	18	2,601	136	6,266	200	139	105	1,846	22
Geometry	Minimally Proficient	27,896	13,394	14,502	1,196	345	72	14,946	2,217	8,713	403	4,231	2,177	11,896	824
	Partially Proficient	17,178	9,232	7,946	545	420	46	7,786	876	7,272	232	746	652	5,727	119
	Proficient	18,943	9,762	9,181	420	880	49	6,417	522	10,361	293	396	350	4,728	44
	Highly Proficient	7,030	3,137	3,893	89	717	18	1,457	98	4,569	81	95	94	1,108	9
Algebra II	Minimally Proficient	28,747	14,375	14,372	1,072	432	87	15,821	1,817	9,107	408	3,209	1,851	10,740	442
	Partially Proficient	13,684	7,395	6,289	400	361	41	6,135	605	5,909	232	442	451	4,150	47
	Proficient	16,389	8,684	7,705	377	810	45	5,507	423	8,956	271	274	343	3,571	34
	Highly Proficient	5,799	2,598	3,201	90	800	17	1,097	58	3,649	88	58	65	701	9

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

Exhibit 4.5.3 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based ELA

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	43	39	47	54	23	53	53	66	31	34	77	89	54	84
	Partially Proficient	13	13	13	15	11	10	14	13	13	15	7	6	14	7
	Proficient	30	31	28	24	38	26	25	17	36	34	12	4	24	8
	Highly Proficient	14	16	12	7	29	11	8	4	21	17	4	0	7	2
	At or Above Proficient	43	47	40	31	67	37	33	21	57	51	16	5	32	9
4	Minimally Proficient	40	37	44	51	18	32	51	64	27	30	77	87	51	81
	Partially Proficient	12	13	12	13	8	13	13	13	12	13	7	7	13	8
	Proficient	34	35	32	29	41	41	28	19	41	39	13	5	28	9
	Highly Proficient	14	15	12	8	32	14	7	4	21	18	3	0	7	1
	At or Above Proficient	47	51	44	36	73	55	36	23	62	57	16	6	35	10
5	Minimally Proficient	30	26	34	41	13	33	38	52	19	19	72	78	40	75
	Partially Proficient	22	23	22	23	14	24	25	23	20	22	14	17	25	15
	Proficient	33	35	31	28	41	31	28	21	39	39	11	5	28	8
	Highly Proficient	15	16	13	8	32	12	8	3	22	20	3	0	7	1
	At or Above Proficient	47	51	44	37	73	43	36	25	61	59	14	5	35	9
6	Minimally Proficient	38	33	43	50	14	39	48	60	26	31	79	89	49	82
	Partially Proficient	23	24	23	25	18	25	25	23	22	24	12	9	25	12
	Proficient	33	36	30	22	52	29	25	16	43	37	8	2	24	6
	Highly Proficient	6	7	4	2	17	7	2	1	9	8	1	0	2	1
	At or Above Proficient	39	43	34	24	68	37	27	17	52	45	9	2	27	6
7	Minimally Proficient	36	30	42	45	13	33	45	61	25	27	80	89	47	83
	Partially Proficient	19	20	19	21	13	18	21	19	18	19	11	8	21	10
	Proficient	35	39	32	29	45	39	29	18	43	41	8	3	28	7
	Highly Proficient	9	11	8	5	28	10	5	2	14	13	1	0	4	0
	At or Above Proficient	45	50	39	34	74	49	34	21	57	54	9	3	32	7
8	Minimally Proficient	40	34	45	50	17	43	49	65	27	33	83	87	51	84
	Partially Proficient	22	23	20	21	18	25	23	20	21	19	10	8	23	9
	Proficient	29	32	27	24	40	24	23	14	37	35	6	4	22	6
	Highly Proficient	9	11	8	5	25	7	5	2	14	13	1	1	4	1
	At or Above Proficient	39	43	35	29	65	31	28	16	52	48	7	5	26	7
9	Minimally Proficient	35	28	41	45	16	37	43	54	24	29	78	78	45	82
	Partially Proficient	24	25	24	25	16	28	26	26	23	24	14	14	26	11

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	Proficient	28	31	26	22	34	23	24	17	34	32	6	8	23	4
	Highly Proficient	13	15	10	8	33	12	6	3	19	15	1	1	6	1
	At or Above Proficient	41	47	36	30	68	35	31	20	53	47	8	9	29	5
10	Minimally Proficient	53	50	57	63	27	52	64	75	40	49	88	86	66	92
	Partially Proficient	14	15	13	14	13	17	14	11	15	15	5	6	13	3
	Proficient	20	21	18	16	27	20	16	11	25	21	5	6	15	3
	Highly Proficient	13	15	11	7	33	12	6	3	20	15	2	2	6	2
	At or Above Proficient	33	36	30	23	60	31	22	14	45	36	7	7	21	5
11	Minimally Proficient	53	48	57	61	26	54	64	74	41	45	88	81	65	90
	Partially Proficient	18	20	16	17	17	16	18	15	19	19	7	10	18	4
	Proficient	20	22	18	16	31	19	14	10	26	24	4	8	14	4
	Highly Proficient	9	10	8	6	26	11	4	2	14	12	1	2	4	1
	At or Above Proficient	29	32	26	22	57	30	18	11	40	36	5	9	18	5

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

Exhibit 4.5.4 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based Mathematics

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	23	22	24	33	7	27	29	40	15	19	54	58	30	58
	Partially Proficient	24	25	22	28	12	20	27	29	20	22	22	25	27	23
	Proficient	31	32	30	27	32	36	29	23	34	33	16	13	29	15
	Highly Proficient	22	21	24	12	50	18	15	8	31	27	8	3	14	4
	At or Above Proficient	53	52	54	39	81	54	44	31	65	60	24	16	43	19
4	Minimally Proficient	26	26	27	38	9	18	33	47	17	21	62	65	34	64
	Partially Proficient	27	27	26	28	15	28	30	29	23	25	21	25	30	23
	Proficient	34	35	34	28	41	42	30	21	41	37	14	9	29	11
	Highly Proficient	13	12	14	7	36	13	7	4	19	17	4	1	7	2
	At or Above Proficient	47	47	47	34	76	55	37	25	60	54	17	10	36	12



Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
5	Minimally Proficient	28	26	29	41	9	29	35	46	19	20	66	65	37	67
	Partially Proficient	25	27	24	28	14	26	28	28	22	24	19	24	28	20
	Proficient	27	29	26	21	29	27	25	20	31	29	10	9	24	10
	Highly Proficient	19	18	21	10	48	18	12	6	28	26	5	2	11	3
	At or Above Proficient	47	47	47	31	77	45	37	26	59	56	15	11	35	13
6	Minimally Proficient	37	36	37	54	11	42	46	58	25	30	77	80	47	78
	Partially Proficient	20	21	19	20	12	14	22	20	19	21	12	13	22	12
	Proficient	27	27	26	19	31	28	23	17	32	29	9	6	23	8
	Highly Proficient	16	16	17	7	45	16	9	5	24	20	3	1	9	2
	At or Above Proficient	43	43	44	27	77	44	32	22	56	48	12	7	31	10
7	Minimally Proficient	45	44	45	61	16	43	56	69	31	37	85	88	57	87
	Partially Proficient	19	20	18	18	14	23	19	16	20	20	8	8	19	8
	Proficient	19	19	19	12	24	16	15	10	24	21	4	3	15	4
	Highly Proficient	17	17	18	9	46	18	10	5	25	22	2	1	9	1
	At or Above Proficient	36	36	36	21	70	34	25	16	49	43	7	4	24	5
8	Minimally Proficient	49	47	50	61	23	50	57	68	37	44	86	85	58	86
	Partially Proficient	20	21	19	19	17	24	20	18	22	19	8	9	20	8
	Proficient	19	20	19	13	26	17	16	11	25	21	4	4	15	4
	Highly Proficient	12	11	12	7	34	9	7	4	17	15	2	1	7	2
	At or Above Proficient	31	32	31	20	60	26	23	14	41	36	6	5	22	6
Algebra I	Minimally Proficient	42	40	45	53	17	41	52	63	31	33	83	77	55	84
	Partially Proficient	18	20	17	20	12	20	19	18	18	16	9	13	18	10
	Proficient	27	29	26	21	37	32	22	16	34	32	6	9	21	5
	Highly Proficient	12	12	12	7	34	8	6	3	17	19	2	2	6	1
	At or Above Proficient	39	41	38	28	71	39	28	19	51	51	8	11	27	6
Geometry	Minimally Proficient	39	38	41	53	15	39	49	60	28	40	77	67	51	83
	Partially Proficient	24	26	22	24	18	25	25	24	24	23	14	20	24	12
	Proficient	27	27	26	19	37	26	21	14	34	29	7	11	20	4
	Highly Proficient	10	9	11	4	30	10	5	3	15	8	2	3	5	1
	At or Above Proficient	37	36	37	23	68	36	26	17	48	37	9	14	25	5
Algebra II	Minimally Proficient	44	43	46	55	18	46	55	63	33	41	81	68	56	83
	Partially Proficient	21	22	20	21	15	22	21	21	21	23	11	17	22	9
	Proficient	25	26	24	19	34	24	19	15	32	27	7	13	19	6
	Highly Proficient	9	8	10	5	33	9	4	2	13	9	1	2	4	2

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	<b>At or Above Proficient</b>	34	34	35	24	67	33	23	17	46	36	8	15	22	8

**Note:** Alaskan = Alaskan Native; Hawaiian/Pacific = Native 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 a student is 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 mathematics 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.<sup>25</sup>

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.<sup>26</sup>

### 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 2018 AzMERIT ELA and mathematics assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

<sup>25</sup> 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.

<sup>26</sup> 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 4.6.1.1 Internal Consistency Reliabilities for AzMERIT Scores**

Grade	ELA		Mathematics	
	Reliability	Variance	Reliability	Variance
<b>G3</b>	0.90	1117	0.92	2241
<b>G4</b>	0.90	1078	0.93	1974
<b>G5</b>	0.90	1233	0.93	2193
<b>G6</b>	0.90	1053	0.93	2170
<b>G7</b>	0.91	1181	0.93	1859
<b>G8</b>	0.91	1085	0.92	1552
<b>G9E / Algebra I</b>	0.90	1021	0.91	1364
<b>G10E / Geometry</b>	0.91	1106	0.91	1538
<b>G11E / Algebra II</b>	0.89	885	0.88	1190

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 mathematics 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.<sup>27</sup> 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 4.6.2.1 Average Standard Errors of Measurement at Performance Level Spring 2018: ELA**

Grade	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
<b>3</b>	10.08	9.64	10.32	12.95	10.54
<b>4</b>	9.96	9.00	9.81	13.39	10.33
<b>5</b>	10.16	9.22	10.54	14.46	10.84
<b>6</b>	10.54	9.23	10.39	13.61	10.40
<b>7</b>	10.38	9.19	10.40	13.82	10.53
<b>8</b>	9.95	9.00	10.14	12.19	10.04
<b>9</b>	9.89	9.00	9.71	12.77	10.01
<b>10</b>	9.92	9.00	10.06	12.34	10.15

<sup>27</sup> 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 near each cut score.

Grade	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
11	9.82	9.51	10.24	12.18	10.06

Exhibit 4.6.2.2 Average Standard Errors of Measurement at Performance Level Spring 2018: Mathematics

Grade	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
3	12.26	10.00	11.77	17.47	13.00
4	12.25	10.00	11.34	16.96	12.11
5	13.21	10.00	10.53	15.66	12.32
6	13.01	10.00	10.39	14.08	11.98
7	11.77	10.00	10.42	13.84	11.61
8	11.96	10.00	10.20	13.44	11.45
Algebra I	11.58	10.00	10.07	13.06	11.11
Geometry	13.40	10.49	10.00	12.81	11.83
Algebra II	13.66	10.84	10.00	10.98	12.02

### 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).<sup>28</sup> This index considers the consistency of classifications for the percentage of test takers 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 taken and the classifications that would be made based on 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 taken and the classifications that would be made based on 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 in the following sections. The true score is an expected value of the test score with measurement error.

<sup>28</sup> 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.

#### 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  $\theta$  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 based on the true score agrees with the decision made based on the form 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 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)

#### 4.6.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

Exhibit 4.6.6.1 shows the classification accuracy and consistency indexes for spring 2018 administration of the AzMERIT. Exhibit 4.6.6.2 and 4.6.6.3 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 SPED, FRL, 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
<b>ELA</b>						
<b>3</b>	0.92	0.92	0.94	0.88	0.88	0.92
<b>4</b>	0.92	0.92	0.95	0.88	0.89	0.93
<b>5</b>	0.93	0.91	0.94	0.90	0.88	0.92
<b>6</b>	0.92	0.92	0.97	0.88	0.88	0.96
<b>7</b>	0.92	0.91	0.96	0.89	0.88	0.94
<b>8</b>	0.92	0.92	0.95	0.89	0.89	0.93
<b>9</b>	0.92	0.91	0.95	0.89	0.88	0.93
<b>10</b>	0.92	0.93	0.95	0.89	0.90	0.93
<b>11</b>	0.91	0.92	0.96	0.87	0.89	0.94
<b>Mathematics</b>						
<b>3</b>	0.95	0.93	0.93	0.93	0.90	0.90
<b>4</b>	0.95	0.92	0.95	0.92	0.89	0.92
<b>5</b>	0.94	0.93	0.94	0.92	0.91	0.93
<b>6</b>	0.94	0.93	0.95	0.91	0.91	0.93
<b>7</b>	0.94	0.93	0.95	0.91	0.91	0.93
<b>8</b>	0.92	0.94	0.96	0.89	0.91	0.95
<b>Algebra I</b>	0.92	0.93	0.96	0.88	0.90	0.94
<b>Geometry</b>	0.91	0.93	0.97	0.87	0.91	0.96
<b>Algebra II</b>	0.89	0.92	0.97	0.85	0.89	0.96

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

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
<b>G3E</b>	Overall	0.92	0.92	0.94	0.88	0.88	0.92
	Female	0.92	0.92	0.94	0.88	0.88	0.91
	Male	0.92	0.92	0.95	0.89	0.89	0.93

Grade	Subgroup	Accuracy			Consistency			
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient	
	African American	0.91	0.92	0.96	0.87	0.89	0.95	
	Hispanic/Latino	0.91	0.92	0.96	0.88	0.89	0.95	
	Asian	0.93	0.92	0.91	0.90	0.88	0.88	
	White	0.92	0.91	0.92	0.89	0.88	0.89	
	Hawaiian/Pacific	0.93	0.92	0.95	0.89	0.89	0.93	
	American Indian	0.91	0.93	0.98	0.88	0.91	0.97	
	Multiple Ethnicities	0.91	0.91	0.93	0.88	0.87	0.91	
	LEP	0.95	0.97	1	0.93	0.96	0.99	
	SPED	0.95	0.96	0.98	0.93	0.94	0.97	
	FRL	0.91	0.92	0.96	0.87	0.89	0.95	
	Accommodations	0.95	0.97	0.99	0.93	0.95	0.99	
	Overall	0.92	0.92	0.95	0.88	0.89	0.93	
	Female	0.92	0.91	0.94	0.88	0.88	0.92	
	Male	0.92	0.92	0.95	0.88	0.89	0.94	
<b>G4E</b>	African American	0.91	0.92	0.97	0.87	0.89	0.96	
	Hispanic/ Latino	0.91	0.92	0.97	0.87	0.89	0.95	
	Asian	0.95	0.93	0.91	0.93	0.90	0.87	
	White	0.93	0.91	0.93	0.90	0.88	0.90	
	Hawaiian/Pacific	0.92	0.89	0.95	0.88	0.86	0.94	
	American Indian	0.90	0.93	0.98	0.86	0.90	0.97	
	Multiple Ethnicities	0.93	0.91	0.94	0.90	0.87	0.92	
	LEP	0.93	0.97	1	0.91	0.96	1	
	SPED	0.94	0.96	0.99	0.91	0.94	0.98	
	FRL	0.91	0.92	0.97	0.87	0.89	0.96	
	Accommodations	0.94	0.96	0.99	0.91	0.95	0.99	
	Overall	0.93	0.91	0.94	0.90	0.88	0.92	
	Female	0.93	0.91	0.94	0.90	0.88	0.91	
	Male	0.93	0.92	0.95	0.90	0.89	0.93	
<b>G5E</b>	African American	0.92	0.92	0.96	0.89	0.89	0.94	
	Hispanic/ Latino	0.91	0.92	0.96	0.88	0.88	0.95	
	Asian	0.96	0.92	0.90	0.94	0.90	0.86	
	White	0.94	0.91	0.92	0.92	0.88	0.89	
	Hawaiian/Pacific	0.93	0.91	0.95	0.90	0.88	0.93	
	American Indian	0.91	0.93	0.98	0.88	0.91	0.97	
	Multiple Ethnicities	0.94	0.91	0.92	0.92	0.87	0.89	
	LEP	0.91	0.97	1	0.88	0.96	1	
	SPED	0.94	0.96	0.99	0.91	0.95	0.98	
	FRL	0.91	0.92	0.96	0.88	0.89	0.95	
	Accommodations	0.93	0.97	0.99	0.90	0.95	0.99	
	Overall	0.92	0.92	0.97	0.88	0.88	0.96	
	<b>G6E</b>	Female	0.92	0.91	0.96	0.88	0.88	0.95
		Male	0.92	0.92	0.97	0.88	0.89	0.96

Grade	Subgroup	Accuracy			Consistency			
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient	
	African American	0.90	0.93	0.99	0.86	0.90	0.98	
	Hispanic/ Latino	0.91	0.92	0.98	0.87	0.89	0.98	
	Asian	0.95	0.91	0.93	0.92	0.88	0.91	
	White	0.93	0.91	0.95	0.90	0.87	0.93	
	Hawaiian/Pacific	0.90	0.93	0.97	0.85	0.89	0.96	
	American Indian	0.90	0.94	0.99	0.87	0.92	0.99	
	Multiple Ethnicities	0.92	0.91	0.96	0.89	0.88	0.94	
	LEP	0.94	0.99	1	0.91	0.98	1	
	SPED	0.94	0.97	0.99	0.91	0.96	0.99	
	FRL	0.91	0.92	0.98	0.87	0.89	0.98	
	Accommodations	0.94	0.98	1	0.91	0.96	0.99	
	Overall	0.92	0.91	0.96	0.89	0.88	0.94	
	Female	0.92	0.91	0.95	0.89	0.87	0.93	
	Male	0.92	0.92	0.96	0.89	0.89	0.95	
<b>G7E</b>	African American	0.91	0.91	0.97	0.88	0.88	0.96	
	Hispanic/Latino	0.91	0.92	0.97	0.88	0.89	0.96	
	Asian	0.95	0.91	0.92	0.93	0.88	0.89	
	White	0.93	0.91	0.94	0.91	0.87	0.92	
	Hawaiian/Pacific	0.94	0.90	0.96	0.91	0.87	0.94	
	American Indian	0.91	0.94	0.99	0.88	0.92	0.98	
	Multiple Ethnicities	0.92	0.90	0.95	0.90	0.87	0.93	
	LEP	0.95	0.98	1	0.93	0.98	1	
	SPED	0.94	0.97	0.99	0.92	0.96	0.99	
	FRL	0.91	0.92	0.98	0.88	0.89	0.96	
	Accommodations	0.95	0.97	1	0.92	0.96	1	
	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	0.93	0.92	0.96	0.89	0.89	0.94	
<b>G8E</b>	African American	0.92	0.93	0.97	0.89	0.90	0.95	
	Hispanic/ Latino	0.92	0.92	0.97	0.88	0.89	0.96	
	Asian	0.95	0.92	0.90	0.92	0.88	0.87	
	White	0.93	0.91	0.93	0.90	0.87	0.90	
	Hawaiian/Pacific	0.91	0.93	0.95	0.87	0.89	0.94	
	American Indian	0.92	0.95	0.99	0.88	0.92	0.98	
	Multiple Ethnicities	0.93	0.91	0.93	0.90	0.88	0.91	
	LEP	0.96	0.98	0.99	0.94	0.98	0.99	
	SPED	0.95	0.98	0.99	0.93	0.97	0.99	
	FRL	0.92	0.93	0.97	0.88	0.90	0.96	
	Accommodations	0.96	0.98	0.99	0.94	0.97	0.99	
	Overall	0.92	0.91	0.95	0.89	0.88	0.93	
	<b>G9E</b>	Female	0.92	0.91	0.94	0.89	0.87	0.92
		Male	0.92	0.92	0.96	0.89	0.89	0.94



Grade	Subgroup	Accuracy			Consistency			
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient	
	African American	0.92	0.93	0.96	0.88	0.90	0.95	
	Hispanic/Latino	0.91	0.91	0.97	0.88	0.88	0.95	
	Asian	0.95	0.92	0.91	0.93	0.88	0.87	
	White	0.93	0.91	0.93	0.91	0.87	0.90	
	Hawaiian/Pacific	0.91	0.93	0.95	0.87	0.90	0.93	
	American Indian	0.90	0.93	0.98	0.87	0.90	0.97	
	Multiple Ethnicities	0.92	0.91	0.94	0.90	0.87	0.92	
	LEP	0.94	0.96	0.99	0.92	0.95	0.99	
	SPED	0.94	0.97	0.99	0.91	0.96	0.99	
	FRL	0.91	0.92	0.97	0.88	0.89	0.95	
	Accommodations	0.94	0.98	0.99	0.92	0.97	0.99	
	G10E	Overall	0.92	0.93	0.95	0.89	0.90	0.93
		Female	0.92	0.93	0.95	0.89	0.90	0.93
		Male	0.93	0.93	0.95	0.90	0.91	0.94
African American		0.92	0.94	0.96	0.89	0.91	0.95	
Hispanic/ Latino		0.92	0.94	0.97	0.89	0.91	0.95	
Asian		0.93	0.92	0.91	0.90	0.88	0.88	
White		0.92	0.92	0.93	0.89	0.88	0.90	
Hawaiian/Pacific		0.91	0.93	0.95	0.88	0.90	0.94	
American Indian		0.93	0.95	0.98	0.91	0.93	0.97	
Multiple Ethnicities		0.92	0.93	0.94	0.89	0.90	0.92	
LEP		0.96	0.97	0.99	0.94	0.96	0.99	
SPED		0.96	0.98	0.99	0.95	0.97	0.99	
FRL		0.93	0.94	0.97	0.89	0.91	0.96	
Accommodations		0.97	0.99	0.99	0.96	0.98	0.99	
G11E	Overall	0.91	0.92	0.96	0.87	0.89	0.94	
	Female	0.90	0.91	0.95	0.86	0.87	0.93	
	Male	0.91	0.93	0.96	0.88	0.90	0.94	
	African American	0.91	0.93	0.97	0.87	0.90	0.96	
	Hispanic/ Latino	0.90	0.93	0.97	0.87	0.90	0.96	
	Asian	0.92	0.89	0.91	0.88	0.85	0.88	
	White	0.91	0.90	0.94	0.87	0.86	0.91	
	Hawaiian/Pacific	0.91	0.93	0.95	0.87	0.90	0.93	
	American Indian	0.91	0.95	0.99	0.87	0.93	0.98	
	Multiple Ethnicities	0.91	0.91	0.94	0.87	0.87	0.92	
	LEP	0.93	0.96	0.99	0.90	0.95	0.98	
	SPED	0.96	0.98	0.99	0.94	0.97	0.99	
	FRL	0.91	0.93	0.97	0.87	0.91	0.96	
	Accommodations	0.97	0.98	0.99	0.96	0.98	0.99	

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

Exhibit 4.6.6.3 Classification Accuracy and Consistency Estimates for Performance Standards across Subgroups: Mathematics

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
<b>G3M</b>	Overall	0.95	0.93	0.93	0.93	0.90	0.90
	Female	0.95	0.92	0.93	0.93	0.90	0.90
	Male	0.95	0.93	0.92	0.94	0.91	0.90
	African American	0.94	0.93	0.95	0.92	0.90	0.93
	Hispanic/Latino	0.94	0.92	0.94	0.92	0.90	0.93
	Asian	0.98	0.94	0.88	0.97	0.92	0.84
	White	0.96	0.93	0.90	0.95	0.90	0.87
	Hawaiian/Pacific	0.94	0.92	0.94	0.92	0.90	0.92
	American Indian	0.93	0.93	0.96	0.91	0.90	0.95
	Multiple Ethnicities	0.96	0.93	0.91	0.94	0.90	0.88
	LEP	0.93	0.95	0.99	0.90	0.94	0.98
	SPED	0.94	0.95	0.97	0.92	0.94	0.96
	FRL	0.94	0.93	0.95	0.92	0.90	0.93
	Accommodations	0.94	0.96	0.98	0.91	0.94	0.98
<b>G4M</b>	Overall	0.95	0.92	0.95	0.92	0.89	0.92
	Female	0.94	0.92	0.95	0.92	0.89	0.93
	Male	0.95	0.93	0.95	0.93	0.90	0.92
	African American	0.93	0.93	0.96	0.91	0.91	0.95
	Hispanic/Latino	0.93	0.93	0.96	0.91	0.89	0.95
	Asian	0.97	0.94	0.90	0.96	0.91	0.86
	White	0.96	0.92	0.93	0.94	0.89	0.89
	Hawaiian/Pacific	0.94	0.92	0.93	0.92	0.88	0.91
	American Indian	0.93	0.93	0.98	0.90	0.91	0.97
	Multiple Ethnicities	0.95	0.92	0.93	0.94	0.89	0.91
	LEP	0.92	0.96	0.99	0.89	0.95	0.99
	SPED	0.94	0.96	0.98	0.92	0.94	0.98
	FRL	0.93	0.93	0.96	0.91	0.89	0.95
	Accommodations	0.93	0.96	0.99	0.91	0.94	0.99
<b>G5M</b>	Overall	0.94	0.93	0.94	0.92	0.91	0.93
	Female	0.94	0.93	0.94	0.92	0.90	0.93
	Male	0.94	0.94	0.94	0.92	0.91	0.93
	African American	0.93	0.94	0.97	0.90	0.92	0.95
	Hispanic/Latino	0.93	0.93	0.96	0.90	0.91	0.95
	Asian	0.97	0.95	0.91	0.95	0.93	0.88
	White	0.95	0.93	0.93	0.94	0.91	0.90
	Hawaiian/Pacific	0.93	0.94	0.95	0.91	0.92	0.94
	American Indian	0.92	0.94	0.97	0.89	0.92	0.96
	Multiple Ethnicities	0.95	0.94	0.93	0.93	0.91	0.90
	LEP	0.92	0.96	0.99	0.89	0.95	0.99
	SPED	0.94	0.97	0.99	0.92	0.95	0.98

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	FRL	0.93	0.93	0.96	0.90	0.91	0.95
	Accommodations	0.94	0.97	0.99	0.91	0.95	0.99
<b>G6M</b>	Overall	0.94	0.93	0.95	0.91	0.91	0.93
	Female	0.93	0.93	0.95	0.91	0.90	0.93
	Male	0.94	0.93	0.95	0.91	0.91	0.93
	African American	0.93	0.94	0.97	0.91	0.92	0.96
	Hispanic/Latino	0.93	0.93	0.97	0.90	0.91	0.95
	Asian	0.97	0.94	0.92	0.95	0.91	0.89
	White	0.95	0.93	0.93	0.92	0.90	0.91
	Hawaiian/Pacific	0.95	0.94	0.95	0.92	0.92	0.93
	American Indian	0.93	0.94	0.98	0.89	0.92	0.97
	Multiple Ethnicities	0.94	0.93	0.94	0.91	0.90	0.92
	LEP	0.94	0.97	0.99	0.91	0.96	0.99
	SPED	0.95	0.97	0.99	0.93	0.96	0.98
	FRL	0.93	0.93	0.97	0.90	0.91	0.95
	Accommodations	0.95	0.97	0.99	0.92	0.96	0.99
	<b>G7M</b>	Overall	0.94	0.93	0.95	0.91	0.91
Female		0.93	0.93	0.95	0.90	0.91	0.93
Male		0.94	0.94	0.95	0.91	0.91	0.93
African American		0.93	0.95	0.97	0.90	0.92	0.96
Hispanic/Latino		0.93	0.94	0.97	0.90	0.92	0.95
Asian		0.96	0.93	0.91	0.94	0.91	0.87
White		0.94	0.92	0.93	0.91	0.89	0.90
Hawaiian/Pacific		0.95	0.93	0.93	0.92	0.90	0.91
American Indian		0.94	0.96	0.98	0.91	0.94	0.97
Multiple Ethnicities		0.94	0.93	0.93	0.91	0.90	0.91
LEP		0.96	0.98	1.00	0.95	0.98	0.99
SPED		0.96	0.98	0.99	0.95	0.97	0.99
FRL		0.93	0.94	0.97	0.90	0.92	0.95
Accommodations		0.97	0.98	0.99	0.95	0.98	0.99
<b>G8M</b>		Overall	0.92	0.94	0.96	0.89	0.91
	Female	0.92	0.94	0.96	0.89	0.91	0.95
	Male	0.93	0.94	0.97	0.90	0.92	0.95
	African American	0.92	0.95	0.98	0.89	0.93	0.97
	Hispanic/Latino	0.92	0.95	0.97	0.89	0.92	0.96
	Asian	0.95	0.93	0.94	0.92	0.90	0.91
	White	0.93	0.93	0.95	0.89	0.90	0.93
	Hawaiian/Pacific	0.90	0.93	0.97	0.87	0.91	0.96
	American Indian	0.92	0.96	0.98	0.89	0.94	0.98
	Multiple Ethnicities	0.93	0.94	0.96	0.90	0.91	0.94
	LEP	0.95	0.98	1.00	0.93	0.97	0.99
	SPED	0.96	0.98	0.99	0.94	0.97	0.99

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

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	FRL	0.88	0.93	0.98	0.83	0.90	0.98
	Accommodations	0.91	0.97	0.99	0.87	0.96	0.99

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 individualized education plans [IEPs] SPED<sup>29</sup>, FRL, 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.

**Exhibit 4.6.7.1 Internal Consistency Reliability by Subgroup: ELA**

Grade	Statistic	Subgroups													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
3	Reliability	0.90	0.90	0.90	0.89	0.89	0.88	0.89	0.86	0.89	0.90	0.88	0.76	0.89	0.82
	Variance	1117	1101	1111	941	1110	909	969	769	1105	1078	924	462	934	621
4	Reliability	0.90	0.90	0.90	0.89	0.89	0.89	0.89	0.86	0.89	0.90	0.87	0.74	0.88	0.81
	Variance	1078	1050	1088	930	1172	978	901	734	1063	1081	826	396	873	564
5	Reliability	0.90	0.90	0.91	0.90	0.89	0.90	0.90	0.89	0.89	0.89	0.88	0.79	0.90	0.84
	Variance	1233	1185	1253	1095	1266	1181	1057	906	1188	1175	915	491	1033	700
6	Reliability	0.90	0.89	0.90	0.88	0.89	0.88	0.88	0.86	0.89	0.90	0.84	0.70	0.88	0.81
	Variance	1053	1027	1046	904	1066	918	880	758	1036	1052	718	384	872	613
7	Reliability	0.91	0.90	0.91	0.90	0.90	0.90	0.90	0.88	0.90	0.90	0.85	0.78	0.90	0.83
	Variance	1181	1104	1206	1060	1224	1167	1021	886	1139	1148	786	524	1021	670
8	Reliability	0.91	0.90	0.91	0.90	0.90	0.90	0.90	0.88	0.90	0.91	0.85	0.83	0.90	0.85
	Variance	1085	1009	1119	1040	1093	1002	957	783	1026	1093	706	629	952	710
9	Reliability	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.87	0.90	0.90	0.85	0.86	0.89	0.82
	Variance	1021	967	1027	993	1127	1020	891	748	998	985	681	736	893	597
10	Reliability	0.91	0.90	0.91	0.89	0.91	0.90	0.89	0.87	0.91	0.90	0.84	0.85	0.89	0.82
	Variance	1106	1069	1122	962	1209	1018	914	754	1125	1063	692	728	923	631
11	Reliability	0.89	0.87	0.89	0.88	0.88	0.90	0.87	0.84	0.88	0.89	0.83	0.85	0.87	0.82
	Variance	885	791	952	823	884	1008	748	620	892	916	617	708	746	596

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

<sup>29</sup> 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 4.6.7.2 Internal Consistency Reliability by Subgroup: Math

Grade	Statistic	Subgroups													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
3	Reliability	0.92	0.92	0.93	0.93	0.86	0.92	0.93	0.93	0.91	0.92	0.94	0.92	0.93	0.92
	Variance	2241	2093	2380	2211	1688	1896	2109	1868	2051	2274	2633	1740	2120	2056
4	Reliability	0.93	0.92	0.93	0.93	0.89	0.92	0.92	0.92	0.91	0.92	0.92	0.90	0.92	0.91
	Variance	1974	1819	2123	2097	1767	1920	1817	1665	1795	1894	2160	1408	1809	1695
5	Reliability	0.93	0.93	0.93	0.93	0.91	0.93	0.93	0.92	0.92	0.93	0.91	0.88	0.92	0.89
	Variance	2193	1942	2434	1992	2028	2237	1929	1725	2110	2155	1955	1303	1905	1678
6	Reliability	0.93	0.93	0.94	0.93	0.92	0.93	0.93	0.91	0.93	0.93	0.90	0.86	0.92	0.89
	Variance	2170	2033	2301	2106	2062	2204	1894	1738	2046	2003	1859	1259	1884	1640
7	Reliability	0.93	0.93	0.93	0.92	0.91	0.93	0.92	0.90	0.92	0.93	0.87	0.83	0.92	0.84
	Variance	1859	1766	1948	1662	1820	1879	1593	1403	1764	1867	1236	951	1590	1055
8	Reliability	0.92	0.91	0.92	0.90	0.93	0.90	0.90	0.87	0.92	0.92	0.81	0.81	0.9	0.81
	Variance	1552	1449	1645	1314	2088	1309	1354	1056	1567	1620	842	797	1297	849
Alg I	Reliability	0.91	0.91	0.91	0.89	0.92	0.90	0.89	0.85	0.91	0.92	0.80	0.83	0.89	0.77
	Variance	1364	1273	1445	1138	1639	1201	1103	875	1421	1571	705	782	1094	615
Geo	Reliability	0.91	0.90	0.91	0.87	0.93	0.91	0.88	0.84	0.92	0.90	0.78	0.84	0.88	0.67
	Variance	1538	1397	1679	1109	1945	1569	1173	969	1653	1421	827	1003	1210	569
Alg II	Reliability	0.88	0.87	0.89	0.84	0.92	0.88	0.83	0.78	0.89	0.88	0.73	0.81	0.83	0.71
	Variance	1190	1066	1319	957	1594	1196	925	738	1253	1172	716	886	901	667

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native 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 2018 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

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 3	0.75	0.77	0.75
Grade 4	0.78	0.74	0.74
Grade 5	0.78	0.76	0.73
Grade 6	0.75	0.76	0.75
Grade 7	0.81	0.75	0.72
Grade 8	0.80	0.71	0.77

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.78	0.73	0.76
Grade 10	0.79	0.76	0.75
Grade 11	0.75	0.68	0.75

Exhibit 4.6.8.2 Subscale Reliabilities: Mathematics Grades 3–5

	Numbers & Operations-Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
Grade 3	0.68	0.73	0.86
Grade 4	0.79	0.60	0.86
Grade 5	0.80	0.75	0.81

Exhibit 4.6.8.3 Subscale Reliabilities: Mathematics Grades 6 & 7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
Grade 6	0.80	0.78	0.72	0.60
Grade 7	0.73	0.67	0.71	0.75

Exhibit 4.6.8.4 Subscale Reliabilities: Mathematics Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability & the Number System
Grade 8	0.81	0.66	0.66	0.57

Exhibit 4.6.8.5 Subscale Reliabilities: Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.82	0.78	0.59
Algebra II	0.72	0.70	0.63

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	0.60	0.70	0.55	0.75

## 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.<sup>30</sup> 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. The disattenuated correlation equals 1 when the disattenuated correlation is greater than 1.

**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	0.74		0.98	
	Writing & Language	0.65	0.65	0.86	0.86
4	Literature	0.73		0.96	
	Writing & Language	0.67	0.65	0.88	0.88
5	Literature	0.77		1.00	
	Writing & Language	0.68	0.66	0.90	0.89
6	Literature	0.73		0.97	
	Writing & Language	0.67	0.66	0.89	0.88
7	Literature	0.75		0.97	
	Writing & Language	0.67	0.65	0.88	0.89
8	Literature	0.72		0.95	
	Writing & Language	0.70	0.64	0.89	0.86
9	Literature	0.74		0.98	
	Writing & Language	0.66	0.65	0.86	0.87
10	Literature	0.74		0.95	
	Writing & Language	0.67	0.61	0.87	0.81
11	Literature	0.66		0.93	
	Writing & Language	0.67	0.63	0.89	0.88

**Exhibit 4.7.2 Subscale Intercorrelations– Mathematics Grades 3–5**

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.73		1.00	
	OAT_NBT	0.75	0.81	0.94	1.00
4	MDG	0.73		1.00	
	OAT_NBT	0.81	0.76	1.00	1.00
5	MDG	0.77		0.99	
	OAT_NBT	0.81	0.78	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.

<sup>30</sup> 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 4.7.3 Subscale Intercorrelations: Mathematics Grade 6 & 7**

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.81			1.00		
	RP	0.79	0.79		1.00	1.00	
	GSP	0.72	0.71	0.65	1.00	1.00	0.99
7	NS	0.79			1.00		
	RP	0.79	0.77		1.00	1.00	
	GSP	0.77	0.77	0.76	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: Mathematics Grade 8**

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
8	Functions (F)	0.75			1.00		
	Geometry(G)	0.74	0.65		1.00	0.99	
	SPNS	0.73	0.67	0.63	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	0.80		1.00	
	Statistics	0.70	0.69	1.00	1.00
Algebra II	Functions	0.71		1.00	
	Statistics	0.70	0.71	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	0.68			1.00		
	MG	0.68	0.66		1.00	1.00	
	SRTT	0.73	0.73	0.71	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) hand-scored 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 2018 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 non-scorable condition codes (Mismatch NS), or a non-scorable/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 2018 Administration**

Grade	Dimension	Total Read	Second Read	Non Adj Low	Adj Low	Equal	Adj High	Non Adj High	Mismatch NS	MM NS/Score
3	Purpose/Organization	15,667	2,750	1.2	18.5	59.7	18.5	1.2	0.0	0.8
	Evidence/Elaboration	15,668	2,750	1.3	18.1	60.4	18.1	1.3	0.0	0.8
	Conventions	16,043	2,750	0.9	14.0	69.5	14.0	0.9	0.0	0.8
4	Purpose/Organization	15,952	2,812	0.6	16.2	66.3	16.2	0.6	0.0	0.0
	Evidence/Elaboration	15,946	2,812	0.4	17.0	65.1	17.0	0.4	0.0	0.0
	Conventions	16,416	2,812	1.1	16.0	65.7	16.0	1.1	0.0	0.0
5	Purpose/Organization	15,940	2,788	0.5	18.1	62.8	18.1	0.5	0.0	0.0
	Evidence/Elaboration	15,933	2,788	0.3	17.7	64.1	17.7	0.3	0.0	0.0
	Conventions	16,371	2,788	0.9	15.0	68.1	15.0	0.9	0.0	0.0
6	Purpose/Organization	16,038	2,822	0.7	19.5	58.6	19.5	0.7	0.0	0.9
	Evidence/Elaboration	16,050	2,822	1.2	17.5	61.7	17.5	1.2	0.0	0.9
	Conventions	16,352	2,822	0.1	11.8	75.3	11.8	0.1	0.0	0.9
7	Purpose/Organization	16,661	2,956	1.6	19.3	58.1	19.3	1.6	0.0	0.2
	Evidence/Elaboration	16,667	2,956	1.8	19.7	56.9	19.7	1.8	0.0	0.2
	Conventions	17,138	2,956	0.7	16.9	64.4	16.9	0.7	0.0	0.2
8	Purpose/Organization	16,088	2,892	1.4	21.5	53.9	21.5	1.4	0.0	0.3
	Evidence/Elaboration	16,091	2,892	1.5	21.4	53.9	21.4	1.5	0.0	0.3
	Conventions	16,358	2,892	0.8	9.5	79.1	9.5	0.8	0.0	0.3
9	Purpose/Organization	14,292	2,536	0.8	17.2	63.9	17.2	0.8	0.0	0.2
	Evidence/Elaboration	14,289	2,536	0.7	18.6	61.2	18.6	0.7	0.0	0.2
	Conventions	14,505	2,536	0.3	8.9	81.4	8.9	0.3	0.0	0.2
10	Purpose/Organization	12,760	2,226	0.4	16.5	64.9	16.5	0.4	0.0	1.3
	Evidence/Elaboration	12,760	2,226	0.4	18.9	60.2	18.9	0.4	0.0	1.3
	Conventions	13,040	2,226	0.0	12.9	72.8	12.9	0.0	0.0	1.3
11	Purpose/Organization	11,339	1,994	0.6	17.5	63.6	17.5	0.6	0.0	0.4
	Evidence/Elaboration	11,344	1,994	0.8	16.9	64.2	16.9	0.8	0.0	0.4
	Conventions	11,578	1,994	0.3	12.2	74.5	12.2	0.3	0.0	0.4

## 5. ITEM DEVELOPMENT AND TEST CONSTRUCTION

The Arizona’s Measurement of Educational Readiness to Inform Teaching (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 State Standards. 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 Arizona State Standards.<sup>31</sup>

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 Arizona Department of Education (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 Arizona State Standards were used. To supplement the AzMERIT pool of items, a few previously developed Arizona items that also aligned to the Arizona State Standards were used.

Items used to develop the spring 2018 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 ADE, Arizona content experts and educators, and Arizona community members prior to field testing in spring 2016 and spring 2017, and subsequent operational test administration in spring 2017 and spring 2018. Only items that were found to align well with the Arizona State Standards 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 Arizona State Standards was evaluated, alignment of test blueprints with the content standards was critical. The English language arts (ELA) and mathematics blueprints are provided as an attachment in Appendix B.

With the desired alignment of test blueprints to Arizona State Standards, 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 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 (DOK),

---

<sup>31</sup> 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.

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 PROCESS<sup>32</sup>

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-pencil 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 the American Institutes for Research (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 (TDS). 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.<sup>33</sup> These item specifications, developed by content experts at AIR and the ADE, strategically guide the item-development process. 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.

---

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

<sup>33</sup> Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended test-taker 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).

Additionally, the item specifications provide models for item writers. The models include item samples that target different 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.
- **DOK:** 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, 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 mathematics 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.

---

## MATHEMATICS

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 several 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.

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 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-pencil 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-pencil forms, responses were handscored. For other MSCR items that could not readily be rendered for paper-based testing (PBT) 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 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, but, 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 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.<sup>34</sup> Items used to develop the spring 2018 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 ADE, Arizona content experts and educators, and Arizona community members, prior to field testing in spring 2016 and spring 2017, and subsequent operational test administration in spring 2017 and spring 2018. Only items that were found to align well with the Arizona State Standards 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 Arizona State Standards. 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 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 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.

---

<sup>34</sup> 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 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 clearly and appropriately online
- 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, spring 2017, and spring 2018 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. Several 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:<sup>35</sup>

- 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 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 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.

---

<sup>35</sup> 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.

In addition, a fixed block of field-test items was also embedded in paper-pencil 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 mathematics assessments were calibrated following the online administration to establish the AzMERIT bank scale. Following the spring 2016 and spring 2017 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 the ELA and mathematics tests. However, only the ELA spring 2016 operational tests were scored using the post-equated item parameters. In the spring 2018 test administration, the pre-equated parameters calibrated and equated following spring 2016 and spring 2017 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.<sup>36</sup> Analysis of field-test items includes classical item statistics as well as the item response theory (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 test takers in the 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

---

<sup>36</sup> 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 test taker 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.

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 test takers 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 ITEM RESPONSE THEORY STATISTICS

Rasch and Masters' Partial Credit Model are used to estimate the 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 in Exhibit 5.5.3. 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 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 DOK 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.<sup>37</sup>

AIR’s test developers used Form Builder software to help construct operational forms. Form Builder interfaces with AIR’s ITS to extract test information and interactively create test characteristics curves (TCCs), test information curves, 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, Form Builder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, Form Builder 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.

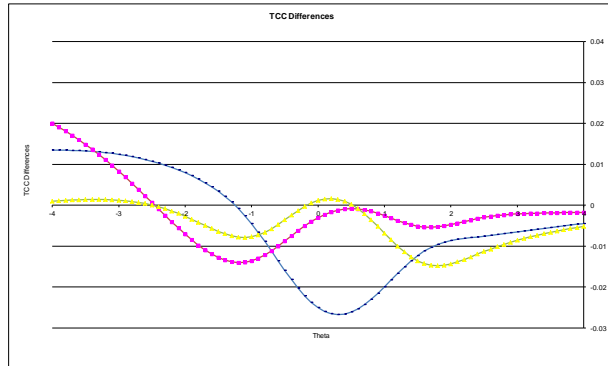
As test developers built forms, the, Form Builder-generated TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 5.6.1.1 shows a sample graph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

<sup>37</sup> 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.



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.

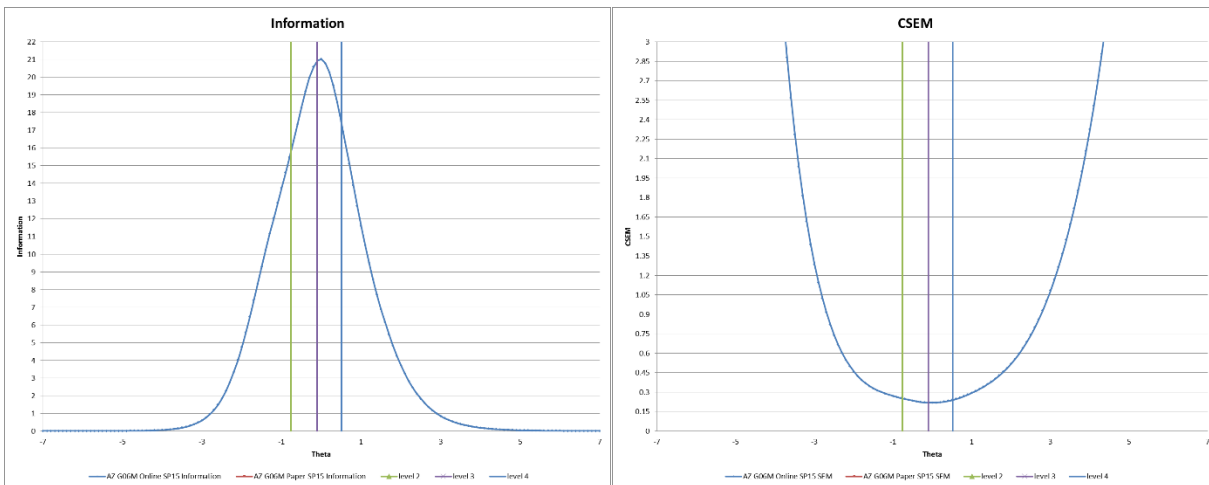
**Exhibit 5.6.1.1 Test Characteristics Curve Differences**



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 2018 operational test forms are presented in Appendix I.

In addition, although paper-pencil test forms were developed to be as nearly identical to the online forms, there were some items that could not readily be rendered for PBT administration. In those instances, replacement items were identified and TCCs and SEMCs were evaluated to ensure equivalence between online and paper-pencil 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 ITS. 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 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 Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) testing.<sup>38</sup> 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 mathematics at the grade level in which they were enrolled. Students, in any grade, who are enrolled in high school-level ELA courses (freshman English, sophomore English, junior English, or their equivalents) or high school-level mathematics 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 mathematics were not required to take the grade 8 mathematics test.

Students with significant cognitive disabilities and whose current individualized education program (IEP) designates them eligible for the alternate assessment for ELA and mathematics were excluded from AzMERIT and instead took the Multi-State Alternate Assessment.

### 6.2 ADMINISTRATION PROCEDURES

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

A secure browser developed by the American Institutes for Research (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-based testing (PBT) and computer-based testing (CBT) administrations. The training also provided an overview of the test delivery system (TDS), Online Reporting System (ORS), and the 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, TAs were required to complete the online TA Certification Course before CBT administration.<sup>39</sup> District Test Coordinators and School Test Coordinators were responsible ensuring that all test administration personnel (PBT and CBT) were properly trained using the various resources prior to the start of testing.

---

<sup>38</sup> 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.

<sup>39</sup> Standard 6.1 – TAs 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.<sup>40</sup> The Test Administrator User Guide was designed to familiarize test administrators with the test delivery system (TDS) and contains tips and screenshots throughout the text. The guide provides enough how-to information to enable TAs to access and navigate the TDS. 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 CBT
- 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 CBT and PBT 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.<sup>41</sup> Similar guidance is provided for the PBT environment, including instructions for the PBT 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 the Arizona Department of Education (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 TAs. 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.

---

<sup>40</sup> 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.

<sup>41</sup> 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.

TAs 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*.<sup>42</sup> 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.

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

TAs 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 mathematics tests for grades 3–6. Scientific calculators were permitted in AzMERIT Mathematics Part 1 for grades 7 and 8. Graphing calculators were permitted in AzMERIT Mathematics EOC Parts 1 and 2 (Algebra I, Geometry, and Algebra II). Online calculators were provided as embedded tools within the appropriate CBT 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-pencil 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-pencil AzMERIT tests, TAs 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 mathematics 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 TAs 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 CBTs or pass out and collect test materials for paper-pencil tests.

---

### 6.2.1 MANAGING TESTING

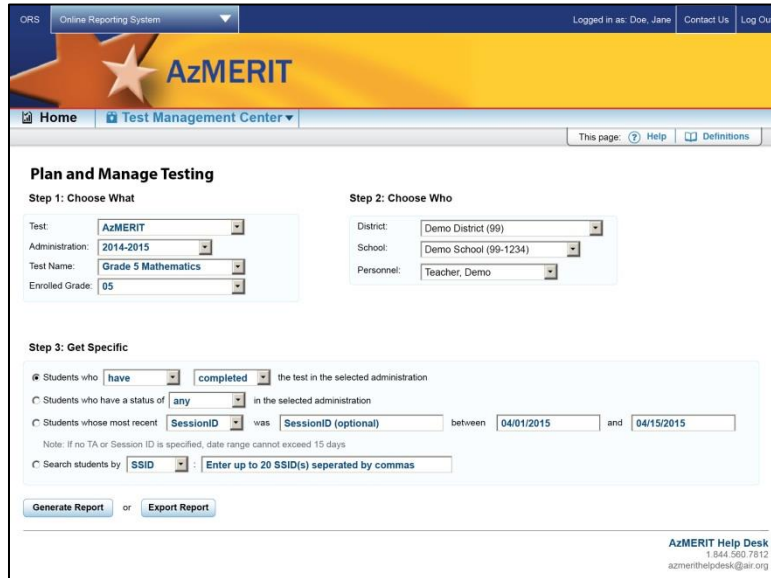
To help schools manage their test schedule, allocate testing resources, and prioritize testing, the AzMERIT ORS, which is described in detail later in this chapter, offered participation reports for online testers. Within the ORS, educators can

---

<sup>42</sup> Standard 6.1 – TAs should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

<sup>43</sup> 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.

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.



## 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

TAs 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.<sup>44</sup> Universal test administration conditions are available for both PBT and CBT. Universal test administration conditions include:

- 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 TA
- 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

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

<sup>44</sup> 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.

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

- 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 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 TESTING

The AzMERIT CBT platform offers numerous testing tools. All tools are available in the AzMERIT Sample Tests, which are available to TAs and students prior to each test administration. TAs 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 TAs.

**Exhibit 6.3.2.1 Universal Testing Tools for CBT Available to All Students**

Universal Test Tool	Description
<b>Area Boundaries</b>	Click anywhere on the selected-response text or button for multiple-choice options
<b>Expand/Collapse Passage</b>	Expand a passage for easier readability. Expanded passages can also be collapsed.
<b>Help</b>	View the on-screen <i>Test Instructions and Help</i> .
<b>Highlighter</b>	Highlight text in a passage or item.
<b>Line Reader</b>	This allows student to track the line he or she is reading.
<b>Mark (Flag) for Review</b>	Mark an item for review so that it can be easily found later.
<b>Notes/Comments</b>	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 mathematics, comments are attached to a specific test item and available throughout the session.
<b>Pause and Restart</b>	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.
<b>Review Test</b>	This allows student to review the test before ending it.
<b>Strikethrough</b>	Cross out answer options for multiple-choice and multi-select items.
<b>System Settings</b>	Adjust audio (volume) during the test.
<b>Text-to-Speech for Instructions</b>	Listen to test instructions.
<b>Tutorial</b>	View a short video about each item type and how to respond.



Universal Test Tool	Description
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 Mathematics	<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>	Mathematics	<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 (EL) or a Fluent English Proficient (FEP) 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 (SPED) need, or language need and the accommodation(s) provided to the student during educational activities, including assessment. TAs 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.<sup>45</sup> 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 IEP calls for a testing accommodation that is not listed, TAs are instructed to contact the ADE for guidance.

Allowable accommodations are described in the following pages.<sup>46</sup>

## 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
<b>Adult Transcription</b>	<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-pencil 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>
<b>Assistive Technology</b>	<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>
<b>Rest/Breaks</b>	<p>Students may take breaks during testing sessions to rest.</p>

<sup>45</sup> 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.

<sup>46</sup> 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 test takers’ 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 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
<b>Read Aloud Test Content</b>	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the mathematics test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test upon student request.</p> <p>Reading aloud the content of the Reading portion of the ELA test is prohibited.</p>
<b>Rest/Breaks</b>	Provide students with breaks during testing sessions to rest.
<b>Simplified Directions</b>	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
<b>Translate Directions</b>	<p>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.</p> <p>Translations that paraphrase, simplify, or clarify directions are not permitted.</p> <p>Written translations are not permitted.</p> <p>Translation of test content is not permitted.</p>
<b>Translation Dictionary</b>	<p>Provide a word-for-word published, paper translation dictionary.</p> <p>Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.</p>

## 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
<b>Abacus</b>	Students with a visual impairment may use an abacus without restrictions for any AzMERIT mathematics test.
<b>Adult Transcription</b>	<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-pencil 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-pencil booklet.</p>
<b>ASL and Closed Caption</b>	In CBTs, this is available for the listening items on the Reading ELA test.
<b>Assistive Technology</b>	<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>
<b>Braille Test Booklet</b>	Provide a paper braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
<b>Large Print Test Booklet</b>	<p>CBT – Either increase default zoom settings when a 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>
<b>Paper-Pencil Test Booklet</b>	CBT – Student’s IEP must indicate that student cannot enter their own responses on the computer and requires a paper-pencil 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-pencil booklet and then entered into the DEI or entered directly into the DEI. See above for rules on Adult Transcription.
<b>Read Aloud Test Content</b>	<p>CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the mathematics test.</p> <p>PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test.</p> <p>Reading aloud the content of the Reading portion of the ELA test.</p>
<b>Rest/Breaks</b>	Provide students with breaks during testing sessions to rest.
<b>Sign Test Content</b>	<p>Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the mathematics test.</p> <p>Signing the content of the Reading portion of the ELA test.</p>
<b>Simplified Directions</b>	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 TIDE, the test delivery system (TDS), and the 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 always recoverable. 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 the hardware used by our systems.

### 6.4.2 SYSTEM SECURITY COMPONENTS

AIR has built-in security controls in all its data stores and transmissions.<sup>47</sup> 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.

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.

---

<sup>47</sup> 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.

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 TDS 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 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 can do. Because AzMERIT was administered both as a PBT and a CBT assessment, test security procedures must guard against item exposure, cheating on the part of TAs 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

TAs are trained on test security procedures, and both test security policies and procedures are clearly presented with the *AzMERIT Test Administration Directions*.<sup>48</sup>

---

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

## 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 mathematics 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 because this instruction may only be given before the students take the test;
- review students' responses;
- read or review students' scratch paper; or
- participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures.

Additional security violations for PBT 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

---

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.

- Erasing any stray marks in test booklet
- Failing to return all test booklets and other test materials

TAs 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, TAs 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’sTIDE. 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 TA training, and tools to identify possible irregularities in test administrations.

For online administrations, quality assurance 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 TDS captures all this information.



Unlike with paper-pencil 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, TA, 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, TA, 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 TA). 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 TA); 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 TA 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 test takers 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 TA.

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  $l_z$  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  $l_z$  for systematic flagging of aberrant response patterns. Students with  $|l_z|$  values greater than 3 are flagged. Aggregate units are flagged with  $|t|$  greater than 3, where  $t$  is calculated by

$$t = \frac{\text{Average } l_z \text{ values}}{\sqrt{(s^2 + 1)/n}},$$

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

### Response Change in Paper-Pencil Tests

Erasure patterns on paper-pencil tests are also examined for unusual patterns of response changes. For paper-pencil 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 because 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

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		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 Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) assessment for the 2017–2018 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 online and paper-pencil. The ORS allows users to compare score data between individual students and the school, district, or overall state, and 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-pencil 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.<sup>49</sup> 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.

---

<sup>49</sup> 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

**FAMILY SCORE REPORT**

**Maria A. Doe**  
Birth Date: 04/17/2007      ABC School (123654)  
SAIS ID: 99999123      ABC District (987456)

**AzMERIT**  
**SPRING 2018**

**Grade 5 English Language Arts (ELA) Assessment**

---

**About This Assessment**

Maria took the AzMERIT Grade 5 ELA assessment in spring 2018. The questions in this assessment measure the knowledge and skills taught in this grade and subject area.

Maria's score shows how well she understands Grade 5 ELA content. A student who scores **Level 3** (Proficient) or **Level 4** (Highly Proficient) on AzMERIT is likely to be ready for the next grade level of ELA.

**About This Report**

Front:

- Maria's overall score for this assessment includes a numeric score and a proficiency level.
- Her numeric score can be compared with the school, district, and state averages.
- The proficiency level shows how well students understand current grade-level material and how likely they are to be ready for the next grade.

Back:

- Maria's level of mastery is shown for each scoring category.
- Scoring categories represent specific knowledge and skills included in this assessment.
- There is a detailed description of the mastery level for each scoring category.

---

**Maria's Performance on the ELA Assessment**

Maria's score in ELA is **2590**, which is **Level 4** (Highly Proficient).

**Level 4** (Highly Proficient):  
Advanced understanding, highly likely to be ready

**Level 3** (Proficient):  
Strong understanding, likely to be ready

**Level 2** (Partially Proficient):  
Partial understanding, likely to need support to be ready

**Level 1** (Minimally Proficient):  
Minimal understanding, highly likely to need support to be ready

School Average: 2555  
District Average: 2550  
State Average: 2543

Maria's score is **Level 4** (Highly Proficient).  
She shows an **advanced** understanding of the expectations for her tested grade. She is highly likely to be ready for ELA in the next grade.

AZED.GOV      ARIZONA DEPARTMENT OF EDUCATION

Spring 2018 987456-1

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 English language arts (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 Arizona Department of Education (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 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 in 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.



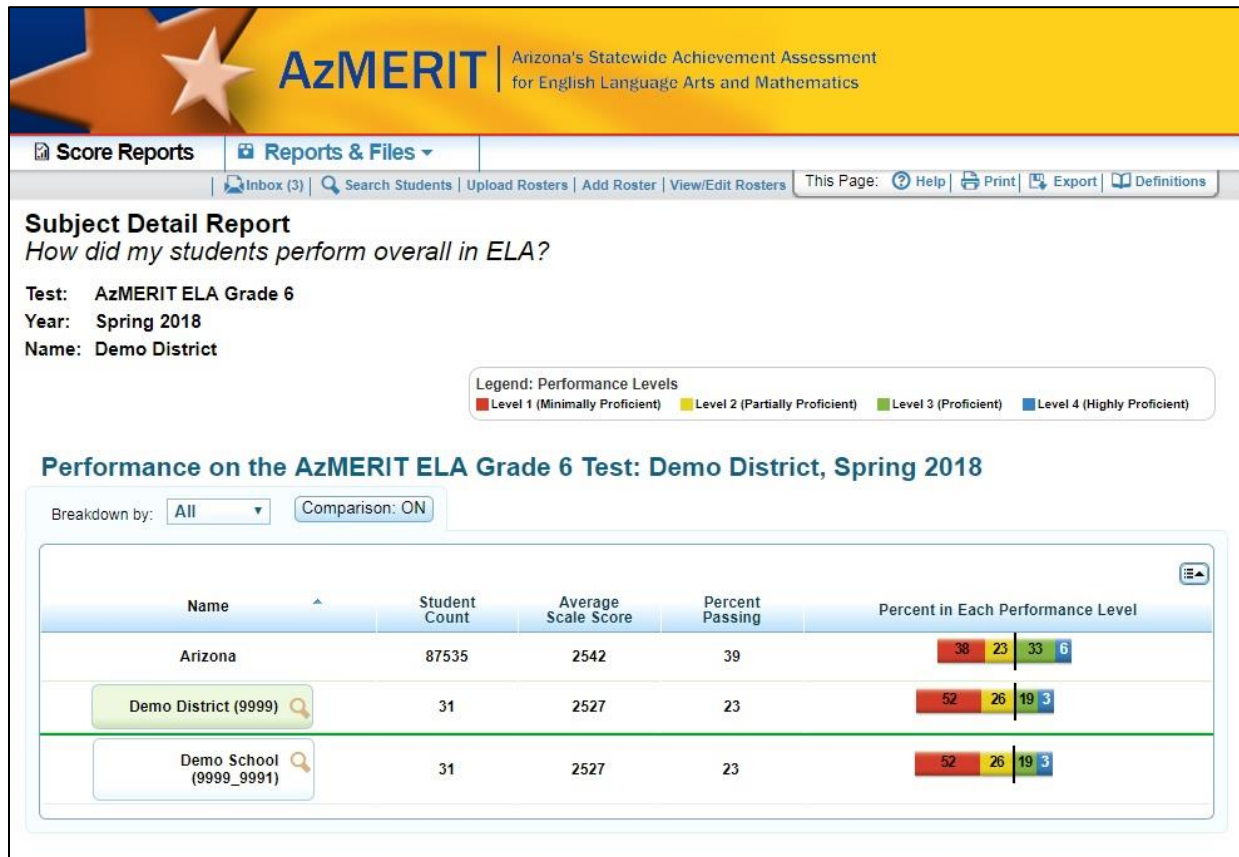
On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district 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**

<b>Type of Report Page</b>	<b>Level of Aggregation</b>	<b>Description</b>
<b>Home Page Dashboard</b>	District, school, and teacher	Summary of performance and participation (Number Tested and Percentage Passing) across grades and subjects or course
<b>Subject Detail</b>	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
<b>Scoring Category Detail</b>	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
<b>Student Roster</b>	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
<b>Individual Student Report</b>	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

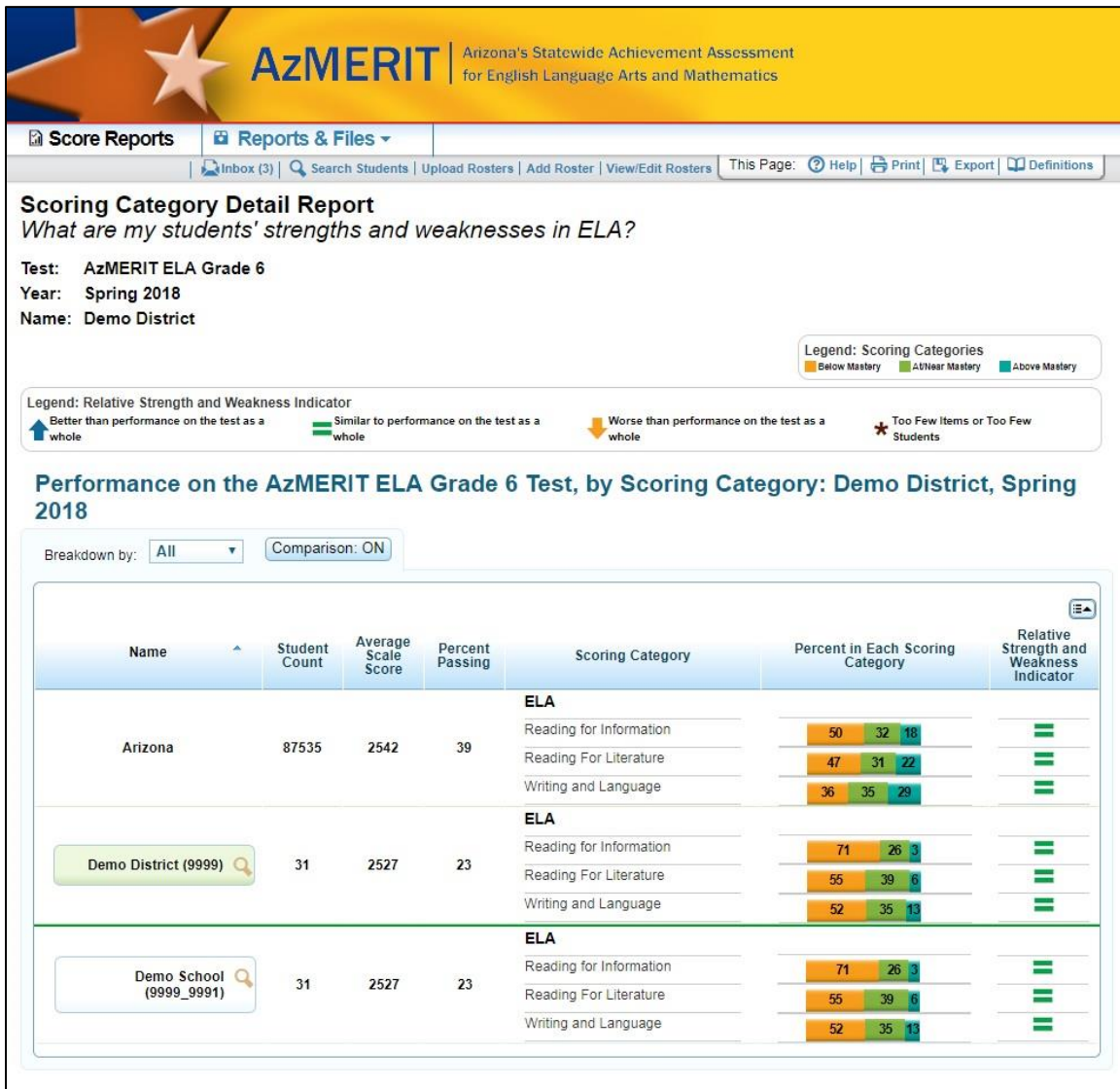


Aggregated subject reports show average performance for the state, districts, schools, teachers, and classes. Bar charts show the distribution of students' performance levels. These reports provide users with rosters of schools, teachers, and classes, allowing for simple comparisons across smaller groups.

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

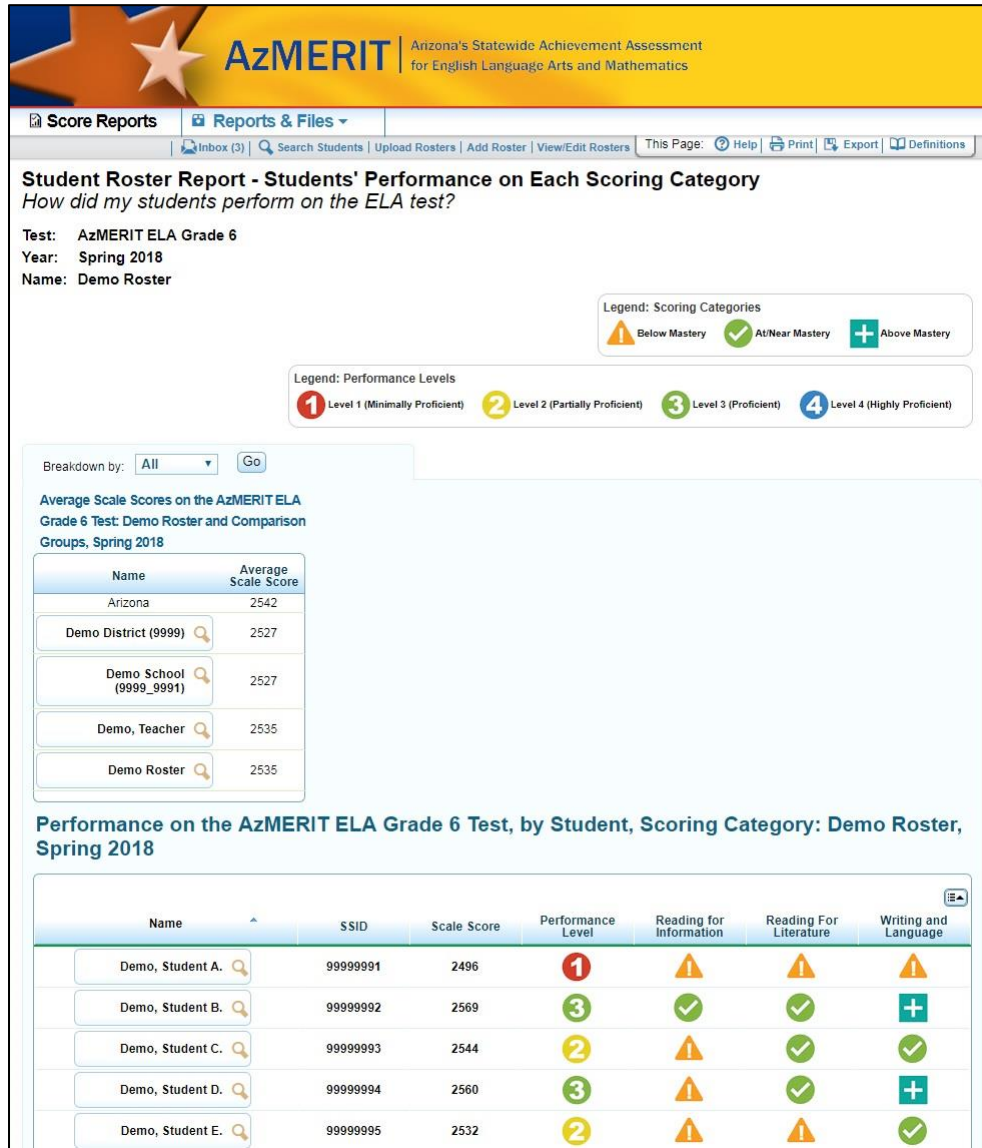
## SCORING CATEGORY DETAIL REPORTS



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: AzMERIT ELA Grade 6  
 Year: Spring 2018  
 Name: Demo Roster

Legend: Scoring Categories  
 ⚠ Below Mastery   ✓ At/Near Mastery   + Above Mastery

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

Breakdown by: All   Go

**Average Scale Scores on the AzMERIT ELA Grade 6 Test: Demo Roster and Comparison Groups, Spring 2018**

Name	Average Scale Score
Arizona	2542
Demo District (9999)	2527
Demo School (9999_9991)	2527
Demo, Teacher	2535
Demo Roster	2535

**Performance on the AzMERIT ELA Grade 6 Test, by Student, Scoring Category: Demo Roster, Spring 2018**

Name	SSID	Scale Score	Performance Level	Reading for Information	Reading For Literature	Writing and Language
Demo, Student A.	99999991	2496	1	⚠	⚠	⚠
Demo, Student B.	99999992	2569	3	✓	✓	+
Demo, Student C.	99999993	2544	2	⚠	✓	✓
Demo, Student D.	99999994	2560	3	⚠	✓	+
Demo, Student E.	99999995	2532	2	⚠	⚠	✓

Student roster reports provide users with performance data for a group of students associated with a teacher or a school, as defined in the Test Information Distribution Engine (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 State Standards (4, representing Highly Proficient, to 1, representing Minimally Proficient)
- **Scoring Categories:** Represents levels of scoring category mastery with respect to the Arizona State Standards, characterizing achievement at "above," "at or near," or "below" mastery on each scoring category

# INDIVIDUAL STUDENT REPORTS

Arizona's Statewide Achievement Assessment  
for English Language Arts and Mathematics

Score Reports | Reports & Files

Inbox (3) | Search Students | Upload Rosters | Add Roster | View/Edit Rosters | This Page | Help | Print | Definitions

### Individual Student Report

*How did my student perform on the ELA test?*

**Test:** AzMERIT ELA Grade 6  
**Year:** Spring 2018  
**Name:** Demo, Student A.

**Legend: Scoring Categories**

⚠ Below Mastery | ✔ At/Near Mastery | + Above Mastery

**Legend: Performance Levels**

1 Level 1 (Minimally Proficient) | 2 Level 2 (Partially Proficient) | 3 Level 3 (Proficient) | 4 Level 4 (Highly Proficient)

**Overall Performance on the AzMERIT ELA Grade 6 Test: Demo, Student A., Spring 2018**

Name	SSID	Birth Date	Scale Score	Performance Level
Demo, Student A.	99999991	02/26/2006	2603	4

**Scale Score and Performance on the AzMERIT ELA Grade 6 Test: Demo, Student A., Spring 2018**

**Average Scale Scores on the AzMERIT ELA Grade 6 Test: Demo Roster and Comparison Groups, Spring 2018**

Name	Average Scale Score
Arizona	2542
Demo District (9999)	2527
Demo School (9999_9991)	2527
Demo, Teacher	2535
Demo Roster	2535

**Performance on the AzMERIT ELA Grade 6 Test, by Scoring Category: Demo, Student A., Spring 2018**

Scoring Categories	Performance
Reading for Information	<p><b>What was assessed?</b> Students find the central idea and supporting details of a text. They tell about the author's purpose and point of view. They show how ideas are developed and supported in a text. They use media to understand a text, and they compare and contrast different texts on the same topic.</p> <p><b>What do these results mean?</b> Your student almost always finds relationships between people or events in a text; finds the author's purpose and point of view in a text; finds details an author uses to support a claim; finds similarities and differences between the ways two authors describe the same event.</p>
Reading For Literature	<p><b>What was assessed?</b> Students find the theme and supporting details of a text. They show how the story develops and how characters change throughout a text. They tell how an author uses organization and point of view to tell a story. They compare and contrast two texts with the same theme.</p> <p><b>What do these results mean?</b> Your student almost always tells how changes in characters move the plot of a story forward; tells how one part of a story fits into the overall text; tells the way a word or phrase affects the feeling of a text; finds similarities and differences in two texts with the same theme.</p>
Writing and Language	<p><b>What was assessed?</b> Students write to argue an opinion and give information using supporting details. They use pronouns correctly. They use clues in a text, dictionaries, and relationships between words in a text to find the meaning of figurative language and new words or phrases.</p> <p><b>What do these results mean?</b> Your student almost always organizes writing for a purpose (like to give information or make an argument) and uses supporting details; uses pronouns and punctuation correctly; tells the differences between the definition of a word and the feeling it gives the text.</p>

**Writing Essay Performance**

Statement of Purpose, Focus & Organization	Evidence & Elaboration	Conventions & Editing
Your student earned 4 out of 4 possible points. Your student's essay is on topic and focused. The response is well organized and develops claims that use details from supporting sources. Transitions are used consistently to vary sentences and explain relationships between ideas. Ideas are logically developed and are strongly connected from beginning to end.	Your student earned 3 out of 4 possible points. Your student's essay is well supported and uses facts and details as evidence to support the claim. It uses citations and some evidence from sources. It uses transitions to make connections between ideas. It includes general and specific vocabulary appropriate for the audience and fits the purpose of the task.	Your student earned 2 out of 2 possible points. Your student's essay shows a strong understanding of sentence formation and other conventions. The response is clear but has some minor mistakes. It correctly uses punctuation, capitalization, and spelling rules.

Individual Student Reports, which closely mirror the Family Reports, are also available through the ORS.

### 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 ORS and understanding paper family reports.<sup>50</sup> 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 State Standards 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 test takers 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 2016–2017 school year indicates the same level of achievement as an average scale score of 2450 for grade 3 students in the 2017–2018 school year, even though the test may include a slightly different set of items.

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 mathematics assessment, the scale score reported can range from 3395 to 3839. Overall scale scores for ELA and mathematics 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.

---

<sup>50</sup> 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.

Exhibit 7.3.1 AzMERIT Scale Score Ranges

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

ELA and mathematics 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, the American Institutes for Research (AIR) convened panels of Arizona educators to recommend performance standards on each of the Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) assessments. Details of the panels, procedures, and outcomes are documented in the “Recommending AzMERIT Performance Standards” technical report, which is available from the Arizona Department of Education (ADE).<sup>51</sup> 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 State Standards. The cut score establishing the Proficient level of performance is the most critical because it indicates that students are meeting grade-level expectations for achievement of the Arizona State 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.

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 Arizona State Standards. 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 Arizona State Standards and student achievement to map the performance-level descriptors adopted by the Arizona State Board of Education to an ordered-item booklet (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 English language arts (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 National Assessment of Educational Progress (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 mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous Arizona’s Instrument to Measure Standards (AIMS) performance standards. Panelists were asked to consider the location of these

---

<sup>51</sup> 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.



benchmark locations when making their content-based cut-score recommendations. When panelists can 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:

1. Minimally Proficient
2. Partially Proficient
3. Proficient
4. 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 Arizona State Standards. 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](http://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
<b>ELA</b>						
<b>3</b>	-0.09	56	0.29	41	1.36	10
<b>4</b>	0.14	57	0.6	39	1.8	5

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% at or Above	Theta	% at or Above	Theta	% at or Above
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
<b>Mathematics</b>						
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 mathematics 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
<b>ELA</b>				
<b>3</b>	44	15	31	10
<b>4</b>	43	19	33	5
<b>5</b>	37	33	27	3
<b>6</b>	39	27	30	4
<b>7</b>	41	26	29	4
<b>8</b>	40	27	26	6
<b>9</b>	47	26	21	6
<b>10</b>	49	21	22	8
<b>11</b>	54	20	17	8
<b>Mathematics</b>				
<b>3</b>	27	31	27	15
<b>4</b>	29	29	32	10
<b>5</b>	29	31	27	13
<b>6</b>	38	30	21	11
<b>7</b>	48	22	18	13
<b>8</b>	43	24	20	13
<b>Algebra I</b>	45	23	23	9
<b>Geometry</b>	47	24	24	6
<b>Algebra II</b>	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
<b>ELA</b>				
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
<b>Mathematics</b>				
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  $\theta$  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 Arizona's Measurement of Educational Readiness to Inform Teaching (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  $\theta$  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 English language arts (ELA) and mathematics. 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 mathematics 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.<sup>52</sup> This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by the American Institutes for Research (AIR), the Arizona Department of Education (ADE), and HumRRO, which acts as a third-party quality assurance (QA) contractor.

Within AzMERIT, students can skip items in both the online and paper-pencil tests. 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 2018, pre-equated item parameters were used to score student test records for the mathematics assessments. For ELA, because two 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.

---

### 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  $\theta$  can be written as

---

<sup>52</sup> 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 -test taker 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.

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

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  $\theta$  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. Because the log-likelihood is a monotonic function of the likelihood, the following derivatives based on the log-likelihood function are used:

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

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

The maximum likelihood estimates of  $\theta$  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 MATHEMATICS

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 ( $g - 1$ ).<sup>53</sup> 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

Because 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 ELA in high school continue as end-of-course (EOC) or grade-level measures of student achievement of the Arizona State Standards, each assessment can be linked to the grade above using all available operational items.

However, AzMERIT assessments of high school mathematics 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 mathematics 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 mathematics 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 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 because 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 mathematics scale to both the Algebra I and Geometry EOC scales because 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 IRT

---

<sup>53</sup> 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.



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 mathematics 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**

Exhibit 9.2.2.1 shows the number of items dropped and remaining in the final vertical 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 mathematics, 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 mathematics 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	Mathematics Dropped Items	Mathematics 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

on 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 mathematics and ELA and proceeded upwards. For mathematics EOC assessments, the grade 8 mathematics 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	Mathematics Chain Linked	Mathematics 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	Mathematics Chain Linked	Mathematics 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 ELA and mathematics, respectively, which plot the linking constants across grade-level assessments. As the linking constants indicate, for mathematics 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: ELA

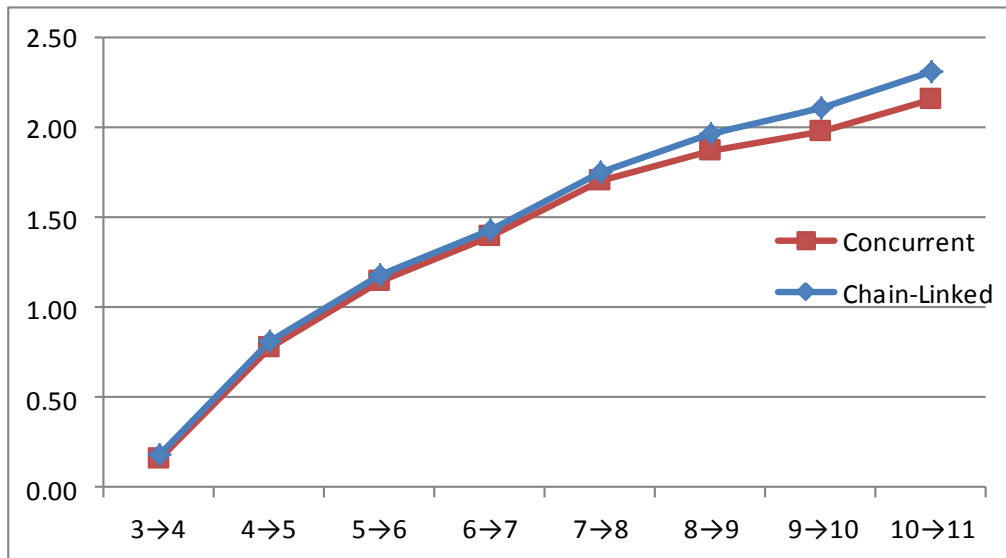
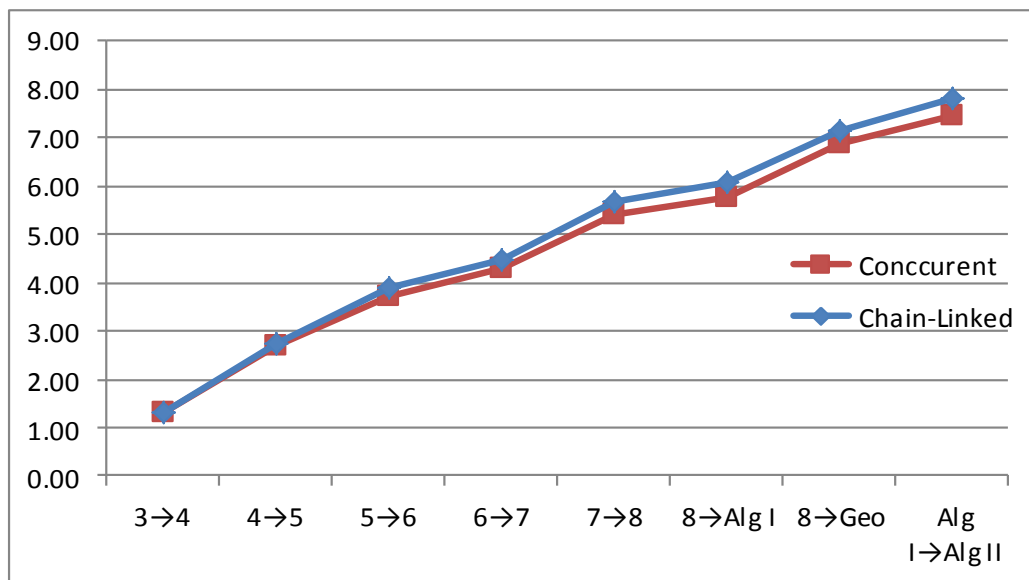


Exhibit 9.2.2.5 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: Mathematics



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 recommending performance standards. We note that ordered-item booklets (OIBs) 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 mathematics than is observed for ELA. In mathematics, growth is on the order of about one standard deviation per year, except for 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 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 many 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 ELA and mathematics assessments are presented in Exhibits 9.2.2.6 9.2.2.7.

**Exhibit 9.2.2.6 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

Exhibit 9.2.2.7 Vertical Linking Constants Resulting from Chain-Linking and Concurrent Calibration: Mathematics

Mathematics	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
<b>G3M</b>	0	0	0	0	0	0
<b>G4M</b>	1.73	1.62	1.81	1.65	1.68	1.64
<b>G5M</b>	3.24	3.06	3.37	3.13	3.30	3.29
<b>G6M</b>	4.45	4.16	4.57	4.23	4.52	4.47
<b>G7M</b>	5.12	4.75	5.30	4.86	5.19	5.10
<b>G8M</b>	6.15	5.70	6.40	5.89	6.23	6.17
<b>Alg I</b>	7.04	6.26	7.32	6.48	7.17	7.04
<b>Geometry</b>	7.49	6.63	7.84	6.94	7.73	7.54
<b>Alg II</b>	8.03	6.83	8.34	7.08	8.05	7.85

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 ELA and mathematics, 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 mathematics, although the vertical linking constants by grade in 2017 are uniformly higher than those in 2015, the difference between the 2015 and 2017 mathematics linking constant for each grade is not larger than one standard deviation, except for 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: ELA

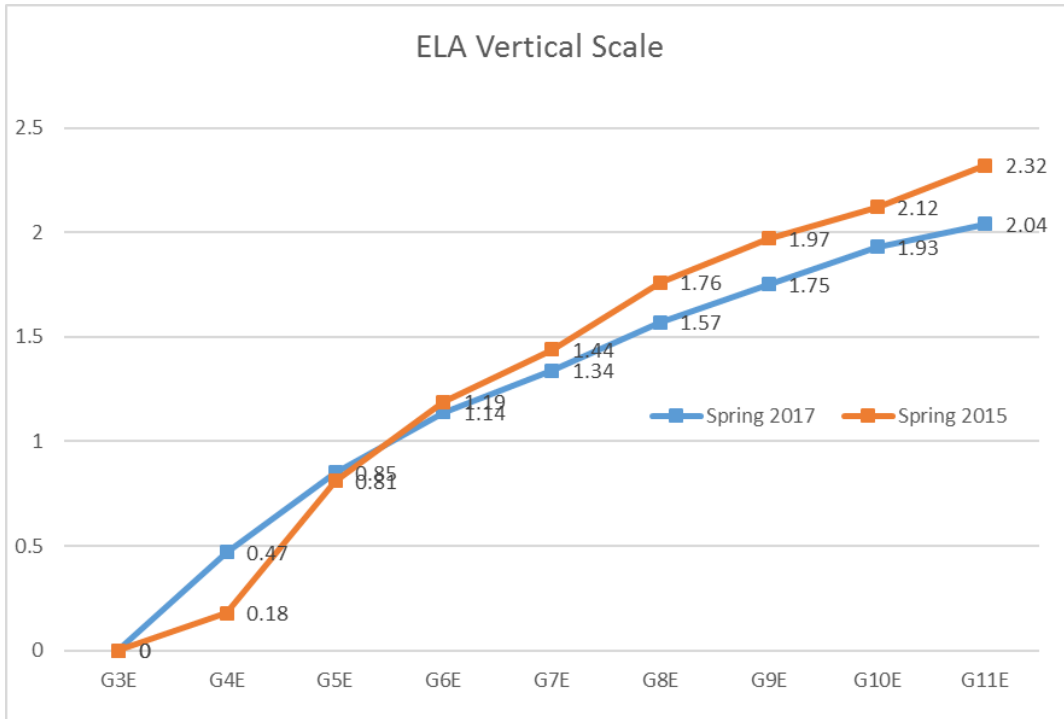
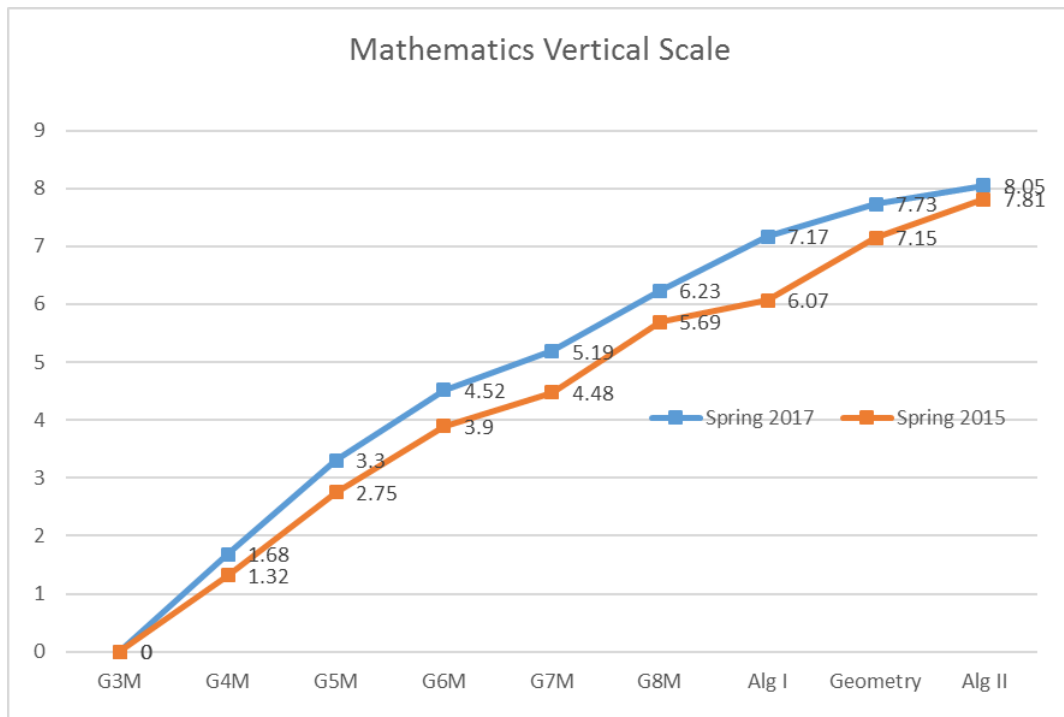


Exhibit 9.2.2.9 Comparison of 2015 and 2017 Vertical Linking Constants Estimated from Chain-Linking Calibrations: Mathematics



### 9.3 AZMERIT REPORTING SCALE (SCALE SCORES)

The AzMERIT assessments are reported on common scales within each subject (ELA and mathematics). 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,<sup>54</sup>

$$SS = a * \theta_v + d$$

where  $a = 30$ ,  $d = 2500$  for ELA tests, and  $a = 30$ ,  $d = 3500$  for mathematics tests.  $\theta_v = \theta + c$ , where  $\theta$  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  $\pm 3.5$ .

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.

**Exhibit 9.3.1 Scale Score Ranges for Performance Levels**

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
<b>ELA</b>				
<b>Grade 3</b>	2395–2496	2497–2508	2509–2540	2541–2605
<b>Grade 4</b>	2400–2509	2510–2522	2523–2558	2559–2610
<b>Grade 5</b>	2419–2519	2520–2542	2543–2577	2578–2629
<b>Grade 6</b>	2431–2531	2532–2552	2553–2596	2597–2641
<b>Grade 7</b>	2438–2542	2543–2560	2561–2599	2600–2648
<b>Grade 8</b>	2448–2550	2551–2571	2572–2603	2604–2658
<b>Grade 9</b>	2454–2554	2555–2576	2577–2605	2606–2664
<b>Grade 10</b>	2458–2566	2567–2580	2581–2605	2606–2668
<b>Grade 11</b>	2465–2568	2569–2584	2585–2607	2608–2675
<b>Mathematics</b>				
<b>Grade 3</b>	3395–3494	3495–3530	3531–3572	3573–3605
<b>Grade 4</b>	3435–3529	3530–3561	3562–3605	3606–3645
<b>Grade 5</b>	3478–3562	3563–3594	3595–3634	3635–3688
<b>Grade 6</b>	3512–3601	3602–3628	3629–3662	3663–3722
<b>Grade 7</b>	3529–3628	3629–3651	3652–3679	3680–3739

<sup>54</sup> Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
<b>Mathematics</b>				
<b>Grade 8</b>	3566–3649	3650–3672	3673–3704	3705–3776
<b>Algebra I</b>	3577–3660	3661–3680	3681–3719	3720–3787
<b>Geometry</b>	3609–3672	3673–3696	3697–3742	3743–3819
<b>Algebra II</b>	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.<sup>55</sup>

### 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-pencil AzMERIT assessment from previous year achievement on Arizona’s Instrument to Measure Standards (AIMS), covarying demographic variables that included gender, ethnicity, income level status, English Learner (EL) status, and individualized education program (IEP) in the development of the prediction equation. A 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-based testing (PBT) administration mode samples. The linking constant necessary to bring the matched sample paper-pencil item parameters on 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-pencil 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-pencil test scores were used to predict student performance on the spring 2016 paper-pencil 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-pencil 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 mathematics test to identify matched samples

<sup>55</sup> 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.



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 a paper-pencil form. 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 mathematics multiple-choice (MC) 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 – Mathematics 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 – Mathematics 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 – Mathematics 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 mathematics assessments. As observed for ELA, in the grade 4–8, and Algebra I mathematics 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 a paper-pencil form 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 mathematics assessments were somewhat more difficult online than on a paper-pencil form. 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 mathematics 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 Mathematics Assessments**

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
<b>G3M</b>	2015	-0.71	-0.77	0.06	1.80
	2016 – Mathematics MC Match	-0.84	-0.57	-0.27	-8.10
<b>G4M</b>	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 – Mathematics MC Match	-0.41	-0.24	-0.17	-5.10
<b>G5M</b>	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 – Mathematics MC Match	-0.07	-0.06	0.00	0.00
<b>G6M</b>	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
<b>G7M</b>	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
<b>G8M</b>	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
<b>Alg I</b>	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
<b>Geo</b>	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
<b>Alg II</b>	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 mathematics assessment, as with grade 3 ELA, samples were matched based on student performance on the mathematics 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 a paper-pencil form.

As with ELA, the identified mode effects varied across test administrations. The advantage of online over paper-pencil 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 on paper-pencil forms and then examined the two samples of schools for any differences in performance on the spring 2014 PBT 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 their students. This increased level of matching was appropriate because 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-pencil forms, 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, AND PISA

To facilitate comparisons of Arizona achievement to other national and international benchmarks, several 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.<sup>56</sup> 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.

After the 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 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

---

<sup>56</sup> 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 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 booklet (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 was 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 on 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 many 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: Test takers 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:

$$\hat{Y} = -290.65 + 0.12 * X1 + 0.26 * X2 - 2.35 * X3 - 0.79 * X4 + 0.57 * X5 - 2.32 * X6 - 1.79 * X7 - 2.40 * X8 - 1.82 * X9 - 2.07 * X10$$

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.

Mathematics: 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 Mathematics 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 ( $F(10, 13768) = 1764.13, p < .0001$ ; 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, mathematics 3727) as the ones from the full data model. Then the predictive model was applied to 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 equipercntile 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 Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) assessments in English language arts (ELA) and mathematics utilize a variety of item types to assess students’ mastery of the Arizona State Standards. The Arizona Department of Education (ADE) leverages the American Institutes for Research’s (AIR) 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-pencil 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, mathematics assessments that were administered on paper-pencil forms 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-pencil, and the mathematics 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 rescoring 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 2018 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 two 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 autoscoring 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 TDS to score all remaining essay responses.<sup>57</sup>

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-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

---

<sup>57</sup> Standard 4.19 – When automated algorithms are to be used to score complex test taker responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

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 hand-scored, 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 2018 administration.<sup>58</sup> Please see the 2016 AzMERIT Technical Report, available at [www.azed.gov](http://www.azed.gov), for the values for the complete list of prompts. Indicators include percentage exact agreement, Pearson’s correlation, a quadratic 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. Because 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 2018, Exhibit 10.1.2.2 presents the correlations among the dimension scores.

---

<sup>58</sup> 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.

Exhibit 10.1.2.1 Summary of Human and Machine Scores for Spring 2018 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 κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
3	13021	Conventions	2	2092	1.43	1.55	0.75	0.71	0.69	0.65	0.65	0.03	0.72	0.71	0.70	0.16
		Evidence	4		1.93	1.90	0.78	0.61	0.65	0.64	0.64	0.02	0.65	0.65	0.63	0.05
		Organization	4		1.93	2.00	0.76	0.66	0.66	0.67	0.67	0.00	0.67	0.66	0.65	0.10
3	13024	Conventions	2	2096	1.44	1.53	0.70	0.67	0.71	0.66	0.66	0.01	0.76	0.68	0.67	0.13
		Evidence	4		1.93	1.90	0.76	0.64	0.63	0.64	0.64	0.04	0.64	0.63	0.62	0.04
		Organization	4		1.96	1.96	0.80	0.65	0.63	0.66	0.66	0.05	0.64	0.64	0.63	0.00
4	13118	Conventions	2	2096	1.15	1.16	0.71	0.65	0.64	0.60	0.60	0.01	0.67	0.63	0.63	0.01
		Evidence	4		1.33	1.29	0.49	0.48	0.76	0.55	0.55	0.01	0.84	0.64	0.64	0.07
		Organization	4		1.56	1.53	0.61	0.56	0.71	0.59	0.59	0.03	0.77	0.67	0.67	0.04
4	13121	Conventions	2	2096	1.10	1.08	0.69	0.59	0.67	0.65	0.65	0.03	0.68	0.61	0.60	0.02
		Evidence	4*		1.34	1.27	0.54	0.49	0.77	0.60	0.60	0.03	0.81	0.65	0.64	0.14
		Organization	4*		1.53	1.45	0.58	0.54	0.72	0.61	0.61	0.03	0.74	0.59	0.59	0.13
5	13237	Conventions	2	2095	1.30	1.40	0.74	0.67	0.73	0.72	0.71	0.04	0.73	0.69	0.68	0.13
		Evidence	4		1.59	1.53	0.60	0.53	0.73	0.61	0.61	0.04	0.76	0.62	0.62	0.09
		Organization	4		1.75	1.75	0.66	0.57	0.72	0.66	0.66	0.01	0.72	0.64	0.64	0.01
5	13238	Conventions	2	2099	1.47	1.51	0.62	0.61	0.72	0.65	0.65	0.00	0.75	0.65	0.64	0.06
		Evidence	4		1.87	1.88	0.64	0.53	0.69	0.63	0.63	0.01	0.75	0.63	0.62	0.02
		Organization	4		1.95	1.99	0.68	0.56	0.70	0.65	0.65	0.01	0.74	0.62	0.61	0.06
6	13305	Conventions	2	2095	1.45	1.59	0.68	0.61	0.66	0.57	0.57	0.00	0.76	0.69	0.67	0.22
		Evidence	4		1.53	1.43	0.60	0.55	0.70	0.58	0.58	0.01	0.74	0.61	0.60	0.17
		Organization	4		1.62	1.60	0.68	0.62	0.65	0.59	0.59	0.01	0.70	0.62	0.62	0.02
6	13309	Conventions	2	2093	1.39	1.48	0.65	0.56	0.68	0.58	0.58	0.06	0.76	0.68	0.67	0.15
		Evidence	4		1.69	1.60	0.73	0.67	0.65	0.59	0.59	0.02	0.72	0.71	0.70	0.13
		Organization	4		1.84	1.83	0.78	0.69	0.61	0.62	0.62	0.01	0.70	0.71	0.71	0.01
7	13400	Conventions	2	2082	1.35	1.45	0.66	0.63	0.70	0.67	0.67	0.02	0.74	0.70	0.69	0.14
		Evidence	4		1.84	1.83	0.61	0.53	0.66	0.60	0.60	0.07	0.77	0.65	0.65	0.03
		Organization	4		1.92	1.90	0.64	0.54	0.65	0.62	0.62	0.02	0.74	0.61	0.60	0.03
7	13405	Conventions	2	2093	1.46	1.48	0.61	0.62	0.75	0.63	0.63	0.03	0.77	0.68	0.68	0.02
		Evidence	4		1.63	1.66	0.59	0.62	0.74	0.70	0.70	0.03	0.79	0.72	0.72	0.04
		Organization	4		1.83	1.80	0.62	0.56	0.73	0.68	0.68	0.03	0.79	0.70	0.69	0.05

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted $\kappa^*$	SMD*	% Exact	Pearson r	Weighted $\kappa^*$	SMD*
8	13438	Conventions	2	2631	2.01	1.95	0.77	0.71	0.79	0.70	0.70	0.01	0.73	0.75	0.74	0.08
		Evidence	4		2.11	2.08	0.80	0.76	0.77	0.78	0.78	0.00	0.71	0.75	0.74	0.04
		Organization	4		1.55	1.57	0.63	0.59	0.73	0.76	0.76	0.01	0.79	0.72	0.72	0.03
8	13453	Conventions	2	2538	1.53	1.57	0.64	0.60	0.76	0.68	0.68	0.01	0.78	0.71	0.71	0.06
		Evidence	4		1.99	1.99	0.78	0.74	0.76	0.78	0.78	0.01	0.73	0.76	0.76	0.00
		Organization	4		2.14	2.12	0.79	0.73	0.75	0.79	0.79	0.02	0.74	0.77	0.77	0.03
9	13554	Conventions	2	2751	1.61	1.68	0.59	0.55	0.81	0.71	0.71	0.02	0.80	0.69	0.68	0.13
		Evidence	4		1.89	1.92	0.62	0.53	0.82	0.76	0.76	0.01	0.79	0.68	0.67	0.04
		Organization	4		2.02	2.03	0.65	0.60	0.79	0.76	0.76	0.02	0.80	0.74	0.73	0.01
9	13565	Conventions	2	2869	1.52	1.56	0.62	0.61	0.80	0.75	0.75	0.02	0.78	0.71	0.71	0.06
		Evidence	4		1.92	1.92	0.67	0.60	0.81	0.80	0.80	0.03	0.79	0.74	0.73	0.01
		Organization	4		2.11	2.11	0.72	0.66	0.79	0.80	0.80	0.01	0.78	0.76	0.76	0.00
10	13635	Conventions	2	2436	1.61	1.65	0.55	0.53	0.71	0.60	0.60	0.02	0.77	0.61	0.61	0.07
		Evidence	4		2.04	2.08	0.77	0.71	0.69	0.73	0.73	0.01	0.75	0.76	0.76	0.05
		Organization	4		2.25	2.26	0.76	0.69	0.70	0.73	0.73	0.04	0.72	0.73	0.72	0.02
10	13636	Conventions	2	2344	1.69	1.78	0.49	0.45	0.72	0.58	0.57	0.01	0.83	0.63	0.62	0.19
		Evidence	4		1.99	1.96	0.74	0.66	0.74	0.73	0.73	0.00	0.76	0.76	0.76	0.04
		Organization	4		2.06	2.08	0.75	0.72	0.72	0.74	0.74	0.00	0.79	0.81	0.81	0.03
11	13723	Conventions	2	2095	1.60	1.63	0.59	0.57	0.75	0.61	0.61	0.01	0.77	0.65	0.65	0.05
		Evidence	4		2.24	2.24	0.83	0.74	0.62	0.70	0.70	0.02	0.70	0.74	0.73	0.00
		Organization	4		2.47	2.47	0.74	0.68	0.64	0.69	0.69	0.00	0.74	0.74	0.73	0.01
11	13725	Conventions	2	2085	1.45	1.50	0.67	0.63	0.71	0.64	0.64	0.04	0.78	0.74	0.73	0.08
		Evidence	4		2.21	2.26	0.83	0.78	0.64	0.72	0.72	0.02	0.74	0.79	0.79	0.05
		Organization	4		2.36	2.34	0.81	0.71	0.66	0.73	0.73	0.00	0.74	0.78	0.77	0.02

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 2018 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
3	13021	Conventions Evidence Organization	2 4 4	2092	0.61 0.60	0.79
3	13024	Conventions Evidence Organization	2 4 4	2096	0.72 0.61	0.89
4	13118	Conventions Evidence Organization	2 4 4	2096	0.51 0.67	0.61
4	13121	Conventions Evidence Organization	2 4 4	2096	0.49 0.57	0.66
5	13237	Conventions Evidence Organization	2 4 4	2095	0.59 0.63	0.71
5	13238	Conventions Evidence Organization	2 4 4	2099	0.57 0.53	0.78
6	13305	Conventions Evidence Organization	2 4 4	2095	0.53 0.62	0.77
6	13309	Conventions Evidence Organization	2 4 4	2093	0.74 0.71	0.77
7	13400	Conventions Evidence Organization	2 4 4	2082	0.63 0.65	0.73
7	13405	Conventions Evidence Organization	2 4 4	2093	0.58 0.62	0.76
8	13438	Conventions Evidence Organization	2 4 4	2631	0.89 0.61	0.49
8	13453	Conventions Evidence Organization	2 4 4	2538	0.60 0.61	0.85
9	13554	Conventions Evidence Organization	2 4 4	2751	0.47 0.50	0.79
9	13565	Conventions Evidence Organization	2 4 4	2869	0.63 0.59	0.81
10	13635	Conventions Evidence Organization	2 4 4	2436	0.52 0.51	0.87
10	13636	Conventions Evidence Organization	2 4 4	2344	0.46 0.51	0.79
11	13723	Conventions Evidence Organization	2 4 4	2095	0.59 0.63	0.77

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
11	13725	Conventions	2	2085	0.65	0.87
		Evidence	4			
		Organization	4			

### 10.1.3 MACHINE-IDENTIFIED CONDITION CODES

#### Verifications with Machine-Identified Condition Codes:

The Autoscore models have been expanded to include limited identification of condition codes. It should be noted that machine assigned condition codes are not the same as those previously assigned by human readers. A general, non-specific condition code category is estimated by a statistical scoring model based on responses in the training set that were assigned condition codes by human readers. In addition, a set of rule-based condition codes is also computed.

The available condition codes include:

- **NO\_RESPONSE:** No non-blank characters are detected in the response.
- **NOT\_ENOUGH\_DATA:** Student response is less than 11 words.
- **PROMPT\_COPY\_MATCH:** Student response is substantially copied from the passage or item prompt (flagged when more than 50% of response text matches the prompt or when the response includes more than 70% sequential match with prompt).
- **DUPLICATE\_TEXT:** Student response is substantially comprised of repeated text copied over and over (flagged when ratio of unique text is less than 70% of total response).
- **NONSPECIFIC:** Essay scoring engine predicts the assignment of a condition code.

Responses receiving the **NO\_RESPONSE** condition code are considered not attempted and do not receive a score. All other condition codes imply an attempt and receive the lowest possible dimension score for purposes of ability estimation.

All responses assigned the **NONSPECIFIC** condition code for human verification:

- If the verification reader confirms that a condition code should be assigned, the verification reader returns the **NONSPECIFIC** condition code.
- If the verification reader would not assign a condition code to the response, then the verification reader provides a dimension score.

For score reporting, **NO\_RESPONSE** will be reported as Blank. All other condition codes will be reported as non-scorable responses (e.g., NS). Please note the responses receiving machine assigned condition codes should not be routed for human verification with exception of **NONSPECIFIC**. Exhibit 10.1.3.1 presents percentage of the machine assigned condition code for spring 2017 administrations and Exhibit 10.1.3.2 presents percentage of the machine assigned condition code for spring 2018 administrations.

**Exhibit 10.1.3.1 Frequency of Machine Assigned Condition Codes for Spring 2017 Writing Prompts**

Machine Assigned Condition Code	Percentage of Condition Code						
	PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension	ALL	ALL	ALL	ALL	C	E	O

<b>G3E</b>	13023	9	0	0	1	0	0	0
	13026	13	0	0	1	0	0	0
<b>G4E</b>	13094	26	0	0	1	0	0	0
	13095	9	0	0	1	0	0	0
<b>G5E</b>	13236	7	0	0	0	0	0	0
	13239	10	0	0	0	0	0	0
<b>G6E</b>	13304	9	0	0	0	0	0	0
	13308	6	0	0	0	0	0	0
<b>G7E</b>	13402	12	0	0	0	0	0	0
	13403	3	0	0	0	0	0	0
<b>G8E</b>	13437	7	0	0	0	2	0	2
	13452	4	0	0	0	0	0	0
<b>G9E</b>	13557	4	0	0	0	1	3	3
	13566	4	0	0	0	0	0	0
<b>G10E</b>	13639	4	0	0	0	0	6	6
	13640	1	0	0	0	0	3	0
<b>G11E</b>	13722	1	0	0	0	0	0	0
	13724	2	0	0	1	0	0	0

Note: The machine-identified condition code except NONSPECIFIC should be assigned across all three dimensions.

#### Exhibit 10.1.3.2 Frequency of Machine Assigned Condition Codes for Spring 2018 Writing Prompts

Machine Assigned Condition Code	Percentage of Condition Code							
	PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC			
Dimension	ALL	ALL	ALL	ALL	C	E	O	
<b>G3E</b>	13021	12	0	0	1	0	0	0
	13024	10	0	0	1	0	0	0
<b>G4E</b>	13118	6	0	0	1	0	0	0
	13121	5	0	0	1	0	0	0
<b>G5E</b>	13237	9	0	0	0	0	0	0
	13238	4	0	0	0	0	0	0
<b>G6E</b>	13305	8	0	0	0	0	0	0
	13309	5	0	0	0	0	0	0
<b>G7E</b>	13400	8	0	0	0	0	0	0
	13405	5	0	0	0	0	0	0
<b>G8E</b>	13438	4	0	0	0	1	1	1
	13453	4	0	0	0	2	0	2
<b>G9E</b>	13554	5	0	0	0	2	2	2
	13565	2	0	0	0	1	0	0
<b>G10E</b>	13635	2	0	0	0	0	0	0
	13636	5	0	0	0	0	0	0
<b>G11E</b>	13723	1	0	0	0	0	0	0
	13725	3	0	0	1	0	0	0

## 10.2 HANDSCORING

Handscoring of online essay responses for statistical model building, as well as handscoring of all essay responses from paper-based testing (PBT) 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 PBT administrations were handscored, with 10 percent of all paper responses receiving a second reading (Reader 2) to monitor and maintain sufficient inter-rater reliability, as discussed in the following sections. 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 mathematics 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 mathematics was based on the resolution score, when present, or else on the initial read.

In spring 2018, 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, non-scorable 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 PBT 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 on to 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 the responses in a set, the scorer reviewed all 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 mathematics constructed-response items.

---

### 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.<sup>59</sup>

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 mathematics. 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

---

<sup>59</sup> 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.

writing conventions, and solve a series of mathematics 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 mathematics 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 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.<sup>60</sup>

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 mathematics 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 of scoring student responses, and assess the scorers' understanding of the scoring criteria before they can 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. Using 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

---

<sup>60</sup> 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.

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 mathematics, MI trained scorers to handscore a limited number of mathematics items from the paper-pencil assessment that could not be machine-scored. Scoring leadership reviewed all handscored mathematics 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. Mathematics items were trained and scored individually with the use of the provided scoring rubrics. Qualified mathematics scorers received training that included all possible answers to each individual item.

Mathematics 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 (QA) procedures are enforced throughout all stages of Arizona’s Measurement of Educational Readiness to Inform Teaching (AzMERIT) test development, administration, and scoring and reporting of results. This section describes QA procedures associated with the following:

- Test construction
- Test production
- Answer document processing
- Data preparation
- Equating and scaling
- Scoring and reporting

Because QA procedures pervade all aspects of test development, we note that discussion of QA 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 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 (DOK) 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 Form Builder software, which automates important form construction activities to ensure development of equated test forms. Form Builder interfaces with AIR’s Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves, 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 Form Builder generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, Form Builder 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 the Arizona Department of Education (ADE), The American Institutes for Research (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 Form Builder 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-pencil 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 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 QA 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 QA process:

1. Content for online delivery shares some processes with paper-pencil versions, but also requires additional, unique steps.
2. Online test delivery system (TDS) 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 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-pencil 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 blue-line review of items much earlier in the process, removing it from the critical path.

The production of computer-based tests (CBT) 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 Form Builder 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 (UAT) and final review.

#### 11.3.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The ITS integrates directly with the 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. In the following sections, 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 Form Builder (a component of 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 UAT. UAT of the TDS 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 to the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document QA 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 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. Measurement, Inc. (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 QA 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 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 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 Local Educational Agency (LEA) 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 TDS 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 is the tool that is used to pull data

from the DOR for delivery to ADE and their QA 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 QA 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 Secure File Transfer Protocol site.

Prior to operational work, AIR produces simulated datasets for 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 (VSC) software, 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 if 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 QA 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 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 test taker or those around him or her. 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 QA 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, TA, and school level that may indicate cheating. The QA 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 QA reports.

### Exhibit 11.7.2.1 Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
<b>Item Analysis Report</b>	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)
<b>Forensic Analysis</b>	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 item response theory- (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  $\delta_{ij}$  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, TA, 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 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 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 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 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. 1-28–1-40), Retrieved from [http://www.azed.gov/assessment/files/2014/11/spring-2015-azmerit-standard-setting\\_091415-full-report.pdf](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  $\pi$ .

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.

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	13%	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	13%	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	13%	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%	13%-19% (Writing)

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	28%	32%
Expressions and Equations	29%	33%
Geometry, Statistics and Probability	15%	19%

Algebra I		
Conceptual Categories	Min.	Max.
Algebra	33%	39%
Functions	37%	43%
Statistics	23%	28%

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%

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	28%	32%
Similarity, Right Triangles and Trigonometry	30%	34%
Circles, Geometric Measurement and Geometric Properties with Equations	15%	19%
Modeling with Geometry	19%	23%

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

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	29%	33%
Functions	21%	25%
Geometry	17%	21%
Statistics and Probability and The Number System	19%	27%

Algebra II		
Conceptual Categories	Min.	Max.
Algebra	34%	38%
Functions	30%	34%
Statistics	30%	34%

For more information go to [www.azed.gov/AzMERIT](http://www.azed.gov/AzMERIT)



**Appendix C.1a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 ELA**

Invariance Model	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	13733.772	1804				
Metric	13956.001	1847	Configural	222.230 (43)	< .01	.000
Scalar	14638.886	1890	Metric	682.885 (43)	< .01	.000
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	7117.742	1804				
Metric	7271.625	1847	Configural	153.883 (43)	< .01	.000
Scalar	7347.689	1890	Metric	76.064 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	11819.287	1804				
Metric	12378.423	1847	Configural	559.136 (43)	< .01	.000
Scalar	12648.918	1890	Metric	270.495 (43)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	7034.814	1804				
Metric	7278.311	1847	Configural	243.498 (43)	< .01	.000
Scalar	7436.232	1890	Metric	157.921 (43)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	13509.610	1804				
Metric	13867.551	1847	Configural	357.941 (43)	< .01	.000
Scalar	14304.595	1890	Metric	437.045 (43)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
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
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	13549.468	1804				
Metric	14031.524	1847	Configural	482.057 (43)	< .01	.000
Scalar	14284.069	1890	Metric	252.545 (43)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	14638.886	1890	< .01	0.945	0.039
<b>Model B-1</b>	7347.689	1890	< .01	0.954	0.035
<b>Model B-2</b>	12648.918	1890	< .01	0.941	0.039
<b>Model B-3</b>	7101.063	1890	< .01	0.956	0.033
<b>Model B-4</b>	7436.232	1890	< .01	0.971	0.023
<b>Model B-5</b>	7172.388	1890	< .01	0.954	0.034
<b>Model C</b>	14304.595	1890	< .01	0.944	0.036
<b>Model D</b>	14103.428	1890	< .01	0.946	0.039
<b>Model E</b>	14284.069	1890	< .01	0.966	0.025
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	10163.750	1804				
Metric	10328.986	1847	Configural	165.236 (43)	< .01	.000
Scalar	10978.449	1890	Metric	649.463 (43)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	5563.647	1804				
Metric	5758.299	1847	Configural	194.652 (43)	< .01	.000
Scalar	5871.122	1890	Metric	112.823 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	8716.399	1804				
Metric	9243.675	1847	Configural	527.276 (43)	< .01	.001
Scalar	9913.604	1890	Metric	669.929 (43)	< .01	.001
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	5405.843	1804				
Metric	5488.063	1847	Configural	82.220 (43)	< .01	.001
Scalar	5701.817	1890	Metric	213.754 (43)	< .01	.001
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	5509.142	1804				
Metric	5806.752	1847	Configural	297.610 (43)	< .01	.001
Scalar	6041.940	1890	Metric	235.189 (43)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	5468.639	1804				
Metric	5521.446	1847	Configural	52.807	0.15	.001
Scalar	5556.291	1890	Metric	34.845	0.81	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	10234.397	1804				
Metric	10695.816	1847	Configural	461.419 (43)	< .01	.000
Scalar	11194.600	1890	Metric	498.784 (43)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	10133.666	1804				
Metric	10554.761	1847	Configural	421.096 (43)	< .01	.001
Scalar	10709.722	1890	Metric	154.961 (43)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	10034.335	1804				
Metric	10648.370	1847	Configural	614.035 (43)	< .01	.001
Scalar	10960.887	1890	Metric	312.516 (43)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	10978.449	1890	< .01	0.947	0.036
<b>Model B-1</b>	5871.122	1890	< .01	0.986	0.016
<b>Model B-2</b>	9913.604	1890	< .01	0.968	0.024
<b>Model B-3</b>	5701.817	1890	< .01	0.947	0.031
<b>Model B-4</b>	6041.940	1890	< .01	0.984	0.016
<b>Model B-5</b>	5556.291	1890	< .01	0.974	0.021
<b>Model C</b>	11194.600	1890	< .01	0.943	0.033
<b>Model D</b>	10709.722	1890	< .01	0.946	0.036
<b>Model E</b>	10960.887	1890	< .01	0.986	0.016
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	11011.824	1804				
Metric	11305.868	1847	Configural	294.044 (43)	< .01	.000
Scalar	11780.179	1890	Metric	474.311 (43)	< .01	.000
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	6128.336	1804				
Metric	6312.812	1847	Configural	184.476 (43)	< .01	.000
Scalar	6406.820	1890	Metric	94.008 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	9646.302	1804				
Metric	10044.148	1847	Configural	397.846 (43)	< .01	.001
Scalar	10327.027	1890	Metric	282.879 (43)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	6160.021	1804				
Metric	6479.098	1847	Configural	319.077 (43)	< .01	.001
Scalar	6674.439	1890	Metric	195.341 (43)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	10971.222	1804				
Metric	11588.948	1847	Configural	617.727 (43)	< .01	.000
Scalar	12189.102	1890	Metric	600.153 (43)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	11119.735	1804				
Metric	11347.250	1847	Configural	227.515 (43)	< .01	.000
Scalar	11426.483	1890	Metric	79.233 (43)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	11049.637	1804				
Metric	11494.150	1847	Configural	444.512 (43)	< .01	.000
Scalar	11680.610	1890	Metric	186.460 (43)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	11780.179	1890	< .01	0.970	0.031
<b>Model B-1</b>	6406.820	1890	< .01	0.986	0.016
<b>Model B-2</b>	10327.027	1890	< .01	0.964	0.032
<b>Model B-3</b>	6278.654	1890	< .01	0.969	0.027
<b>Model B-4</b>	6674.439	1890	< .01	0.984	0.016
<b>Model B-5</b>	6129.819	1890	< .01	0.970	0.027
<b>Model C</b>	12189.102	1890	< .01	0.966	0.029
<b>Model D</b>	11426.483	1890	< .01	0.970	0.031
<b>Model E</b>	11680.610	1890	< .01	0.987	0.015
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	9698.728	1804				
Metric	10093.239	1847	Configural	394.510 (43)	< .01	.001
Scalar	10841.059	1890	Metric	747.820 (43)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	5611.090	1804				
Metric	5743.655	1847	Configural	132.565 (43)	< .01	.000
Scalar	5860.740	1890	Metric	117.085 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	8409.068	1804				
Metric	8938.388	1847	Configural	529.319 (43)	< .01	.001
Scalar	9380.459	1890	Metric	442.072 (43)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	5466.053	1804				
Metric	5516.666	1847	Configural	50.614	0.20	.000
Scalar	5655.420	1890	Metric	138.753	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	5652.413	1804				
Metric	5960.932	1847	Configural	308.519 (43)	< .01	.000
Scalar	6181.560	1890	Metric	220.628 (43)	< .01	.001
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	9547.579	1804				
Metric	10051.098	1847	Configural	503.520 (43)	< .01	.001
Scalar	10564.126	1890	Metric	513.028 (43)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	9673.726	1804				
Metric	10094.442	1847	Configural	420.715 (43)	< .01	.001
Scalar	10208.083	1890	Metric	113.641 (43)	< .01	.001
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	9627.312	1804				
Metric	9925.714	1847	Configural	298.403 (43)	< .01	.000
Scalar	10293.927	1890	Metric	368.213 (43)	< .01	.001
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	10841.059	1890	< .01	0.880	0.032
<b>Model B-1</b>	5860.740	1890	< .01	0.966	0.030
<b>Model B-2</b>	9380.459	1890	< .01	0.961	0.033
<b>Model B-3</b>	5655.420	1890	< .01	0.989	0.013
<b>Model B-4</b>	6181.560	1890	< .01	0.986	0.016
<b>Model B-5</b>	5526.887	1890	< .01	0.969	0.027
<b>Model C</b>	10564.126	1890	< .01	0.870	0.031
<b>Model D</b>	10208.083	1890	< .01	0.890	0.030
<b>Model E</b>	10293.927	1890	< .01	0.880	0.031
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	7386.057	1804				
Metric	7573.876	1847	Configural	187.819 (43)	< .01	.000
Scalar	8502.762	1890	Metric	928.886 (43)	< .01	.002
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	4703.545	1804				
Metric	4856.004	1847	Configural	152.459 (43)	< .01	.000
Scalar	4973.729	1890	Metric	117.725 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	6684.172	1804				
Metric	7103.969	1847	Configural	419.798 (43)	< .01	.000
Scalar	7451.795	1890	Metric	347.826 (43)	< .01	.001
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	4623.233	1804				
Metric	4784.118	1847	Configural	160.885 (43)	< .01	.000
Scalar	4959.031	1890	Metric	174.913 (43)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	7391.785	1804				
Metric	7699.688	1847	Configural	307.903 (43)	< .01	.001
Scalar	8152.133	1890	Metric	452.445 (43)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
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
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	7541.658	1804				
Metric	7791.811	1847	Configural	250.154 (43)	< .01	.000
Scalar	8127.541	1890	Metric	335.730 (43)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	8502.762	1890	< .01	0.967	0.030
<b>Model B-1</b>	4973.729	1890	< .01	0.988	0.013
<b>Model B-2</b>	7451.795	1890	< .01	0.961	0.029
<b>Model B-3</b>	4939.561	1890	< .01	0.966	0.024
<b>Model B-4</b>	4959.031	1890	< .01	0.988	0.013
<b>Model B-5</b>	4769.442	1890	< .01	0.969	0.023
<b>Model C</b>	8152.133	1890	< .01	0.989	0.013
<b>Model D</b>	7923.163	1890	< .01	0.966	0.028
<b>Model E</b>	8127.541	1890	< .01	0.990	0.012
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	14027.579	1804				
Metric	14325.678	1847	Configural	298.099 (43)	< .01	.000
Scalar	15084.733	1890	Metric	759.055 (43)	< .01	.000
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	7738.802	1804				
Metric	7889.631	1847	Configural	150.829 (43)	< .01	.000
Scalar	8035.171	1890	Metric	145.540 (43)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	12202.711	1804				
Metric	12716.273	1847	Configural	513.562 (43)	< .01	.000
Scalar	12922.089	1890	Metric	205.816 (43)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	7776.336	1804				
Metric	7970.933	1847	Configural	194.597 (43)	< .01	.000
Scalar	8130.334	1890	Metric	159.402 (43)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	13905.703	1804				
Metric	14240.091	1847	Configural	334.388 (43)	< .01	.000
Scalar	14864.115	1890	Metric	624.024 (43)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	14078.981	1804				
Metric	14327.866	1847	Configural	248.885 (43)	< .01	.000
Scalar	14421.238	1890	Metric	93.372 (43)	< .01	.001
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	14095.956	1804				
Metric	14211.952	1847	Configural	115.996 (43)	< .01	.000
Scalar	14414.906	1890	Metric	202.954 (43)	< .01	.001
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	15084.733	1890	< .01	0.948	0.044
<b>Model B-1</b>	8035.171	1890	< .01	0.957	0.041
<b>Model B-2</b>	12922.089	1890	< .01	0.951	0.044
<b>Model B-3</b>	7793.346	1890	< .01	0.957	0.037
<b>Model B-4</b>	8130.334	1890	< .01	0.958	0.040
<b>Model B-5</b>	7616.296	1890	< .01	0.961	0.036
<b>Model C</b>	14864.115	1890	< .01	0.947	0.038
<b>Model D</b>	14421.238	1890	< .01	0.947	0.042
<b>Model E</b>	14414.906	1890	< .01	0.950	0.035
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	9785.018	1978				
Metric	10049.839	2023	Configural	264.821 (45)	< .01	.000
Scalar	10769.979	2068	Metric	720.140 (45)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	5881.394	1978				
Metric	5997.772	2023	Configural	116.378 (45)	< .01	.000
Scalar	6098.463	2068	Metric	100.691 (45)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	8602.375	1978				
Metric	9013.777	2023	Configural	411.402 (45)	< .01	.000
Scalar	9358.174	2068	Metric	344.397 (45)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	5876.617	1978				
Metric	6018.198	2023	Configural	141.581 (45)	< .01	.000
Scalar	6266.719	2068	Metric	248.521 (45)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	9819.145	1978				
Metric	10020.292	2023	Configural	201.147 (45)	< .01	.000
Scalar	10344.222	2068	Metric	323.930 (45)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	9860.151	1978				
Metric	10147.583	2023	Configural	287.432 (45)	< .01	.000
Scalar	10258.257	2068	Metric	110.674 (45)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	9970.516	1978				
Metric	10173.098	2023	Configural	202.582 (45)	< .01	.000
Scalar	10314.595	2068	Metric	141.497 (45)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	10769.979	2068	< .01	0.965	0.023
<b>Model B-1</b>	6098.463	2068	< .01	0.972	0.019
<b>Model B-2</b>	9358.174	2068	< .01	0.963	0.023
<b>Model B-3</b>	5961.451	2068	< .01	0.974	0.017
<b>Model B-4</b>	6266.719	2068	< .01	0.973	0.019
<b>Model B-5</b>	5776.327	2068	< .01	0.978	0.015
<b>Model C</b>	10344.222	2068	< .01	0.970	0.019
<b>Model D</b>	10258.257	2068	< .01	0.967	0.022
<b>Model E</b>	10314.595	2068	< .01	0.973	0.018
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	11398.967	1978				
Metric	11643.610	2023	Configural	244.644 (45)	< .01	.000
Scalar	12454.369	2068	Metric	810.759 (45)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	6787.701	1978				
Metric	6887.953	2023	Configural	100.253 (45)	< .01	.000
Scalar	7000.945	2068	Metric	112.992 (45)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	9962.149	1978				
Metric	10396.724	2023	Configural	434.575 (45)	< .01	.001
Scalar	10754.618	2068	Metric	357.894 (45)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	6718.523	1978				
Metric	6794.553	2023	Configural	76.030 (45)	< .01	.000
Scalar	6955.160	2068	Metric	160.607 (45)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	6677.657	1978				
Metric	6909.453	2023	Configural	231.797 (45)	< .01	.001
Scalar	7122.427	2068	Metric	212.974 (45)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	11516.948	1978				
Metric	11743.401	2023	Configural	226.452 (45)	< .01	.000
Scalar	12092.078	2068	Metric	348.678 (45)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	11482.165	1978				
Metric	11785.845	2023	Configural	303.680 (45)	< .01	.000
Scalar	11878.384	2068	Metric	92.539 (45)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	11538.971	1978				
Metric	11651.848	2023	Configural	112.877 (45)	< .01	.000
Scalar	11768.221	2068	Metric	116.373 (45)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	12454.369	2068	< .01	0.962	0.036
<b>Model B-1</b>	7000.945	2068	< .01	0.960	0.021
<b>Model B-2</b>	10754.618	2068	< .01	0.947	0.025
<b>Model B-3</b>	6955.160	2068	< .01	0.958	0.021
<b>Model B-4</b>	7122.427	2068	< .01	0.960	0.021
<b>Model B-5</b>	6837.360	2068	< .01	0.963	0.018
<b>Model C</b>	12092.078	2068	< .01	0.965	0.031
<b>Model D</b>	11878.384	2068	< .01	0.965	0.035
<b>Model E</b>	11768.221	2068	< .01	0.978	0.028
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	7351.566	1978				
Metric	7591.812	2023	Configural	240.246 (45)	< .01	.000
Scalar	8303.823	2068	Metric	712.011 (45)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	5120.192	1978				
Metric	5209.060	2023	Configural	88.868 (45)	< .01	.000
Scalar	5319.957	2068	Metric	110.897 (45)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	6619.298	1978				
Metric	6981.561	2023	Configural	362.263 (45)	< .01	.001
Scalar	7298.593	2068	Metric	317.032 (45)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
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
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	5244.742	1978				
Metric	5422.163	2023	Configural	177.422 (45)	< .01	.001
Scalar	5688.003	2068	Metric	265.839 (45)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
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
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	7513.769	1978				
Metric	7678.313	2023	Configural	164.544 (45)	< .01	.000
Scalar	8058.493	2068	Metric	380.180 (45)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	7488.459	1978				
Metric	7754.618	2023	Configural	266.159 (45)	< .01	.001
Scalar	7914.265	2068	Metric	159.647 (45)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	7717.892	1978				
Metric	7812.260	2023	Configural	94.368 (45)	< .01	.000
Scalar	7923.082	2068	Metric	110.822 (45)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	8303.823	2068	< .01	0.959	0.034
<b>Model B-1</b>	5319.957	2068	< .01	0.973	0.020
<b>Model B-2</b>	7298.593	2068	< .01	0.964	0.023
<b>Model B-3</b>	5129.101	2068	< .01	0.975	0.018
<b>Model B-4</b>	5688.003	2068	< .01	0.972	0.021
<b>Model B-5</b>	5145.823	2068	< .01	0.977	0.016
<b>Model C</b>	8058.493	2068	< .01	0.969	0.027
<b>Model D</b>	7914.265	2068	< .01	0.960	0.032
<b>Model E</b>	7923.082	2068	< .01	0.985	0.014
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	178020.127	1890				
Metric	178975.724	1934	Configural	955.597 (44)	< .01	.001
Scalar	183551.597	1978	Metric	4575.874 (44)	< .01	.000
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	74981.430	1890				
Metric	76809.764	1934	Configural	1828.334 (44)	< .01	.000
Scalar	77843.804	1978	Metric	1034.040 (44)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	143550.580	1890				
Metric	149168.011	1934	Configural	5617.431 (44)	< .01	.000
Scalar	154151.396	1978	Metric	4983.385 (44)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	71303.946	1890				
Metric	71673.151	1934	Configural	369.205 (44)	< .01	.001
Scalar	72140.060	1978	Metric	466.909 (44)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	74961.905	1890				
Metric	77173.889	1934	Configural	2211.984 (44)	< .01	.000
Scalar	78790.231	1978	Metric	1616.342 (44)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	71635.066	1890				
Metric	71792.403	1934	Configural	157.337 (44)	< .01	.000
Scalar	71909.228	1978	Metric	116.825 (44)	< .01	.001
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	170307.577	1890				
Metric	177157.840	1934	Configural	6850.263 (44)	< .01	.001
Scalar	180389.601	1978	Metric	3231.762 (44)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	175711.661	1890				
Metric	179583.439	1934	Configural	3871.778 (44)	< .01	.000
Scalar	181201.055	1978	Metric	1617.616 (44)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	174317.255	1890				
Metric	180824.968	1934	Configural	6507.712 (44)	< .01	.000
Scalar	182097.212	1978	Metric	1272.244 (44)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	183551.597	1978	< .01	0.917	0.058
<b>Model B-1</b>	77843.804	1978	< .01	0.913	0.056
<b>Model B-2</b>	154151.396	1978	< .01	0.909	0.056
<b>Model B-3</b>	72140.060	1978	< .01	0.922	0.050
<b>Model B-4</b>	78790.231	1978	< .01	0.917	0.055
<b>Model B-5</b>	71909.228	1978	< .01	0.915	0.054
<b>Model C</b>	180389.601	1978	< .01	0.905	0.058
<b>Model D</b>	181201.055	1978	< .01	0.914	0.057
<b>Model E</b>	182097.212	1978	< .01	0.916	0.057
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	78513.118	1890				
Metric	79737.405	1934	Configural	1224.287 (44)	< .01	.000
Scalar	84155.308	1978	Metric	4417.903 (44)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	31709.452	1890				
Metric	33546.915	1934	Configural	1837.464 (44)	< .01	.000
Scalar	34341.647	1978	Metric	794.731 (44)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	62665.155	1890				
Metric	66540.482	1934	Configural	3875.328 (44)	< .01	.001
Scalar	69648.954	1978	Metric	3108.472 (44)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	29285.910	1890				
Metric	29504.006	1934	Configural	218.096 (44)	< .01	.000
Scalar	30032.709	1978	Metric	528.703 (44)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	31111.359	1890				
Metric	32984.692	1934	Configural	1873.333 (44)	< .01	.001
Scalar	34001.235	1978	Metric	1016.544 (44)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	29303.472	1890				
Metric	29364.693	1934	Configural	61.220 (44)	.044	.000
Scalar	29494.527	1978	Metric	129.835 (44)	< .01	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	74429.878	1890				
Metric	80370.099	1934	Configural	5940.221 (44)	< .01	.001
Scalar	84016.074	1978	Metric	3645.976 (44)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	77747.392	1890				
Metric	80662.164	1934	Configural	2914.772 (44)	< .01	.000
Scalar	81460.601	1978	Metric	798.437 (44)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	76537.210	1890				
Metric	82372.857	1934	Configural	5835.648 (44)	< .01	.001
Scalar	83624.630	1978	Metric	1251.772 (44)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	84155.308	1978	< .01	0.962	0.035
<b>Model B-1</b>	34341.647	1978	< .01	0.961	0.032
<b>Model B-2</b>	69648.954	1978	< .01	0.957	0.033
<b>Model B-3</b>	30032.709	1978	< .01	0.960	0.029
<b>Model B-4</b>	34001.235	1978	< .01	0.961	0.031
<b>Model B-5</b>	29494.527	1978	< .01	0.964	0.029
<b>Model C</b>	84016.074	1978	< .01	0.959	0.033
<b>Model D</b>	81460.601	1978	< .01	0.962	0.034
<b>Model E</b>	83624.630	1978	< .01	0.964	0.033
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	72135.653	1890				
Metric	73067.921	1934	Configural	932.268 (44)	< .01	.000
Scalar	76282.010	1978	Metric	3214.089 (44)	< .01	.000
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	29168.642	1890				
Metric	31046.373	1934	Configural	1877.730 (44)	< .01	.001
Scalar	31621.282	1978	Metric	574.909 (44)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	57946.466	1890				
Metric	62395.316	1934	Configural	4448.850 (44)	< .01	.001
Scalar	64315.473	1978	Metric	1920.157 (44)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	26768.137	1890				
Metric	27112.072	1934	Configural	343.935 (44)	< .01	.000
Scalar	27481.486	1978	Metric	369.413 (44)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	29519.527	1890				
Metric	32104.916	1934	Configural	2585.389 (44)	< .01	.001
Scalar	32872.831	1978	Metric	767.915 (44)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	26763.882	1890				
Metric	26846.021	1934	Configural	82.140 (44)	< .01	.001
Scalar	26943.191	1978	Metric	97.169 (44)	< .01	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	68366.428	1890				
Metric	74117.853	1934	Configural	5751.424 (44)	< .01	.001
Scalar	79295.211	1978	Metric	5177.358 (44)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	71225.495	1890				
Metric	74139.995	1934	Configural	2914.500 (44)	< .01	.000
Scalar	74628.714	1978	Metric	488.719 (44)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	70880.121	1890				
Metric	74815.092	1934	Configural	3934.970 (44)	< .01	.000
Scalar	77058.496	1978	Metric	2243.404 (44)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	76282.010	1978	< .01	0.968	0.032
<b>Model B-1</b>	31621.282	1978	< .01	0.968	0.029
<b>Model B-2</b>	64315.473	1978	< .01	0.964	0.031
<b>Model B-3</b>	27481.486	1978	< .01	0.968	0.027
<b>Model B-4</b>	32872.831	1978	< .01	0.968	0.029
<b>Model B-5</b>	26943.191	1978	< .01	0.970	0.027
<b>Model C</b>	79295.211	1978	< .01	0.963	0.031
<b>Model D</b>	74628.714	1978	< .01	0.967	0.031
<b>Model E</b>	77058.496	1978	< .01	0.968	0.030
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	90030.261	2068				
Metric	91811.680	2114	Configural	1781.419 (46)	< .01	.000
Scalar	98752.423	2160	Metric	6940.743 (46)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	37453.505	2068				
Metric	40103.361	2114	Configural	2649.856 (46)	< .01	.001
Scalar	40679.803	2160	Metric	576.442 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	71421.466	2068				
Metric	78474.546	2114	Configural	7053.080 (46)	< .01	.002
Scalar	80280.423	2160	Metric	1805.877 (46)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	35345.538	2068				
Metric	35560.721	2114	Configural	215.182 (46)	< .01	.001
Scalar	35866.572	2160	Metric	305.851 (46)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	37208.036	2068				
Metric	39896.741	2114	Configural	2688.705 (46)	< .01	.001
Scalar	40737.236	2160	Metric	840.495 (46)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	34775.418	2068				
Metric	34927.089	2114	Configural	151.671 (46)	< .01	.000
Scalar	34986.399	2160	Metric	59.310 (46)	.090	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	80371.915	2068				
Metric	87601.812	2114	Configural	7229.897 (46)	< .01	.001
Scalar	96412.066	2160	Metric	8810.254 (46)	< .01	.002
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	88516.533	2068				
Metric	93104.084	2114	Configural	4587.551 (46)	< .01	.001
Scalar	93599.712	2160	Metric	495.628 (46)	< .01	.001
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	86690.679	2068				
Metric	91830.076	2114	Configural	5139.397 (46)	< .01	.000
Scalar	94418.244	2160	Metric	2588.169 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	98752.423	2160	< .01	0.958	0.037
<b>Model B-1</b>	40679.803	2160	< .01	0.960	0.033
<b>Model B-2</b>	80280.423	2160	< .01	0.953	0.036
<b>Model B-3</b>	35866.572	2160	< .01	0.957	0.031
<b>Model B-4</b>	40737.236	2160	< .01	0.960	0.033
<b>Model B-5</b>	34986.399	2160	< .01	0.962	0.031
<b>Model C</b>	96412.066	2160	< .01	0.952	0.034
<b>Model D</b>	93599.712	2160	< .01	0.957	0.036
<b>Model E</b>	94418.244	2160	< .01	0.964	0.033
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	37314.410	2068				
Metric	38506.401	2114	Configural	1191.991 (46)	< .01	.000
Scalar	44483.010	2160	Metric	5976.609 (46)	< .01	.002
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	18571.581	2068				
Metric	20221.959	2114	Configural	1650.378 (46)	< .01	.001
Scalar	20875.847	2160	Metric	653.888 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	30441.823	2068				
Metric	35078.471	2114	Configural	4636.648 (46)	< .01	.001
Scalar	36819.342	2160	Metric	1740.870 (46)	< .01	.001
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	17534.361	2068				
Metric	17791.556	2114	Configural	257.195 (46)	< .01	.000
Scalar	18399.719	2160	Metric	608.162 (46)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	18001.839	2068				
Metric	19957.793	2114	Configural	1955.954 (46)	< .01	.001
Scalar	21012.517	2160	Metric	1054.724 (46)	< .01	.001
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	17639.958	2068				
Metric	17735.461	2114	Configural	95.503 (46)	< .01	.000
Scalar	17822.776	2160	Metric	87.315 (46)	< .01	.001
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	35767.378	2068				
Metric	39249.591	2114	Configural	3482.213 (46)	< .01	.001
Scalar	44955.521	2160	Metric	5705.930 (46)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	36685.569	2068				
Metric	39848.742	2114	Configural	3163.173 (46)	< .01	.001
Scalar	40105.854	2160	Metric	257.113 (46)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	36944.840	2068				
Metric	39359.374	2114	Configural	2414.535 (46)	< .01	.001
Scalar	41609.312	2160	Metric	2249.938 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	44483.010	2160	< .01	0.980	0.022
<b>Model B-1</b>	20875.847	2160	< .01	0.981	0.021
<b>Model B-2</b>	36819.342	2160	< .01	0.977	0.022
<b>Model B-3</b>	18399.719	2160	< .01	0.982	0.020
<b>Model B-4</b>	21012.517	2160	< .01	0.982	0.020
<b>Model B-5</b>	17822.776	2160	< .01	0.983	0.019
<b>Model C</b>	44955.521	2160	< .01	0.980	0.020
<b>Model D</b>	40105.854	2160	< .01	0.981	0.021
<b>Model E</b>	41609.312	2160	< .01	0.985	0.019
<b>Model F</b>	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	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	51923.973	2068				
Metric	53374.054	2114	Configural	1450.081 (46)	< .01	.000
Scalar	57215.968	2160	Metric	3841.913 (46)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	24553.604	2068				
Metric	25354.154	2114	Configural	800.550 (46)	< .01	.000
Scalar	26162.288	2160	Metric	808.134 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	43941.130	2068				
Metric	45642.765	2114	Configural	1701.635 (46)	< .01	.000
Scalar	46780.427	2160	Metric	1137.662 (46)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	23020.720	2068				
Metric	23325.360	2114	Configural	304.640 (46)	< .01	.000
Scalar	23646.081	2160	Metric	320.721 (46)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	24766.121	2068				
Metric	25703.123	2114	Configural	937.002 (46)	< .01	.000
Scalar	26643.128	2160	Metric	940.004 (46)	< .01	.001
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	23056.447	2068				
Metric	23120.850	2114	Configural	64.402 (46)	.038	.000
Scalar	23224.440	2160	Metric	103.590 (46)	< .01	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	48778.924	2068				
Metric	52531.621	2114	Configural	3752.697 (46)	< .01	.001
Scalar	57853.129	2160	Metric	5321.508 (46)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	52081.658	2068				
Metric	53241.940	2114	Configural	1160.282 (46)	< .01	.000
Scalar	53724.454	2160	Metric	482.513 (46)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	51790.117	2068				
Metric	52710.250	2114	Configural	920.133 (46)	< .01	.000
Scalar	54441.917	2160	Metric	1731.667 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	57215.968	2160	< .01	0.965	0.028
<b>Model B-1</b>	26162.288	2160	< .01	0.964	0.026
<b>Model B-2</b>	46780.427	2160	< .01	0.964	0.027
<b>Model B-3</b>	23646.081	2160	< .01	0.964	0.026
<b>Model B-4</b>	26643.128	2160	< .01	0.962	0.027
<b>Model B-5</b>	23224.440	2160	< .01	0.966	0.025
<b>Model C</b>	57853.129	2160	< .01	0.965	0.024
<b>Model D</b>	53724.454	2160	< .01	0.966	0.026
<b>Model E</b>	54441.917	2160	< .01	0.971	0.023
<b>Model F</b>	58394.447	2160	< .01	0.966	0.024

Appendix C.17a. Global Model Fit Indices of Measurement Invariance Tests for Algebra I

Invariance Model	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	31880.312	2068				
Metric	32905.318	2114	Configural	1025.006 (46)	< .01	.000
Scalar	36406.798	2160	Metric	3501.480 (46)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	16142.158	2068				
Metric	17121.893	2114	Configural	979.735 (46)	< .01	.001
Scalar	17604.115	2160	Metric	482.221 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	26017.305	2068				
Metric	29072.903	2114	Configural	3055.599 (46)	< .01	.001
Scalar	30186.879	2160	Metric	1113.975 (46)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	15857.962	2068				
Metric	16107.596	2114	Configural	249.634 (46)	< .01	.000
Scalar	16536.207	2160	Metric	428.611 (46)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	16758.206	2068				
Metric	18161.297	2114	Configural	1403.091 (46)	< .01	.001
Scalar	18954.656	2160	Metric	793.359 (46)	< .01	.001
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	15465.159	2068				
Metric	15518.118	2114	Configural	52.959 (46)	.224	.001
Scalar	15611.302	2160	Metric	93.185 (46)	< .01	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	30806.225	2068				
Metric	32582.106	2114	Configural	1775.881 (46)	< .01	.000
Scalar	34778.963	2160	Metric	2196.858 (46)	< .01	.001
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	31432.467	2068				
Metric	33610.533	2114	Configural	2178.066 (46)	< .01	.001
Scalar	34009.880	2160	Metric	399.346 (46)	< .01	.001
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	31591.441	2068				
Metric	32369.885	2114	Configural	778.443 (46)	< .01	.000
Scalar	33674.958	2160	Metric	1305.073 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	36406.798	2160	< .01	0.929	0.024
<b>Model B-1</b>	17604.115	2160	< .01	0.981	0.021
<b>Model B-2</b>	30186.879	2160	< .01	0.974	0.023
<b>Model B-3</b>	16536.207	2160	< .01	0.981	0.021
<b>Model B-4</b>	18954.656	2160	< .01	0.979	0.022
<b>Model B-5</b>	15611.302	2160	< .01	0.983	0.020
<b>Model C</b>	34778.963	2160	< .01	0.927	0.024
<b>Model D</b>	34009.880	2160	< .01	0.930	0.023
<b>Model E</b>	33674.958	2160	< .01	0.936	0.023
<b>Model F</b>	33945.363	2160	< .01	0.982	0.020

Appendix C.16a. Global Model Fit Indices of Measurement Invariance Tests for Geometry

Invariance Model	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	53224.910	2068				
Metric	54454.502	2114	Configural	1229.591 (46)	< .01	.000
Scalar	57940.176	2160	Metric	3485.674 (46)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	27938.686	2068				
Metric	28828.529	2114	Configural	889.843 (46)	< .01	.000
Scalar	29338.108	2160	Metric	509.579 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	43389.350	2068				
Metric	47107.528	2114	Configural	3718.178 (46)	< .01	.001
Scalar	48016.657	2160	Metric	909.129 (46)	< .01	.000
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	27092.053	2068				
Metric	27384.954	2114	Configural	292.902 (46)	< .01	.000
Scalar	27629.871	2160	Metric	244.916 (46)	< .01	.001
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	27989.160	2068				
Metric	29872.199	2114	Configural	1883.040 (46)	< .01	.001
Scalar	30830.067	2160	Metric	957.867 (46)	< .01	.000
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	26391.366	2068				
Metric	26454.861	2114	Configural	63.495 (46)	.045	.001
Scalar	26525.935	2160	Metric	71.074 (46)	.010	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	52492.263	2068				
Metric	53783.151	2114	Configural	1290.888 (46)	< .01	.000
Scalar	55970.200	2160	Metric	2187.049 (46)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	51894.256	2068				
Metric	54836.635	2114	Configural	2942.380 (46)	< .01	.001
Scalar	55244.890	2160	Metric	408.255 (46)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	53475.540	2068				
Metric	54059.554	2114	Configural	584.014 (46)	< .01	.000
Scalar	54692.029	2160	Metric	632.476 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	57940.176	2160	< .01	0.930	0.038
<b>Model B-1</b>	29338.108	2160	< .01	0.950	0.033
<b>Model B-2</b>	48016.657	2160	< .01	0.929	0.035
<b>Model B-3</b>	27629.871	2160	< .01	0.947	0.034
<b>Model B-4</b>	30830.067	2160	< .01	0.951	0.032
<b>Model B-5</b>	26525.935	2160	< .01	0.954	0.031
<b>Model C</b>	55970.200	2160	< .01	0.942	0.031
<b>Model D</b>	55244.890	2160	< .01	0.934	0.036
<b>Model E</b>	54692.029	2160	< .01	0.944	0.032
<b>Model F</b>	55431.388	2160	< .01	0.955	0.027

**Appendix C.18a. Global Model Fit Indices of Measurement Invariance Tests for Algebra II**

Invariance Model	$\chi^2$	df	$\chi^2$ Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	p value	
<b>Model A: Students' Gender (Female vs. Male)</b>						
Configural	17718.329	2068				
Metric	18874.760	2114	Configural	1156.431 (46)	< .01	.000
Scalar	20911.565	2160	Metric	2036.806 (46)	< .01	.001
<b>Model B-1: Students' Ethnicity (African American vs. White)</b>						
Configural	10607.551	2068				
Metric	11002.537	2114	Configural	394.986 (46)	< .01	.000
Scalar	11497.152	2160	Metric	494.615 (46)	< .01	.000
<b>Model B-2: Students' Ethnicity (Hispanics vs. White)</b>						
Configural	15096.010	2068				
Metric	16321.688	2114	Configural	1225.678 (46)	< .01	.001
Scalar	17382.791	2160	Metric	1061.103 (46)	< .01	.001
<b>Model B-3: Students' Ethnicity (Asian vs. White)</b>						
Configural	10235.537	2068				
Metric	10522.133	2114	Configural	286.595 (46)	< .01	.001
Scalar	10828.678	2160	Metric	306.545 (46)	< .01	.000
<b>Model B-4: Students' Ethnicity (American Indian vs. White)</b>						
Configural	10485.655	2068				
Metric	11186.459	2114	Configural	700.803 (46)	< .01	.001
Scalar	12270.158	2160	Metric	1083.699 (46)	< .01	.001
<b>Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)</b>						
Configural	10048.537	2068				
Metric	10133.402	2114	Configural	84.865 (46)	< .01	.001
Scalar	10216.730	2160	Metric	83.328 (46)	< .01	.000
<b>Model C: Students' SPED Status (Special Education vs. Non-SPED)</b>						
Configural	17849.740	2068				
Metric	18266.012	2114	Configural	416.272 (46)	< .01	.000
Scalar	19649.191	2160	Metric	1383.179 (46)	< .01	.000
<b>Model D: Students' Low Income Status (Low Income vs. Non-Low Income)</b>						
Configural	17496.271	2068				
Metric	18725.869	2114	Configural	1229.598 (46)	< .01	.001
Scalar	19370.546	2160	Metric	644.677 (46)	< .01	.000
<b>Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)</b>						
Configural	17883.903	2068				
Metric	18249.789	2114	Configural	365.886 (46)	< .01	.000
Scalar	18641.474	2160	Metric	391.685 (46)	< .01	.000
<b>Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)</b>						
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		
<b>Model A</b>	20911.565	2160	< .01	0.979	0.017
<b>Model B-1</b>	11497.152	2160	< .01	0.986	0.014
<b>Model B-2</b>	17382.791	2160	< .01	0.978	0.017
<b>Model B-3</b>	10828.678	2160	< .01	0.985	0.015
<b>Model B-4</b>	12270.158	2160	< .01	0.986	0.015
<b>Model B-5</b>	10216.730	2160	< .01	0.988	0.013
<b>Model C</b>	19649.191	2160	< .01	0.986	0.013
<b>Model D</b>	19370.546	2160	< .01	0.982	0.016
<b>Model E</b>	18641.474	2160	< .01	0.986	0.013
<b>Model F</b>	19474.733	2160	< .01	0.989	0.011

**Appendix D. Regression Model Parameter Estimates of Differential Growth across Subgroups-ELA**

Parameter	2017_G3E to 2018_G4E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2523.66	0.14	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	1.14	0.13	<.0001	0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-6.49	0.26	<.0001	-0.06
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-13.49	0.44	<.0001	-0.12
<b>Low income vs. Non-Low Income(<math>\beta_{04}</math>)</b>	-2.28	0.14	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	2.81	0.43	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-3.35	0.15	<.0001	-0.05
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-4.12	0.42	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-0.29	1.46	0.8447	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-7.01	0.34	<.0001	-0.05
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.46	0.46	0.3183	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.80	0.00	<.0001	0.79
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.00	0.00	0.7160	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.01	0.01	0.3218	0.00
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.25	0.01	<.0001	-0.09
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.01	0.01	0.6557	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.01	0.00	0.0584	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.02	0.01	0.0715	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.04	0.05	0.4176	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.03	0.01	0.0101	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.01	0.6921	0.00

Parameter	2017_G4E to 2018_G5E			
	Unstandardized	SE	p value	Standardized
<b>Intercept (<math>\beta_{00}</math>)</b>	2542.43	0.14	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	1.33	0.13	<.0001	0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-8.85	0.28	<.0001	-0.08
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-13.29	0.48	<.0001	-0.11
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.54	0.14	<.0001	-0.04
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	2.76	0.44	<.0001	0.01
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.11	0.15	<.0001	-0.03
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-3.76	0.43	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-3.34	1.34	0.0123	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-6.56	0.35	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	0.45	0.47	0.3373	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.86	0.00	<.0001	0.78
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.01	0.00	0.0127	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	0.03	0.01	0.0001	0.01
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.19	0.01	<.0001	-0.06
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.04	0.01	0.0015	-0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	0.00	0.00	0.8268	0.00
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.02	0.01	0.1024	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.02	0.04	0.6231	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	0.00	0.01	0.9016	0.00
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.00	0.01	0.8749	0.00

Parameter	2017_G5E to 2018_G6E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2544.17	0.13	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	2.81	0.12	<.0001	0.04
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-5.81	0.29	<.0001	-0.06
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-13.15	0.61	<.0001	-0.10
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-1.35	0.13	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	5.85	0.42	<.0001	0.03
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.58	0.14	<.0001	-0.04
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-3.22	0.42	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-1.80	1.36	0.1873	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-4.76	0.34	<.0001	-0.03
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-1.11	0.46	0.0148	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.76	0.00	<.0001	0.80
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.01	0.00	0.0675	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	0.00	0.01	0.5279	0.00
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.22	0.01	<.0001	-0.07
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.01	0.01	0.4419	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.01	0.00	0.0065	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.00	0.01	0.7153	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.04	0.04	0.2950	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.01	0.01	0.5782	0.00
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.00	0.01	0.7603	0.00

Parameter	2017_G6E to 2018_G7E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2555.46	0.13	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	3.78	0.13	<.0001	0.06
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-9.25	0.33	<.0001	-0.08
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-12.29	0.65	<.0001	-0.08
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.39	0.14	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	4.22	0.45	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.37	0.15	<.0001	-0.03
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-2.43	0.44	<.0001	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-1.18	1.54	0.4461	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-6.72	0.35	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.76	0.51	0.1410	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.85	0.00	<.0001	0.80
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.01	0.00	0.0996	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.06	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.17	0.01	<.0001	-0.05
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.01	0.01	0.3706	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.00	0.01	0.9338	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.00	0.05	0.9453	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.04	0.01	0.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.02	0.5675	0.00

Parameter	2017_G7E to 2018_G8E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2561.85	0.13	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	1.90	0.12	<.0001	0.03
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-8.80	0.33	<.0001	-0.08
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-7.04	0.55	<.0001	-0.05
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-1.68	0.13	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	3.00	0.41	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.08	0.14	<.0001	-0.03
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-2.39	0.40	<.0001	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-4.41	1.34	0.0010	-0.01
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-6.02	0.34	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.57	0.50	0.2610	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.80	0.00	<.0001	0.81
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.01	0.00	0.0010	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.05	0.01	<.0001	-0.02
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.11	0.01	<.0001	-0.03
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.01	0.00	0.0917	0.00
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.00	0.01	0.9724	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	0.00	0.00	0.8689	0.00
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.00	0.01	0.9987	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.01	0.04	0.7702	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.05	0.01	<.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.04	0.02	0.0039	0.01

Parameter	2017_G8E to 2018_G9E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2571.00	0.13	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	2.13	0.13	<.0001	0.03
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-8.80	0.38	<.0001	-0.08
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-3.52	0.48	<.0001	-0.02
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-1.60	0.15	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	3.44	0.42	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-3.08	0.15	<.0001	-0.05
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-2.83	0.40	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-2.16	1.35	0.1087	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-2.35	0.39	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.61	0.69	0.3717	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.76	0.00	<.0001	0.80
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.01	0.00	0.0554	0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.02	0.01	0.0059	-0.01
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	0.02	0.01	0.0635	0.01
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.04	0.01	0.0010	0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.01	0.01	0.3288	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.00	0.04	0.9175	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.02	0.01	0.0845	0.00
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.02	0.02	0.3906	0.00

Parameter	2017_G9E to 2018_G10E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2569.74	0.14	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	-0.80	0.15	<.0001	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-6.83	0.43	<.0001	-0.05
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-3.86	0.53	<.0001	-0.02
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.18	0.17	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	0.70	0.49	0.1521	0.00
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-3.97	0.17	<.0001	-0.06
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-4.10	0.45	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-2.44	1.69	0.1493	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-5.05	0.42	<.0001	-0.03
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-1.25	0.60	0.0368	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.94	0.00	<.0001	0.82
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.00	0.01	0.4564	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.09	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.06	0.02	<.0001	-0.01
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.01	<.0001	-0.01
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.03	0.01	0.0225	-0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.02	0.01	<.0001	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.05	0.02	0.0034	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.01	0.06	0.8897	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.07	0.01	<.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	-0.01	0.02	0.4897	0.00



Parameter	2017_G10E to 2018_G11E			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	2568.80	0.14	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	2.81	0.15	<.0001	0.05
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-7.00	0.43	<.0001	-0.06
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-0.95	0.51	0.0649	-0.01
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.01	0.17	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	3.40	0.47	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.68	0.17	<.0001	-0.04
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-1.54	0.45	0.0006	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-2.38	1.47	0.1040	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-2.59	0.43	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	0.01	0.59	0.9818	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.85	0.00	<.0001	0.83
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.04	0.01	<.0001	-0.03
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.07	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	0.01	0.01	0.6065	0.00
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.05	0.01	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.00	0.01	0.9355	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.05	0.01	<.0001	-0.03
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.04	0.02	0.0118	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.03	0.05	0.5152	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.06	0.01	<.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	-0.05	0.02	0.0269	-0.01

**Appendix D. Regression Model Parameter Estimates of Differential Growth across Subgroups-MATH**

Parameter	2017_G3M to 2018_G4M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3560.68	0.19	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	0.34	0.17	0.0491	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-8.56	0.33	<.0001	-0.06
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-11.94	0.46	<.0001	-0.08
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.97	0.18	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	5.93	0.62	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-3.22	0.20	<.0001	-0.04
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-4.68	0.57	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-0.29	1.95	0.8823	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-7.36	0.46	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.31	0.62	0.6203	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.73	0.00	<.0001	0.78
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	0.05	0.01	<.0001	0.02
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.07	0.01	<.0001	-0.03
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.00	0.01	0.8212	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.01	0.00	0.2350	0.00
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.03	0.01	0.0041	0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.05	0.04	0.2114	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	0.01	0.01	0.5326	0.00
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.01	0.3809	0.00

Parameter	2017_G4M to 2018_G5M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3593.73	0.18	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	0.25	0.17	0.1518	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-9.91	0.35	<.0001	-0.07
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-9.44	0.48	<.0001	-0.06
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-4.07	0.18	<.0001	-0.04
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	7.48	0.61	<.0001	0.03
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.61	0.20	<.0001	-0.03
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-5.34	0.58	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	0.87	1.79	0.6262	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-7.26	0.45	<.0001	-0.03
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	0.16	0.63	0.8031	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.89	0.00	<.0001	0.86
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.05	0.00	<.0001	-0.03
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.05	0.01	<.0001	-0.02
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.13	0.01	<.0001	-0.04
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.04	0.00	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.02	0.01	0.1874	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.01	0.01	0.5284	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.00	0.04	0.9645	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.08	0.01	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.01	0.5155	0.00

Parameter	2017_G5M to 2018_G6M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3624.19	0.18	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	-1.22	0.17	<.0001	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-9.04	0.37	<.0001	-0.06
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-12.10	0.57	<.0001	-0.06
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-3.35	0.18	<.0001	-0.04
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	4.59	0.62	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-5.05	0.19	<.0001	-0.05
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-6.53	0.58	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-5.14	1.87	0.0059	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-8.82	0.44	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-1.84	0.62	0.0030	-0.01
<b>Slope (<math>\beta_{10}</math>)</b>	0.85	0.00	<.0001	0.82
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.00	0.00	0.5793	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	0.00	0.01	0.6500	0.00
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.13	0.01	<.0001	-0.04
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.02	0.01	0.0367	0.00
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.02	0.00	<.0001	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.01	0.01	0.2822	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.03	0.04	0.4424	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.04	0.01	<.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.01	0.3261	0.00

Parameter	2017_G6M to 2018_G7M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3639.77	0.16	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	-0.58	0.15	<.0001	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-10.27	0.36	<.0001	-0.07
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-12.25	0.61	<.0001	-0.06
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-2.55	0.16	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	4.71	0.55	<.0001	0.02
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-4.10	0.17	<.0001	-0.05
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-4.00	0.53	<.0001	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-1.90	1.78	0.2874	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-7.40	0.40	<.0001	-0.04
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.78	0.59	0.1846	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.82	0.00	<.0001	0.84
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.01	0.00	0.1246	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.08	0.01	<.0001	-0.04
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.15	0.01	<.0001	-0.05
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.00	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	-0.03	0.01	0.0123	-0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	0.00	0.00	0.7954	0.00
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.01	0.01	0.3682	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	0.03	0.04	0.4381	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.02	0.01	0.0325	0.00
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.04	0.01	0.0038	0.01

Parameter	2017_G7M to 2018_G8M			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3657.23	0.17	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	-0.88	0.16	<.0001	-0.01
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-10.82	0.38	<.0001	-0.09
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-9.44	0.63	<.0001	-0.05
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-1.74	0.17	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	2.37	0.63	0.0002	0.01
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-1.88	0.19	<.0001	-0.02
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-3.05	0.51	<.0001	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-1.07	1.79	0.5504	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-5.37	0.41	<.0001	-0.03
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-1.13	0.68	0.0976	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.78	0.00	<.0001	0.86
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.02	0.00	<.0001	-0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.16	0.01	<.0001	-0.07
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.19	0.01	<.0001	-0.05
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.04	0.00	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.07	0.01	<.0001	0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.02	0.00	<.0001	-0.02
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.03	0.01	0.0039	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.04	0.04	0.3633	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.07	0.01	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.02	0.7084	0.00

Parameter	2017_G8M to 2018_Algl			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3669.78	0.17	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	1.55	0.17	<.0001	0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-7.47	0.42	<.0001	-0.07
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-5.58	0.52	<.0001	-0.04
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-3.62	0.19	<.0001	-0.06
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	2.89	0.64	<.0001	0.01
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-1.87	0.19	<.0001	-0.03
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-0.74	0.50	0.1421	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-1.42	1.81	0.4325	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-2.53	0.45	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-0.46	0.91	0.6136	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.71	0.00	<.0001	0.79
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	0.00	0.00	0.5769	0.00
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.08	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.11	0.01	<.0001	-0.03
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.01	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.04	0.01	0.0053	0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.02	0.01	0.0005	-0.01
<b>African American vs. White (<math>\beta_{17}</math>)</b>	0.00	0.01	0.9344	0.00
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.04	0.06	0.4335	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.03	0.01	0.0239	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	0.01	0.02	0.6920	0.00

Parameter	2017_AlgI to 2018_Geo			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3692.38	0.18	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	-2.28	0.19	<.0001	-0.03
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-6.73	0.54	<.0001	-0.05
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	-1.63	0.63	0.0100	-0.01
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-1.60	0.22	<.0001	-0.02
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	3.30	0.65	<.0001	0.01
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-4.38	0.22	<.0001	-0.06
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-5.00	0.59	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	-4.15	1.95	0.0330	-0.01
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-3.74	0.54	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-3.17	0.82	0.0001	-0.01
<b>Slope (<math>\beta_{10}</math>)</b>	0.91	0.00	<.0001	0.84
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.03	0.01	<.0001	-0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.12	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	-0.08	0.02	<.0001	-0.02
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.03	0.01	<.0001	-0.01
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.05	0.01	0.0002	0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.06	0.01	<.0001	-0.04
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.10	0.02	<.0001	-0.02
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.07	0.05	0.1957	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.07	0.02	<.0001	-0.01
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	-0.06	0.02	0.0091	-0.01



Parameter	2017_Geo to 2018_Algl			
	Unstandardized Estimate	SE	p value	Standardized Estimate
<b>Intercept (<math>\beta_{00}</math>)</b>	3700.27	0.18	<.0001	0.00
<b>Female vs. Male (<math>\beta_{01}</math>)</b>	1.10	0.18	<.0001	0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{02}</math>)</b>	-6.95	0.57	<.0001	-0.05
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{03}</math>)</b>	2.18	0.60	0.0003	0.01
<b>Low income vs. Non-Low Income (<math>\beta_{04}</math>)</b>	-3.33	0.22	<.0001	-0.04
<b>Asian vs. White (<math>\beta_{05}</math>)</b>	5.43	0.60	<.0001	0.03
<b>Hispanic vs. White (<math>\beta_{06}</math>)</b>	-2.76	0.21	<.0001	-0.04
<b>African American vs. White (<math>\beta_{07}</math>)</b>	-1.08	0.60	0.0707	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{08}</math>)</b>	0.08	1.76	0.9632	0.00
<b>American Indian vs. White (<math>\beta_{09}</math>)</b>	-4.87	0.52	<.0001	-0.03
<b>Multiple vs. White (<math>\beta_{010}</math>)</b>	-1.42	0.76	0.0637	0.00
<b>Slope (<math>\beta_{10}</math>)</b>	0.79	0.00	<.0001	0.84
<b>Female vs. Male (<math>\beta_{11}</math>)</b>	-0.02	0.01	<.0001	-0.02
<b>Special Education Status vs. Non-SPED (<math>\beta_{12}</math>)</b>	-0.10	0.01	<.0001	-0.03
<b>Limited English Proficiency vs. Non-LEP (<math>\beta_{13}</math>)</b>	0.04	0.02	0.0093	0.01
<b>Low income vs. Non-Low Income (<math>\beta_{14}</math>)</b>	-0.05	0.01	<.0001	-0.03
<b>Asian vs. White (<math>\beta_{15}</math>)</b>	0.05	0.01	<.0001	0.01
<b>Hispanic vs. White (<math>\beta_{16}</math>)</b>	-0.06	0.01	<.0001	-0.04
<b>African American vs. White (<math>\beta_{17}</math>)</b>	-0.06	0.02	0.0002	-0.01
<b>Hawaiian/Pacific Islander vs. White (<math>\beta_{18}</math>)</b>	-0.03	0.05	0.4825	0.00
<b>American Indian vs. White (<math>\beta_{19}</math>)</b>	-0.11	0.01	<.0001	-0.02
<b>Multiple vs. White (<math>\beta_{110}</math>)</b>	-0.03	0.02	0.1222	0.00

## Appendix E. Equations and Formula for Estimating Reliability

### E.1 Standard Error Formula

For the AzMERIT assessments scored using MLE, according to Masters (1982), the asymptotic estimate of the standard error for ability  $\theta$  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  $\theta$  to the reporting scale. For both ELA and Mathematics tests,  $a = 30$ .

### E.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  $\theta$  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  $\theta_c$  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

$$P(\theta < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta} - \theta_c}{se(\hat{\theta})}\right).$$

#### E.2.1 Classification Accuracy Formula

The probability of a student with true ability  $\theta$  being classified *at or above* the cut score  $\theta_c$ , 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:

$$\frac{N_{11} + N_{00}}{N_1 + N_0}.$$

### E.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  $\theta_c$  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—Spring 18 Operational Item Parameter Estimates — Grade 3 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13021_E	HTML , HTML , textEntryExtendedResponse	-1.13009	2.10828	4.15506	1.711083
2	13021_O	HTML , HTML , textEntryExtendedResponse	-1.10203	1.71104	4.1758	1.594937
3	13021_C	HTML , HTML , textEntryExtendedResponse	-1.31327	-0.98811		-1.15069
4	13024_E	HTML , HTML , textEntryExtendedResponse	-1.10833	2.26727	4.05905	1.73933
5	13024_O	HTML , HTML , textEntryExtendedResponse	-1.24303	2.04797	4.13136	1.645433
6	13024_C	HTML , HTML , textEntryExtendedResponse	-1.6353	-0.94878		-1.29204
7	11951	HTML , multipleChoice	-0.74234			-0.74234
8	11953	HTML , multipleChoice	-0.42539			-0.42539
9	11948	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	-0.11163			-0.11163
10	11950	HTML , multipleChoice	-0.26458			-0.26458
11	11945	HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	-0.62994	1.02765		0.198855
12	9691	HTML , multipleChoice	-0.37757			-0.37757
13	9698	HTML , multipleChoice	-1.11745			-1.11745
14	9700	HTML , multipleChoice	1.23623			1.23623
15	9690	HTML , multipleChoice	-1.209			-1.209
16	9692	HTML , multipleChoice	-0.60073			-0.60073
17	9699	HTML , multipleChoice	-0.30132			-0.30132
18	8708	editTaskWithChoice	-1.85419			-1.85419
19	8709	editTaskWithChoice , editTaskWithChoice	-1.02511	-0.31661		-0.67086
20	8710	editTaskWithChoice	-1.27986			-1.27986
21	9414	HTML , multipleChoice	-0.03304			-0.03304
22	9422	HTML , multipleChoice	-0.87555			-0.87555
23	9419	hotTextCustom	-0.05709			-0.05709
24	10632	HTML , multipleChoice	0.48179			0.48179
25	9410	hotTextCustom	1.68151			1.68151
26	10628	HTML , multipleChoice	-1.10023			-1.10023
27	10630	HTML , multipleChoice	-0.90745			-0.90745
28	10634	HTML , multipleChoice	1.17725			1.17725
29	9373	hotTextCustom	1.26088			1.26088
30	9330	HTML , multipleChoice	-0.2646			-0.2646
31	9338	hotTextCustom	3.07537			3.07537
32	10268	HTML , multipleChoice	0.2005			0.2005

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
33	9337	hotTextCustom	-0.07027			-0.07027
34	9331	HTML , multipleChoice	0.62573			0.62573
35	9335	HTML , multipleChoice	-0.64679			-0.64679
36	9333	HTML , multipleChoice	0.30718			0.30718
37	9336	HTML , multipleChoice	-0.0956			-0.0956
38	14675	HTML , multipleChoice	-1.79687			-1.79687
39	14681	HTML , multipleChoice	0.30828			0.30828
40	14685	HTML , multipleChoice	0.22023			0.22023
41	14688	hotTextCustom	0.02792			0.02792
42	14679	HTML , multipleChoice	0.03203			0.03203
43	14684	HTML , multipleChoice	0.49866			0.49866
44	14682	HTML , multipleSelect	2.07256			2.07256
45	12979	editTaskWithChoice	0.50861			0.50861
46	12981	editTaskWithChoice	-2.23435			-2.23435
47	12982	editTaskWithChoice	-0.69194			-0.69194

**Appendix F.2—Spring 18 Operational Item Parameter Estimates — Grade 4 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13121_E	HTML , HTML , textEntryExtendedResponse	1.87054	4.48532	7.08457	4.480143
2	13121_O	HTML , HTML , textEntryExtendedResponse	0.80218	4.25587	8.11115	4.389733
3	13121_C	HTML , HTML , textEntryExtendedResponse	-1.70071	1.64402		-0.02835
4	13118_E	HTML , HTML , textEntryExtendedResponse	1.69157	4.56961	4.98225	3.74781
5	13118_O	HTML , HTML , textEntryExtendedResponse	0.73557	4.29715	4.84718	3.2933
6	13118_C	HTML , HTML , textEntryExtendedResponse	-1.60239	1.2628		-0.1698
7	9598	HTML , multipleChoice	-0.0095			-0.0095
8	9604	hotTextCustom	1.79072			1.79072
9	9595	HTML , multipleChoice	-0.42531			-0.42531
10	9597	HTML , multipleChoice	-0.09234			-0.09234
11	9596	HTML , multipleChoice	0.65108			0.65108
12	9425	HTML , multipleChoice	-0.64738			-0.64738
13	10263	HTML , multipleChoice	-0.72314			-0.72314
14	9382	HTML , multipleChoice	-1.1843			-1.1843
15	9387	HTML , multipleChoice	1.48043			1.48043
16	9386	HTML , multipleSelect	0.46331			0.46331
17	11695	HTML , multipleChoice	-0.1009			-0.1009
18	11684	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.05055	-0.00818		0.521185
19	11689	HTML , multipleChoice	0.52739			0.52739
20	11703	HTML , multipleChoice	0.48761			0.48761
21	11701	HTML , multipleChoice	0.46287			0.46287
22	11704	HTML , multipleChoice	0.82454			0.82454
23	11697	HTML , multipleChoice	-0.42274			-0.42274
24	10644	editTaskWithChoice	-0.05707			-0.05707
25	10647	editTaskWithChoice	-0.85463			-0.85463
26	13044	HTML , multipleChoice	-0.284			-0.284
27	13073	hotTextCustom	-0.20656			-0.20656
28	13042	HTML , multipleChoice	-1.39269			-1.39269
29	19259	hotTextCustom	0.22957			0.22957
30	13043	HTML , multipleChoice	-0.80847			-0.80847
31	13070	HTML , multipleChoice	0.52289			0.52289
32	13071	HTML , multipleChoice	-0.31909			-0.31909

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
33	13046	HTML , multipleChoice	0.01302			0.01302
34	11837	HTML , multipleChoice	-0.34149			-0.34149
35	12567	HTML , multipleChoice	-0.08581			-0.08581
36	11840	HTML , multipleChoice	0.09196			0.09196
37	11844	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.01626	0.92279		0.969525
38	11842	hotTextCustom	2.16926			2.16926
39	11841	HTML , multipleChoice	0.87491			0.87491
40	11846	HTML , multipleChoice	0.86215			0.86215
41	11838	HTML , multipleChoice	1.0915			1.0915
42	11967	hotTextCustom	0.32005			0.32005
43	11847	HTML , multipleChoice	0.29982			0.29982
44	16080	editTaskWithChoice	-0.62177			-0.62177
45	16081	editTaskWithChoice , editTaskWithChoice	-1.07187	0.37369		-0.34909
46	16084	editTaskWithChoice	-0.33913			-0.33913
47	16085	editTaskWithChoice	-0.28574			-0.28574



**Appendix F.3—Spring 18 Operational Item Parameter Estimates — Grade 5 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13238_E	HTML , HTML , textEntryExtendedResponse	-0.90788	3.36594	4.06761	2.175223
2	13238_O	HTML , HTML , textEntryExtendedResponse	-1.30216	2.84958	4.12669	1.89137
3	13238_C	HTML , HTML , textEntryExtendedResponse	-2.13488	0.01614		-1.05937
4	13237_E	HTML , HTML , textEntryExtendedResponse	0.76049	4.48983	5.72602	3.65878
5	13237_O	HTML , HTML , textEntryExtendedResponse	-0.3275	3.62724	5.71883	3.00619
6	13237_C	HTML , HTML , textEntryExtendedResponse	-1.61986	0.18987		-0.715
7	9305	HTML , multipleChoice	-1.06814			-1.06814
8	9310	hotTextCustom	0.74289			0.74289
9	9301	HTML , multipleChoice	0.26042			0.26042
10	9304	HTML , multipleSelect	0.5267			0.5267
11	9302	HTML , multipleChoice	0.397			0.397
12	12069	HTML , multipleChoice	-0.68003			-0.68003
13	12068	HTML , multipleChoice	-0.40849			-0.40849
14	12072	HTML , multipleChoice	-0.63304			-0.63304
15	12067	HTML , multipleSelect	-0.09943			-0.09943
16	12065	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.7085			0.7085
17	13216	HTML , multipleChoice	0.62654			0.62654
18	13228	HTML , multipleChoice	-0.37981			-0.37981
19	13217	HTML , multipleSelect	0.35086			0.35086
20	13186	HTML , multipleChoice	-0.42893			-0.42893
21	13187	HTML , multipleChoice	0.4539			0.4539
22	13189	hotTextCustom	-0.59918			-0.59918
23	10659	editTaskWithChoice	-0.82784			-0.82784
24	10661	editTaskWithChoice , editTaskWithChoice	-2.20002	-0.66798		-1.434
25	16209	hotTextCustom , hotTextSelectable	-0.94997			-0.94997
26	16211	HTML , multipleChoice	-0.79003			-0.79003
27	16217	HTML , multipleChoice	0.17403			0.17403
28	16214	HTML , multipleChoice	-0.41677			-0.41677
29	16213	HTML , multipleChoice	1.10977			1.10977
30	16216	HTML , multipleChoice	1.65472			1.65472
31	16215	HTML , multipleChoice	-0.29232			-0.29232

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
32	16039	HTML , multipleChoice	0.0773			0.0773
33	16044	HTML , multipleChoice	1.35992			1.35992
34	16043	hotTextCustom	-0.2509			-0.2509
35	16045	HTML , multipleChoice	-1.37714			-1.37714
36	16041	hotTextCustom	0.35149			0.35149
37	16035	HTML , multipleChoice	0.52592			0.52592
38	16042	HTML , multipleChoice	0.79995			0.79995
39	12687	HTML , multipleChoice	-0.28139			-0.28139
40	12894	HTML , multipleChoice	0.42103			0.42103
41	12690	HTML , multipleChoice	-0.12644			-0.12644
42	12671	HTML , multipleChoice	0.00268			0.00268
43	12865	HTML , multipleChoice	-0.57988			-0.57988
44	12649	HTML , multipleChoice	0.45115			0.45115
45	9286	editTaskWithChoice	-0.23947			-0.23947
46	9287	editTaskWithChoice	1.18837			1.18837
47	9288	editTaskWithChoice , editTaskWithChoice	-1.29281	0.66758		-0.31262

**Appendix F.4—Spring 18 Operational Item Parameter Estimates — Grade 6 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13309_E	HTML , HTML , textEntryExtendedResponse	0.33264	2.78999	4.58425	2.56896
2	13309_O	HTML , HTML , textEntryExtendedResponse	-0.4541	2.19863	4.51841	2.087647
3	13309_C	HTML , HTML , textEntryExtendedResponse	-2.8389	0.01884		-1.41003
4	13305_E	HTML , HTML , textEntryExtendedResponse	0.79496	3.84657	4.56174	3.067757
5	13305_O	HTML , HTML , textEntryExtendedResponse	0.0947	3.01629	4.70018	2.603723
6	13305_C	HTML , HTML , textEntryExtendedResponse	-1.96936	-0.65918		-1.31427
7	16027	HTML , multipleChoice	1.02018			1.02018
8	16032	HTML , multipleChoice	-0.0518			-0.0518
9	16028	HTML , multipleChoice	0.24012			0.24012
10	16033	HTML , multipleChoice	0.33379			0.33379
11	16029	HTML , multipleChoice	-0.00757			-0.00757
12	16030	HTML , multipleChoice	0.70326			0.70326
13	16031	HTML , multipleChoice	-0.01495			-0.01495
14	16138	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.62183			0.62183
15	16114	HTML , multipleChoice	0.08609			0.08609
16	16108	HTML , multipleChoice	-0.47937			-0.47937
17	16110	HTML , multipleChoice	-0.69747			-0.69747
18	16060	HTML , multipleChoice	-1.19216			-1.19216
19	16109	HTML , multipleChoice	0.24005			0.24005
20	16106	HTML , multipleChoice	0.75801			0.75801
21	16112	HTML , multipleChoice	0.22782			0.22782
22	16113	hotTextCustom	1.36279			1.36279
23	16075	editTaskWithChoice	-2.40634			-2.40634
24	16077	editTaskWithChoice	-0.82828			-0.82828
25	9137	HTML , multipleChoice	-1.63221			-1.63221
26	9138	HTML , multipleChoice	-1.48057			-1.48057
27	9134	HTML , multipleChoice	-0.24354			-0.24354
28	9135	HTML , multipleSelect	0.52718			0.52718
29	9154	HTML , multipleSelect	2.33626			2.33626
30	9266	HTML , multipleChoice	-1.04177			-1.04177
31	9273	HTML , multipleChoice	-0.43509			-0.43509
32	9267	hotTextCustom	-0.59143			-0.59143

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
33	9263	HTML , multipleChoice	-0.51352			-0.51352
34	9268	hotTextCustom	2.08178			2.08178
35	14723	HTML , multipleChoice	0.32472			0.32472
36	14721	hotTextCustom	-0.30326			-0.30326
37	14752	HTML , multipleSelect	1.1886			1.1886
38	14751	HTML , multipleChoice	-0.25551			-0.25551
39	14748	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	1.35145			1.35145
40	12410	HTML , multipleChoice	0.85336			0.85336
41	12409	HTML , multipleChoice	0.36443			0.36443
42	12415	HTML , multipleChoice	0.24204			0.24204
43	12407	HTML , multipleChoice	-0.06404			-0.06404
44	12895	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.79934	0.88536		0.84235
45	9107	editTaskWithChoice	-1.86994			-1.86994
46	9108	editTaskWithChoice , editTaskWithChoice	-1.32061	1.24437		-0.03812
47	9109	editTaskWithChoice , editTaskWithChoice	-1.49476	0.62001		-0.43738

**Appendix F.5—Spring 18 Operational Item Parameter Estimates — Grade 7 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13405_E	HTML , HTML , textEntryExtendedResponse	0.02255	3.38944	4.52755	2.646513
2	13405_O	HTML , HTML , textEntryExtendedResponse	-0.88948	3.04194	4.92252	2.358327
3	13405_C	HTML , HTML , textEntryExtendedResponse	-2.87726	-0.07362		-1.47544
4	13400_E	HTML , HTML , textEntryExtendedResponse	-0.89746	3.17022	4.70894	2.327233
5	13400_O	HTML , HTML , textEntryExtendedResponse	-1.10875	2.81094	4.67186	2.124683
6	13400_C	HTML , HTML , textEntryExtendedResponse	-1.93478	0.24135		-0.84672
7	16197	HTML , multipleChoice	-0.12794			-0.12794
8	16198	HTML , multipleChoice	0.00146			0.00146
9	16118	HTML , multipleChoice	-0.99956			-0.99956
10	16154	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	2.39285			2.39285
11	16119	HTML , multipleChoice	0.16421			0.16421
12	16201	HTML , multipleChoice	-0.22346			-0.22346
13	16098	HTML , multipleChoice	-0.49359			-0.49359
14	16100	HTML , multipleChoice	-0.45653			-0.45653
15	16103	HTML , multipleChoice	-0.66761			-0.66761
16	16101	HTML , multipleChoice	0.58414			0.58414
17	16099	HTML , multipleChoice	-0.04159			-0.04159
18	16104	HTML , multipleChoice	-0.23917			-0.23917
19	16128	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	-0.5101			-0.5101
20	16120	editTaskWithChoice	-1.16575			-1.16575
21	16121	editTaskWithChoice	-0.77536			-0.77536
22	16122	editTaskWithChoice	-1.17838			-1.17838
23	11706	HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	0.42177	0.86239		0.64208
24	11718	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.21192	0.94393		1.077925
25	11714	HTML , multipleChoice	1.34846			1.34846
26	11750	HTML , multipleChoice	0.75945			0.75945
27	11739	hotTextCustom	0.80398			0.80398
28	11696	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.13052			1.13052
29	9711	HTML , multipleChoice	0.29644			0.29644

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
30	9713	HTML , multipleChoice	0.2301			0.2301
31	9614	HTML , multipleChoice	-1.02121			-1.02121
32	10695	HTML , multipleChoice	-0.51732			-0.51732
33	9709	HTML , multipleChoice	-0.24287			-0.24287
34	9750	HTML , multipleSelect	-0.17847			-0.17847
35	9740	HTML , multipleChoice	-1.52845			-1.52845
36	9743	HTML , multipleChoice	0.07478			0.07478
37	9741	HTML , multipleSelect	1.30368			1.30368
38	9742	HTML , multipleChoice	-0.99456			-0.99456
39	9747	HTML , multipleSelect	1.6429			1.6429
40	9845	HTML , multipleChoice	1.05787			1.05787
41	13357	HTML , multipleChoice	-0.38788			-0.38788
42	13356	hotTextCustom	0.07253			0.07253
43	13359	HTML , multipleChoice	-0.6065			-0.6065
44	13358	HTML , multipleChoice	0.23406			0.23406
45	13354	hotTextCustom	1.78903			1.78903
46	13335	editTaskWithChoice	-1.96148			-1.96148
47	13337	editTaskWithChoice , editTaskWithChoice	-1.38768	0.76648		-0.3106

**Appendix F.6—Spring 18 Operational Item Parameter Estimates — Grade 8 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13438_E	HTML , HTML , textEntryExtendedResponse	-1.30982	1.60181	3.36366	1.21855
2	13438_O	HTML , HTML , textEntryExtendedResponse	-1.25185	1.38879	3.19409	1.110343
3	13438_C	HTML , HTML , textEntryExtendedResponse	-2.48946	-0.96722		-1.72834
4	13453_E	HTML , HTML , textEntryExtendedResponse	-1.06493	1.5231	3.24717	1.235113
5	13453_O	HTML , HTML , textEntryExtendedResponse	-1.23977	0.91325	3.22926	0.96758
6	13453_C	HTML , HTML , textEntryExtendedResponse	-1.99191	-0.7686		-1.38026
7	16390	HTML , multipleChoice	0.19272			0.19272
8	16389	hotTextCustom	0.00309			0.00309
9	16386	HTML , multipleChoice	-0.83635			-0.83635
10	16385	HTML , multipleChoice	0.92208			0.92208
11	16384	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.00655			0.00655
12	16387	HTML , multipleChoice	0.61754			0.61754
13	16393	HTML , multipleChoice	0.16575			0.16575
14	16394	HTML , multipleSelect	1.37373			1.37373
15	16218	hotTextCustom	-0.49091			-0.49091
16	16324	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	0.12971			0.12971
17	16219	HTML , multipleChoice	0.78245			0.78245
18	16226	HTML , multipleSelect	0.35972			0.35972
19	16222	HTML , multipleChoice	0.53465			0.53465
20	16223	HTML , multipleChoice	-0.95205			-0.95205
21	9100	editTaskWithChoice , editTaskWithChoice	-2.0418	-0.81173		-1.42677
22	9101	editTaskWithChoice	-0.12448			-0.12448
23	10262	editTaskWithChoice	-1.75213			-1.75213
24	12427	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.93677			1.93677
25	11815	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.62015			0.62015
26	11820	hotTextCustom	-0.7631			-0.7631
27	11819	HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	0.40419	1.84388		1.124035
28	11816	HTML , multipleChoice	-0.07166			-0.07166

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
29	11811	HTML , multipleChoice	-0.25816			-0.25816
30	12429	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	-0.57203	2.62188		1.024925
31	11812	HTML , multipleChoice	-0.25561			-0.25561
32	11810	HTML , HTML , multipleChoice , HTML , HTML , multipleSelect	0.87267			0.87267
33	9026	HTML , multipleChoice	0.5053			0.5053
34	10626	hotTextCustom	1.26297			1.26297
35	10627	hotTextCustom	0.47781			0.47781
36	9024	HTML , multipleChoice	-0.53744			-0.53744
37	9020	HTML , multipleSelect	1.09383			1.09383
38	9017	HTML , multipleChoice	-0.09267			-0.09267
39	9014	HTML , multipleChoice	-0.9944			-0.9944
40	9019	HTML , multipleChoice	0.04892			0.04892
41	9230	HTML , multipleChoice	-1.15492			-1.15492
42	9018	HTML , multipleChoice	0.80965			0.80965
43	9046	HTML , multipleChoice	-0.82499			-0.82499
44	9015	HTML , multipleSelect	1.20929			1.20929
45	9727	editTaskWithChoice	-2.20531			-2.20531
46	9728	editTaskWithChoice , editTaskWithChoice	-1.65694	-0.20084		-0.92889
47	9729	editTaskWithChoice , editTaskWithChoice	-1.73003	0.26659		-0.73172



**Appendix F.7—Spring 18 Operational Item Parameter Estimates — Grade 9 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13565_E	HTML , HTML , textEntryExtendedResponse	-1.41497	1.96395	3.80007	1.449683
2	13565_O	HTML , HTML , textEntryExtendedResponse	-1.69345	1.40948	3.82485	1.180293
3	13565_C	HTML , HTML , textEntryExtendedResponse	-2.20831	-0.65189		-1.4301
4	13554_E	HTML , HTML , textEntryExtendedResponse	-1.58337	2.72154	4.17335	1.770507
5	13554_O	HTML , HTML , textEntryExtendedResponse	-1.49156	1.80507	4.60922	1.64091
6	13554_C	HTML , HTML , textEntryExtendedResponse	-2.17867	-1.14465		-1.66166
7	9734	editTaskWithChoice	-0.38567			-0.38567
8	9735	editTaskWithChoice , editTaskWithChoice	-1.01417	1.57881		0.28232
9	9736	editTaskWithChoice	-0.28792			-0.28792
10	16494	HTML , multipleChoice	-0.2389			-0.2389
11	16491	HTML , multipleChoice	0.55173			0.55173
12	16487	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.2467			0.2467
13	16496	HTML , multipleChoice	0.01893			0.01893
14	16488	HTML , multipleChoice	0.26425			0.26425
15	16485	HTML , multipleChoice	0.45266			0.45266
16	16464	HTML , multipleSelect	0.33373			0.33373
17	16492	HTML , multipleChoice	-0.44001			-0.44001
18	16493	HTML , multipleChoice	-0.03566			-0.03566
19	16239	HTML , multipleChoice	0.07774			0.07774
20	16240	HTML , multipleChoice	-0.65636			-0.65636
21	16238	HTML , multipleChoice	0.44799			0.44799
22	16236	HTML , multipleChoice	-0.08862			-0.08862
23	16237	HTML , multipleChoice	0.27251			0.27251
24	8987	HTML , multipleChoice	-0.57348			-0.57348
25	8989	HTML , multipleChoice	-1.00653			-1.00653
26	9043	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.37237			0.37237
27	9003	HTML , multipleChoice	0.30221			0.30221
28	8990	HTML , multipleChoice	0.50835			0.50835
29	9004	HTML , multipleChoice	0.0071			0.0071
30	13560	HTML , multipleChoice	0.13487			0.13487
31	13559	HTML , multipleChoice	-1.26254			-1.26254

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
32	13561	HTML , multipleChoice	0.79566			0.79566
33	13563	HTML , multipleChoice	0.34102			0.34102
34	13562	HTML , multipleChoice	0.241			0.241
35	13543	hotTextCustom	0.49797			0.49797
36	13545	HTML , multipleChoice	-0.43877			-0.43877
37	13547	HTML , multipleChoice	-0.39906			-0.39906
38	13535	HTML , multipleChoice	-0.8022			-0.8022
39	13539	HTML , multipleChoice	-0.01696			-0.01696
40	13549	hotTextCustom	1.31296			1.31296
41	13567	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.47861			0.47861
42	13541	HTML , multipleChoice	-0.37461			-0.37461
43	13518	HTML , multipleSelect	0.23285			0.23285
44	13516	HTML , multipleChoice	-0.71545			-0.71545
45	13551	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.69902			0.69902
46	13534	HTML , multipleChoice	-0.67176			-0.67176
47	13515	HTML , multipleChoice	-0.23416			-0.23416
48	13455	editTaskWithChoice	0.34843			0.34843
49	13456	editTaskWithChoice , editTaskWithChoice	0.10449	1.39644		0.750465

**Appendix F.8—Spring 18 Operational Item Parameter Estimates — Grade 10 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13635_E	HTML , HTML , textEntryExtendedResponse	-1.64858	1.09969	3.31762	0.92291
2	13635_O	HTML , HTML , textEntryExtendedResponse	-2.00564	0.76273	3.12769	0.62826
3	13635_C	HTML , HTML , textEntryExtendedResponse	-3.09387	-1.09958		-2.09673
4	13636_E	HTML , HTML , textEntryExtendedResponse	-1.14411	1.51623	4.30412	1.558747
5	13636_O	HTML , HTML , textEntryExtendedResponse	-1.25137	0.71571	3.49201	0.98545
6	13636_C	HTML , HTML , textEntryExtendedResponse	-2.77526	-1.56572		-2.17049
7	9822	HTML , multipleChoice	-0.54876			-0.54876
8	9819	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	-0.48425			-0.48425
9	9826	HTML , multipleChoice	-0.43866			-0.43866
10	9814	HTML , multipleChoice	-0.3295			-0.3295
11	9825	HTML , multipleChoice	-0.52709			-0.52709
12	9813	hotTextCustom	0.53987			0.53987
13	15110	HTML , multipleChoice	0.66313			0.66313
14	15105	HTML , multipleChoice	0.11533			0.11533
15	15103	HTML , multipleChoice	0.28246			0.28246
16	15104	HTML , multipleChoice	-0.30243			-0.30243
17	15111	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.06335			1.06335
18	15138	HTML , multipleChoice	0.00325			0.00325
19	15112	HTML , multipleChoice	1.06531			1.06531
20	15113	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.65012			1.65012
21	15142	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.62565			1.62565
22	15144	HTML , HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	2.29253			2.29253
23	8757	editTaskWithChoice	-1.36906			-1.36906
24	8758	editTaskWithChoice	0.32886			0.32886
25	8763	editTaskWithChoice	-1.27617			-1.27617
26	16232	HTML , multipleChoice	-1.01634			-1.01634

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
27	16227	HTML , multipleChoice	-0.61308			-0.61308
28	16230	HTML , multipleChoice	-0.30017			-0.30017
29	16234	HTML , multipleChoice	0.59022			0.59022
30	16228	HTML , multipleChoice	-1.27458			-1.27458
31	16233	HTML , multipleChoice	-0.39653			-0.39653
32	16231	HTML , multipleChoice	-0.48274			-0.48274
33	8813	hotTextCustom	-0.22198			-0.22198
34	8812	HTML , multipleChoice	-0.48167			-0.48167
35	10155	hotTextCustom	0.27894			0.27894
36	8810	HTML , multipleChoice	-0.8087			-0.8087
37	8852	HTML , multipleChoice	0.40325			0.40325
38	8811	HTML , multipleChoice	0.14036			0.14036
39	12270	hotTextCustom	-0.06917	0.97503		0.45293
40	12688	HTML , multipleChoice	-1.11024			-1.11024
41	12260	HTML , multipleChoice	0.09296			0.09296
42	12249	HTML , multipleChoice	-0.23706			-0.23706
43	12204	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.41177			1.41177
44	12257	hotTextCustom	0.20616			0.20616
45	12709	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	0.55044			0.55044
46	12268	HTML , multipleSelect	0.22758			0.22758
47	8765	editTaskWithChoice	0.48513			0.48513
48	8767	editTaskWithChoice	-0.78139			-0.78139
49	8768	editTaskWithChoice	-0.44958			-0.44958

**Appendix F.9—Spring 18 Operational Item Parameter Estimates — Grade 11 ELA**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13723_E	HTML , HTML , textEntryExtendedResponse	-1.9275	0.81378	2.56933	0.485203
2	13723_O	HTML , HTML , textEntryExtendedResponse	-3.05795	0.1766	2.29503	-0.19544
3	13723_C	HTML , HTML , textEntryExtendedResponse	-2.93586	-1.14637		-2.04112
4	13725_E	HTML , HTML , textEntryExtendedResponse	-1.76303	0.57642	2.89886	0.57075
5	13725_O	HTML , HTML , textEntryExtendedResponse	-2.23547	0.75545	2.49751	0.339163
6	13725_C	HTML , HTML , textEntryExtendedResponse	-2.62307	-0.55033		-1.5867
7	13678	HTML , multipleChoice	-0.62025			-0.62025
8	13677	HTML , multipleChoice	-0.8214			-0.8214
9	13682	HTML , multipleSelect	-0.51676			-0.51676
10	13669	HTML , multipleChoice	-1.71667			-1.71667
11	13670	HTML , multipleChoice	0.23515			0.23515
12	13679	HTML , multipleChoice	-1.59956			-1.59956
13	9981	HTML , multipleChoice	-0.56402			-0.56402
14	9968	hotTextCustom	1.02523			1.02523
15	9963	hotTextCustom	0.66039			0.66039
16	9977	hotTextCustom	0.38207			0.38207
17	9971	HTML , multipleChoice	-0.46973			-0.46973
18	9964	HTML , multipleChoice	-0.55565			-0.55565
19	9969	HTML , multipleChoice	0.14667			0.14667
20	9966	hotTextCustom	2.03844			2.03844
21	8778	editTaskWithChoice	0.49041			0.49041
22	8779	editTaskWithChoice , editTaskWithChoice	-1.89965	-0.47048		-1.18507
23	8780	editTaskWithChoice	-0.3298			-0.3298
24	12818	HTML , multipleChoice	-0.84126			-0.84126
25	12833	HTML , multipleChoice	-0.38774			-0.38774
26	12877	HTML , multipleChoice	-0.25171			-0.25171
27	12834	HTML , multipleChoice	-0.55063			-0.55063
28	12910	HTML , multipleChoice	-0.462			-0.462
29	8864	HTML , multipleSelect	1.95843			1.95843
30	8859	HTML , multipleChoice	-0.99284			-0.99284
31	8861	HTML , multipleChoice	0.25879			0.25879
32	8860	hotTextCustom	1.67605			1.67605
33	8867	HTML , multipleChoice	0.18188			0.18188
34	8869	HTML , multipleChoice	-0.15979			-0.15979

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
35	8870	HTML , multipleChoice	0.05671			0.05671
36	8871	HTML , multipleChoice	0.03863			0.03863
37	13662	hotTextCustom	0.64934			0.64934
38	13658	HTML , multipleChoice	-1.17131			-1.17131
39	13659	HTML , multipleSelect	0.91927			0.91927
40	13667	HTML , HTML , multipleChoice , HTML , HTML , multipleChoice	1.0877			1.0877
41	16533	HTML , multipleChoice	-0.58372			-0.58372
42	16535	HTML , multipleSelect	1.20024			1.20024
43	16536	HTML , multipleChoice	-0.17979			-0.17979
44	16534	HTML , multipleChoice	-0.10232			-0.10232
45	16529	HTML , multipleChoice	-1.0748			-1.0748
46	16530	HTML , multipleChoice	0.59604			0.59604
47	13644	editTaskWithChoice	-1.40824			-1.40824
48	13646	editTaskWithChoice	-1.88481			-1.88481
49	13647	editTaskWithChoice	-1.47527			-1.47527

**Appendix F.10—Spring 18 Operational Item Parameter Estimates — Grade 3 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10434	HTML , equation	-0.75061			-0.75061
2	10409	HTML , multipleChoice	-1.87291			-1.87291
3	10681	HTML , multipleChoice	-1.12835			-1.12835
4	13765	HTML , equation	-0.9491			-0.9491
5	10460	HTML , multipleChoice	-0.44542			-0.44542
6	12054	HTML , equation	0.46601			0.46601
7	12569	HTML , multipleChoice	0.72111			0.72111
8	13767	HTML , equation	0.78005			0.78005
9	13965	HTML , equation	-1.00408			-1.00408
10	13978	HTML , equation	0.54289			0.54289
11	13746	HTML , equation	-0.1961			-0.1961
12	12902	HTML , equation	-0.70461			-0.70461
13	11644	HTML , equation	-1.01497			-1.01497
14	12085	HTML , equation	-0.62452			-0.62452
15	10462	HTML , multipleChoice	0.8652			0.8652
16	13882	HTML , multipleSelect	1.87056			1.87056
17	15420	HTML , equation	0.80698			0.80698
18	10683	HTML , multipleChoice	1.42686			1.42686
19	10424	HTML , equation	1.17507			1.17507
20	15370	HTML , equation	0.78401			0.78401
21	13982	HTML , equation	-0.16461			-0.16461
22	10442	HTML , equation	1.47025			1.47025
23	12921	HTML , multipleChoice	-1.04976			-1.04976
24	13749	HTML , multipleChoice	-1.80241			-1.80241
25	12281	HTML , equation	-1.24975			-1.24975
26	12943	HTML , equation	-0.81857			-0.81857
27	10435	HTML , multipleChoice	0.18402			0.18402
28	10396	HTML , equation	0.38434			0.38434
29	15914	HTML , multipleChoice	0.75837			0.75837
30	12421	HTML , equation	1.09421			1.09421
31	13970	HTML , multipleChoice	1.04176			1.04176
32	12053	HTML , equation	0.84834			0.84834
33	15397	HTML , equation	-0.99683			-0.99683
34	12087	HTML , equation	-0.18891			-0.18891
35	15384	HTML , equation	-1.19186			-1.19186
36	10411	HTML , multipleChoice	-1.61799			-1.61799
37	12941	HTML , equation	-0.5795			-0.5795
38	10448	HTML , equation	1.09544			1.09544
39	10415	HTML , equation	1.81916			1.81916
40	15383	HTML , equation	2.32606			2.32606

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10687	HTML , multipleChoice	0.89578			0.89578
42	10393	HTML , equation	1.1567			1.1567
43	15413	HTML , multipleSelect	0.93393			0.93393
44	15376	HTML , equation	0.94886			0.94886
45	11761	HTML , multipleChoice	-1.51748			-1.51748



**Appendix F.11—Spring 18 Operational Item Parameter Estimates — Grade 4 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13786	HTML , multipleChoice	-1.5972			-1.5972
2	13772	HTML , equation	-1.24222			-1.24222
3	10718	HTML , equation	0.03231			0.03231
4	14041	HTML , tableMatch	-1.06093			-1.06093
5	15450	HTML , equation	0.80359			0.80359
6	10737	HTML , equation	1.403			1.403
7	13782	HTML , tableMatch	-1.05896			-1.05896
8	15431	HTML , equation	0.00268			0.00268
9	10735	HTML , equation	-0.72229			-0.72229
10	10777	HTML , multipleChoice	0.71091			0.71091
11	10716	HTML , equation	1.10308			1.10308
12	13796	HTML , equation	1.79789			1.79789
13	13766	HTML , equation	-1.1843			-1.1843
14	12266	HTML , equation	-1.49821			-1.49821
15	15579	HTML , equation	-0.00102			-0.00102
16	12035	HTML , equation	1.3573			1.3573
17	13757	HTML , multipleSelect	0.8628			0.8628
18	12279	HTML , equation	1.35297			1.35297
19	15438	HTML , equation	-0.75252			-0.75252
20	14030	HTML , equation	-1.88215			-1.88215
21	10710	HTML , multipleSelect	-0.44114			-0.44114
22	15426	HTML , multipleSelect	0.4846			0.4846
23	10708	HTML , multipleChoice	-0.36231			-0.36231
24	12276	HTML , equation	-0.57774			-0.57774
25	10774	HTML , equation	1.86572			1.86572
26	10779	HTML , equation	0.34347			0.34347
27	13992	HTML , equation	-1.42162			-1.42162
28	10783	HTML , multipleChoice	-0.58052			-0.58052
29	15428	HTML , multipleChoice	-1.58475			-1.58475
30	14116	HTML , equation	0.62635			0.62635
31	13320	HTML , equation	0.01192			0.01192
32	10750	HTML , equation	-0.58904			-0.58904
33	10746	HTML , equation	0.99022			0.99022
34	10826	HTML , grid	-1.49232	0.80594		-0.34319
35	11608	HTML , equation	0.79089			0.79089
36	15462	HTML , multipleChoice	0.58475			0.58475
37	14020	HTML , equation	-0.16617			-0.16617
38	15441	HTML , multipleChoice	0.70772			0.70772
39	13883	HTML , tableMatch	-1.23055			-1.23055
40	14010	HTML , equation	1.49984			1.49984

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	14068	HTML , equation	1.84998			1.84998
42	14096	HTML , equation	0.10244			0.10244
43	13780	HTML , multipleChoice	-0.17136			-0.17136
44	10760	HTML , equation	-0.20261			-0.20261
45	15434	HTML , equation	-0.05381			-0.05381

**Appendix F.12—Spring 18 Operational Item Parameter Estimates — Grade 5 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	9716	HTML , tableMatch	-2.30144			-2.30144
2	12185	HTML , equation	-1.29281			-1.29281
3	15503	HTML , equation	-0.21214			-0.21214
4	10858	HTML , equation	0.63797			0.63797
5	11379	HTML , multipleChoice	-0.19187			-0.19187
6	10863	HTML , multipleChoice	0.62779			0.62779
7	13062	HTML , equation	0.66283			0.66283
8	10814	HTML , equation	1.37493			1.37493
9	15488	HTML , multipleSelect	0.99219			0.99219
10	10796	HTML , equation	0.28237			0.28237
11	15506	HTML , multipleChoice	-0.19127			-0.19127
12	15472	equation , HTML	-1.93639			-1.93639
13	11107	HTML , multipleChoice	-0.46606			-0.46606
14	10820	HTML , multipleChoice	1.62329			1.62329
15	15919	HTML , equation	0.08916			0.08916
16	12176	HTML , equation	-0.05286			-0.05286
17	15917	HTML , equation	0.46781			0.46781
18	14084	HTML , equation	-0.88864			-0.88864
19	11526	HTML , equation	-0.17516			-0.17516
20	14157	HTML , equation	1.67504			1.67504
21	12374	HTML , equation	0.48942			0.48942
22	15484	HTML , equation	-0.4499			-0.4499
23	11710	HTML , equation	1.23326			1.23326
24	10875	HTML , multipleChoice	-0.69325			-0.69325
25	12373	HTML , equation	-1.06775			-1.06775
26	9486	HTML , equation	0.14711			0.14711
27	15918	HTML , equation	-0.94523			-0.94523
28	11893	HTML , equation	1.54747			1.54747
29	14138	HTML , multipleSelect	-1.18708			-1.18708
30	10829	HTML , equation	-1.59476			-1.59476
31	14155	HTML , equation	-0.01572			-0.01572
32	15463	HTML , equation	0.80098			0.80098
33	15475	HTML , multipleSelect	1.65408			1.65408
34	15485	HTML , equation	0.13277			0.13277
35	13047	HTML , equation	-1.38963			-1.38963
36	14121	HTML , equation	-0.03083			-0.03083
37	12951	HTML , equation	0.34243			0.34243
38	10841	HTML , multipleChoice	0.50079			0.50079
39	11764	HTML , equation	1.33182			1.33182
40	15480	HTML , equation	-0.94707			-0.94707

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	14078	HTML , equation	0.8118			0.8118
42	10795	HTML , equation	-1.39678			-1.39678
43	14171	HTML , equation	-0.60178			-0.60178
44	13329	HTML , equation	0.31116			0.31116
45	11636	HTML , multipleChoice	0.75774			0.75774

**Appendix F.13—Spring 18 Operational Item Parameter Estimates — Grade 6 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10054	HTML , multipleChoice	-3.09329			-3.09329
2	14187	HTML , multipleChoice	-1.27839			-1.27839
3	11728	HTML , equation	-1.56689			-1.56689
4	10056	HTML , equation	1.64577			1.64577
5	10148	HTML , multipleChoice	-0.4154			-0.4154
6	15603	HTML , tableMatch	-0.23201			-0.23201
7	15611	HTML , equation	0.29943			0.29943
8	15605	HTML , equation	0.2235			0.2235
9	12468	HTML , equation	0.48993			0.48993
10	14226	HTML , equation	0.75318			0.75318
11	11531	HTML , equation	0.92185			0.92185
12	15624	HTML , equation	1.32882			1.32882
13	10143	HTML , equation	0.87374			0.87374
14	11375	HTML , equation	0.73262			0.73262
15	13117	HTML , multipleChoice	-0.38515			-0.38515
16	10048	HTML , multipleChoice	0.40685			0.40685
17	13112	HTML , equation	1.07232			1.07232
18	10104	HTML , equation	1.39329			1.39329
19	11361	HTML , equation	0.37871			0.37871
20	12345	HTML , equation	1.08427			1.08427
21	10062	HTML , equation	0.87598			0.87598
22	14224	HTML , multipleChoice	-0.12021			-0.12021
23	11777	HTML , equation	-1.03449			-1.03449
24	13792	HTML , equation	-2.66401			-2.66401
25	13331	HTML , equation	-2.78919			-2.78919
26	10113	HTML , multipleChoice	-1.43848			-1.43848
27	10108	HTML , multipleChoice	-0.6664			-0.6664
28	11643	HTML , equation	-0.4474			-0.4474
29	10064	HTML , equation	-0.12482			-0.12482
30	14222	HTML , multipleChoice	-2.23636			-2.23636
31	11903	HTML , equation	0.97964			0.97964
32	10047	HTML , equation	1.59727			1.59727
33	10139	HTML , multipleSelect	2.24036			2.24036
34	12055	HTML , equation	1.00917			1.00917
35	10119	HTML , tableInput	-0.44667			-0.44667
36	15618	HTML , equation	-0.92681			-0.92681
37	15610	equation , HTML	-0.40317			-0.40317
38	14426	HTML , multipleSelect	0.85707			0.85707
39	10078	HTML , equation	1.05127			1.05127
40	12949	HTML , equation	1.44591			1.44591

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	9498	HTML , equation	2.73279			2.73279
42	10126	HTML , equation	1.10667			1.10667
43	14425	HTML , equation	1.02222			1.02222
44	14210	HTML , equation	0.09258			0.09258
45	9492	HTML , equation	-1.41594			-1.41594
46	13330	HTML , equation	-1.59289			-1.59289
47	10107	HTML , equation	-1.82823			-1.82823

**Appendix F.14—Spring 18 Operational Item Parameter Estimates — Grade 7 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	8698	HTML , multipleChoice	-2.05133			-2.05133
2	14317	HTML , equation	-1.2123			-1.2123
3	11524	HTML , equation	-0.29406			-0.29406
4	10315	HTML , equation	0.98783			0.98783
5	9946	HTML , equation	-0.47645			-0.47645
6	10298	HTML , equation	0.12637			0.12637
7	15666	HTML , equation	-1.02157			-1.02157
8	10363	HTML , equation	0.10543			0.10543
9	15655	HTML , multipleChoice	0.48306			0.48306
10	15682	HTML , equation	0.72006			0.72006
11	10350	HTML , equation	1.52734			1.52734
12	10295	HTML , equation	0.71996			0.71996
13	14314	HTML , multipleChoice	-1.76299			-1.76299
14	15658	HTML , equation	0.55429			0.55429
15	10301	HTML , equation	-0.54092			-0.54092
16	15680	HTML , equation	0.7272			0.7272
17	15653	HTML , equation	0.45519			0.45519
18	12930	HTML , multipleChoice	0.01787			0.01787
19	10344	HTML , multipleSelect	1.05834			1.05834
20	15692	HTML , multipleChoice	-1.97504			-1.97504
21	12929	HTML , equation	0.45298			0.45298
22	12183	HTML , equation	1.91021			1.91021
23	10339	HTML , equation	2.28753			2.28753
24	12472	HTML , multipleChoice	-1.55344			-1.55344
25	10378	HTML , multipleChoice	-2.47548			-2.47548
26	15671	HTML , equation	0.25148			0.25148
27	13132	HTML , equation	1.03065			1.03065
28	15668	HTML , equation	-0.74078			-0.74078
29	14309	HTML , equation	0.82777			0.82777
30	14649	HTML , equation	0.61892			0.61892
31	10299	HTML , equation	-0.49012			-0.49012
32	10362	HTML , tableMatch	-0.5352			-0.5352
33	13798	HTML , equation	1.4237			1.4237
34	11580	HTML , multipleChoice	0.232			0.232
35	14316	HTML , multipleChoice	-0.84689			-0.84689
36	15699	HTML , multipleChoice	-0.88774			-0.88774
37	10303	HTML , multipleChoice	-1.12478			-1.12478
38	13122	HTML , equation	0.33845			0.33845
39	9504	HTML , equation	0.16514			0.16514
40	15656	HTML , multipleChoice	-0.9032			-0.9032

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10371	HTML , multipleChoice	-1.15822			-1.15822
42	12280	HTML , multipleChoice	0.15052			0.15052
43	11742	HTML , multipleChoice	-0.71542			-0.71542
44	15687	HTML , equation	-0.33934			-0.33934
45	14185	HTML , equation	0.74628			0.74628
46	10308	HTML , tableMatch	1.49225			1.49225
47	9508	HTML , equation	1.88124			1.88124



**Appendix F.15—Spring 18 Operational Item Parameter Estimates — Grade 8 Mathematics**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10499	HTML , multipleChoice	-2.01606			-2.01606
2	15715	HTML , equation	-1.23959			-1.23959
3	15728	HTML , multipleSelect	0.08268			0.08268
4	11304	HTML , equation	-0.65743			-0.65743
5	15716	HTML , multipleSelect	0.07104			0.07104
6	15951	HTML , multipleChoice	0.24159			0.24159
7	8251	HTML , equation	0.87585			0.87585
8	14372	HTML , equation	1.15236			1.15236
9	14271	HTML , multipleSelect	1.49935			1.49935
10	10536	HTML , equation	2.18874			2.18874
11	15733	HTML , equation	0.32537			0.32537
12	13152	HTML , equation	0.34924			0.34924
13	10583	HTML , multipleChoice	-0.41648			-0.41648
14	15723	HTML , multipleChoice	-0.85255			-0.85255
15	12460	HTML , equation	-1.85636			-1.85636
16	10585	HTML , equation	1.6013			1.6013
17	10532	HTML , equation	0.16579			0.16579
18	15740	HTML , equation	1.73743			1.73743
19	10543	HTML , multipleChoice	-1.05211			-1.05211
20	9532	HTML , equation	0.73485			0.73485
21	10561	HTML , HTML , multipleChoice	-0.22329			-0.22329
22	10567	HTML , equation	-0.90128			-0.90128
23	10584	HTML , multipleChoice	-2.09132			-2.09132
24	14640	HTML , multipleChoice	-1.6905			-1.6905
25	10554	HTML , multipleChoice	-1.02849			-1.02849
26	12005	HTML , multipleChoice	-0.97905			-0.97905
27	15752	HTML , multipleChoice	-1.64451			-1.64451
28	10517	HTML , tableMatch	-0.60752			-0.60752
29	12476	HTML , multipleChoice	-0.37326			-0.37326
30	10487	HTML , multipleChoice	-0.12623			-0.12623
31	15744	HTML , equation	0.10768			0.10768
32	11527	HTML , equation	0.39793			0.39793
33	11360	HTML , equation	-0.36993			-0.36993
34	10546	HTML , equation	1.42971			1.42971
35	12461	HTML , equation	1.61536			1.61536
36	10510	HTML , multipleChoice	0.71001			0.71001
37	14378	HTML , equation	-0.29899			-0.29899
38	15950	HTML , multipleChoice	-0.81943			-0.81943
39	10478	HTML , multipleChoice	-0.51804			-0.51804
40	15725	HTML , multipleChoice	-0.09246			-0.09246

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12037	HTML , equation	0.92297			0.92297
42	12181	HTML , equation	1.77848			1.77848
43	15738	HTML , equation	1.58339			1.58339
44	14377	HTML , equation	0.46408			0.46408
45	15724	HTML , multipleChoice	-0.55342			-0.55342
46	14390	HTML , tableInput	-1.13771			-1.13771
47	12458	HTML , multipleChoice	-1.52613			-1.52613

**Appendix F.16—Spring 18 Operational Item Parameter Estimates — Algebra I**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	14104	HTML , equation	-1.14858			-1.14858
2	14029	HTML , multipleChoice	-1.62968			-1.62968
3	15957	HTML , multipleChoice	-0.91456			-0.91456
4	10953	HTML , HTML , multipleChoice	-0.75521			-0.75521
5	14063	HTML , equation	-0.41151			-0.41151
6	13164	HTML , multipleChoice	-0.4663			-0.4663
7	10981	HTML , multipleChoice	-0.28524			-0.28524
8	14635	HTML , multipleChoice	-0.0971			-0.0971
9	13973	HTML , multipleSelect	0.70549			0.70549
10	14008	HTML , equation	0.8123			0.8123
11	13994	HTML , equation	0.65012			0.65012
12	10882	HTML , equation	0.8083			0.8083
13	12020	HTML , equation	1.19453			1.19453
14	10966	HTML , multipleChoice	0.15284			0.15284
15	10942	HTML , multipleChoice	0.09573			0.09573
16	10974	HTML , multipleChoice	-0.58968			-0.58968
17	10963	HTML , multipleChoice	-0.87904			-0.87904
18	15771	HTML , multipleChoice	-0.5789			-0.5789
19	13185	HTML , multipleSelect	1.09113			1.09113
20	12898	HTML , equation	0.33053			0.33053
21	9544	HTML , equation	1.38977			1.38977
22	10907	HTML , multipleChoice	-1.12152			-1.12152
23	9535	HTML , equation	-0.04011			-0.04011
24	11537	HTML , tableInput	1.22253			1.22253
25	9705	HTML , grid	-0.96679			-0.96679
26	10934	HTML , equation	-0.94776			-0.94776
27	14176	HTML , multipleChoice	-1.04401			-1.04401
28	15758	HTML , equation	-1.21972			-1.21972
29	9545	HTML , equation	1.46604			1.46604
30	12593	HTML , multipleChoice	-0.11769			-0.11769
31	10990	HTML , multipleChoice	0.65596			0.65596
32	10940	HTML , multipleChoice	0.54039			0.54039
33	11574	HTML , equation	0.52554			0.52554
34	11005	HTML , multipleSelect	1.27454			1.27454
35	14118	HTML , equation	1.71547			1.71547
36	12023	HTML , equation	-0.46606			-0.46606
37	14055	HTML , multipleChoice	-0.60106			-0.60106
38	12150	HTML , multipleChoice	-0.7494			-0.7494
39	13977	HTML , multipleSelect	1.06567			1.06567
40	13155	HTML , multipleChoice	0.87644			0.87644

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	11577	HTML , multipleChoice	1.47585			1.47585
42	14600	HTML , multipleChoice	-0.7925			-0.7925
43	9536	HTML , grid	1.36497			1.36497
44	11664	HTML , equation	1.6623			1.6623
45	10977	HTML , HTML , multipleChoice	-0.42252			-0.42252
46	10905	HTML , multipleChoice	-0.97781			-0.97781
47	11548	HTML , multipleChoice	-1.18555			-1.18555

**Appendix F.17—Spring 18 Operational Item Parameter Estimates — Geometry**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	12342	HTML , equation	-0.52835			-0.52835
2	11070	HTML , multipleChoice	-1.99292			-1.99292
3	10910	HTML , multipleChoice	-0.7847			-0.7847
4	11792	HTML , equation	0.43582			0.43582
5	11734	hotTextGapMatch	-0.93824			-0.93824
6	14278	HTML , multipleChoice	-0.24634			-0.24634
7	14968	HTML , multipleChoice	0.55906			0.55906
8	14936	HTML , equation	1.46698			1.46698
9	11078	HTML , multipleChoice	-0.56802			-0.56802
10	13531	HTML , tableMatch	0.3836			0.3836
11	12947	HTML , equation	1.5331			1.5331
12	15854	HTML , equation	1.39183			1.39183
13	10932	HTML , multipleChoice	-1.44689			-1.44689
14	11040	HTML , equation	-0.73442			-0.73442
15	11545	HTML , equation	1.96788			1.96788
16	11087	HTML , multipleSelect	-0.11232			-0.11232
17	13500	HTML , multipleChoice	-0.35376			-0.35376
18	13538	HTML , equation	1.10559			1.10559
19	11062	HTML , multipleSelect	0.92341			0.92341
20	11035	HTML , multipleChoice	-0.58779			-0.58779
21	15837	HTML , equation	-0.29917			-0.29917
22	11068	HTML , multipleChoice	-0.64576			-0.64576
23	15804	HTML , multipleChoice	-1.95846			-1.95846
24	12579	HTML , HTML , multipleChoice	-0.59389			-0.59389
25	14972	HTML , equation	-1.5177			-1.5177
26	15815	HTML , multipleChoice	-0.46105			-0.46105
27	14298	HTML , equation	-0.74605			-0.74605
28	15820	hotTextCustom	0.50878			0.50878
29	10929	HTML , multipleChoice	-0.9701			-0.9701
30	14282	HTML , multipleChoice	-0.60042			-0.60042
31	14305	HTML , equation	0.74206			0.74206
32	14134	HTML , multipleSelect	0.76369			0.76369
33	14246	HTML , equation	1.5708			1.5708
34	12344	HTML , multipleSelect	1.46466			1.46466
35	15169	HTML , equation	1.5933			1.5933
36	9564	HTML , equation	0.43952			0.43952
37	9547	HTML , equation	-1.26142			-1.26142
38	12564	HTML , multipleChoice	-1.0368			-1.0368
39	11448	HTML , multipleChoice	-1.87122			-1.87122
40	14942	HTML , multipleChoice	-0.11675			-0.11675

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	12925	HTML , equation	1.31561			1.31561
42	14874	HTML , equation	0.90565			0.90565
43	12152	HTML , equation	1.23693			1.23693
44	12028	HTML , equation	0.62314			0.62314
45	11041	HTML , multipleChoice	-1.59467			-1.59467
46	11018	HTML , multipleChoice	-0.61172			-0.61172
47	14285	HTML , multipleChoice	-0.76083			-0.76083

**Appendix F.18—Spring 18 Operational Item Parameter Estimates — Algebra II**

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	15859	HTML , multipleChoice	-1.10733			-1.10733
2	14266	HTML , multipleChoice	-1.86537			-1.86537
3	11121	HTML , multipleChoice	-1.0416			-1.0416
4	9573	HTML , equation	0.31592			0.31592
5	10166	HTML , multipleChoice	-1.94739			-1.94739
6	10223	HTML , equation	0.48221			0.48221
7	14772	HTML , multipleSelect	1.16796			1.16796
8	15909	HTML , multipleSelect	0.2565			0.2565
9	10252	HTML , equation	0.50929			0.50929
10	14366	HTML , equation	0.72513			0.72513
11	10220	HTML , multipleChoice	0.5445			0.5445
12	14265	HTML , multipleChoice	-1.49409			-1.49409
13	14350	HTML , multipleSelect	0.76145			0.76145
14	10249	HTML , equation	1.06528			1.06528
15	11541	HTML , multipleChoice	-1.39444			-1.39444
16	14676	HTML , equation	0.93692			0.93692
17	15887	HTML , multipleChoice	-1.31843			-1.31843
18	10193	HTML , multipleChoice	-0.57364			-0.57364
19	8252	equation , HTML	-1.4959			-1.4959
20	15866	HTML , multipleSelect	0.87468			0.87468
21	10192	HTML , multipleChoice	-1.21466			-1.21466
22	13404	HTML , grid	0.32394			0.32394
23	12158	HTML , multipleSelect	1.05562			1.05562
24	10204	hotTextCustom	-1.37366			-1.37366
25	10200	HTML , multipleChoice	-1.46483			-1.46483
26	10214	HTML , multipleChoice	-1.65318			-1.65318
27	14288	HTML , equation	1.94923			1.94923
28	10187	HTML , multipleChoice	-0.17866			-0.17866
29	11804	HTML , equation	1.19746			1.19746
30	14355	HTML , equation	0.68733			0.68733
31	14652	HTML , equation	-0.27312	0.41036		0.06862
32	12611	HTML , equation	1.48557			1.48557
33	9576	HTML , equation	-0.8424			-0.8424
34	15883	HTML , equation	0.87271			0.87271
35	11394	HTML , multipleSelect	0.53748			0.53748
36	11936	HTML , multipleChoice	-0.51723			-0.51723
37	11401	HTML , equation	0.64249			0.64249
38	14855	HTML , equation	1.27157			1.27157
39	12076	HTML , multipleSelect	1.05206			1.05206
40	10236	HTML , multipleChoice	-0.89747			-0.89747

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10228	HTML , tableMatch	0.55259			0.55259
42	10210	HTML , multipleChoice	-1.79638			-1.79638
43	11836	HTML , equation	1.19312			1.19312
44	9567	HTML , equation	0.98222			0.98222
45	15869	HTML , equation	-1.12221			-1.12221
46	11604	HTML , equation	1.65144			1.65144
47	13231	HTML , multipleChoice	-1.90022			-1.90022



**Appendix G.1 – Number of Participating Students by Demographic Subgroups – ELA Online**

<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>	<b>Grade 9</b>	<b>Grade 10</b>	<b>Grade 11</b>
<b>All students</b>	71,562	74,397	74,970	73,437	71,436	70,990	69,227	63,174	59,553
<b>Female</b>	34,759	36,537	36,762	36,018	35,234	34,389	34,131	31,496	29,897
<b>Male</b>	36,803	37,860	38,208	37,419	36,202	36,601	35,096	31,678	29,656
<b>African American</b>	1,850	1,968	1,953	1,871	1,816	1,994	2,231	1,982	1,792
<b>Asian</b>	1,996	2,017	1,910	1,927	1,878	2,015	2,073	1,833	1,829
<b>Native Hawaiian/Pacific</b>	144	136	184	150	129	156	183	122	163
<b>Hispanic/Latino</b>	33,202	34,883	34,612	33,393	32,069	31,649	30,220	27,469	26,087
<b>American Indian or Alaskan</b>	3,738	3,887	3,937	3,786	3,995	3,868	3,865	3,386	2,970
<b>White</b>	29,058	29,875	30,733	30,839	30,194	30,127	29,887	27,384	25,748
<b>Multiple</b>	1,571	1,625	1,633	1,471	1,354	1,181	758	992	956
<b>Limited English Proficiency</b>	6,102	6,720	6,453	5,083	3,782	3,565	4,364	3,335	2,837
<b>Special Education</b>	8,503	9,248	9,304	8,525	7,980	7,431	5,987	5,156	4,592
<b>Free/Reduced Lunch</b>	32,984	34,755	34,912	33,484	31,838	30,704	21,799	19,272	16,901
<b>Accommodation</b>	4,289	4,724	4,537	3,811	3,205	2,924	1,533	1,068	853

**Appendix G.2 – Number of Participating Students by Demographic Subgroups – ELA Paper**


<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>	<b>Grade 9</b>	<b>Grade 10</b>	<b>Grade 11</b>
<b>All students</b>	14,155	14,439	14,442	14,509	14,926	14,460	12,605	11,246	9,852
<b>Female</b>	6,915	7,246	7,181	7,178	7,470	7,140	6,289	5,538	4,949
<b>Male</b>	7,240	7,193	7,261	7,331	7,456	7,320	6,316	5,708	4,903
<b>African American</b>	397	384	390	416	366	400	381	376	354
<b>Asian</b>	860	888	972	924	943	828	750	665	623
<b>Native Hawaiian/Pacific</b>	7	35	18	28	27	29	29	25	27
<b>Hispanic/Latino</b>	6,217	6,591	6,491	6,504	6,738	6,698	6,224	5,536	4,780
<b>American Indian or Alaskan</b>	762	673	697	735	528	504	295	241	187
<b>White</b>	5,683	5,675	5,692	5,667	6,150	5,807	4,854	4,222	3,732
<b>Multiple</b>	229	193	182	234	174	194	71	181	148
<b>Limited English Proficiency</b>	1,089	1,312	1,237	912	876	837	623	255	148
<b>Special Education</b>	1,380	1,511	1,508	1,526	1,414	1,393	1,025	887	733
<b>Free/Reduced Lunch</b>	6,630	6,978	6,747	6,668	6,735	6,535	4,887	4,486	3,814
<b>Accommodation</b>	1,354	1,351	1,106	926	975	828	174	47	37

**Appendix G.3 – Number of Participating Students by Demographic Subgroups – Mathematics Online**



<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>	<b>Algebra I</b>	<b>Geometry</b>	<b>Algebra II</b>
<b>All students</b>	71,852	74,754	75,221	73,708	71,605	61,535	74,288	59,877	54,351
<b>Female</b>	34,861	36,675	36,864	36,154	35,303	29,662	36,355	30,049	27,798
<b>Male</b>	36,991	38,079	38,357	37,554	36,302	31,873	37,933	29,828	26,553
<b>African American</b>	1,880	1,996	1,984	1,899	1,841	1,885	2,380	1,873	1,580
<b>Asian</b>	2,001	2,029	1,920	1,930	1,887	1,506	2,093	1,817	1,741
<b>Native Hawaiian/Pacific</b>	146	139	185	152	129	134	198	157	167
<b>Hispanic/Latino</b>	33,334	35,044	34,747	33,516	32,169	28,165	33,662	24,957	23,533
<b>American Indian or Alaskan</b>	3,746	3,921	3,955	3,824	4,016	3,699	3,879	3,463	2,710
<b>White</b>	29,163	29,988	30,786	30,913	30,209	25,129	31,131	26,754	23,767
<b>Multiple</b>	1,582	1,635	1,643	1,474	1,354	1,017	945	852	850
<b>Limited English Proficiency</b>	6,152	6,772	6,493	5,121	3,813	3,179	4,638	2,783	2,400
<b>Special Education</b>	8,599	9,329	9,361	8,605	8,045	7,230	6,489	4,593	3,333
<b>Free/Reduced Lunch</b>	33,111	34,946	35,023	33,607	31,898	27,318	25,456	18,857	15,308
<b>Accommodation</b>	4,232	4,692	4,443	3,748	3,138	2,782	1,465	960	512

**Appendix G.4 – Number of Participating Students by Demographic Subgroups – Mathematics Paper**

<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>	<b>Algebra I</b>	<b>Geometry</b>	<b>Algebra II</b>
<b>All students</b>	14,190	14,448	14,495	14,486	14,827	12,133	13,058	11,167	10,268
<b>Female</b>	6,923	7,245	7,198	7,170	7,412	5,947	6,452	5,475	5,254
<b>Male</b>	7,267	7,203	7,297	7,316	7,415	6,186	6,606	5,692	5,014
<b>African American</b>	398	385	392	416	365	367	361	377	359
<b>Asian</b>	865	887	974	893	825	397	831	544	662
<b>Native Hawaiian/Pacific</b>	7	35	18	28	26	25	30	28	23
<b>Hispanic/Latino</b>	6,225	6,597	6,513	6,511	6,779	6,172	6,484	5,647	5,027
<b>American Indian or Alaskan</b>	769	681	697	736	533	485	325	250	193
<b>White</b>	5,697	5,671	5,719	5,666	6,134	4,552	4,900	4,161	3,854
<b>Multiple</b>	229	192	182	235	165	135	127	157	149
<b>Limited English Proficiency</b>	1,091	1,313	1,243	913	893	838	695	489	310
<b>Special Education</b>	1,387	1,522	1,521	1,533	1,423	1,361	1,178	874	650
<b>Free/Reduced Lunch</b>	6,648	6,988	6,763	6,671	6,776	6,082	5,210	4,600	3,854
<b>Accommodation</b>	1,434	1,348	1,106	893	775	614	61	36	20




**Statistical Review  
Training for ADE**




### Statistical Review of Items

- 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 differently across subgroups?




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 **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, 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



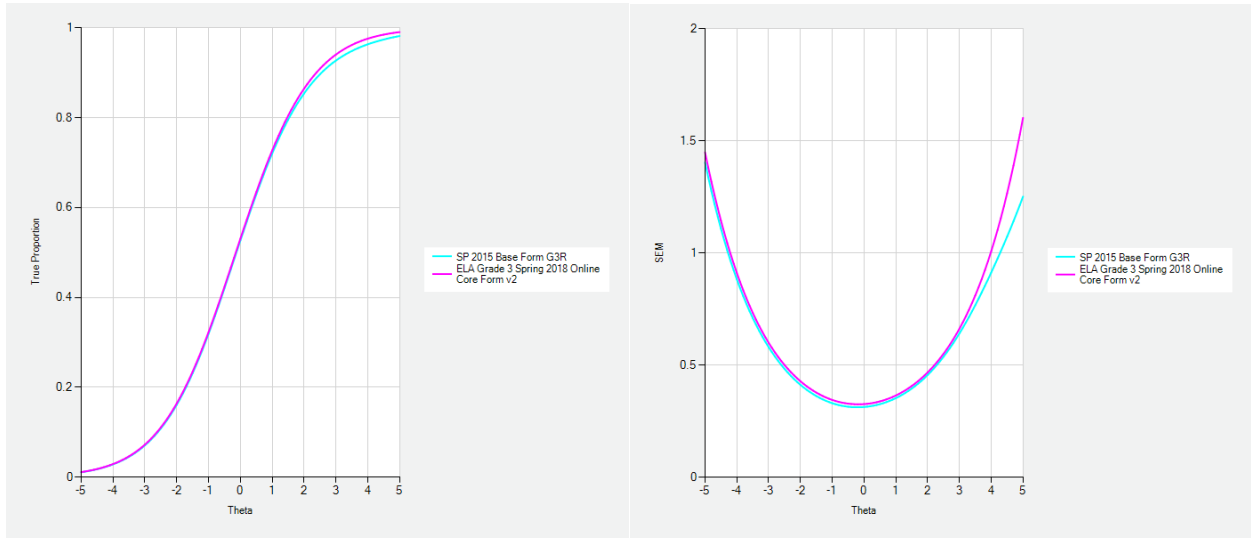
## Logistics

- 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



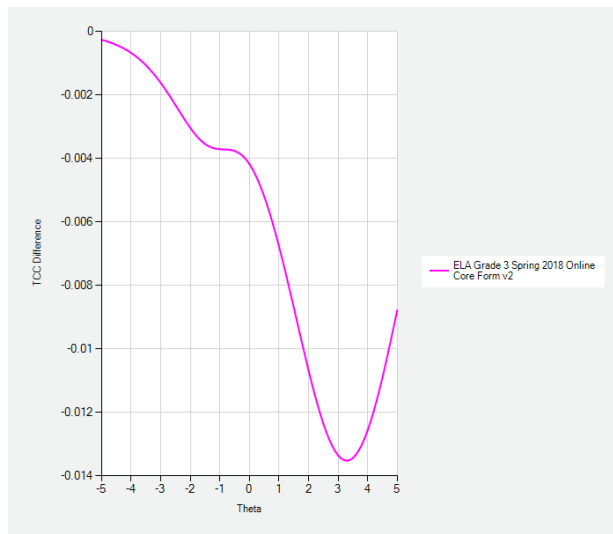
15

Appendix I.1 – Spring 2018 ELA Grade 3



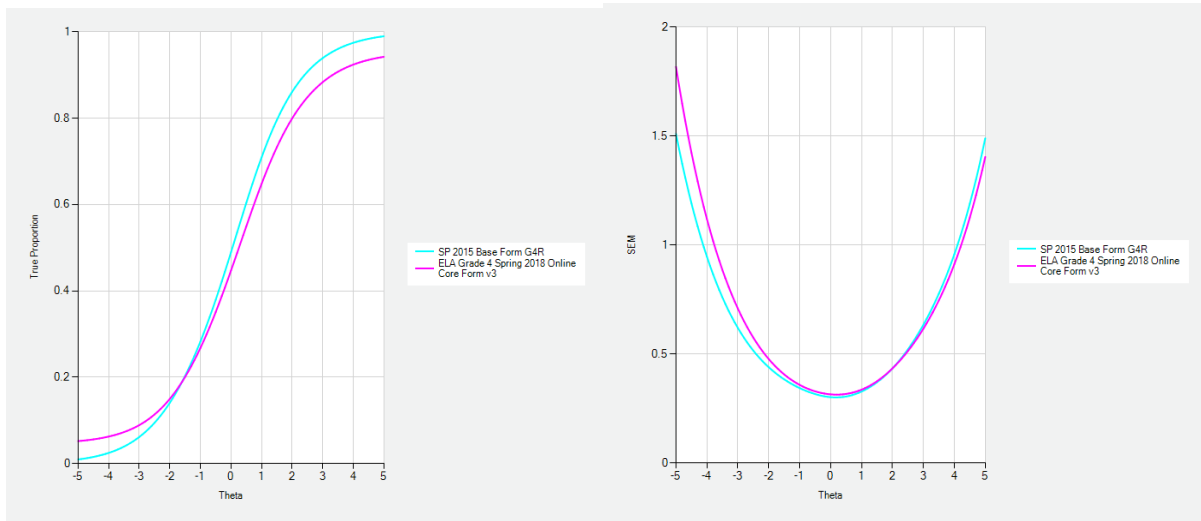
Test Characteristic Curves

Standard Errors of Measurement



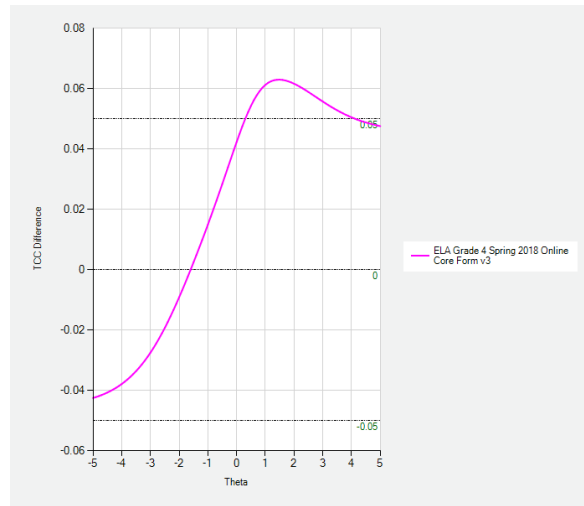
TCC Differences

Appendix I.2 – Spring 2018 ELA Grade 4



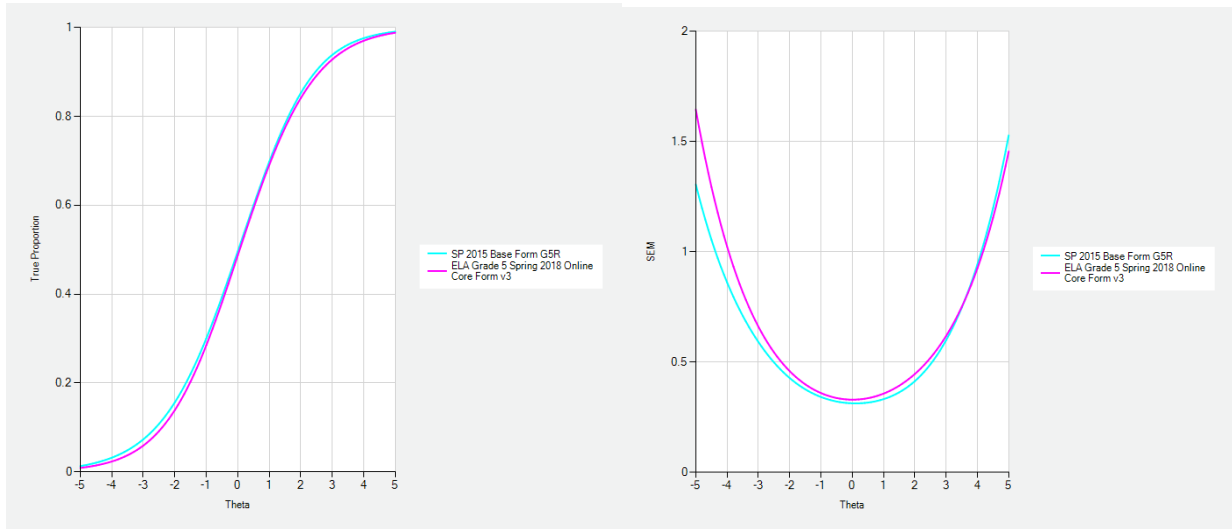
Test Characteristic Curves

Standard Errors of Measurement



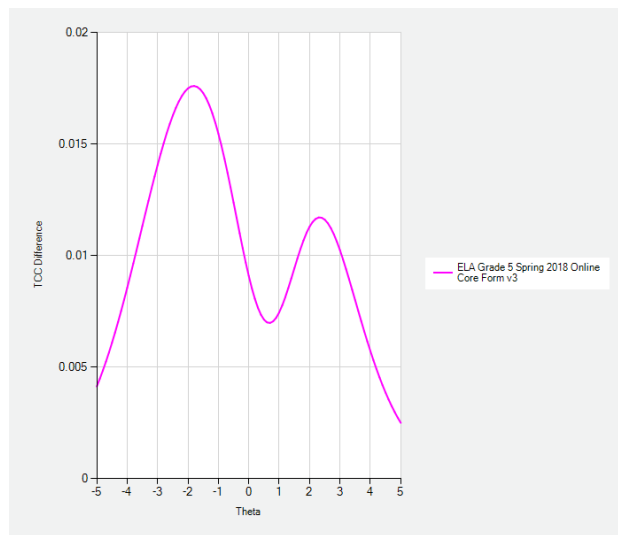
TCC Differences

Appendix I.3 – Spring 2018 ELA Grade 5



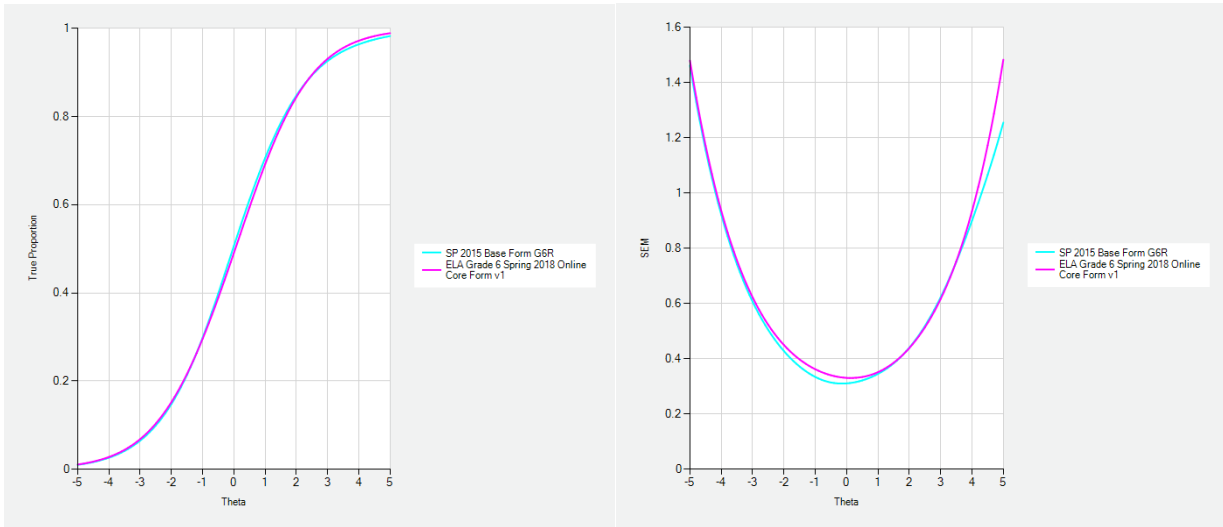
Test Characteristic Curves

Standard Errors of Measurement



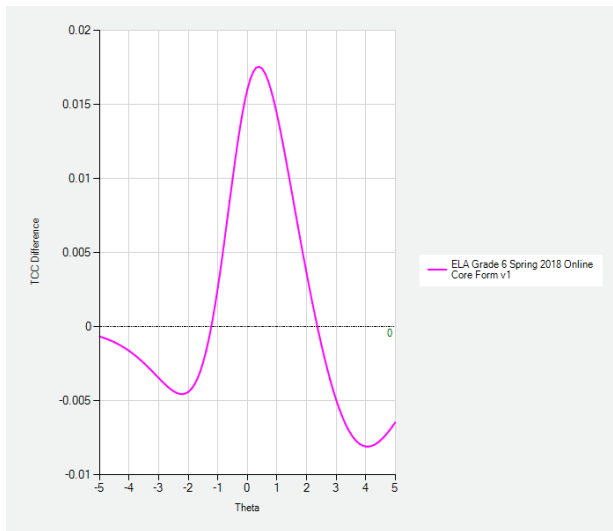
TCC Differences

Appendix I.4 – Spring 2018 ELA Grade 6



Test Characteristic Curves

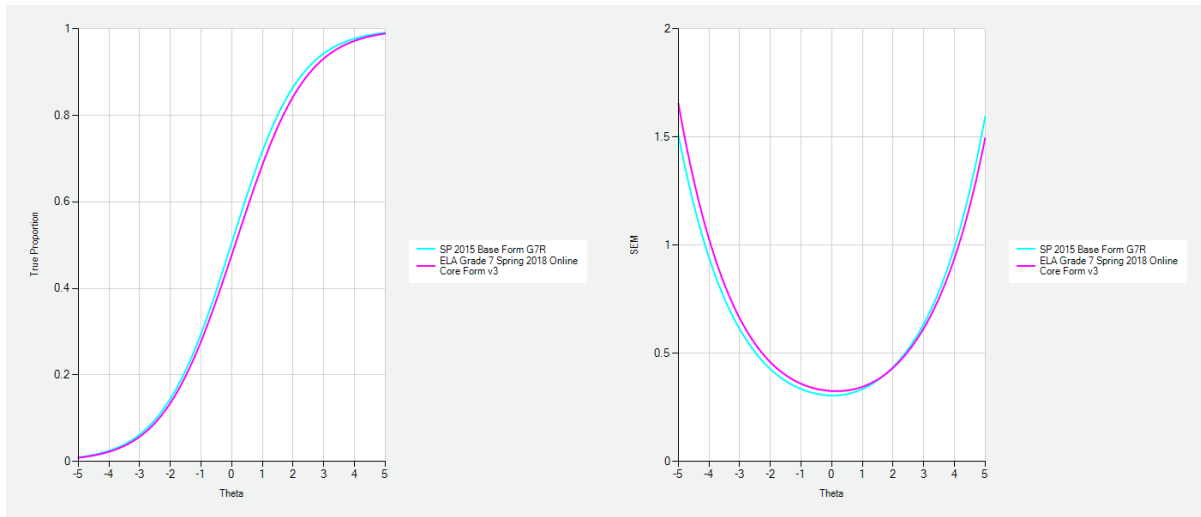
Standard Errors of Measurement



TCC Differences

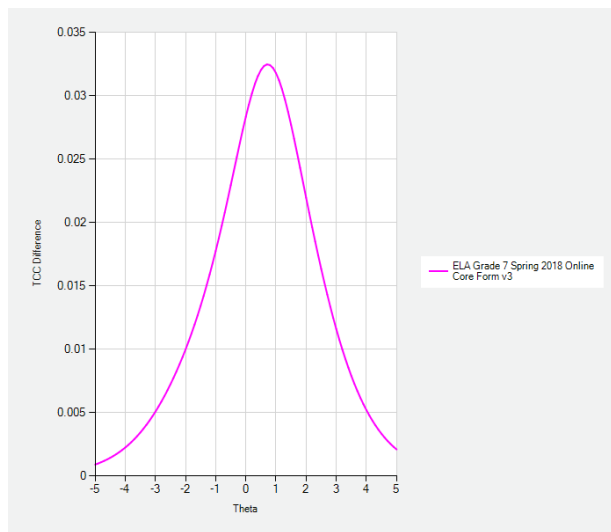


## Appendix I.5 – Spring 2018 ELA Grade 7



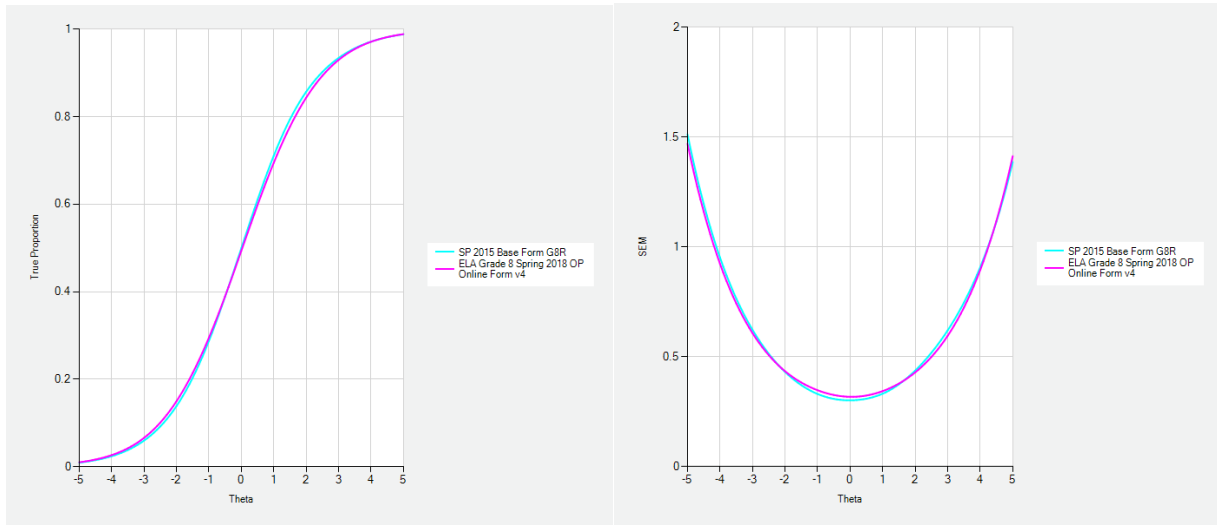
Test Characteristic Curves

Standard Errors of Measurement



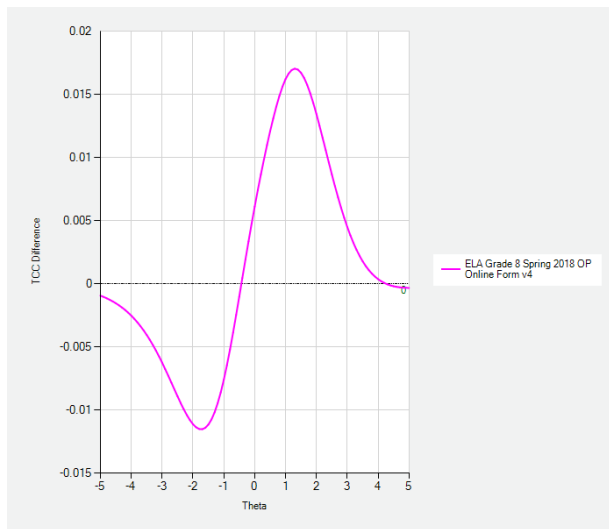
TCC Differences

## Appendix I.6 – Spring 2018 ELA Grade 8



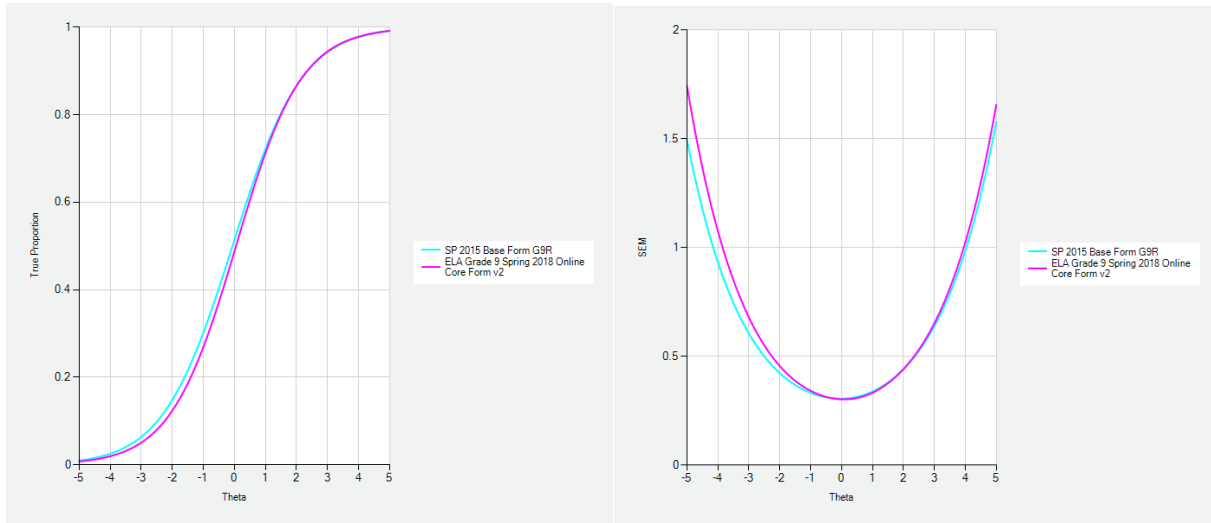
Test Characteristic Curves

Standard Errors of Measurement



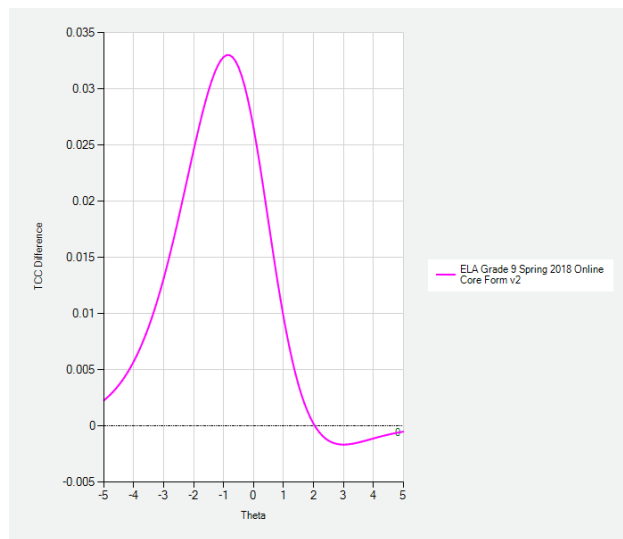
TCC Differences

Appendix I.7 – Spring 2018 ELA Grade 9



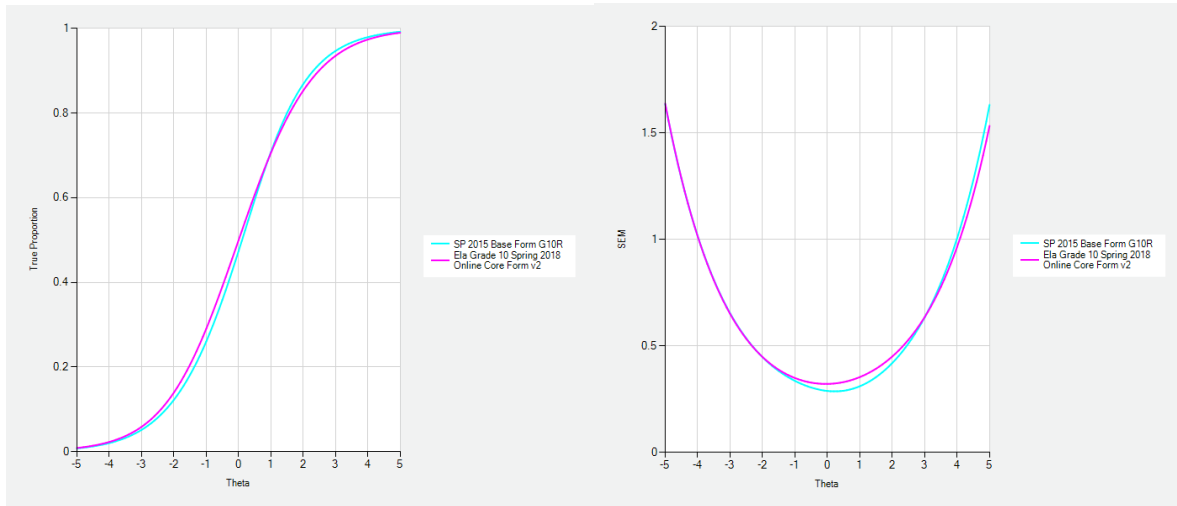
Test Characteristic Curves

Standard Errors of Measurement



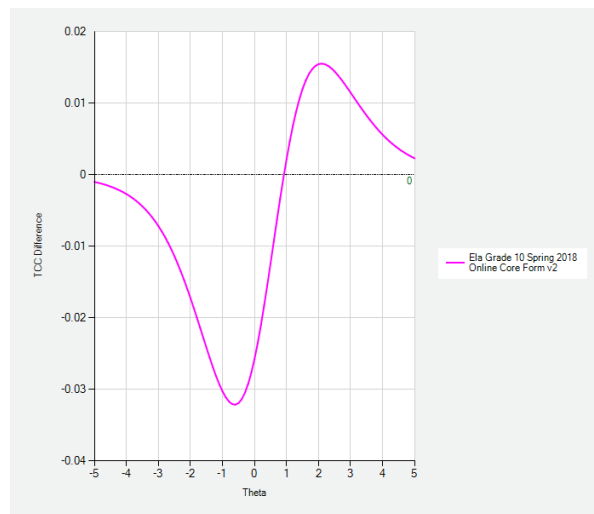
TCC Differences

## Appendix I.8 – Spring 2018 ELA Grade 10



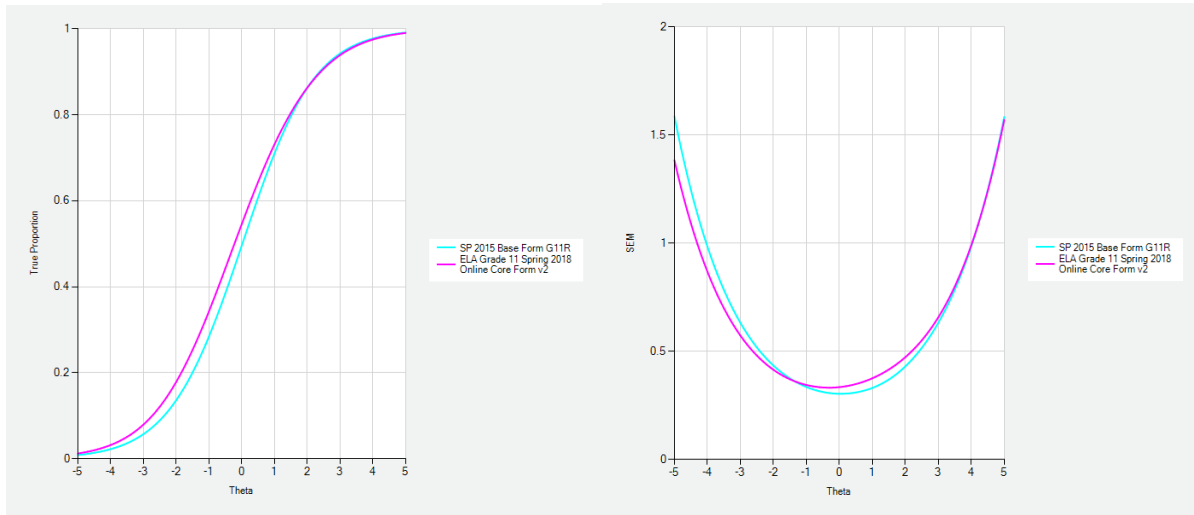
Test Characteristic Curves

Standard Errors of Measurement



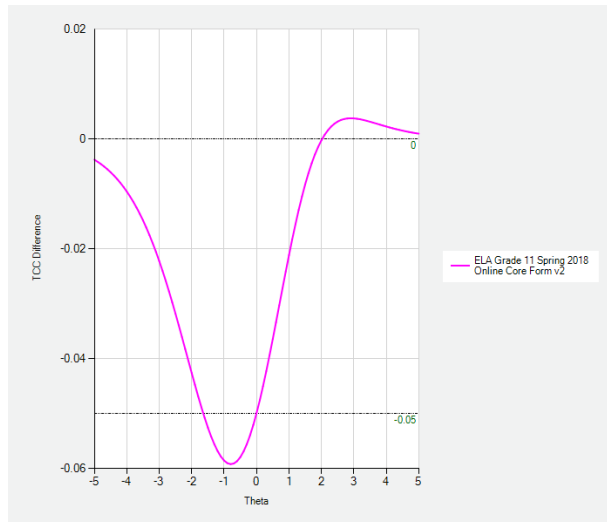
TCC Differences

Appendix I.9 – Spring 2018 ELA Grade 11



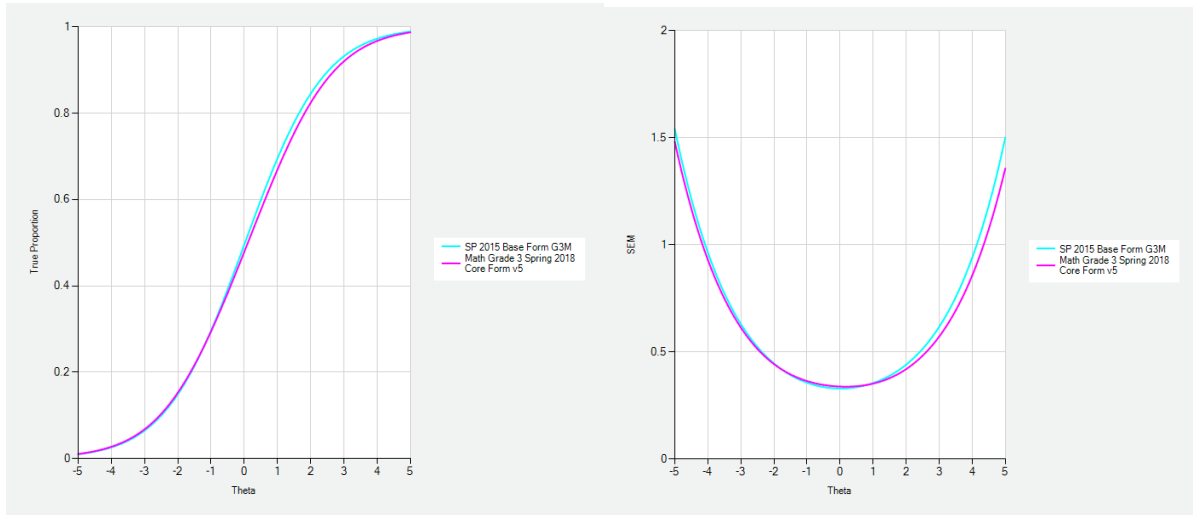
Test Characteristic Curves

Standard Errors of Measurement



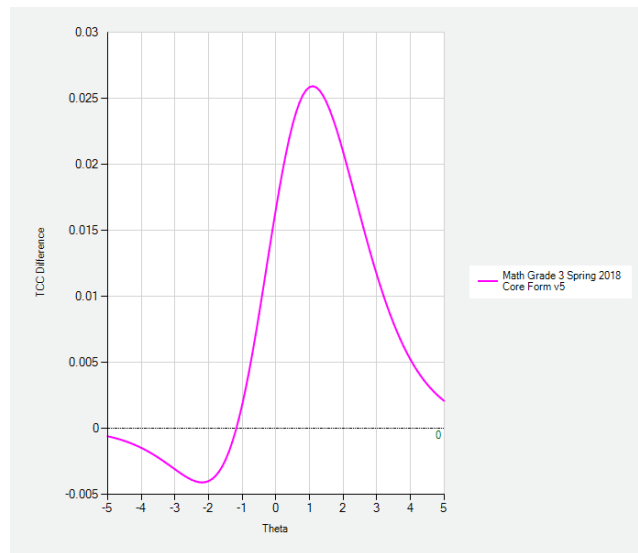
TCC Differences

Appendix I.10 – Spring 2018 Math Grade 3



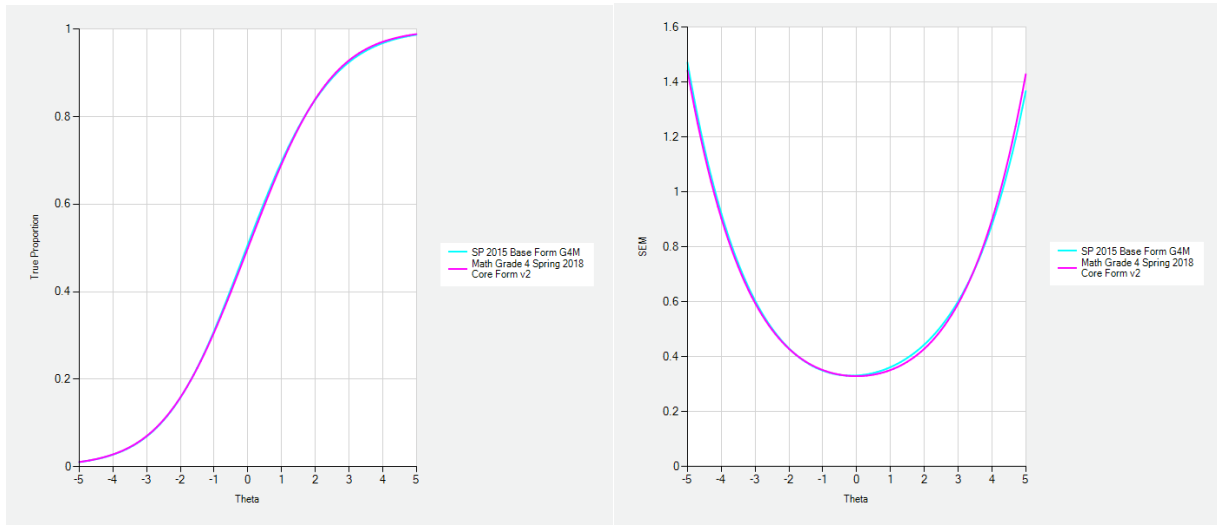
Test Characteristic Curves

Standard Errors of Measurement



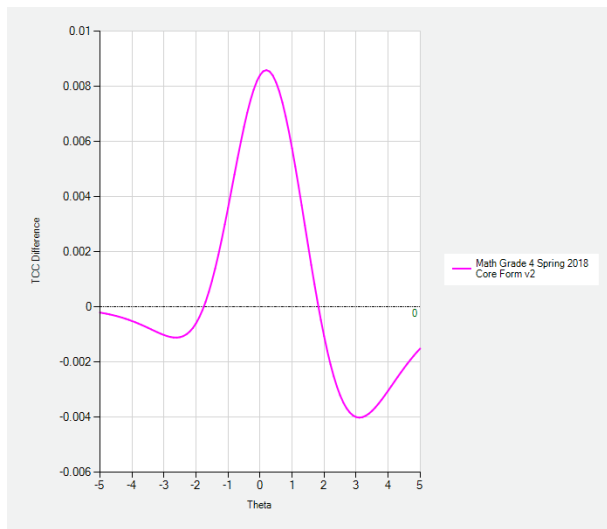
TCC Differences

Appendix I.11 – Spring 2018 Math Grade 4



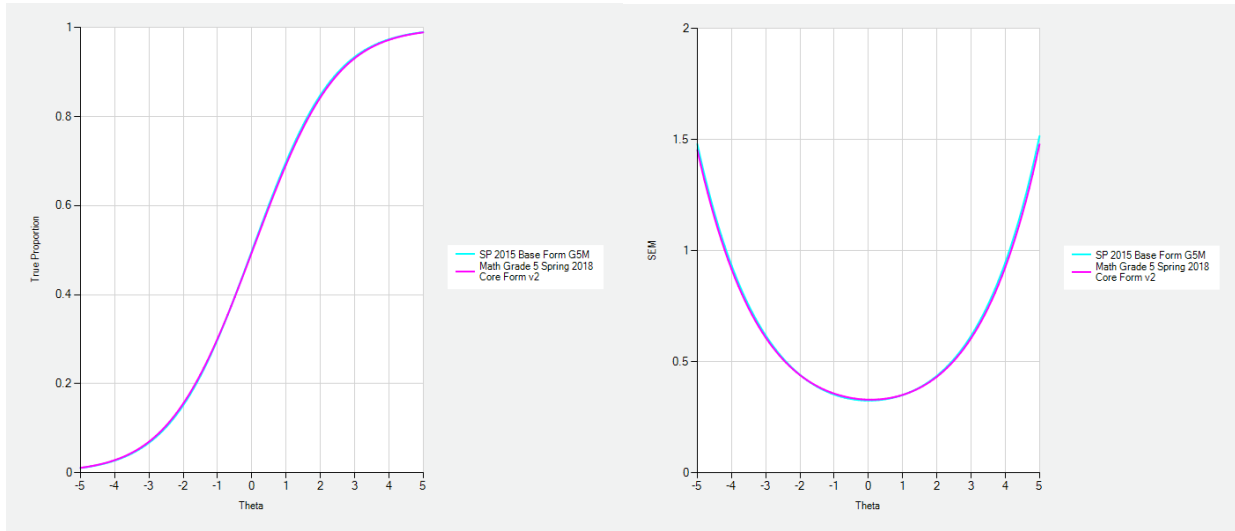
Test Characteristic Curves

Standard Errors of Measurement



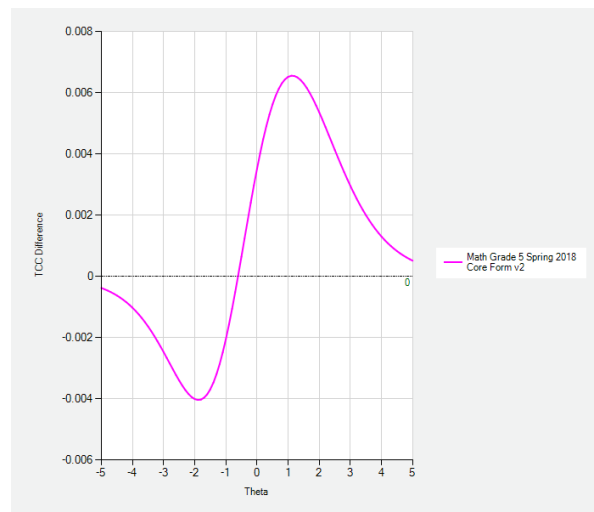
TCC Differences

Appendix I.12 – Spring 2018 Math Grade 5



Test Characteristic Curves

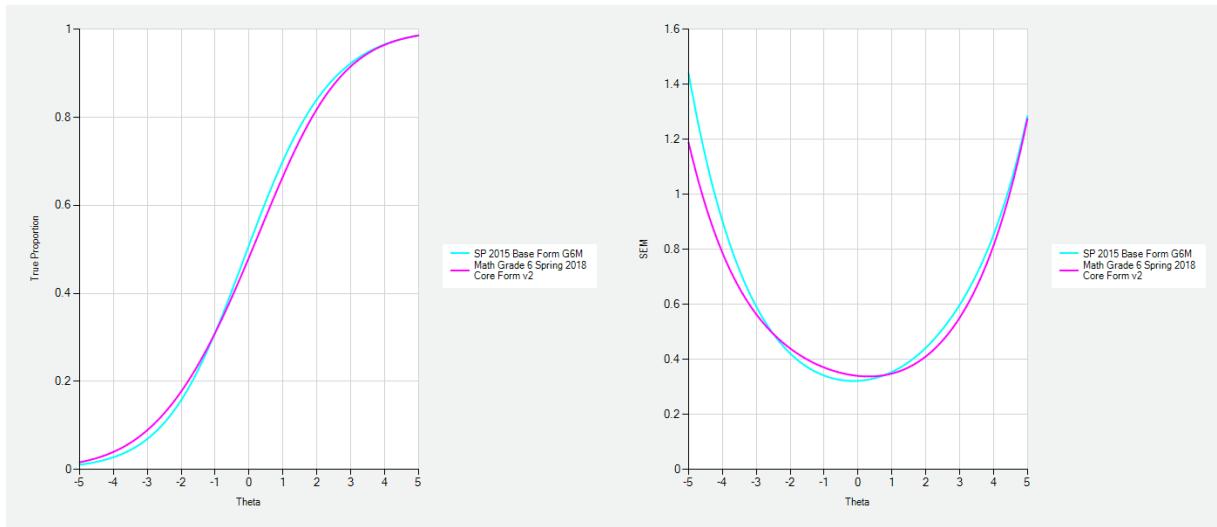
Standard Errors of Measurement



TCC Differences

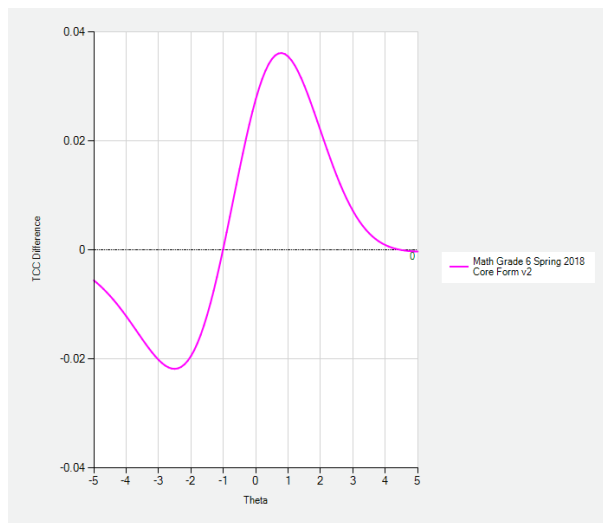


Appendix I.13 – Spring 2018 Math Grade 6



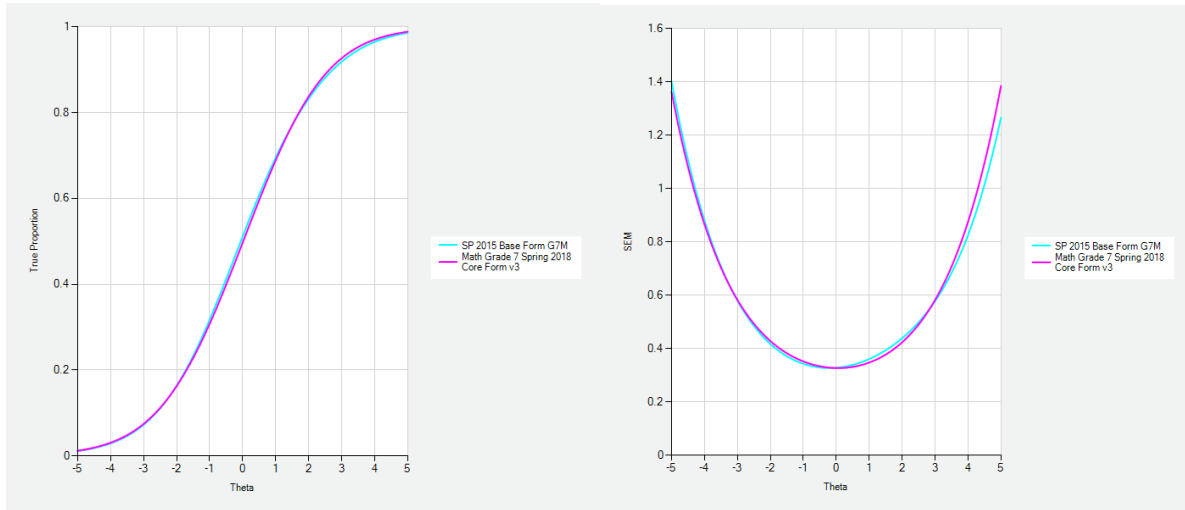
Test Characteristic Curves

Standard Errors of Measurement



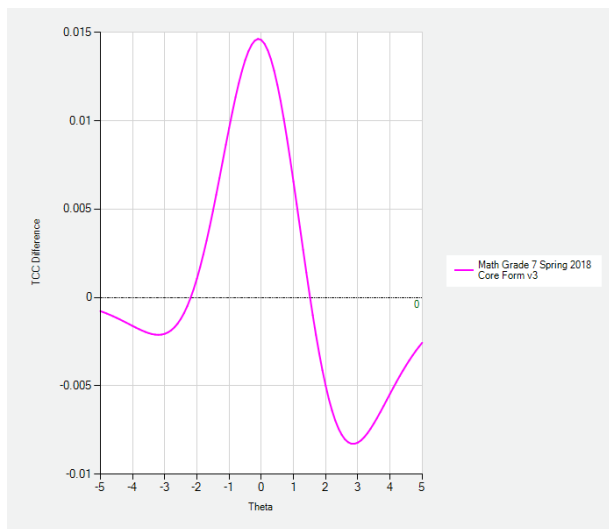
TCC Differences

Appendix I.14 – Spring 2018 Math Grade 7



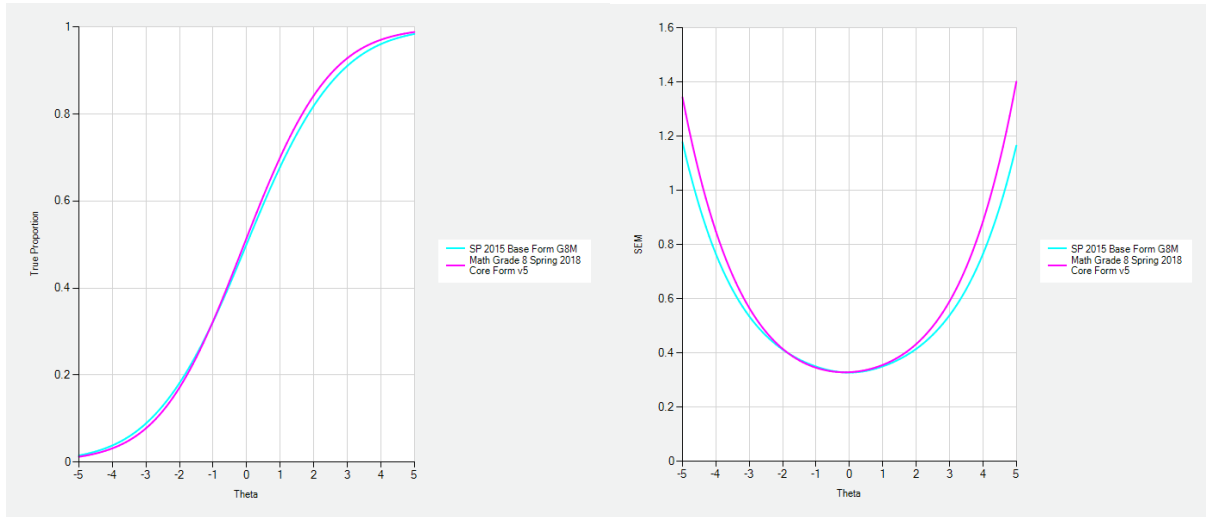
Test Characteristic Curves

Standard Errors of Measurement



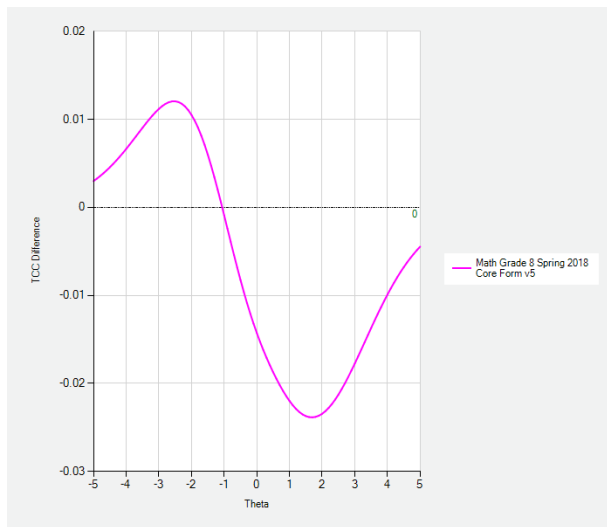
TCC Differences

Appendix I.15 – Spring 2018 Math Grade 8



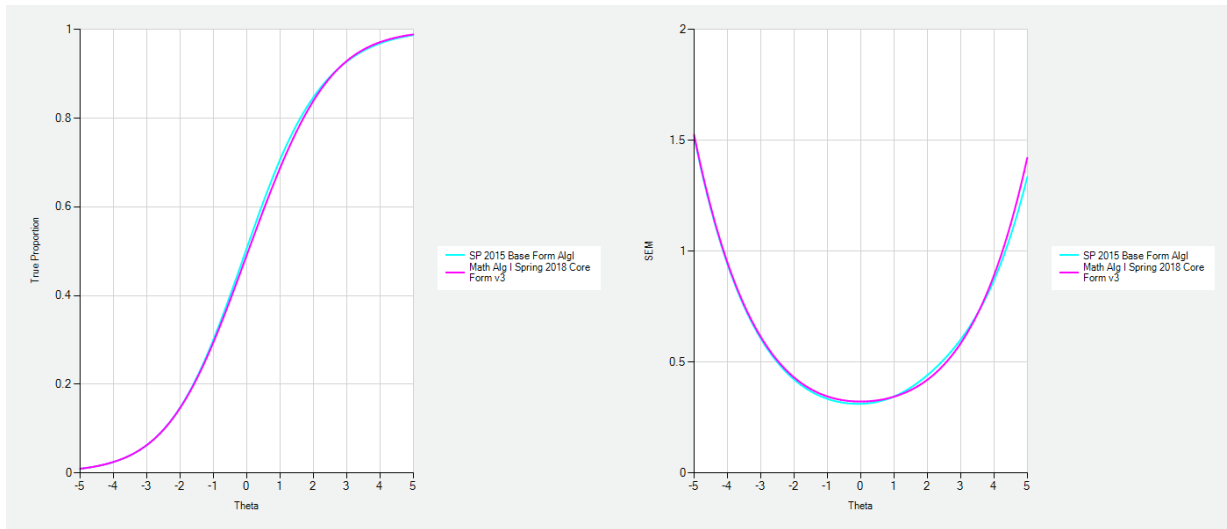
Test Characteristic Curves

Standard Errors of Measurement



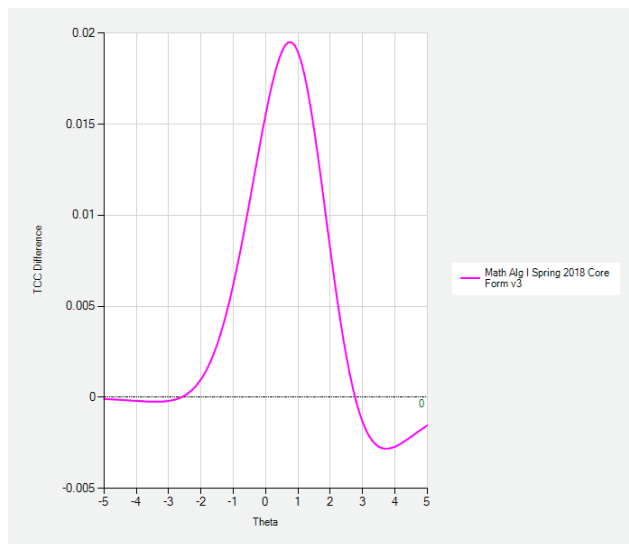
TCC Differences

Appendix I.16 – Spring 2018 Math Algebra I



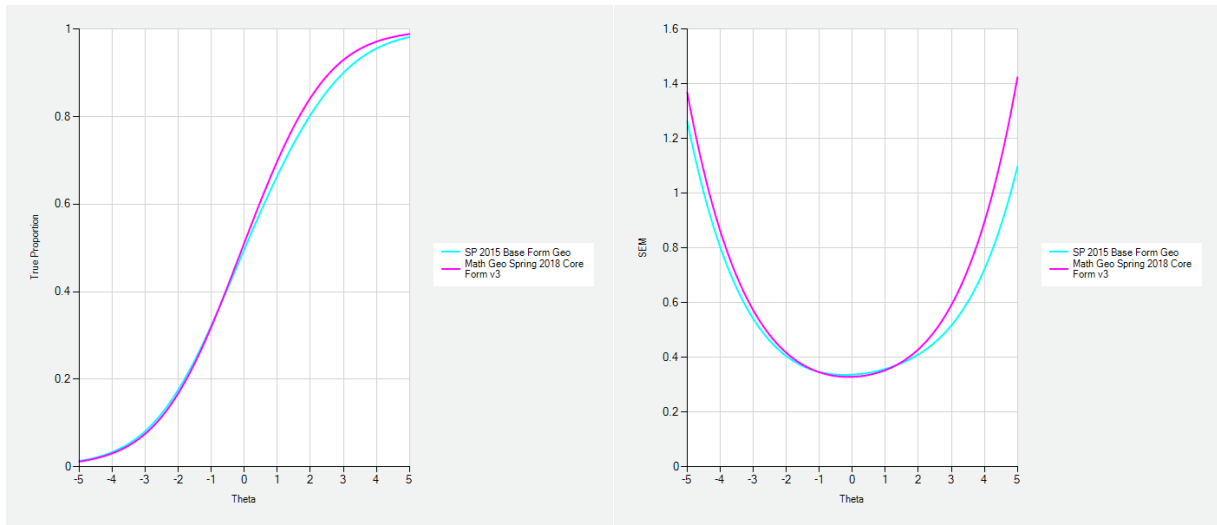
Test Characteristic Curves

Standard Errors of Measurement



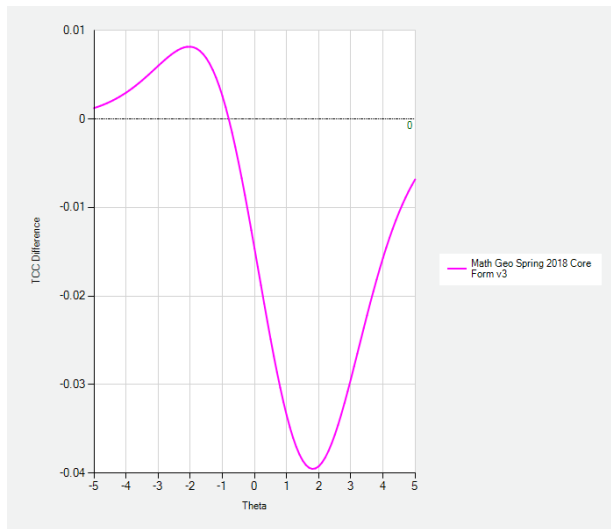
TCC Differences

Appendix I.17 – Spring 2018 Math Geometry



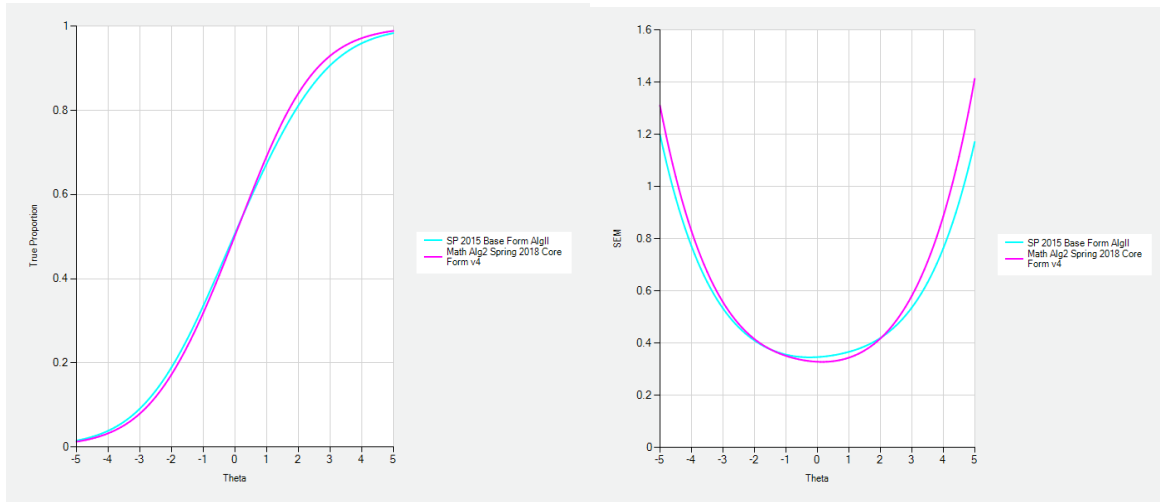
Test Characteristic Curves

Standard Errors of Measurement



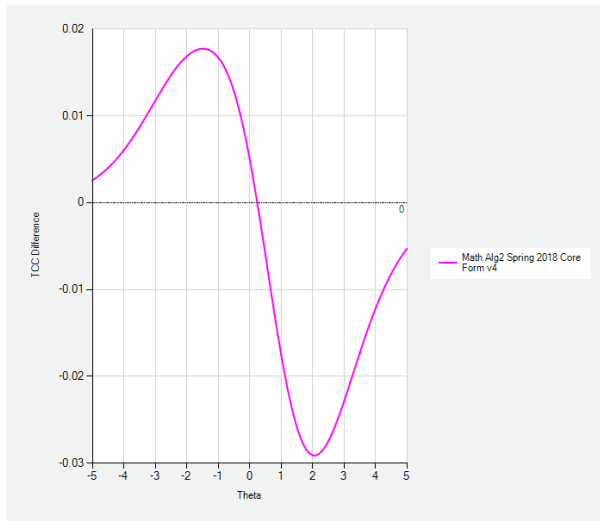
TCC Differences

Appendix I.17 – Spring 2018 Math Algebra II



Test Characteristic Curves

Standard Errors of Measurement



TCC Differences