AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Mathematics

2018–2019 School Year

April 2020

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

2018–2019 ANNUAL TECHNICAL REPORT

APRIL 2020

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

TABLE OF CONTENTS

APPENDICES

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 2018 and fall 2018 administrations of the high school end-of-course (EOC) assessments, and Chapter 4 shows the results of the spring 2019 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 handscored 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.[1] 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.[2] 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.[3] Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards, in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the 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.

---

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

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

[3] Standard 4.1: Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended 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.[4] Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

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

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

In addition, the AzMERIT test delivery system (TDS) provides a range of accessibility tools and accommodations to virtually all students for reducing construct-irrelevant barriers to accessing test content.[5] 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

---

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

[5] Standard 3.1: Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
Standard 3.2: Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.
Standard 12.3: Those responsible for the development and use of educational assessments should design all relevant steps of the testing process to promote access to the construct for all individuals and subgroups for whom the assessment is intended.

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
- Providing extended time (the testing session must be competed 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
Area Boundaries	The student may click anywhere on the selected-response text or button for multiple-choice options.
Expand/Collapse Passage	The student may expand a passage for easier readability. Expanded passages can also be collapsed.
Help	The student may view the on-screen *Test Instructions and Help*.
Highlighter	The student may highlight text in a passage or item.
Line Reader	The student may track the line he or she is reading.
Mark (Flag) for Review	The student may mark an item for review so that it can be easily found later.
Notes/Comments	The student may open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and throughout the session. In mathematics, comments are attached to a specific test item and available throughout the session.
Pause and Restart	The student may pause the session at any time and restart the test if taken over a one-day period. For test security purposes, visibility of past items is not allowed when the test is paused longer than 20 minutes.
Review Test	The student may review the test before ending it.
Strikethrough	The student may cross out answer options for multiple-choice and multi-select items.
System Settings	The student may adjust the audio volume during the test.
Text-to-Speech for Instructions	The student may listen to test instructions.
Tutorial	The student may view a short video about each item type and how to respond.
Writing Tools	The student may use editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italics) for extended-response items.
Zoom In/Zoom Out	The student may zoom in to enlarge the font and images in the test and 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
Adult Transcription	If a student with an injury is testing at a CBT school and cannot enter his or her own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided, 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.
Assistive Technology	Assistive technology may be used for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copies must be shredded. Any electronic copies must be deleted. This accommodation also requires adult transcription (see above for rules on adult transcription).
Rest/Breaks	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
Read Aloud Test Content	CBT: Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and for the mathematics test. 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. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Students may take breaks during testing sessions.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Provide exact oral translation, in the student's native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word, published paper translation dictionary. Students with a visual impairment may use an electronic, word-for-word translation dictionary with other features turned off.

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

Exhibit 1.2.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus for any AzMERIT mathematics test without restrictions.
Adult Transcription	If a student testing at a CBT school has an IEP indicating that he or she cannot enter their own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided 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.
Assistive Technology	This is the use of assistive technology for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copies must be shredded. Any electronic copies must be deleted. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Braille Test Booklet	Provide a paper braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Large Print Test Booklet	CBT: Either increase default zoom settings 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).
Paper-Pencil Test Booklet	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.[6]

Alignment of content standards is achieved through a rigorous test-development process that proceeds from the content standards and refers to those standards in a highly iterative 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

[6] Standard 12.4: When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in enough 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.[7] Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the 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
- 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.

[7] 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.[8]

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.[9]

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 American College Testing (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 grades 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

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

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

performance standards for the grades 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 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			
	AzMERIT Proficient	Arizona ACT College-Ready	Arizona NAEP Proficient	Projected SBAC
ELA				
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%
Mathematics				
Grade 3	42%			39%
Grade 4	42%		42%	38%
Grade 5	40%			33%
Grade 6	32%			33%
Grade 7	31%			33%
Grade 8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

Although AIR previously identified ACT college-ready cut scores on the AzMERIT ELA and 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 Locations of the ACT College-Ready Cut Scores on the AzMERIT Scales

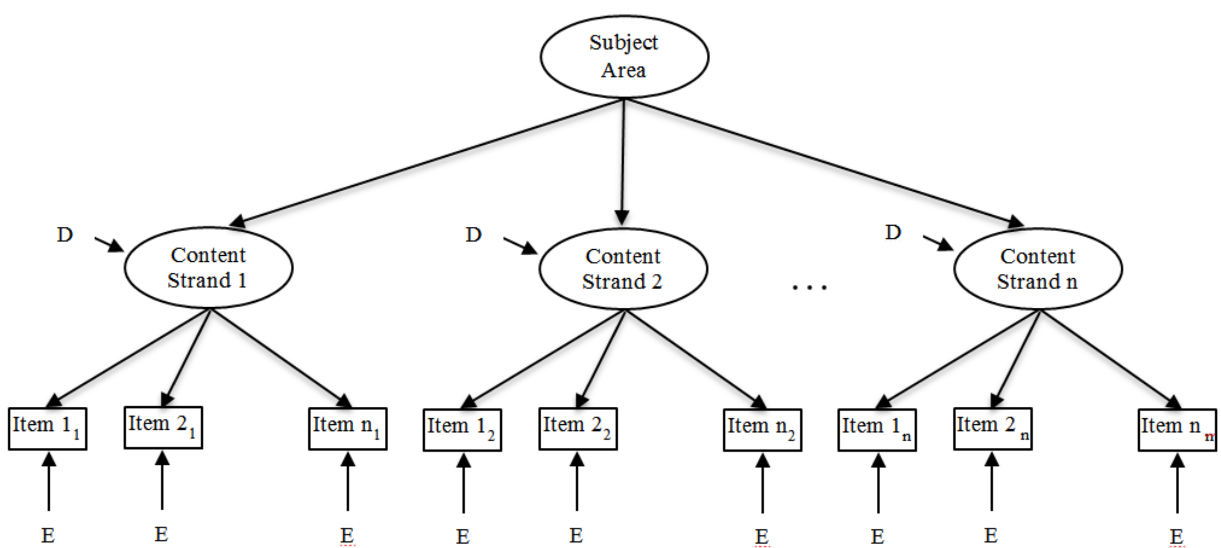
	Location of ACT College-Ready Cut on AzMERIT Scale		AzMERIT Meets Performance Standard
	Via Indirect Linkage Through AIMS	Via Direct Linkage with AzMERIT	
Algebra II	3704	3727	3711
Grade 11 ELA	2579	2585	2585

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 operational test administration in spring 2019, confirmatory factor analysis (CFA) was used to evaluate the fit of this structural model to student response data.[10] For each of the test forms administered in spring 2019, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size, however; even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices, such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (Tucker & Lewis, 1973), and the Root Mean Square Error of Approximation (RMSEA) were also used to evaluate model fit. The guidelines for evaluating goodness of fit is presented in Exhibit in 1.5.2.

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.[11] 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 2019 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.[12]

1.5.1 ELA CONTENT MODEL

We began by evaluating the fit of the first-order, general-achievement model in which all items are indicators of a common subject-area factor. This model importantly evaluates the assumption of unidimensionality of the subject-area assessments, and it provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the first-order, general-achievement models in ELA are shown in Exhibit 1.5.1.1. All the statistics

[10] Standard 1.13: If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.
[11] Standard 1.12: If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.
[12] Standard 1.14: When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

indicate that the general-achievement model fits the data well. This pattern was true across all grades. The CFI and TLI values were all greater than 0.95, 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	0.97	0.96	0.04
4	0.97	0.97	0.03
5	0.97	0.97	0.03
6	0.97	0.97	0.03
7	0.97	0.97	0.03
8	0.97	0.97	0.03
9	0.96	0.96	0.03
10	0.96	0.96	0.03
11	0.98	0.98	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.95, with RMSEA values well below the .05 threshold used to indicate good fit.

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

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

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

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

Grade	χ^2	*df*	*p* value
3	11104.664	3	p < .001
4	8710.343	3	p < .001
5	11209.327	3	p < .001
6	7277.245	3	p < .001
7	6496.039	3	p < .001
8	9434.533	3	p < .001
9	8908.423	3	p < .001
10	2080.738	3	p < .001
11	6686.436	3	p < .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 .98 to .99. RMSEA values were all .02.

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

Grade	CFI	TLI	RMSEA
3	0.99	0.99	0.02
4	0.99	0.99	0.02
5	0.99	0.99	0.02
6	0.99	0.99	0.02
7	0.99	0.99	0.02
8	0.98	0.98	0.02
9	0.99	0.99	0.02
10	0.99	0.99	0.02
11	0.99	0.99	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 χ^2_{Diff} *p*-values were less than .001 across all grade levels.

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

Grade	χ^2	*df*	*p* value
3	10941.713	4	$p < .001$
4	9541.961	4	$p < .001$
5	9820.848	4	$p < .001$
6	8350.609	4	$p < .001$
7	6979.488	4	$p < .001$
8	10244.295	4	$p < .001$
9	9743.542	4	$p < .001$
10	5643.834	4	$p < .001$
11	7237.696	4	$p < .001$

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	0.98	0.03
4	0.95	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.03
7	0.99	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.03
Algebra II	0.98	0.98	0.02
Geometry	0.97	0.97	0.03

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 .95, 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	0.03
4	0.96	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.02
7	0.99	0.99	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.03
Algebra II	0.98	0.98	0.02
Geometry	0.98	0.97	0.03

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 χ^2_{Diff} *p*-values less than .001 across grade levels.

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

Grade	χ^2	*df*	*p* value
3	4858.475	2	$p < .001$
4	7470.266	2	$p < .001$
5	6475.997	3	$p < .001$
6	2124.797	4	$p < .001$
7	1269.169	4	$p < .001$
8	6948.457	3	$p < .001$
Algebra I	350.264	3	$p < .001$
Algebra II	1423.305	3	$p < .001$
Geometry	2981.361	3	$p < .001$

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 .95 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	0.95	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.02
7	0.99	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.03
Algebra II	0.98	0.98	0.02
Geometry	0.97	0.97	0.03

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 χ^2_{Diff} *p*-values less than .001 across grade levels.

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

Grade	χ^2	*df*	*p* value
3	276.254	3	$p < .001$
4	1296.511	3	$p < .001$
5	1064.235	3	$p < .001$
6	2275.704	3	$p < .001$
7	127.198	3	$p < .001$
8	2819.923	3	$p < .001$
Algebra I	943.054	2	$p < .001$
Algebra II	231.444	3	$p < .001$
Geometry	764.109	3	$p < .001$

1.6 EVIDENCE FOR RELATIONSHIPS WITH CONCEPTUALLY RELATED CONSTRUCTS

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

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

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

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.[14] The parameters of interest in measurement invariance testing are the factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that it is not necessary to demonstrate invariance across groups on these parameters. In general, measurement invariance testing can be conducted using a series of multiple-group 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 2019 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

[14] 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 ≤ 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 “b” series tables in Appendix C (e.g., tables C.1b, C.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 ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.947 to 0.989, and RMSEA ranged from 0.013 to 0.035. For mathematics, CFI values ranged from 0.943 to 0.991, and RMSEA ranged from 0.014 to 0.043.

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. 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 status, economic disadvantaged status, limited English proficiency status, and accommodation test forms.

1.8 DIFFERENTIAL MODE EFFECTS ACROSS SUBGROUPS

To explore the possibility that mode of test administration may exert differential effects across subgroups, we began by identifying matched samples of students participating online 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 or reduced-price 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 grogram. 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
ELA											
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			
Mathematics											
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.[15] 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).

[15] 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 2018 and the spring 2019 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 2019 Algebra I performance. In spring 2018, grade 8 students enrolled in Algebra I were required to participate in both assessments, but in spring 2019, 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 2018→2019	N	Spring 2018 Scale Score		Spring 2019 Scale Score		Change from 2018 to 2019		Percentage Scoring Lower		Correlation
		Mean	Std. Dev	Mean	Std. Dev	Mean	IRT based Standard Error	Expected	Observed	
ELA										
G3E→G4E	80581	2503	33.09	2524	32.15	21	14.52	0.19	0.12	0.83
G4E→G5E	84041	2520	32.77	2542	37.08	22	14.96	0.18	0.13	0.84
G5E→G6E	84141	2539	34.85	2546	32.40	7	14.78	0.38	0.34	0.84
G6E→G7E	82148	2543	32.29	2552	34.43	10	14.97	0.34	0.29	0.84
G7E→G8E	81000	2554	34.28	2560	36.00	6	14.57	0.40	0.36	0.85
G8E→G9E	59363	2561	31.82	2567	31.53	6	13.97	0.40	0.36	0.83
G9E→G10E	54169	2571	30.87	2567	31.45	-4	13.92	0.57	0.57	0.83
G10E→G11E	48461	2568	32.65	2571	32.79	3	14.42	0.44	0.41	0.82
Mathematics										
G3M→G4M	81007	3529	47.20	3557	45.22	28	17.37	0.19	0.14	0.83
G4M→G5M	84379	3556	44.22	3588	42.48	33	16.53	0.14	0.09	0.83
G5M→G6M	84393	3590	46.62	3617	44.04	28	16.59	0.17	0.12	0.85
G6M→G7M	82384	3618	46.25	3637	43.16	19	16.42	0.25	0.19	0.87
G7M→G8M	72491	3631	41.29	3656	39.78	25	15.77	0.18	0.13	0.84
G8M→Algebra I	43155	3654	35.28	3668	32.51	14	15.51	0.29	0.24	0.79
Algebra I→Geometry	50064	3678	36.71	3689	37.16	11	15.87	0.34	0.30	0.82
Geometry→Algebra II II	41750	3693	38.50	3705	38.62	13	16.29	0.32	0.28	0.81

older adults—test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Exhibit 1.9.2 Achievement Gap Trends between Spring 2018 and Spring 2019 for ELA

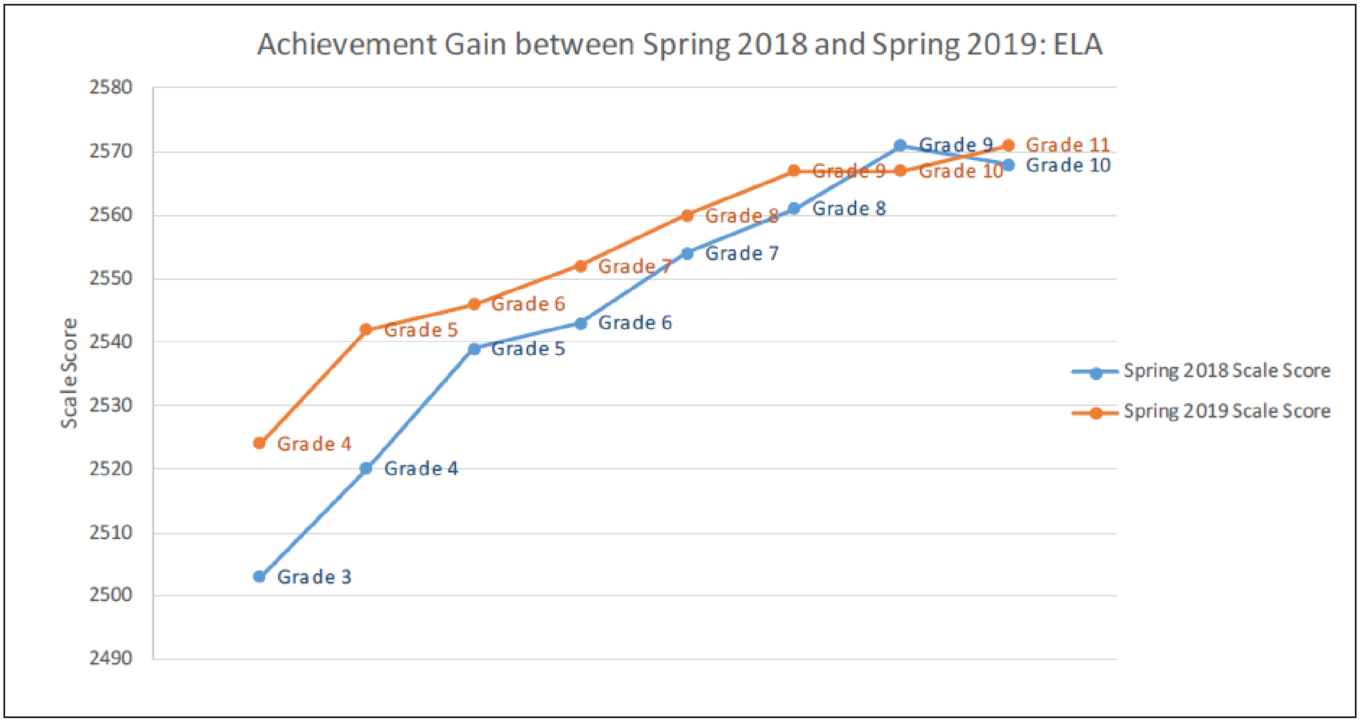
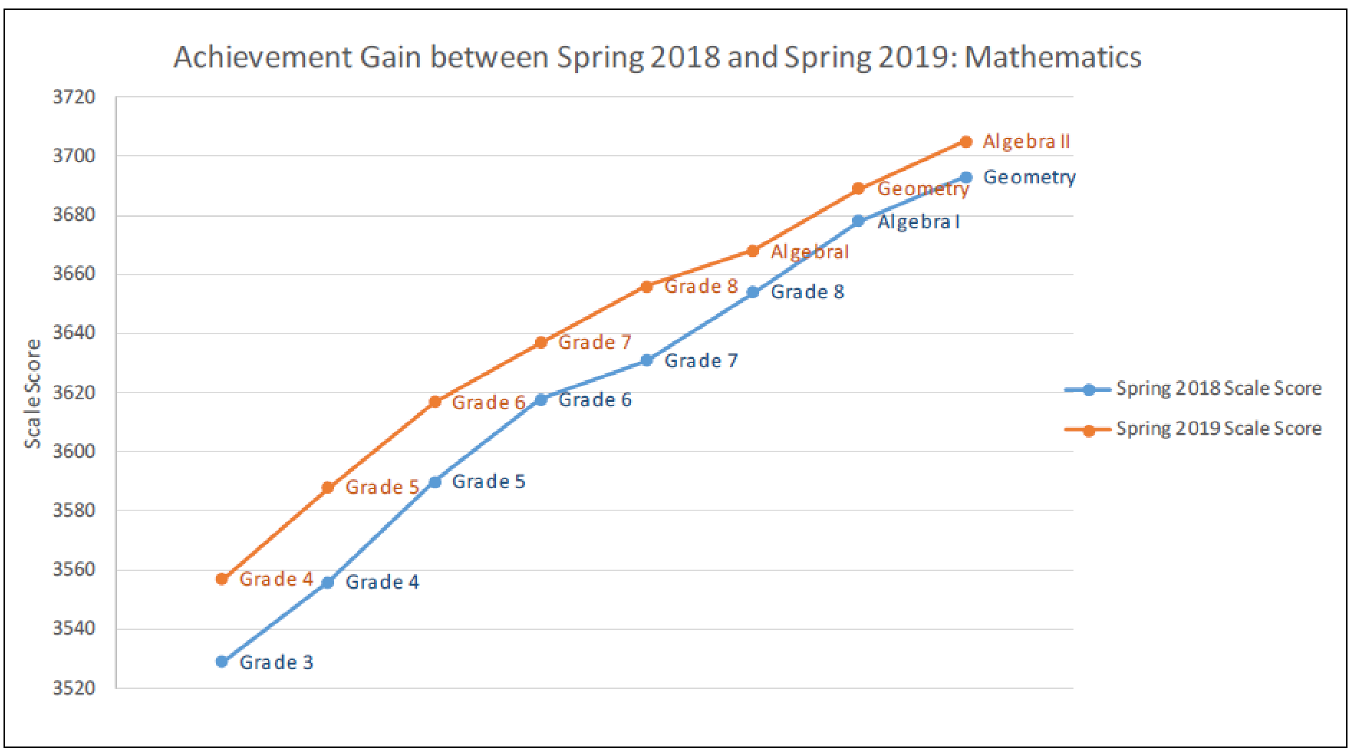


Exhibit 1.9.3 Achievement Gap Trends between Spring 2018 and Spring 2019 for Mathematics



The exhibit 1.9.2 and exhibit 1.9.3 also show 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 about three-quarters standard deviation (e.g., average gain of 33 scale score points in grade 4 mathematics is 78% of the 43-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 to one-fifth standard deviation and drop to about one- fifth 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 2019 test scores from 2018 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 1.9.3 shows the standardized regression coefficient estimates of the differential effect on student's growth rate from 2018 to 2019 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 2019 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 2018 and 2019 test scores, controlling for demographic subgroups. The overall slope parameter indicates the rate of growth in test scores between 2018 and 2019. 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 ELA and 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, Hispanic, and American Indian students showed lower gain rates than whites in mathematics assessments.

Exhibit 1.9.4.1 Standardized Regression Coefficient of Differential Growth from 2018 to 2019 Administration Across Subgroups: ELA

2018 Administration **2019 Administration**	**G3E** **G4E**	**G4E** **G5E**	**G5E** **G6E**	**G6E** **G7E**	**G7E** **G8E**	**G8E** **G9E**	**G9E** **G10E**	**G10E** **G11E**
Intercept								
Female	0.01	0.03	0.03	0.03	0.04	0.04	0.01	0.04
SPED	-0.06	-0.08	-0.06	-0.07	-0.07	-0.07	-0.04	-0.06
LEP	-0.09	-0.08	-0.06	-0.06	-0.05	-0.04	-0.02	
Low Income	-0.04	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03
Asian	0.02	0.01	0.03	0.02	0.02	0.03	0.02	0.01
Hispanic	-0.06	-0.03	-0.03	-0.02	-0.04	-0.05	-0.04	-0.04
African American	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03	-0.01	-0.02
Native Hawaiian/Pacific Islander								
American Indian	-0.05	-0.04	-0.03	-0.03	-0.04	-0.02	-0.03	-0.03
Multiple Ethnicities		-0.01	-0.01					-0.01
Slope	0.77	0.74	0.83	0.80	0.81	0.82	0.79	0.82
Female				-0.02		0.01		-0.01
SPED	0.02	0.02	-0.01		-0.02	-0.05	-0.02	-0.02
LEP	-0.05	-0.04	-0.05	-0.03	-0.04	-0.02		
Low Income	-0.01	0.01	-0.02	-0.02	-0.01	-0.02		-0.01
Asian		-0.01				0.01		0.01
Hispanic		0.02	-0.02		0.01	-0.03		-0.02
African American		0.01				-0.02		
Native Hawaiian/Pacific Islander								
American Indian		0.01	-0.02		-0.01	-0.02		-0.02
Multiple Ethnicities							0.01	

Note: Only significant effects from the multiple regression models are presented in the table. Intercept (β00): Standardized average test score in 2019 administration. Slope (β10): Rate of gain from 2018 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.

Exhibit 1.9.4.2 Standardized Regression Coefficient of Differential Growth from 2018 to 2019 Administration Across Subgroups: Mathematics

2018 Administration 2019 Administration	G3M G4M	G4M G5M	G5M G6M	G6M G7M	G7M G8M	G8M Alg I	Alg I Geo	Geo Alg II
Intercept								
Female		0.01		−0.01	0.02	0.05	−0.01	0.03
SPED	−0.05	−0.06	−0.05	−0.06	−0.05	−0.06	−0.04	−0.04
LEP	−0.05	−0.04	−0.05	−0.06	−0.02	−0.05	−0.01	
Low Income	−0.04	−0.02	−0.01	−0.02	−0.01	−0.02	−0.01	−0.02
Asian	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.02
Hispanic	−0.05	−0.02	−0.05	−0.05	−0.01	−0.02	−0.03	−0.03
African American	−0.04	−0.03	−0.03	−0.02		−0.01	−0.04	−0.02
Native Hawaiian/Pacific Islander			−0.01					
American Indian	−0.04	−0.03	−0.03	−0.03	−0.01	−0.02	−0.01	−0.03
Multiple Ethnicities			−0.01					
Slope	0.80	0.85	0.85	0.89	0.89	0.80	0.84	0.84
Female	−0.01			−0.01	−0.02			
SPED	0.01	−0.02	−0.02	−0.06	−0.06	−0.02	−0.04	−0.02
LEP	−0.03	−0.05	−0.04	−0.06	−0.04	−0.02	−0.02	
Low Income	−0.02	−0.02	−0.02	−0.02	−0.01			−0.02
Asian			0.01	0.01	0.01		0.02	
Hispanic	−0.01	−0.03	−0.02	−0.02	−0.02	−0.03	−0.04	−0.04
African American		−0.02	−0.01	−0.01	−0.01	−0.01	−0.02	−0.01
Native Hawaiian/Pacific Islander								
American Indian	−0.01	−0.02	−0.02	−0.02	−0.02	−0.01	−0.01	−0.03
Multiple Ethnicities								

Note: Only significant effects from the multiple regression models are presented in the table. Intercept (β00): Standardized average test score in 2019 administration. Slope (β10): Rate of gain from 2018 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
ELA					
Grade 4 ELA		0.02	–0.01	0.03	–0.01
Grade 5 ELA	–0.01	0.01	-0.01	0.02	
Grade 6 ELA	0.02		0.01		
Grade 7 ELA	0.01	0.03	–0.01	–0.01	0.01
Grade 8 ELA		0.02	–0.01		0.02
Grade 9 ELA		0.01	–0.06	0.02	0.01
Grade 10 ELA	–0.02		–0.08	0.03	0.01
Grade 11 ELA	–0.03		–0.08	0.05	0.01
Mathematics					
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	1667	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

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

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

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

β_j effect of item j,

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

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

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

The term β_j is the fixed effect controlling the differences in difficulty across items. The term u_i is a random effect capturing the difference in achievement across students. The coefficient αs 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 the a = 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
Coefficient Estimates									
EL mean of random intercept	-0.98	-0.59	-0.69	-0.64	-0.68	-0.67	-0.66	-0.64	-0.56
ENG main effect	-0.04	-0.02	-0.01	0.00	-0.01	0.00	-0.01	0.00	0.00
ENG SP main effect	-0.03	-0.03	-0.01	-0.01	-0.01	0.00	0.00	0.01	0.00
EL by ENG interaction	0.10	0.05	0.08	0.10	0.10	0.11	0.16	0.10	0.21
EL BY ENG SP interaction	0.04	0.08	0.09	0.08	0.08	0.11	0.10	0.11	0.19
ENG effect (main + interaction)	0.05	0.03	0.07	0.10	0.09	0.11	0.15	0.10	0.21
ENG SP effect (main + interaction)	0.01	0.05	0.08	0.06	0.07	0.12	0.10	0.11	0.20
Standard Errors									
EL mean of random intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
ENG main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ENG SP main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
EL by ENG interaction	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
EL BY ENG SP interaction	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04
ENG effect (main + interaction)	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05
ENG SP effect (main + interaction)	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04
Z Statistics									
EL mean of random intercept	-179.59	-107.86	-117.29	-85.30	-85.37	-74.61	-72.90	-56.74	-33.35
ENG main effect	-6.86	-3.43	-1.26	-0.04	-1.69	-0.11	-2.06	0.32	-0.66
ENG SP main effect	-4.89	-5.30	-1.30	-2.08	-1.82	0.62	0.34	0.83	0.44
EL by ENG interaction	6.76	3.95	4.76	5.62	5.50	5.42	6.02	2.88	4.61
EL BY ENG SP interaction	2.79	5.97	5.67	4.27	4.88	5.67	3.68	3.26	4.61
ENG effect (main + interaction)	3.70	2.43	4.28	5.62	4.96	5.40	5.54	2.94	4.51
ENG SP effect (main + interaction)	0.64	3.61	5.17	3.58	4.27	5.86	3.76	3.43	4.68

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

Effect	G3M	G4M	G5M	G6M	G7M	G8M	Algebra I	Geometry	Algebra II
Coefficient Estimates									
EL mean of random intercept	−0.83	−0.79	−0.86	−0.82	−0.83	−0.60	−0.70	−0.67	−0.44
ENG main effect	0.00	−0.01	0.00	0.00	0.01	0.01	0.01	0.03	−0.02
ENG SP main effect	−0.01	−0.01	−0.01	0.00	0.01	−0.01	0.01	0.02	−0.02
EL by ENG interaction	0.11	0.05	0.01	0.09	0.09	0.18	0.42	0.21	−0.04
EL BY ENG SP interaction	0.11	0.14	0.04	0.06	0.12	0.17	0.48	0.06	0.13
ENG effect (main + interaction)	0.12	0.04	0.01	0.08	0.10	0.19	0.43	0.24	−0.07
ENG SP effect (main + interaction)	0.10	0.12	0.03	0.06	0.13	0.16	0.48	0.08	0.11
Standard Errors									
EL mean of random intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
ENG main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ENG SP main effect	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
EL by ENG interaction	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.10
EL BY ENG SP interaction	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.09
ENG effect (main + interaction)	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.10
ENG SP effect (main + interaction)	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.07	0.09
Z Statistics									
EL mean of random intercept	−85.51	−84.31	−82.73	−70.90	−70.91	−53.80	−62.32	−37.45	−21.00
ENG main effect	0.50	−1.00	0.00	−0.29	0.62	1.20	0.88	2.29	−1.56
ENG SP main effect	−0.82	−1.27	−0.77	0.30	0.63	−0.81	0.74	1.17	−1.12
EL by ENG interaction	5.58	2.31	0.31	2.66	2.87	5.28	8.25	2.93	−0.42
EL BY ENG SP interaction	5.33	5.99	1.41	1.90	3.84	5.01	9.67	0.87	1.41
ENG effect (main + interaction)	5.82	1.91	0.31	2.58	3.06	5.65	8.45	3.36	−0.64
ENG SP effect (main + interaction)	5.01	5.48	1.13	1.99	4.04	4.77	9.85	1.09	1.24

1.12 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score 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[16]:

- 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

[16] Standard 7.1: The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When 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 handscored, 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 handscored (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 handscored. In addition, handscoring 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 2018 AND FALL 2018 OPERATIONAL TEST ADMINISTRATION

The following tests were administered in summer and fall 2018:

- 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 2018 administration of Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) occurred from June 4 to August 2, 2018, and the fall 2018 administration occurred from November 5 to November 30, 2018.

The scoring and reporting of the summer and fall 2018 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 2018 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 2018 and Exhibits 3.1.3 and 3.1.4 for fall 2018.[17]

Exhibit 3.1.1 Number of Students Participating in ELA Assessments by Subgroups: Summer 2018

Group	ELA 9	ELA 10	ELA 11
All Students	952	514	300
Female	425	238	137
Male	527	276	163
African American	53	37	23
Asian	16	8	4
Native Hawaiian/Pacific Islander	1	2	0
Hispanic/Latino	651	304	139
American Indian or Alaskan	46	40	40
White	177	117	89
Multiple Ethnicities	8	6	5
Limited English Proficiency	41	16	3
Special Education	98	42	25
Free or Reduced-Price Lunch	516	199	101

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

Exhibit 3.1.2 Number of Students Participating in Mathematics Assessments by Subgroups: Summer 2018

Group	Algebra I	Geometry	Algebra II
All Students	1321	1167	776
Female	576	599	388
Male	745	568	388
African American	113	99	45
Asian	21	38	18
Native Hawaiian/Pacific Islander	6	4	0
Hispanic/Latino	762	644	505
American Indian or Alaskan	67	48	20
White	333	309	175
Multiple Ethnicities	19	25	13
Limited English Proficiency	135	67	19
Special Education	96	57	90
Free or Reduced-Price Lunch	342	267	206

Exhibit 3.1.3 Number of Students Participating in ELA Assessments by Subgroups: Fall 2018

Group	ELA 9	ELA 10	ELA 11
All Students	3703	4598	4688
Female	1639	2187	2268
Male	2064	2411	2420
African American	202	224	256
Asian	73	86	90
Native Hawaiian/Pacific Islander	13	18	20
Hispanic/Latino	1755	2058	2139
American Indian or Alaskan	211	256	311
White	1305	1774	1705
Multiple Ethnicities	144	182	167
Limited English Proficiency	143	122	110
Special Education	274	414	368
Free or Reduced-Price Lunch	1220	1420	1410

Exhibit 3.1.4 Number of Students Participating in Mathematics Assessments by Subgroups: Fall 2018

Group	Algebra I	Geometry	Algebra II
All Students	4990	5632	4476
Female	2355	2728	2247
Male	2635	2904	2229
African American	282	311	217
Asian	106	103	101
Native Hawaiian/Pacific Islander	23	23	15

Group	Algebra I	Geometry	Algebra II
Hispanic/Latino	2217	2472	1854
American Indian or Alaskan	301	269	316
White	1834	2248	1855
Multiple Ethnicities	227	206	118
Limited English Proficiency	359	253	152
Special Education	300	389	233
Free or Reduced-Price Lunch	1760	1661	1352

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 2018 and in Exhibit 3.2.2 for fall 2018.

Exhibit 3.2.1 Test Score Summary Statistics: Summer 2018

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
ELA					
9	952	2550	27.45	2664	2485
10	514	2547	30.07	2641	2479
11	300	2552	30.35	2647	2465
Mathematics					
Algebra I	1321	3656	27.56	3787	3579
Geometry	1167	3678	35.37	3798	3609
Algebra II	776	3690	31.39	3828	3629

Exhibit 3.2.2 Test Score Summary Statistics: Fall 2018

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
ELA					
9	3703	2561	31.67	2664	2455
10	4598	2558	33.76	2668	2479
11	4688	2558	27.91	2656	2484
Mathematics					
Algebra I	4990	3670	38.6	3787	3577
Geometry	5632	3678	34.57	3819	3609
Algebra II	4476	3695	36.19	3839	3629

The percentages 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 2018 and in Exhibit 3.2.4 for fall 2018.

Exhibit 3.2.3 Percentage of Students in Performance Levels: Summer 2018

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
9	952	60	23	13	4	17
10	514	76	8	10	5	16
11	300	72	12	12	4	16
Mathematics						
Algebra I	1321	65	20	13	3	15
Geometry	1167	50	23	21	6	27
Algebra II	776	52	23	21	4	25

Exhibit 3.2.4 Percentage of Students in Performance Levels: Fall 2018

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
9	3703	43	23	25	9	34
10	4598	60	13	17	10	27
11	4688	64	17	15	4	19
Mathematics						
Algebra I	4990	50	15	22	13	35
Geometry	5632	48	25	22	5	27
Algebra II	4476	49	19	22	9	32

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 eligible for free or reduced-price lunch (FRL)—for summer 2018. Exhibits 3.3.3 and 3.3.4 show the same information for fall 2018.

Exhibit 3.3.1 Number of Students at Each Performance Level by Subgroups: Summer 2018

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	567	240	327	34	4	1	410	35	79	4	37	86	322
	Partially Proficient	222	107	115	11	7	0	150	7	46	1	3	10	120
	Proficient	127	62	65	8	4	0	78	3	32	2	1	2	61
	Highly Proficient	36	16	20	0	1	0	13	1	20	1	0	0	13
10	Minimally Proficient	391	176	215	28	3	2	261	36	57	4	15	39	165
	Partially Proficient	43	22	21	4	2	0	20	4	12	1	1	1	19
	Proficient	53	26	27	5	2	0	16	0	30	0	0	2	10
	Highly Proficient	27	14	13	0	1	0	7	0	18	1	0	0	5
11	Minimally Proficient	217	95	122	19	2	0	105	36	51	4	3	23	75
	Partially Proficient	35	15	20	2	0	0	16	2	15	0	0	2	15
	Proficient	35	21	14	2	1	0	15	2	14	1	0	0	7
	Highly Proficient	13	6	7	0	1	0	3	0	9	0	0	0	4
Algebra I	Minimally Proficient	858	365	493	86	12	6	541	47	154	12	113	72	272
	Partially Proficient	261	123	138	18	3	0	140	12	85	3	16	17	41
	Proficient	166	72	94	8	3	0	74	8	72	1	6	7	27
	Highly Proficient	36	16	20	1	3	0	7	0	22	3	0	0	2
Geometry	Minimally Proficient	581	292	289	52	9	2	356	29	123	10	41	41	196
	Partially Proficient	272	131	141	23	8	1	148	13	75	4	18	12	51
	Proficient	242	136	106	17	15	1	109	6	85	9	5	3	17
	Highly Proficient	72	40	32	7	6	0	31	0	26	2	3	1	3
Algebra II	Minimally Proficient	405	192	213	23	3	0	289	9	76	5	13	31	114
	Partially Proficient	176	93	83	13	3	0	106	9	42	3	3	13	33
	Proficient	161	84	77	8	6	0	98	2	43	4	2	37	49
	Highly Proficient	34	19	15	1	6	0	12	0	14	1	1	9	10

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

Exhibit 3.3.2 Percentage of Students at Each Performance Level by Subgroups: Summer 2018

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	56	62	64	25	100	63	76	45	50	90	88	62
	Partially Proficient	23	25	22	21	44	0	23	15	26	13	7	10	23
	Proficient	13	15	12	15	25	0	12	7	18	25	2	2	12
	Highly Proficient	4	4	4	0	6	0	2	2	11	13	0	0	3
	At or Above Proficient	17	18	16	15	31	0	14	9	29	38	2	2	14
10	Minimally Proficient	76	74	78	76	38	100	86	90	49	67	94	93	83
	Partially Proficient	8	9	8	11	25	0	7	10	10	17	6	2	10
	Proficient	10	11	10	14	25	0	5	0	26	0	0	5	5
	Highly Proficient	5	6	5	0	13	0	2	0	15	17	0	0	3
	At or Above Proficient	16	17	14	14	38	0	8	0	41	17	0	5	8
11	Minimally Proficient	72	69	75	83	50	0	76	90	57	80	100	92	74
	Partially Proficient	12	11	12	9	0	0	12	5	17	0	0	8	15
	Proficient	12	15	9	9	25	0	11	5	16	20	0	0	7
	Highly Proficient	4	4	4	0	25	0	2	0	10	0	0	0	4
	At or Above Proficient	16	20	13	9	50	0	13	5	26	20	0	0	11
Algebra I	Minimally Proficient	65	63	66	76	57	100	71	70	46	63	84	75	80
	Partially Proficient	20	21	19	16	14	0	18	18	26	16	12	18	12
	Proficient	13	13	13	7	14	0	10	12	22	5	4	7	8
	Highly Proficient	3	3	3	1	14	0	1	0	7	16	0	0	1
	At or Above Proficient	15	15	15	8	29	0	11	12	28	21	4	7	8
Geometry	Minimally Proficient	50	49	51	53	24	50	55	60	40	40	61	72	73
	Partially Proficient	23	22	25	23	21	25	23	27	24	16	27	21	19
	Proficient	21	23	19	17	39	25	17	13	28	36	7	5	6
	Highly Proficient	6	7	6	7	16	0	5	0	8	8	4	2	1
	At or Above Proficient	27	29	24	24	55	25	22	13	36	44	12	7	7
Algebra II	Minimally Proficient	52	49	55	51	17	0	57	45	43	38	68	34	55
	Partially Proficient	23	24	21	29	17	0	21	45	24	23	16	14	16
	Proficient	21	22	20	18	33	0	19	10	25	31	11	41	24
	Highly Proficient	4	5	4	2	33	0	2	0	8	8	5	10	5
	At or Above Proficient	25	27	24	20	67	0	22	10	33	38	16	51	29

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

Exhibit 3.3.3 Number of Students at Each Performance Level by Subgroups: Fall 2018

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	1592	590	1002	101	17	6	953	121	350	44	106	176	649
	Partially Proficient	859	416	443	58	9	1	399	61	292	39	28	37	298
	Proficient	909	451	458	40	25	5	317	26	454	42	7	40	230
	Highly Proficient	343	182	161	3	22	1	86	3	209	19	2	21	43
10	Minimally Proficient	2750	1234	1516	149	40	13	1444	191	811	102	113	327	972
	Partially Proficient	600	318	282	32	12	2	260	31	234	29	6	29	194
	Proficient	775	390	385	31	20	1	255	24	410	34	3	36	177
	Highly Proficient	473	245	228	12	14	2	99	10	319	17	0	22	77
11	Minimally Proficient	3002	1369	1633	182	37	13	1570	252	858	90	102	333	1003
	Partially Proficient	787	417	370	41	27	4	313	41	324	37	6	23	221
	Proficient	707	388	319	26	16	2	216	17	398	32	2	12	159
	Highly Proficient	192	94	98	7	10	1	40	1	125	8	0	0	27
Algebra I	Minimally Proficient	2501	1123	1378	187	20	12	1385	175	625	97	295	246	1051
	Partially Proficient	758	368	390	38	12	8	331	55	269	45	42	23	293
	Proficient	1099	576	523	40	36	0	361	51	562	49	18	22	305
	Highly Proficient	632	288	344	17	38	3	140	20	378	36	4	9	111
Geometry	Minimally Proficient	2702	1312	1390	196	34	11	1425	153	791	92	193	256	920
	Partially Proficient	1419	699	720	66	27	9	562	53	647	55	43	67	451
	Proficient	1256	619	637	42	31	3	414	54	668	44	13	58	260
	Highly Proficient	255	98	157	7	11	0	71	9	142	15	4	8	30
Algebra II	Minimally Proficient	2212	1096	1116	147	16	6	1122	220	644	57	113	162	772
	Partially Proficient	840	444	396	31	21	4	333	64	360	27	25	30	280
	Proficient	1007	506	501	34	35	2	316	26	567	27	14	28	222
	Highly Proficient	417	201	216	5	29	3	83	6	284	7	0	13	78

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

Exhibit 3.3.4 Percentage of Students at Each Performance Level by Subgroups: Fall 2018

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	LEP	SPED	FRL
9	Minimally Proficient	43	36	49	50	23	46	54	57	27	31	74	64	53
	Partially Proficient	23	25	21	29	12	8	23	29	22	27	20	14	24
	Proficient	25	28	22	20	34	38	18	12	35	29	5	15	19
	Highly Proficient	9	11	8	1	30	8	5	1	16	13	1	8	4
	At or Above Proficient	34	39	30	21	64	46	23	14	51	42	6	22	22
10	Minimally Proficient	60	56	63	67	47	72	70	75	46	56	93	79	68
	Partially Proficient	13	15	12	14	14	11	13	12	13	16	5	7	14
	Proficient	17	18	16	14	23	6	12	9	23	19	2	9	12
	Highly Proficient	10	11	9	5	16	11	5	4	18	9	0	5	5
	At or Above Proficient	27	29	25	19	40	17	17	13	41	28	2	14	18
11	Minimally Proficient	64	60	67	71	41	65	73	81	50	54	93	90	71
	Partially Proficient	17	18	15	16	30	20	15	13	19	22	5	6	16
	Proficient	15	17	13	10	18	10	10	5	23	19	2	3	11
	Highly Proficient	4	4	4	3	11	5	2	0	7	5	0	0	2
	At or Above Proficient	19	21	17	13	29	15	12	6	31	24	2	3	13
Algebra I	Minimally Proficient	50	48	52	66	19	52	62	58	34	43	82	82	60
	Partially Proficient	15	16	15	13	11	35	15	18	15	20	12	8	17
	Proficient	22	24	20	14	34	0	16	17	31	22	5	7	17
	Highly Proficient	13	12	13	6	36	13	6	7	21	16	1	3	6
	At or Above Proficient	35	37	33	20	70	13	23	24	51	37	6	10	24
Geometry	Minimally Proficient	48	48	48	63	33	48	58	57	35	45	76	66	55
	Partially Proficient	25	26	25	21	26	39	23	20	29	27	17	17	27
	Proficient	22	23	22	14	30	13	17	20	30	21	5	15	16
	Highly Proficient	5	4	5	2	11	0	3	3	6	7	2	2	2
	At or Above Proficient	27	26	27	16	41	13	20	23	36	29	7	17	17
Algebra II	Minimally Proficient	49	49	50	68	16	40	61	70	35	48	74	70	57
	Partially Proficient	19	20	18	14	21	27	18	20	19	23	16	13	21
	Proficient	22	23	22	16	35	13	17	8	31	23	9	12	16
	Highly Proficient	9	9	10	2	29	20	4	2	15	6	0	6	6
	At or Above Proficient	32	31	32	18	63	33	22	10	46	29	9	18	22

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free or Reduced-Price 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. Test score reliability is traditionally estimated using both classical and IRT approaches. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the IRT framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the test information function represents the standard error of measurement. The standard error of measurement is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale that is at the middle of the test distribution and greater on scaled values farther away from the middle.

The reliability evidence of the AZMERIT test scores is provided with reliability, SEM, and classification accuracy and consistency in each achievement level. [18]

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. Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale, for all students. The marginal reliability coefficients are nearly identical or close to coefficient alpha. For our analysis, the marginal reliability coefficients were computed using operational items.

The marginal reliability ($\bar{\rho}$) is defined as

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i^2$ is the conditional standard error of measurement of the scale score for student i; and $\sigma^2$is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Exhibit 3.4.1.1 and Exhibit 3.4.1.2 shows presents the marginal reliability coefficients for all students. The reliability coefficients for all subjects and grades range from 0.84 to 0.89 for summer 2018 administrations and from 0.87 to 0.91 for fall 2018 administrations.

[18] Standard 2.2: The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures and with the intended interpretations for use of the test scores. Standard 2.3: For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

Exhibit 3.4.1.1 Overall Reliabilities by Subject/Test for AzMERIT Scores: Summer 2018

Grade/Course	ELA		Mathematics	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.87	754	0.84	760
10/Geometry	0.89	904	0.88	1251
11/Algebra II	0.89	921	0.85	985

Note: Reliability ranges from 0 to 1.0 variance is in scale-score metric.

Exhibit 3.4.1.2 Overall Reliabilities by Subject/Test for AzMERIT Scores: Fall 2018

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.90	1003	0.91	1490
10/Geometry	0.91	1140	0.88	1195
11/Algebra II	0.87	779	0.88	1310

Note: Reliability ranges from 0 to 1.0 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.

For the AzMERIT assessments scored using MLE, according to Masters (1982), the asymptotic estimate of the standard error for ability θ is given by

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

which is further placed onto the reporting scale by the following transformation:

$$SE_{vs} = a \times SE(\theta),$$

where a is the slope of the scaling constants that take θ to the reporting scale. For both ELA and Mathematics tests, $a = 30$.

Exhibit 3.4.2.1 shows the conditional standard errors of measurement (CSEMs) for the AzMERIT ELA and mathematics assessments, with respect to the four AzMERIT performance standards for summer 2018 and Exhibit 3.4.2.2 for fall 2018. These tables also include associated CSEM around cut scores. 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.[19] Test scores near the tails of the ability distribution are somewhat less precise, as expected. While these

[19] 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.

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 3.4.2.3 through Exhibit 3.4.2.14 present the CSEMs and corresponding performance levels for each scale score for summer 2018 and fall 2018, respectively.

Exhibit 3.4.2.1 Performance Level and Associated CSEMs: Summer 2018

Grade	CSEM	Proficiency Level				Overall
		Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	
ELA						
Grade 9 ELA	Mean	10	9	10	13	10
	Around Cut Score		9	9	11	
Grade 10 ELA	Mean	10	9	10	12	10
	Around Cut Score		9	10	11	
Grade 11 ELA	Mean	10	9	10	12	10
	Around Cut Score		9	10	11	
Mathematics						
Algebra I	Mean	11	10	10	13	11
	Around Cut Score		10	10	11	
Geometry	Mean	13	11	10	12	12
	Around Cut Score		11	10	11	
Algebra II	Mean	14	11	10	11	12
	Around Cut Score		11	10	10	

Exhibit 3.4.2.2 Performance Level and Associated CSEMs: Fall 2018

Grade	CSEM	Proficiency Level				Overall
		Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	
ELA						
Grade 9 ELA	Mean	10	9	10	13	10
	Around Cut Score		9	9	11	
Grade 10 ELA	Mean	10	9	10	12	10
	Around Cut Score		9	10	11	
Grade 11 ELA	Mean	10	9	10	12	10
	Around Cut Score		9	10	11	
Mathematics						
Algebra I	Mean	12	10	10	13	11
	Around Cut Score		10	10	11	
Geometry	Mean	14	10	10	13	12
	Around Cut Score		11	10	11	
Algebra II	Mean	14	11	10	11	12
	Around Cut Score		11	10	10	

Exhibit 3.4.2.3 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Grade 9 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2485	15	2555	9	2577	9	2606	11
2492	13	2558	9	2579	9	2610	11
2498	13	2560	9	2582	9	2614	12
2503	12	2563	9	2585	9	2619	12
2507	12	2566	9	2588	10	2624	13
2512	11	2568	9	2591	10	2630	14
2516	11	2571	9	2595	10	2636	15
2519	10	2574	9	2598	10	2644	16
2523	10			2602	11	2664	20
2526	10						
2529	10						
2533	10						
2536	9						
2538	9						
2541	9						
2544	9						
2547	9						
2550	9						
2552	9						

Exhibit 3.4.2.4 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Grade 10 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2479	16	2567	9	2582	10	2606	11
2487	15	2570	9	2585	10	2609	11
2493	14	2573	9	2588	10	2613	11
2499	13	2576	9	2591	10	2618	12
2504	12	2579	9	2594	10	2623	12
2509	12			2598	10	2628	13
2513	11			2601	11	2634	14
2517	11					2641	15
2521	11						
2525	10						
2528	10						
2532	10						
2535	10						
2538	10						
2541	10						
2544	9						
2547	9						
2550	9						
2553	9						
2556	9						
2559	9						
2561	9						
2564	9						

Exhibit 3.4.2.5 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Grade 11 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2465	19	2570	9	2585	10	2614	11
2484	15	2573	9	2588	10	2618	12
2491	14	2576	9	2592	10	2623	12
2497	13	2579	10	2595	10	2634	13
2507	12	2582	10	2598	10	2640	14
2512	11			2602	11	2647	15
2516	11			2606	11		
2520	11			2610	11		
2523	10						
2527	10						
2531	10						
2534	10						
2537	10						
2540	10						
2543	10						
2546	10						
2549	9						
2552	9						
2555	9						
2558	9						
2561	9						
2564	9						
2567	9						

Exhibit 3.4.2.6 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Algebra I

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3579	22	3663	10	3682	10	3722	11
3593	18	3666	10	3685	10	3726	11
3602	16	3669	10	3688	10	3730	11
3610	15	3672	10	3691	10	3735	12
3617	14	3676	10	3694	10	3740	12
3623	13	3679	10	3697	10	3745	13
3628	12			3701	10	3751	14
3633	12			3704	10	3758	15
3637	11			3707	10	3766	16
3641	11			3711	10	3775	18
3645	11			3714	10	3787	21
3649	11			3718	11		
3653	10						
3656	10						
3659	10						

Exhibit 3.4.2.7 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Geometry

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3609	21	3673	11	3698	10	3745	11
3619	18	3677	11	3701	10	3749	11
3629	16	3681	11	3705	10	3753	11
3637	15	3684	10	3708	10	3757	11
3644	14	3688	10	3711	10	3761	12
3650	13	3691	10	3714	10	3766	12
3655	12	3695	10	3718	10	3771	13
3660	12			3721	10	3777	13
3665	12			3724	10	3783	14
3669	11			3727	10	3789	15
				3731	10	3798	16
				3734	10		
				3738	10		
				3741	10		

Exhibit 3.4.2.8 Conditional Standard Error of Measurement (CSEM) at Scale Score: Summer 2018 – Algebra II

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3629	20	3691	11	3711	10	3751	10
3636	19	3696	11	3714	10	3754	10
3646	16	3699	11	3717	10	3757	10
3654	15	3703	11	3721	10	3760	10
3661	14	3707	10	3724	10	3764	10
3667	13			3727	10	3767	10
3673	13			3731	10	3771	11
3678	12			3734	10	3775	11
3683	12			3737	10	3779	11
3687	11			3740	10	3783	11
				3744	10	3788	12
				3747	10	3798	13
						3828	18

Exhibit 3.4.2.9 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Grade 9 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2455	22	2555	9	2577	9	2606	11
2478	16	2558	9	2579	9	2610	11
2485	15	2560	9	2582	9	2614	12
2492	13	2563	9	2585	9	2619	12
2498	13	2566	9	2588	10	2624	13
2503	12	2568	9	2591	10	2630	14
2507	12	2571	9	2595	10	2636	15
2512	11	2574	9	2598	10	2644	16
2516	11			2602	11	2653	17
2519	10					2664	20
2523	10						
2526	10						
2529	10						
2533	10						
2536	9						
2538	9						
2541	9						
2544	9						
2547	9						
2550	9						
2552	9						

Exhibit 3.4.2.10 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Grade 10 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2479	16	2567	9	2582	10	2606	11
2487	15	2570	9	2585	10	2609	11
2493	14	2573	9	2588	10	2613	11
2499	13	2576	9	2591	10	2618	12
2504	12	2579	9	2594	10	2623	12
2509	12			2598	10	2628	13
2513	11			2601	11	2634	14
2517	11					2641	15
2521	11					2649	16
2525	10					2659	18
2528	10					2668	20
2532	10						
2535	10						
2538	10						
2541	10						
2544	9						
2547	9						
2550	9						
2553	9						
2556	9						
2559	9						
2561	9						
2564	9						

Exhibit 3.4.2.11 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Grade 11 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2484	15	2570	9	2585	10	2610	11
2491	14	2573	9	2588	10	2614	11
2497	13	2576	9	2592	10	2618	12
2502	12	2579	10	2595	10	2623	12
2507	12	2582	10	2598	10	2628	13
2512	11			2602	11	2634	13
2516	11			2606	11	2640	14
2520	11					2647	15
2523	10					2656	17
2527	10						
2531	10						
2534	10						
2537	10						
2540	10						
2543	10						
2546	10						
2549	9						
2552	9						
2555	9						
2558	9						
2561	9						
2564	9						
2567	9						

Exhibit 3.4.2.12 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Algebra I

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3577	23	3663	10	3682	10	3722	11
3579	22	3666	10	3685	10	3726	11
3593	18	3669	10	3688	10	3730	11
3602	16	3672	10	3691	10	3735	12
3610	15	3676	10	3694	10	3740	12
3617	14	3679	10	3697	10	3745	13
3623	13			3701	10	3751	14
3628	12			3704	10	3758	15
3633	12			3707	10	3766	16
3637	11			3711	10	3775	18
3641	11			3714	10	3787	21
3645	11			3718	11		
3649	11						
3653	10						
3656	10						
3659	10						

Exhibit 3.4.2.13 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Geometry

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3609	21	3673	11	3698	10	3745	11
3619	18	3677	11	3701	10	3749	11
3629	16	3681	11	3705	10	3753	11
3637	15	3684	10	3708	10	3757	11
3644	14	3688	10	3711	10	3761	12
3650	13	3691	10	3714	10	3766	12
3655	12	3695	10	3718	10	3771	13
3660	12			3721	10	3777	13
3665	12			3724	10	3783	14
3669	11			3727	10	3789	15
				3731	10	3798	16
				3734	10	3808	18
				3738	10	3819	22
				3741	10		

Exhibit 3.4.2.14 Conditional Standard Error of Measurement (CSEM) at Scale Score: Fall 2018 – Algebra II

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3629	20	3691	11	3711	10	3751	10
3636	19	3696	11	3714	10	3754	10
3646	16	3699	11	3717	10	3757	10
3654	15	3703	11	3721	10	3760	10
3661	14	3707	10	3724	10	3764	10
3667	13			3727	10	3767	10
3673	13			3731	10	3771	11
3678	12			3734	10	3775	11
3683	12			3737	10	3779	11
3687	11			3740	10	3783	11
				3744	10	3788	12
				3747	10	3793	12
						3798	13
						3804	14
						3810	15
						3818	16
						3828	18
						3839	21

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).[20] 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

[20] 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.

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.

For a student with estimated ability $\hat{\theta}$ and associated standard error $\text{se}(\hat{\theta})$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $\text{se}(\hat{\theta})$, that is, $\hat{\theta} \sim N\left(\theta, \text{se}(\hat{\theta})^2\right)$. The probability of the true score at or above the cut score θ_c is estimated as

$$P(\theta \geq \theta_c) = P\left(\frac{\theta - \hat{\theta}}{\text{se}(\hat{\theta})} \geq \frac{\theta_c - \hat{\theta}}{\text{se}(\hat{\theta})}\right) = P\left(\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} < \frac{\hat{\theta} - \theta_c}{\text{se}(\hat{\theta})}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{\text{se}(\hat{\theta})}\right),$$

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

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

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 θ, given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated 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.

The probability of a student with true ability θ being classified at or above the cut score θ_c, given the student's item scores $\boldsymbol{x} = (x_1, \cdots, x_N)$, can be estimated as

$$P(\theta \geq \theta_c | \boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x}) d\theta},$$

where the likelihood function is

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

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:

$$P(\theta < \theta_c|\boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}$$

Mathematically, we have

$$N_{11} = \sum\nolimits_{i \in N_1} P(\theta_i \geq \theta_c|\boldsymbol{x}),$$

$$N_{01} = \sum\nolimits_{i \in N_1} P(\theta_i < \theta_c|\boldsymbol{x}),$$

$$N_{10} = \sum\nolimits_{i \in N_0} P(\theta_i \geq \theta_c|\boldsymbol{x}) \text{, and}$$

$$N_{00} = \sum\nolimits_{i \in N_0} P(\theta_i < \theta_c|\boldsymbol{x}),$$

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:

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

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

To estimate the consistency, we assume the students are tested twice independently; hence, the probability of the student being classified as at or above the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \ge \theta_c, \theta_2 \ge \theta_c) = P(\theta_1 \ge \theta_c)P(\theta_2 \ge \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

Similarly, the probability of consistency for at or above the cut score is estimated as

$$P(\theta_1 \ge \theta_c, \theta_2 \ge \theta_c|\boldsymbol{x}) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of consistency for below the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of inconsistency is estimated as

$$P(\theta_1 \ge \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta \int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}, \text{ and}$$

$$P(\theta_1 < \theta_c, \theta_2 \ge \theta_c|\boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta \int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}.$$

The consistent index is computed as $\dfrac{N_{11} + N_{00}}{N}$,

$$N_{11} = \sum_{i \in N} P(\theta_{i,1} \ge \theta_c, \theta_{i,2} \ge \theta_c|\boldsymbol{x}),$$

$$N_{01} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} \ge \theta_c|\boldsymbol{x}),$$

$$N_{10} = \sum_{i \in N} P(\theta_i \ge \theta_c, \theta_{i,2} < \theta_c|\boldsymbol{x}),$$

$$N_{00} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} < \theta_c|\boldsymbol{x}), \text{ and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

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 2018 administration of AzMERIT, while Exhibit 3.4.6.2 does the same for the fall 2018 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 2018

Grade	Accuracy			Consistency		
	Partially Proficient	**Proficient**	**Highly Proficient**	**Partially Proficient**	**Proficient**	**Highly Proficient**
ELA						
9	0.91	0.95	0.98	0.87	0.93	0.97
10	0.95	0.96	0.98	0.92	0.94	0.97
11	0.94	0.95	0.97	0.91	0.93	0.96
Mathematics						
Algebra I	0.89	0.95	0.99	0.84	0.92	0.98
Geometry	0.90	0.94	0.98	0.86	0.92	0.97
Algebra II	0.89	0.93	0.98	0.85	0.89	0.97

Exhibit 3.4.6.2 Classification Accuracy and Consistency Indexes for Performance Standards: Fall 2018

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
9	0.92	0.92	0.96	0.89	0.89	0.94
10	0.93	0.94	0.96	0.90	0.91	0.94
11	0.91	0.93	0.97	0.87	0.90	0.96
Mathematics						
Algebra I	0.92	0.95	0.96	0.89	0.93	0.94
Geometry	0.90	0.93	0.98	0.87	0.91	0.97
Algebra II	0.90	0.93	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 marginal 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], special groups [limited English proficiency students], students with individualized education plans [IEPs], special education students [SPED], and students eligible for free or reduced-price lunch [FRL]) for summer 2018; and Exhibits 3.4.7.3 and 3.4.7.4 show this data for fall 2018.[21] Each racial and/or ethnic group was composed of approximately equal numbers of males and females. As the exhibits indicate, reliabilities are consistent across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups. Where reliability estimates are attenuated, there is an associated decrease in variance within the subgroup population, indicating that the decrease in reliability is likely due to a restriction in range.

Exhibit 3.4.7.1 Reliability by Subgroup: ELA Summer 2018

Subgroup	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.87	754	0.89	904	0.89	921
Female	0.87	700	0.88	883	0.89	930
Male	0.88	787	0.88	913	0.89	908
African American	0.86	669	0.85	709	0.85	683
Asian	0.90	958	NA	NA	NA	NA
Native Hawaiian/Pacific Islander*	NA	NA	0.81	481	NA	NA
Hispanic/Latino	0.86	648	0.84	654	0.87	802
American Indian or Alaskan	0.82	510	0.74	393	0.79	510
White	0.90	1012	0.90	1075	0.90	1055
Multiple Ethnicities	NA	NA	NA	NA	NA	NA
Limited English Proficiency	0.67	319	0.68	406	0.28	181
Special Education	0.73	389	0.77	505	0.73	418
Free or Reduced-Price Lunch	0.85	624	0.85	711	0.87	808

*The Native Hawaiian subgroup is not reported due to small sample size (sample size <11).

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

Exhibit 3.4.7.2 Reliability by Subgroup: Mathematics Summer 2018

Subgroup	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.84	760	0.88	1251	0.85	985
Female	0.83	735	0.89	1268	0.85	986
Male	0.84	778	0.88	1224	0.84	984
African American	0.77	562	0.88	1135	0.79	723
Asian	0.91	1417	0.92	1512	0.94	2267
Native Hawaiian/Pacific Islander*	0.43	245	0.93	1765	NA	NA
Hispanic/Latino	0.78	556	0.87	1183	0.82	856
American Indian or Alaskan	0.77	556	0.72	533	0.77	678
White	0.89	1045	0.90	1269	0.87	1087
Multiple Ethnicities	0.92	1715	0.91	1493	0.89	1352
Limited English Proficiency	0.67	417	0.82	878	0.84	1110
Special Education	0.76	558	0.71	590	0.88	1136
Free or Reduced-Price Lunch	0.75	520	0.72	605	0.86	1127

*The Native Hawaiian subgroup is not reported due to small sample size (sample size <11).

Exhibit 3.4.7.3 Marginal Reliability by Subgroup: ELA Fall 2018

Subgroup	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.90	1003	0.91	1140	0.87	779
Female	0.90	962	0.91	1141	0.86	701
Male	0.90	1000	0.91	1124	0.88	832
African American	0.88	747	0.90	1039	0.87	759
Asian	0.91	1238	0.92	1282	0.89	937
Native Hawaiian/Pacific Islander	0.92	1249	0.92	1302	0.83	552
Hispanic/Latino	0.89	866	0.89	915	0.85	662
American Indian or Alaskan	0.84	567	0.87	759	0.81	505
White	0.90	989	0.91	1210	0.88	801
Multiple Ethnicities	0.90	957	0.90	1028	0.89	908
Limited English Proficiency	0.84	633	0.77	520	0.79	513
Special Education	0.91	1134	0.90	1104	0.79	494
Free or Reduced-Price Lunch	0.88	746	0.89	921	0.86	674

Exhibit 3.4.7.4 Marginal Reliability by Subgroup: Mathematics Fall 2018

Subgroup	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.91	1490	0.88	1195	0.88	1310
Female	0.91	1428	0.87	1070	0.88	1199
Male	0.92	1542	0.89	1313	0.89	1422
African American	0.87	976	0.83	917	0.81	940
Asian	0.92	1680	0.91	1577	0.92	1647
Native Hawaiian/Pacific Islander	0.89	1042	0.79	669	0.93	2099
Hispanic/Latino	0.89	1133	0.85	1025	0.84	1021
American Indian or Alaskan	0.89	1081	0.87	1169	0.74	649
White	0.92	1622	0.89	1229	0.90	1386
Multiple Ethnicities	0.91	1436	0.89	1321	0.85	1004
Limited English Proficiency	0.76	567	0.74	671	0.70	588
Special Education	0.84	875	0.84	1035	0.85	1252
Free or Reduced-Price Lunch	0.88	1079	0.83	886	0.85	1057

3.4.8 SUBSCALE RELIABILITY

Marginal reliability estimates associated with the subscales for the summer 2018 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 2018. 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 2018

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.75	0.71	0.67
Grade 10	0.76	0.73	0.68
Grade 11	0.75	0.74	0.70

Exhibit 3.4.8.2 Subscale Reliabilities: Algebra I & II Summer 2018

Grade	Algebra	Functions	Statistics
Algebra I	0.70	0.64	0.45
Algebra II	0.63	0.64	0.56

Exhibit 3.4.8.3 Subscale Reliabilities: Geometry Summer 2018

Grade	Circles, Geometric Measurement & Dimension, and Modeling	Congruence	Geometric Properties with Equations	Similarity, Right Triangles & Trigonometry
Geometry	0.50	0.68	0.40	0.69

Exhibit 3.4.8.4 Subscale Reliabilities: ELA Grades 9–11 Fall 2018

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.80	0.73	0.75
Grade 10	0.79	0.76	0.76
Grade 11	0.70	0.68	0.75

Exhibit 3.4.8.5 Subscale Reliabilities: Algebra I & II Fall 2018

	Algebra	Functions	Statistics
Algebra I	0.83	0.78	0.61
Algebra II	0.73	0.69	0.67

Exhibit 3.4.8.6 Subscale Reliabilities: Geometry Fall 2018

	Circles, Geometric Measurement & Dimension, and Modeling	Congruence	Geometric Properties with Equations	Similarity, Right Triangles & Trigonometry
Geometry	0.45	0.68	0.38	0.67

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 2018 and in Exhibits 3.5.4–3.5.6 for fall 2018. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.[22] The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

[22] 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,

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

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 2018

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.71		0.97	
	Writing & Language	0.54	0.52	0.79	0.76
10	Literature	0.71		0.96	
	Writing & Language	0.60	0.57	0.84	0.81
11	Literature	0.68		0.92	
	Writing & Language	0.66	0.63	0.91	0.87

Exhibit 3.5.2 Subscale Observed and Disattenuated Intercorrelations: Algebra I & Algebra II Summer 2018

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.66		0.99	
	Statistics	0.57	0.60	1.00	1.00
Algebra II	Functions	0.64		1.00	
	Statistics	0.65	0.68	1.00	1.00

Exhibit 3.5.3 Subscale Observed and Disattenuated Intercorrelations: Geometry Summer 2018

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	GP	CGM_GPE	C	GP
Geometry	Congruence(C)	0.66			1.00		
	GP	0.64	0.62		1.00	1.00	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.67	0.69	0.65	1.00	1.00	1.00

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

Exhibit 3.5.4 Subscale Observed and Disattenuated Intercorrelations: ELA Grades 9–11 Fall 2018

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
9	Literature	0.73		0.96	
	Writing & Language	0.65	0.63	0.88	0.85

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 Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
10	Literature	0.74		0.95	
	Writing & Language	0.69	0.63	0.91	0.83
11	Literature	0.64		0.93	
	Writing & Language	0.62	0.61	0.86	0.85

Exhibit 3.5.5 Subscale Observed and Disattenuated Intercorrelations: Algebra I & Algebra II Fall 2018

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.81		1.00	
	Statistics	0.71	0.71	1.00	1.00
Algebra II	Functions	0.73		1.00	
	Statistics	0.73	0.71	1.00	1.00

Exhibit 3.5.6 Subscale Observed and Disattenuated Intercorrelations: Geometry Fall 2018

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	GP	CGM_GPE	C	GP
Geometry	Congruence(C)	0.62			1.00		
	GPGP	0.60	0.60		1.00	1.00	
	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 & Dimension, and Modeling; GP = Geometric Properties with Equations; SRTT = Similarity, Right Triangles, and Trigonometry

4. SUMMARY OF SPRING 2019 OPERATIONAL TEST ADMINISTRATION

The following Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) assessments were administered in spring 2019:

- 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, 2019. The paper-pencil version of the AzMERIT was administered from April 2–10, 2019.

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 2019, 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 2019.

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 2019, 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 2019 final scoring and reporting. This section summarizes the operational test results for the spring 2019 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.[23] We note that some students participated in an end-of-course (EOC) assessment rather than a grade-level assessment, especially in grade 8, where a large number of more-advanced students are enrolled in Algebra I courses. The tables in Appendix F show the demographic composition of test takers by mode of test administration.

Exhibit 4.1.1 Number of Students Participating in ELA Assessments by Subgroups: Spring 2019

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
All Students	82,779	86,693	90,158	90,234	88,623	87,046	69,347	63,288	56,917
Female	40,672	42,176	44,328	44,379	43,555	43,049	33,721	31,424	28,524
Male	42,107	44,517	45,830	45,855	45,068	43,997	35,626	31,864	28,393
African American	4,631	4,871	4,922	4,884	4,913	4,775	3,929	3,531	3,098
Asian	2,431	2,572	2,614	2,575	2,536	2,585	1,982	1,853	1,807
Native Hawaiian/Pacific Islander	335	321	329	368	367	327	303	248	213
Hispanic/Latino	37,845	39,871	42,133	41,519	40,487	39,339	30,983	27,468	24,189
American Indian or Alaskan	3,946	4,218	4,317	4,297	4,272	4,206	3,593	2,994	2,637
White	30,479	31,875	32,809	33,556	33,278	33,304	26,836	25,723	23,641
Multiple Ethnicities	3,112	2,965	3,034	3,035	2,770	2,510	1,721	1,471	1,332
Limited English Proficiency	6,909	7,472	8,240	7,430	6,449	5,160	4,530	2,964	2,012
Special Education	10,357	11,026	11,375	10,929	10,115	9,631	6,684	5,305	4,531
Free or Reduced-Price Lunch	34,529	36,602	38,610	36,383	34,866	33,433	19,101	17,360	15,002
Accommodation	4,506	4,743	4,932	4,560	3,852	3,524	1,223	1,011	714

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

Exhibit 4.1.2 Number of Students Participating in Mathematics Assessments by Subgroups: Spring 2019

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
All Students	83,180	86,919	90,236	90,312	88,751	78,024	76,725	63,327	55,223
Female	40,813	42,275	44,331	44,380	43,589	38,509	37,391	31,380	28,153
Male	42,367	44,644	45,905	45,932	45,162	39,515	39,334	31,947	27,070
African American	4,669	4,896	4,931	4,878	4,933	4,485	4,257	3,435	2,958
Asian	2,434	2,574	2,616	2,574	2,470	1,741	2,421	2,008	1,926
Native Hawaiian/Pacific Islander	336	322	330	370	366	301	348	241	209
Hispanic/Latino	38,029	39,981	42,193	41,545	40,604	36,208	34,580	27,722	23,493
American Indian or Alaskan	3,979	4,237	4,313	4,324	4,298	4,119	3,648	2,961	2,406
White	30,602	31,933	32,817	33,580	33,300	28,946	29,497	25,477	22,962
Multiple Ethnicities	3,131	2,976	3,036	3,041	2,780	2,224	1,974	1,483	1,269
Limited English Proficiency	6,952	7,507	8,257	7,464	6,483	4,940	4,576	3,387	2,118
Special Education	10,492	11,106	11,425	10,957	10,173	9,377	6,975	5,278	3,449
Free or Reduced-Price Lunch	34,653	36,672	38,622	36,358	34,932	31,666	21,697	17,244	13,972
Accommodation	4,507	4,822	4,839	4,318	3,676	3,375	1,186	809	461

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 2019 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 2019 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 *p*-Value	*p*-Value SD	Average Point-Biserial	Point-Biserial SD
ELA				
3	0.48	0.17	0.45	0.13
4	0.54	0.17	0.45	0.1
5	0.56	0.17	0.49	0.11
6	0.53	0.18	0.45	0.12
7	0.52	0.18	0.45	0.11
8	0.52	0.17	0.49	0.12
9	0.52	0.14	0.44	0.12
10	0.5	0.17	0.45	0.11
11	0.5	0.18	0.44	0.13
Mathematics				
3	0.62	0.17	0.51	0.1
4	0.58	0.18	0.52	0.08
5	0.51	0.16	0.51	0.1
6	0.48	0.19	0.51	0.1
7	0.49	0.18	0.51	0.1
8	0.43	0.17	0.49	0.12
Algebra I	0.43	0.19	0.46	0.12
Geometry	0.35	0.15	0.47	0.11
Algebra II	0.34	0.16	0.48	0.1

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
		ELA		
3	0	1	1	1
4	0	0	0	2
5	0	0	0	1
6	0	0	0	1
7	0	0	0	0
8	0	0	0	1
9	0	1	1	1
10	1	0	0	2
11	0	1	0	0
		Mathematics		
3	0	0	0	2
4	0	0	0	0
5	0	0	0	1
6	0	0	0	0
7	0	0	0	0
8	0	0	0	1
Algebra I	0	0	0	0
Geometry	0	0	1	0
Algebra II	0	0	0	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:

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

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

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

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

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

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

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

The tables in Appendix E provide Rasch and Masters' partial credit model item parameter estimates for the spring 2019 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
ELA									
3	39	0.06	0.81	-	-	-	1.59	1.58	-1.16
4	41	0.13	0.61	-	-	-	3.62	4.00	-0.09
5	41	0.10	0.84	-	-	-	2.39	3.07	-0.85
6	41	0.05	0.75	-	-	-	2.28	2.95	-1.21
7	41	0.06	0.86	-	-	-	2.36	2.76	-1.56
8	41	0.06	0.93	-	-	-	0.97	1.16	-1.62
9	43	0.06	0.62	-	-	-	1.27	1.66	-1.82
10	43	0.07	0.83	-	-	-	0.84	1.22	-2.03
11	42	0.00	0.99	1	-0.05	-	0.46	0.99	-1.96
Mathematics									
3	22	-0.11	1.14	23	0.31	1.18	-	-	-
4	12	-0.31	1.31	33	0.16	1.11	-	-	-
5	15	-0.41	0.95	30	0.30	0.84	-	-	-
6	21	-0.34	1.26	26	0.35	0.98	-	-	-

Grade/ Course	SR			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
7	21	-0.58	0.86	26	0.61	0.95	-	-	-
8	25	-0.56	1.09	22	0.33	0.75	-	-	-
Algebra I	29	-0.13	0.96	18	0.64	1.10	-	-	-
Geometry	24	-0.62	0.82	23	0.59	0.77	-	-	-
Algebra II	25	-0.61	0.97	22	0.52	0.57	-	-	-

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
ELA						
3	0	44	1	1	39	5
4	0	46	1	2	43	2
5	0	44	3	0	43	4
6	0	47	0	3	39	5
7	0	46	1	0	44	3
8	0	46	1	3	36	8
9	0	49	0	2	45	2
10	0	48	1	0	47	2
11	0	49	0	2	45	2
Mathematics						
3	1	41	3	2	33	10
4	0	43	2	0	41	4
5	0	43	2	2	35	8
6	0	44	3	2	39	6
7	0	45	2	5	37	5
8	0	47	0	2	36	9
Algebra I	0	47	0	4	38	5
Geometry	0	43	4	4	34	9
Algebra II	0	47	0	2	42	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.
ELA					
3	82,778	2505	31.35	2605	2395
4	86,691	2523	32.36	2610	2400
5	90,158	2541	37.38	2629	2419
6	90,233	2545	32.59	2641	2431
7	88,621	2552	34.68	2648	2438
8	87,046	2559	36.24	2658	2448
9	69,346	2565	31.94	2664	2454
10	63,288	2565	32.05	2668	2458
11	56,917	2569	33.20	2675	2465
Mathematics					
3	83,179	3527	44.60	3605	3395
4	86,916	3557	45.50	3645	3435
5	90,236	3587	42.72	3688	3478
6	90,311	3616	44.24	3722	3512
7	88,749	3636	43.28	3739	3529
8	78,019	3655	39.88	3776	3566
Algebra I	76,725	3675	37.60	3787	3577
Geometry	63,327	3687	37.64	3819	3609
Algebra II	55,223	3704	39.17	3839	3629

Exhibit 4.4.2 Test Score Summary Statistics: Online

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
ELA					
3	73,477	2504	31.29	2605	2395
4	77,032	2522	32.37	2610	2408
5	80,273	2541	37.63	2629	2419
6	80,073	2544	32.27	2641	2431
7	79,539	2551	34.48	2648	2438

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
8	78,657	2558	35.90	2658	2448
9	63,851	2565	31.41	2664	2454
10	58,691	2565	31.92	2668	2458
11	52,827	2569	32.74	2675	2465
Mathematics					
3	73,778	3526	44.55	3605	3395
4	77,198	3556	45.50	3645	3435
5	80,350	3587	42.62	3688	3478
6	80,142	3616	44.07	3722	3512
7	79,779	3635	43.20	3739	3529
8	71,237	3655	39.89	3776	3566
Algebra I	70,501	3675	37.12	3787	3577
Geometry	58,130	3687	37.43	3819	3609
Algebra II	50,749	3704	38.87	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.
ELA					
3	9,302	2510	31.39	2605	2397
4	9,661	2527	31.93	2610	2400
5	9,885	2546	34.92	2629	2420
6	10,161	2551	34.43	2641	2431
7	9,084	2558	35.80	2648	2438
8	8,389	2569	38.01	2658	2448
9	5,496	2562	37.48	2664	2465
10	4,597	2563	33.69	2668	2466
11	4,090	2572	38.59	2675	2465
Math					
3	9,402	3530	44.84	3605	3395
4	9,721	3560	45.34	3645	3435
5	9,886	3593	43.10	3688	3478
6	10,170	3623	45.00	3722	3512
7	8,972	3644	43.15	3739	3529
8	6,787	3658	39.67	3776	3566
Algebra I	6,224	3675	42.65	3787	3577
Geometry	5,197	3683	39.69	3819	3609
Algebra II	4,474	3700	42.22	3839	3629

The percentage of students in each performance level by grade and content area, as well as the percentage of students at or above Proficient are presented in 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
ELA						
3	82778	40	14	32	14	46
4	86691	34	15	37	14	51
5	90158	28	20	32	20	52
6	90233	34	24	34	8	42
7	88621	40	19	31	10	41
8	87046	41	21	25	13	38
9	69346	41	23	24	13	37
10	63288	51	15	24	10	34
11	56917	50	16	20	13	34
Mathematics						
3	83179	23	26	32	18	51
4	86916	27	25	33	15	48
5	90236	27	27	31	15	46
6	90311	38	21	24	16	41
7	88749	44	18	20	18	38
8	78019	49	20	18	13	31
Algebra I	76725	38	18	29	15	43
Geometry	63327	42	21	28	9	37
Algebra II	55223	39	21	26	14	40

Exhibit 4.4.5 Percentage of Students in Performance Levels: Online

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	73477	41	15	31	14	45
4	77032	35	15	36	14	50
5	80273	28	20	31	20	51
6	80073	35	24	34	7	41
7	79539	41	19	30	10	40
8	78657	42	21	25	12	37
9	63851	40	23	24	12	37
10	58691	51	15	24	10	34
11	52827	50	17	21	13	34

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
			Mathematics			
3	73778	23	26	32	18	50
4	77198	28	25	32	15	47
5	80350	28	27	30	15	45
6	80142	39	22	24	16	40
7	79779	45	18	20	17	37
8	71237	49	20	18	13	31
Algebra I	70501	38	18	29	15	44
Geometry	58130	41	21	28	9	38
Algebra II	50749	39	21	27	14	41

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
			ELA			
3	9302	33	12	36	19	54
4	9661	27	15	42	16	58
5	9885	23	20	35	21	57
6	10161	28	22	39	11	50
7	9084	31	20	35	13	49
8	8389	31	21	29	19	48
9	5496	47	20	19	14	33
10	4597	53	15	19	13	32
11	4090	49	14	18	19	37
			Mathematics			
3	9402	21	25	33	20	54
4	9721	25	23	36	17	53
5	9886	23	25	32	20	52
6	10170	33	20	26	21	47
7	8972	37	18	21	23	45
8	6787	45	21	19	14	34
Algebra I	6224	43	15	23	19	42
Geometry	5197	48	18	24	9	33
Algebra II	4474	48	18	19	15	34

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 [female, male] and ethnicity [African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian, White, and Multiple Ethnicities], and by other demographic information, such as special education status (SPED), limited English proficiency (LEP), eligibility for free or reduced-price 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	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	32,906	15,157	17,748	2,338	416	133	18,569	2,643	7,875	931	7,686	5,690	17,542	3,617
	Partially Proficient	11,836	5,961	5,875	671	279	52	5,731	491	4,158	454	919	582	5,269	381
	Proficient	26,345	13,352	12,993	1,292	902	109	10,401	672	11,836	1,133	1,297	565	9,156	421
	Highly Proficient	11,694	6,202	5,492	330	834	41	3,145	140	6,610	594	455	72	2,563	87
4	Minimally Proficient	29,194	12,630	16,563	2,217	349	114	17,127	2,376	6,280	730	7,843	5,978	16,142	3,615
	Partially Proficient	13,183	6,448	6,734	852	255	48	6,724	695	4,176	432	1,171	861	6,291	516
	Proficient	32,080	16,352	15,728	1,459	1,125	117	12,958	996	14,161	1,264	1,604	588	11,600	551
	Highly Proficient	12,238	6,746	5,492	343	843	42	3,062	151	7,258	539	408	45	2,569	61
5	Minimally Proficient	25,156	10,473	14,683	1,915	306	95	14,765	2,180	5,250	645	7,855	5,861	14,032	3,546
	Partially Proficient	18,233	8,801	9,432	1,108	309	49	9,762	1,027	5,429	549	1,698	1,607	9,029	841
	Proficient	28,646	14,937	13,708	1,341	928	125	12,335	874	11,953	1,089	1,339	691	11,072	436
	Highly Proficient	18,125	10,118	8,007	558	1,071	60	5,272	236	10,177	751	483	82	4,477	109
6	Minimally Proficient	30,859	12,871	17,986	2,304	330	130	17,791	2,398	7,067	837	8,396	5,939	15,963	3,569
	Partially Proficient	21,668	11,019	10,649	1,171	429	88	10,805	1,090	7,407	678	1,464	1,093	9,478	639
	Proficient	30,842	16,533	14,309	1,254	1,240	128	11,401	761	14,831	1,227	939	377	9,704	324
	Highly Proficient	6,867	3,956	2,911	155	576	22	1,522	48	4,251	293	130	21	1,238	28

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
7	Minimally Proficient	35,188	14,758	20,430	2,571	360	128	19,906	2,713	8,662	848	8,252	5,572	17,538	3,279
	Partially Proficient	17,175	8,812	8,363	963	342	70	8,244	803	6,246	507	987	560	7,104	312
	Proficient	27,422	14,957	12,465	1,170	1,095	134	10,209	650	13,132	1,032	730	283	8,559	233
	Highly Proficient	8,838	5,028	3,810	209	739	35	2,128	106	5,238	383	146	34	1,665	28
8	Minimally Proficient	35,395	14,440	20,954	2,522	419	118	19,865	2,801	8,854	815	8,027	4,505	17,226	3,017
	Partially Proficient	18,565	9,759	8,805	989	395	79	8,621	773	7,183	524	904	415	7,292	296
	Proficient	22,012	12,182	9,830	944	859	85	8,054	500	10,803	767	549	204	6,753	178
	Highly Proficient	11,076	6,668	4,408	320	912	45	2,799	132	6,464	404	151	36	2,162	33
9	Minimally Proficient	28,133	11,413	16,718	2,148	359	125	15,448	2,239	7,267	545	5,425	3,675	9,808	986
	Partially Proficient	15,853	8,268	7,584	863	316	62	7,473	802	5,925	411	787	549	4,427	134
	Proficient	16,678	9,040	7,637	703	616	78	6,047	450	8,300	483	375	242	3,634	75
	Highly Proficient	8,701	5,001	3,700	220	692	38	2,022	102	5,345	282	101	64	1,236	28
10	Minimally Proficient	32,137	14,408	17,729	2,232	489	137	16,674	2,222	9,740	643	4,573	2,461	10,971	923
	Partially Proficient	9,669	5,047	4,622	467	229	38	4,196	359	4,157	223	344	245	2,479	43
	Proficient	14,969	8,186	6,783	643	601	51	5,128	339	7,778	429	319	218	3,059	40
	Highly Proficient	6,524	3,784	2,739	193	534	22	1,475	75	4,048	176	72	40	858	5

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
11	Minimally Proficient	28,333	12,867	15,464	1,918	473	104	14,520	1,901	8,804	611	3,946	1,618	9,360	631
	Partially Proficient	9,305	4,968	4,337	442	260	38	3,952	351	4,041	221	277	170	2,356	45
	Proficient	11,635	6,373	5,261	500	479	47	4,077	303	5,942	286	215	173	2,271	24
	Highly Proficient	7,651	4,316	3,335	238	595	24	1,644	82	4,854	214	94	51	1,015	14

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL = Free or Reduced-Price 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	Accommodati on
3	Minimally Proficient	19,252	9,379	9,873	1,553	152	77	11,001	1,759	4,158	552	5,770	3,852	10,510	2,668
	Partially Proficient	21,733	11,190	10,543	1,393	336	92	11,443	1,212	6,525	732	2,361	2,006	10,384	1,118
	Proficient	26,949	13,375	13,573	1,294	826	110	11,355	811	11,403	1,149	1,688	905	10,079	583
	Highly Proficient	15,247	6,869	8,378	429	1,120	57	4,230	197	8,516	698	673	189	3,680	138
4	Minimally Proficient	23,666	11,415	12,249	2,023	177	87	13,588	2,021	5,133	635	6,794	4,613	13,034	3,007
	Partially Proficient	21,775	10,969	10,804	1,311	344	71	11,165	1,175	7,007	700	2,263	1,839	10,309	1,045
	Proficient	28,556	14,095	14,461	1,226	987	113	11,722	869	12,533	1,106	1,565	925	10,313	653
	Highly Proficient	12,926	5,796	7,130	336	1,066	51	3,506	172	7,260	535	484	130	3,016	117
5	Minimally Proficient	24,731	11,403	13,326	2,069	202	87	14,237	2,031	5,406	697	7,412	5,004	13,723	3,182
	Partially Proficient	23,932	12,179	11,753	1,415	337	79	12,368	1,279	7,679	775	2,293	2,112	11,341	1,057
	Proficient	27,642	14,153	13,489	1,113	873	108	11,615	816	12,133	984	1,288	981	10,293	503
	Highly Proficient	13,933	6,596	7,337	334	1,204	56	3,973	187	7,599	580	432	160	3,265	97
6	Minimally Proficient	34,344	16,298	18,046	2,677	308	146	19,666	2,526	8,034	987	8,473	5,725	17,503	3,376
	Partially Proficient	19,315	10,160	9,155	991	333	84	9,393	954	6,935	625	1,326	1,101	8,174	548
	Proficient	21,815	11,093	10,722	850	707	92	8,526	653	10,161	826	787	509	7,425	295
	Highly Proficient	14,838	6,829	8,009	360	1,226	48	3,960	191	8,450	603	371	129	3,256	99
7	Minimally Proficient	39,179	19,158	20,019	3,069	359	154	22,350	2,896	9,347	1,002	8,491	5,566	19,489	3,157
	Partially Proficient	16,177	8,294	7,883	826	312	65	7,525	688	6,220	541	844	555	6,475	307
	Proficient	17,668	8,858	8,810	651	559	84	6,694	493	8,564	623	565	258	5,671	151
	Highly Proficient	15,729	7,279	8,450	387	1,240	63	4,035	221	9,169	614	273	104	3,297	61
8	Minimally Proficient	37,891	17,987	19,904	2,800	367	122	20,927	2,860	9,840	975	8,012	4,137	18,447	2,903
	Partially Proficient	15,574	8,132	7,441	838	289	63	6,902	695	6,333	453	750	462	6,126	277
	Proficient	14,220	7,376	6,843	568	410	74	5,193	393	7,114	467	409	222	4,503	134
	Highly Proficient	10,341	5,014	5,327	279	675	42	3,186	171	5,659	329	206	119	2,590	61
Algebra I	Minimally Proficient	29,513	13,105	16,400	2,282	288	129	16,516	2,220	7,486	584	5,523	3,496	10,237	971
	Partially Proficient	13,857	7,337	6,520	787	238	74	6,779	705	4,915	359	748	567	4,119	110
	Proficient	21,919	11,463	10,456	930	775	96	8,506	580	10,384	648	544	432	5,342	77
	Highly Proficient	11,450	5,487	5,963	260	1,120	49	2,783	143	6,712	383	160	82	2,001	28

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodati on
Geometry	Minimally Proficient	26,476	12,711	13,765	2,092	355	102	14,214	1,752	7,414	547	4,232	2,426	9,254	660
	Partially Proficient	13,318	6,873	6,445	667	267	55	6,138	639	5,242	310	628	549	3,723	99
	Proficient	17,794	9,103	8,691	574	794	62	6,210	502	9,237	415	340	366	3,530	45
	Highly Proficient	5,742	2,693	3,049	102	592	22	1,162	68	3,585	211	80	46	737	5
Algebra II	Minimally Proficient	21,674	10,447	11,227	1,595	289	70	11,280	1,376	6,611	453	2,610	1,324	7,026	361
	Partially Proficient	11,357	6,179	5,178	611	235	52	5,125	542	4,524	268	433	364	3,052	57
	Proficient	14,390	7,797	6,593	563	579	52	5,336	397	7,130	333	276	331	2,926	36
	Highly Proficient	7,805	3,731	4,074	190	823	35	1,754	91	4,697	215	131	100	969	7

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL = Free or Reduced-Price 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
	Minimally Proficient	40	37	42	50	17	40	49	67	26	30	74	82	51	80
	Partially Proficient	14	15	14	14	11	16	15	12	14	15	9	8	15	8
3	Proficient	32	33	31	28	37	33	27	17	39	36	13	8	27	9
	Highly Proficient	14	15	13	7	34	12	8	4	22	19	4	1	7	2
	At or Above Proficient	46	48	44	35	71	45	36	21	61	55	17	9	34	11
	Minimally Proficient	34	30	37	46	14	36	43	56	20	25	71	80	44	76
	Partially Proficient	15	15	15	17	10	15	17	16	13	15	11	12	17	11
4	Proficient	37	39	35	30	44	36	33	24	44	43	15	8	32	12
	Highly Proficient	14	16	12	7	33	13	8	4	23	18	4	1	7	1
	At or Above Proficient	51	55	48	37	77	50	40	27	67	61	18	8	39	13
	Minimally Proficient	28	24	32	39	12	29	35	50	16	21	69	71	36	72
	Partially Proficient	20	20	21	23	12	15	23	24	17	18	15	20	23	17
5	Proficient	32	34	30	27	36	38	29	20	36	36	12	8	29	9
	Highly Proficient	20	23	17	11	41	18	13	5	31	25	4	1	12	2
	At or Above Proficient	52	57	47	39	76	56	42	26	67	61	16	9	40	11
	Minimally Proficient	34	29	39	47	13	35	43	56	21	28	77	80	44	78
	Partially Proficient	24	25	23	24	17	24	26	25	22	22	13	15	26	14
6	Proficient	34	37	31	26	48	35	27	18	44	40	9	5	27	7
	Highly Proficient	8	9	6	3	22	6	4	1	13	10	1	0	3	1
	At or Above Proficient	42	46	38	29	71	41	31	19	57	50	10	5	30	8
	Minimally Proficient	40	34	45	52	14	35	49	64	26	31	82	86	50	85
	Partially Proficient	19	20	19	20	13	19	20	19	19	18	10	9	20	8
7	Proficient	31	34	28	24	43	37	25	15	39	37	7	4	25	6
	Highly Proficient	10	12	8	4	29	10	5	2	16	14	1	1	5	1
	At or Above Proficient	41	46	36	28	72	46	30	18	55	51	9	5	29	7
	Minimally Proficient	41	34	48	53	16	36	50	67	27	32	83	87	52	86
	Partially Proficient	21	23	20	21	15	24	22	18	22	21	9	8	22	8
8	Proficient	25	28	22	20	33	26	20	12	32	31	6	4	20	5
	Highly Proficient	13	15	10	7	35	14	7	3	19	16	2	1	6	1
	At or Above Proficient	38	44	32	26	69	40	28	15	52	47	7	5	27	6

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
9	Minimally Proficient	41	34	47	55	18	41	50	62	27	32	81	81	51	81
	Partially Proficient	23	25	21	22	16	20	24	22	22	24	12	12	23	11
	Proficient	24	27	21	18	31	26	20	13	31	28	6	5	19	6
	Highly Proficient	13	15	10	6	35	13	7	3	20	16	2	1	6	2
	At or Above Proficient	37	42	32	23	66	38	26	15	51	44	7	7	25	8
10	Minimally Proficient	51	46	56	63	26	55	61	74	38	44	86	83	63	91
	Partially Proficient	15	16	15	13	12	15	15	12	16	15	6	8	14	4
	Proficient	24	26	21	18	32	21	19	11	30	29	6	7	18	4
	Highly Proficient	10	12	9	5	29	9	5	3	16	12	1	1	5	0
	At or Above Proficient	34	38	30	24	61	29	24	14	46	41	7	9	23	4
11	Minimally Proficient	50	45	54	62	26	49	60	72	37	46	87	80	62	88
	Partially Proficient	16	17	15	14	14	18	16	13	17	17	6	8	16	6
	Proficient	20	22	19	16	27	22	17	11	25	21	5	9	15	3
	Highly Proficient	13	15	12	8	33	11	7	3	21	16	2	3	7	2
	At or Above Proficient	34	37	30	24	59	33	24	15	46	38	7	11	22	5

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL = Free or Reduced-Price 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	23	23	33	6	23	29	44	14	18	55	55	30	59
	Partially Proficient	26	27	25	30	14	27	30	30	21	23	23	29	30	25
	Proficient	32	33	32	28	34	33	30	20	37	37	16	13	29	13
	Highly Proficient	18	17	20	9	46	17	11	5	28	22	6	3	11	3
	At or Above Proficient	51	50	52	37	80	50	41	25	65	59	23	16	40	16

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
4	Minimally Proficient	27	27	27	41	7	27	34	48	16	21	61	61	36	62
	Partially Proficient	25	26	24	27	13	22	28	28	22	24	20	24	28	22
	Proficient	33	33	32	25	38	35	29	21	39	37	14	12	28	14
	Highly Proficient	15	14	16	7	41	16	9	4	23	18	4	2	8	2
	At or Above Proficient	48	47	48	32	80	51	38	25	62	55	18	14	36	16
5	Minimally Proficient	27	26	29	42	8	26	34	47	16	23	65	61	36	66
	Partially Proficient	27	27	26	29	13	24	29	30	23	26	20	26	29	22
	Proficient	31	32	29	23	33	33	28	19	37	32	11	12	27	10
	Highly Proficient	15	15	16	7	46	17	9	4	23	19	4	2	8	2
	At or Above Proficient	46	47	45	29	79	50	37	23	60	52	15	14	35	12
6	Minimally Proficient	38	37	39	55	12	39	47	58	24	32	77	77	48	78
	Partially Proficient	21	23	20	20	13	23	23	22	21	21	12	15	22	13
	Proficient	24	25	23	17	27	25	21	15	30	27	7	7	20	7
	Highly Proficient	16	15	17	7	48	13	10	4	25	20	3	2	9	2
	At or Above Proficient	41	40	41	25	75	38	30	20	55	47	11	9	29	9
7	Minimally Proficient	44	44	44	62	15	42	55	67	28	36	83	86	56	86
	Partially Proficient	18	19	17	17	13	18	19	16	19	19	8	9	19	8
	Proficient	20	20	20	13	23	23	16	11	26	22	6	4	16	4
	Highly Proficient	18	17	19	8	50	17	10	5	28	22	3	2	9	2
	At or Above Proficient	38	37	38	21	73	40	26	17	53	44	8	6	26	6
8	Minimally Proficient	49	47	50	62	21	41	58	69	34	44	85	84	58	86
	Partially Proficient	20	21	19	19	17	21	19	17	22	20	8	9	19	8
	Proficient	18	19	17	13	24	25	14	10	25	21	4	4	14	4
	Highly Proficient	13	13	13	6	39	14	9	4	20	15	2	2	8	2
	At or Above Proficient	31	32	31	19	62	39	23	14	44	36	7	7	22	6
Algebra I	Minimally Proficient	38	35	42	54	12	37	48	61	25	30	79	76	47	82
	Partially Proficient	18	20	17	18	10	21	20	19	17	18	11	12	19	9
	Proficient	29	31	27	22	32	28	25	16	35	33	8	9	25	6
	Highly Proficient	15	15	15	6	46	14	8	4	23	19	2	2	9	2
	At or Above Proficient	43	45	42	28	78	42	33	20	58	52	10	11	34	9
Geometry	Minimally Proficient	42	41	43	61	18	42	51	59	29	37	80	72	54	82
	Partially Proficient	21	22	20	19	13	23	22	22	21	21	12	16	22	12
	Proficient	28	29	27	17	40	26	22	17	36	28	6	11	20	6
	Highly Proficient	9	9	10	3	29	9	4	2	14	14	2	1	4	1

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
	At or Above Proficient	37	38	37	20	69	35	27	19	50	42	8	12	25	6
Algebra II	Minimally Proficient	39	37	41	54	15	33	48	57	29	36	76	63	50	78
	Partially Proficient	21	22	19	21	12	25	22	23	20	21	13	17	22	12
	Proficient	26	28	24	19	30	25	23	17	31	26	8	16	21	8
	Highly Proficient	14	13	15	6	43	17	7	4	20	17	4	5	7	2
	At or Above Proficient	40	41	39	25	73	42	30	20	52	43	12	20	28	9

Note: Alaskan = Alaskan Native; Hawaiian/Pacific = Native Hawaiian/Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL = Free or Reduced-Price 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 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. Test score reliability is traditionally estimated using both classical and IRT approaches. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the IRT framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the test information function represents the standard error of measurement. The standard error of measurement is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale that is at the middle of the test distribution and greater on scaled values farther away from the middle.

The reliability evidence of the AZMERIT test scores is provided with reliability, SEM, and classification accuracy and consistency in each achievement level.

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. Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale, for all students. The marginal reliability coefficients are nearly identical or close to coefficient alpha. For our analysis, the marginal reliability coefficients were computed using operational items.

The marginal reliability ($\bar{\rho}$) is defined as

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i^2$ is the conditional standard error of measurement of the scale score for student i; and $\sigma^2$is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Exhibit 4.6.1.1 shows presents the marginal reliability coefficients for all students. The reliability coefficients for all subjects and grades range from 0.90 to 0.93.

Exhibit 4.6.1.1 Overall Reliabilities by Subject/Test for AzMERIT Scores

Grade	ELA		Mathematics	
	Reliability	Variance	Reliability	Variance
G3	0.90	979	0.92	1985
G4	0.90	1048	0.93	2070
G5	0.91	1416	0.92	1817
G6	0.90	1042	0.93	1942
G7	0.90	1189	0.93	1866
G8	0.92	1289	0.92	1591
G9E / Algebra I	0.90	987	0.91	1378
G10E / Geometry	0.91	1019	0.90	1401
G11E / Algebra II	0.90	1072	0.91	1511

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 conditional standard errors of measurement (CSEM) for the AzMERIT ELA and mathematics assessments with respect to the four AzMERIT performance-level cuts. These tables also include associated CSEM around cut score. 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.[24] 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 4.6.2.3 through Exhibit 4.6.2.29 present the CSEMs and corresponding performance levels for each scale score for the AzMERIT ELA and mathematics assessments.

[24] 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.

Exhibit 4.6.2.1 Performance Level and Associated CSEMs Spring 2019: ELA

Grade	CSEM	Proficiency Level				Overall
		Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	
Grade 3	Mean	10	9	10	12	10
	Around Cut Score		9	9	11	
Grade 4	Mean	10	9	10	13	10
	Around Cut Score		9	9	11	
Grade 5	Mean	11	9	11	14	11
	Around Cut Score		9	10	12	
Grade 6	Mean	10	9	10	14	10
	Around Cut Score		9	9	12	
Grade 7	Mean	11	10	11	14	11
	Around Cut Score		10	10	12	
Grade 8	Mean	10	9	10	13	10
	Around Cut Score		9	9	11	
Grade 9	Mean	10	9	9	12	10
	Around Cut Score		9	9	11	
Grade 10	Mean	10	9	9	11	10
	Around Cut Score		9	9	10	
Grade 11	Mean	10	10	10	12	10
	Around Cut Score		9	10	11	

Exhibit 4.6.2.2 Performance Level and Associated CSEMs Spring 2019: Mathematics

Grade	CSEM	Proficiency Level				Overall
		Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	
Grade 3	Mean	12	10	12	17	13
	Around Cut Score		10	11	14	
Grade 4	Mean	12	10	12	16	12
	Around Cut Score		10	11	13	
Grade 5	Mean	13	10	10	15	12
	Around Cut Score		10	10	12	
Grade 6	Mean	12	10	10	14	12
	Around Cut Score		10	10	11	
Grade 7	Mean	12	10	10	14	12
	Around Cut Score		10	10	11	
Grade 8	Mean	12	9	10	13	11
	Around Cut Score		10	9	11	
Algebra I	Mean	12	10	10	12	11
	Around Cut Score		10	10	10	
Geometry	Mean	13	10	10	13	12
	Around Cut Score		11	10	10	
Algebra II	Mean	14	11	10	11	12
	Around Cut Score		11	10	10	

Exhibit 4.6.2.3 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 3 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2395	22	2497	9	2510	9	2543	11
2408	18	2499	9	2513	9	2547	11
2417	16	2502	9	2516	9	2551	12
2425	15	2505	9	2519	10	2556	12
2432	14	2508	9	2522	10	2561	13
2437	13			2525	10	2567	13
2443	12			2529	10	2573	14
2447	12			2532	10	2580	15
2451	11			2536	10	2588	16
2455	11			2539	11	2598	18
2459	11					2605	20
2463	10						
2466	10						
2470	10						
2473	10						
2476	10						
2479	9						
2482	9						
2485	9						
2488	9						
2491	9						
2493	9						

Note: For Grade 3 ELA, Form 1 = writing prompt 13022 administered

Exhibit 4.6.2.4 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 3 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2395	22	2497	9	2509	9	2541	11
2397	22	2500	9	2511	9	2544	11
2410	18	2503	9	2514	9	2548	11
2419	16	2506	9	2517	9	2552	12
2427	15			2520	10	2557	12
2434	13			2523	10	2562	13
2439	13			2526	10	2568	13
2444	12			2530	10	2574	14
2449	12			2533	10	2581	15
2453	11			2537	10	2589	16

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2457	11					2599	18
2461	10					2605	19
2465	10						
2468	10						
2471	10						
2474	10						
2477	10						
2480	9						
2483	9						
2486	9						
2489	9						
2492	9						
2495	9						

Note: For Grade 3 ELA, Form 2 = writing prompt 13025 administered

Exhibit 4.6.2.5 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 4 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2408	22	2511	9	2524	9	2559	11
2421	18	2514	9	2527	9	2564	12
2431	16	2516	9	2530	9	2568	12
2438	15	2519	9	2533	9	2573	13
2445	13	2522	9	2536	9	2579	14
2451	13			2538	10	2586	15
2456	12			2542	10	2594	16
2460	12			2545	10	2603	18
2465	11			2548	10	2610	19
2469	11			2552	10		
2472	10			2555	11		
2476	10						
2479	10						
2482	10						
2486	10						
2489	9						
2492	9						
2494	9						
2497	9						
2500	9						
2503	9						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2506	9						
2508	9						

Note: For Grade 4 ELA, Form 1 = writing prompt 13119 administered

Exhibit 4.6.2.6 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 4 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2422	18	2510	9	2523	9	2561	11
2432	16	2512	9	2525	9	2565	12
2439	15	2514	9	2528	9	2570	12
2446	13	2517	9	2531	9	2575	13
2451	13	2520	9	2534	9	2581	14
2457	12			2536	9	2588	15
2461	11			2539	10	2596	16
2465	11			2543	10	2605	18
2469	11			2546	10	2610	19
2473	10			2549	10		
2477	10			2553	10		
2480	10			2557	11		
2483	10						
2486	10						
2489	9						
2492	9						
2495	9						
2498	9						
2501	9						
2503	9						
2506	9						

Note: For Grade 4 ELA, Form 2 = writing prompt 13120 administered

Exhibit 4.6.2.7 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 5 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2419	23	2520	9	2544	10	2578	12
2421	22	2522	9	2547	10	2582	13
2435	18	2525	9	2550	10	2588	13
2445	16	2528	9	2553	10	2594	14

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2452	15	2531	9	2557	10	2601	15
2459	14	2534	9	2561	11	2609	16
2465	13	2537	10	2564	11	2618	17
2470	12	2540	10	2569	11	2629	19
2475	12			2573	12		
2479	11						
2483	11						
2487	11						
2491	10						
2494	10						
2498	10						
2501	10						
2504	10						
2508	10						
2511	10						
2514	9						
2517	9						

Note: For Grade 5 ELA, Form 1 = writing prompt 13246 administered

Exhibit 4.6.2.8 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 5 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2420	22	2520	9	2543	10	2578	12
2434	18	2522	9	2546	10	2581	12
2443	16	2524	9	2549	10	2586	13
2451	15	2527	9	2552	10	2591	13
2458	14	2530	9	2556	10	2598	14
2464	13	2533	9	2559	11	2605	15
2469	12	2536	10	2563	11	2612	16
2474	12	2539	10	2567	11	2622	17
2478	11			2571	11	2629	18
2482	11						
2486	11						
2490	10						
2493	10						
2497	10						
2500	10						
2503	10						
2507	10						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2510	10						
2513	10						
2516	9						

Note: For Grade 5 ELA, Form 2 = writing prompt 13247 administered

Exhibit 4.6.2.9 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 6 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2444	18	2532	9	2554	10	2597	13
2454	16	2534	9	2557	10	2601	13
2462	15	2537	9	2560	10	2607	14
2469	14	2539	9	2564	10	2614	15
2474	13	2542	9	2567	10	2623	16
2479	12	2545	9	2570	10	2632	18
2484	12	2548	9	2574	11	2641	20
2489	11	2551	9	2578	11		
2493	11			2582	11		
2496	11			2586	12		
2500	10			2591	12		
2503	10						
2507	10						
2510	10						
2513	10						
2516	10						
2519	9						
2522	9						
2525	9						
2528	9						

Note: For Grade 6 ELA, Form 1 = writing prompt 13306 administered

Exhibit 4.6.2.10 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 6 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2431	22	2532	9	2553	9	2597	12
2445	18	2534	9	2555	10	2601	13
2454	16	2537	9	2558	10	2607	14
2462	15	2540	9	2561	10	2614	14

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2469	14	2543	9	2564	10	2621	15
2475	13	2546	9	2567	10	2630	17
2480	12	2549	9	2571	10	2640	18
2485	12			2575	11	2641	18
2489	11			2578	11		
2493	11			2582	11		
2497	11			2586	11		
2501	10			2591	12		
2504	10						
2507	10						
2511	10						
2514	10						
2517	10						
2520	9						
2523	9						
2526	9						
2529	9						

Note: For Grade 6 ELA, Form 2 = writing prompt 13307 administered

Exhibit 4.6.2.11 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 7 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2438	21	2543	10	2562	10	2600	12
2447	19	2546	10	2566	10	2603	13
2458	16	2549	10	2569	10	2609	13
2466	15	2553	10	2573	10	2615	14
2473	14	2556	10	2576	11	2623	15
2479	13	2559	10	2580	11	2631	16
2484	13			2584	11	2641	18
2489	12			2589	11	2648	19
2494	12			2593	12		
2498	11						
2502	11						
2506	11						
2510	10						
2514	10						
2517	10						
2521	10						
2524	10						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2527	10						
2530	10						
2534	10						
2537	10						
2540	10						

Note: For Grade 7 ELA, Form 1 = writing prompt 13401 administered

Exhibit 4.6.2.12 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 7 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2438	22	2544	10	2561	10	2600	12
2451	19	2548	10	2563	10	2605	13
2461	16	2551	10	2567	10	2611	14
2469	15	2554	10	2570	10	2618	15
2475	14	2557	10	2574	10	2626	16
2481	13			2578	11	2635	18
2487	12			2581	11	2646	19
2492	12			2586	11	2648	20
2496	11			2590	12		
2501	11			2595	12		
2505	11						
2508	11						
2512	10						
2516	10						
2519	10						
2522	10						
2526	10						
2529	10						
2532	10						
2535	10						
2538	10						
2541	10						

Note: For Grade 7 ELA, Form 2 = writing prompt 13406 administered

Exhibit 4.6.2.13 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 8 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2448	20	2551	9	2573	10	2605	11
2453	18	2553	9	2576	10	2610	12
2463	16	2556	9	2580	10	2615	12
2471	15	2559	9	2583	10	2620	13
2477	14	2562	9	2586	10	2626	13
2483	13	2564	9	2590	10	2632	14
2488	12	2567	9	2593	11	2639	15
2493	12	2570	9	2597	11	2648	17
2497	11			2601	11	2658	19
2501	11						
2505	11						
2509	10						
2512	10						
2516	10						
2519	10						
2522	10						
2525	9						
2528	9						
2531	9						
2534	9						
2537	9						
2539	9						
2542	9						
2545	9						
2548	9						

Note: For Grade 8 ELA, Form 1 = writing prompt 13439 administered

Exhibit 4.6.2.14 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 8 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2448	20	2551	9	2572	9	2604	11
2455	18	2554	9	2574	10	2607	12
2465	16	2557	9	2577	10	2611	12
2473	15	2560	9	2580	10	2616	12
2479	14	2562	9	2584	10	2622	13
2485	13	2565	9	2587	10	2627	14
2490	12	2568	9	2591	10	2634	14

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2495	12			2594	11	2641	15
2499	11			2598	11	2650	17
2503	11					2658	18
2507	10						
2511	10						
2514	10						
2517	10						
2520	10						
2523	10						
2526	9						
2529	9						
2532	9						
2535	9						
2538	9						
2541	9						
2543	9						
2546	9						
2549	9						

Note: For Grade 8 ELA, Form 2 = writing prompt 13454 administered

Exhibit 4.6.2.15 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 9 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2466	18	2556	9	2577	9	2606	11
2476	16	2558	9	2579	9	2608	11
2484	15	2561	9	2582	9	2612	11
2490	14	2563	9	2585	9	2617	12
2496	13	2566	9	2588	9	2622	12
2501	12	2569	9	2591	10	2627	13
2506	12	2571	9	2594	10	2633	14
2510	11	2574	9	2597	10	2640	15
2514	11			2601	10	2647	16
2518	10					2657	17
2521	10					2664	19
2525	10						
2528	10						
2531	10						
2534	9						
2537	9						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2540	9						
2543	9						
2545	9						
2548	9						
2551	9						
2553	9						

Note: For Grade 9 ELA, Form 1 = writing prompt 13555 administered

Exhibit 4.6.2.16 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 9 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2454	22	2556	9	2577	9	2606	11
2466	18	2558	9	2579	9	2608	11
2476	16	2561	9	2582	9	2613	11
2483	15	2563	9	2585	9	2617	12
2490	14	2566	9	2588	9	2622	12
2496	13	2569	9	2591	10	2627	13
2501	12	2571	9	2594	10	2633	14
2505	12	2574	9	2597	10	2640	15
2510	11			2601	10	2647	16
2514	11					2656	17
2517	10						
2521	10						
2524	10						
2528	10						
2531	10						
2534	9						
2537	9						
2540	9						
2542	9						
2545	9						
2548	9						
2550	9						
2553	9						

Note: For Grade 9 ELA, Form 2 = writing prompt 13556 administered

Exhibit 4.6.2.17 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 10 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2458	21	2567	9	2581	9	2608	10
2467	19	2569	9	2583	9	2611	11
2477	16	2572	9	2586	9	2615	11
2485	15	2575	9	2589	9	2619	11
2492	14	2577	9	2592	9	2624	12
2498	13			2595	10	2628	12
2503	12			2598	10	2633	13
2508	12			2601	10	2639	13
2513	11			2604	10	2645	14
2517	11					2652	15
2521	11					2660	17
2524	10					2668	18
2528	10						
2531	10						
2535	10						
2538	10						
2541	10						
2544	9						
2547	9						
2550	9						
2553	9						
2555	9						
2558	9						
2561	9						
2564	9						

Note: For Grade 10 ELA, Form 1 = writing prompt 13637 administered

Exhibit 4.6.2.18 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 10 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2458	20	2568	9	2582	9	2607	10
2465	19	2571	9	2585	9	2610	10
2475	17	2574	9	2588	9	2614	11
2484	15	2577	9	2591	9	2618	11
2491	14	2579	9	2594	10	2622	11
2497	13			2597	10	2627	12
2502	12			2600	10	2632	12

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2507	12			2603	10	2637	13
2512	11					2643	14
2516	11					2650	15
2520	11					2659	17
2524	10					2668	19
2527	10						
2531	10						
2534	10						
2537	10						
2540	10						
2543	9						
2546	9						
2549	9						
2552	9						
2555	9						
2557	9						
2560	9						
2563	9						
2566	9						

Note: For Grade 10 ELA, Form 2 = writing prompt 13638 administered

Exhibit 4.6.2.19 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 11 ELA, Form 1

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2465	20	2569	9	2585	10	2608	11
2469	19	2572	9	2587	10	2612	11
2479	16	2575	9	2590	10	2616	11
2487	15	2578	10	2593	10	2620	12
2494	14	2581	10	2597	10	2625	12
2500	13			2600	10	2630	12
2506	12			2604	11	2635	13
2511	12					2641	14
2515	11					2648	14
2520	11					2655	15
2524	11					2664	17
2527	11					2674	19
2531	10					2675	19
2535	10						
2538	10						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2541	10						
2544	10						
2548	10						
2551	10						
2554	10						
2557	9						
2560	9						
2563	9						
2566	9						

Note: For Grade 11 ELA, Form 1 = writing prompt 13720 administered

Exhibit 4.6.2.20 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 11 ELA, Form 2

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2465	19	2570	9	2585	10	2610	11
2476	17	2573	9	2588	10	2614	11
2484	15	2576	10	2592	10	2618	12
2491	14	2579	10	2595	10	2623	12
2498	13	2582	10	2599	10	2628	13
2503	13			2602	10	2633	13
2508	12			2606	11	2639	14
2513	12					2646	15
2517	11					2654	16
2521	11					2663	17
2525	11					2674	19
2529	10					2675	20
2533	10						
2536	10						
2539	10						
2543	10						
2546	10						
2549	10						
2552	10						
2555	10						
2558	9						
2561	9						
2564	9						
2567	9						

Note: For Grade 11 ELA, Form 2 = writing prompt 13721 administered

Exhibit 4.6.2.21 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 3 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3395	22	3495	10	3531	11	3573	14
3408	19	3498	10	3534	11	3580	15
3418	17	3501	10	3537	11	3588	17
3427	15	3505	10	3542	11	3598	19
3434	14	3508	10	3546	12	3605	20
3440	13	3512	10	3550	12		
3446	13	3515	10	3555	12		
3451	12	3519	10	3561	13		
3456	12	3522	10	3566	13		
3460	11	3526	11				
3465	11						
3469	11						
3473	11						
3476	11						
3480	11						
3484	10						
3487	10						
3491	10						

Exhibit 4.6.2.22 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 4 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3435	21	3532	10	3563	11	3608	13
3444	18	3535	10	3567	11	3614	14
3454	16	3538	10	3571	11	3621	15
3462	15	3542	10	3575	11	3629	16
3469	14	3545	10	3579	11	3639	19
3475	13	3549	10	3583	11	3645	20
3480	13	3552	10	3587	12		
3485	12	3556	10	3592	12		
3490	12	3559	10	3597	12		
3494	11			3602	13		
3499	11						
3503	11						
3507	11						
3510	11						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3514	10						
3518	10						
3521	10						
3525	10						
3528	10						

Exhibit 4.6.2.23 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 5 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3478	23	3563	10	3595	10	3635	12
3481	22	3565	10	3597	10	3638	12
3494	18	3568	10	3600	10	3644	13
3504	16	3572	10	3603	10	3650	14
3512	15	3575	10	3607	10	3656	15
3519	14	3578	10	3610	10	3664	16
3525	13	3581	10	3614	10	3674	18
3530	12	3584	10	3617	11	3687	22
3535	12	3587	10	3621	11	3688	22
3539	11	3590	10	3625	11		
3543	11			3629	11		
3547	11						
3551	11						
3555	10						
3558	10						

Exhibit 4.6.2.24 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 6 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3512	21	3602	10	3629	10	3663	11
3521	19	3606	10	3631	10	3668	12
3531	17	3609	10	3635	10	3672	12
3539	15	3612	10	3638	10	3677	12
3546	14	3615	10	3641	10	3683	13
3552	13	3619	10	3645	10	3689	14
3558	13	3622	10	3648	10	3696	15
3563	12	3625	10	3652	10	3704	16
3568	12			3655	11	3714	19

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3573	11			3659	11	3722	21
3577	11						
3581	11						
3585	11						
3588	11						
3592	10						
3596	10						
3599	10						

Exhibit 4.6.2.25 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 7 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3529	22	3629	10	3652	10	3680	11
3543	18	3632	10	3654	10	3685	12
3553	16	3635	10	3658	10	3689	12
3561	15	3638	10	3661	10	3694	13
3567	14	3641	10	3665	10	3700	13
3574	13	3644	10	3668	11	3706	14
3579	12	3648	10	3672	11	3713	15
3584	12			3676	11	3721	16
3589	12					3731	19
3593	11					3739	21
3597	11						
3601	11						
3605	11						
3608	10						
3612	10						
3615	10						
3619	10						
3622	10						
3625	10						

Exhibit 4.6.2.26 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 8 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3566	20	3650	10	3674	9	3705	11
3572	19	3654	10	3677	10	3707	11

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3582	16	3657	10	3681	10	3711	11
3590	15	3660	9	3684	10	3716	11
3597	14	3663	9	3687	10	3720	12
3603	13	3666	9	3690	10	3725	12
3608	12	3669	9	3693	10	3731	13
3613	12	3672	9	3697	10	3737	14
3618	12			3700	10	3744	15
3622	11					3752	16
3626	11					3762	19
3630	11					3776	22
3634	10						
3637	10						
3641	10						
3644	10						
3647	10						

Exhibit 4.6.2.27 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 9 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3578	22	3661	10	3681	10	3720	10
3592	19	3665	10	3684	10	3723	11
3602	16	3668	10	3687	10	3727	11
3610	15	3671	10	3690	10	3731	11
3617	14	3674	10	3693	10	3735	12
3623	13	3678	10	3696	10	3739	12
3629	12			3699	10	3744	12
3634	12			3702	10	3750	13
3639	12			3705	10	3756	14
3643	11			3709	10	3763	15
3647	11			3712	10	3771	16
3651	11			3715	10	3781	18
3654	10					3787	20
3658	10						

Exhibit 4.6.2.28 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 10 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3621	18	3675	11	3697	10	3744	10
3631	16	3679	11	3699	10	3748	11
3639	15	3683	10	3703	10	3752	11
3646	14	3686	10	3706	10	3756	11
3652	13	3690	10	3709	10	3760	12
3658	12	3693	10	3712	10	3765	12
3662	12			3715	10	3770	12
3667	11			3718	10	3775	13
3671	11			3721	10	3781	14
				3724	10	3788	15
				3728	10	3796	16
				3731	10	3806	18
				3734	10	3819	22
				3737	10		
				3741	10		

Exhibit 4.6.2.29 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2019 – Grade 11 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3629	21	3690	11	3711	10	3751	10
3639	18	3693	11	3714	10	3754	10
3649	16	3697	11	3717	10	3757	10
3657	15	3701	10	3720	10	3761	10
3664	14	3704	10	3723	10	3764	10
3670	13	3707	10	3726	10	3767	10
3675	12			3730	10	3771	11
3680	12			3733	9	3775	11
3685	11			3736	9	3779	11
				3739	10	3784	12
				3742	10	3788	12
				3745	10	3794	13
				3748	10	3799	14
						3806	15
						3814	16
						3824	18
						3837	22
						3839	23

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).[25] 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.

For a student with estimated ability $\hat{\theta}$ and associated standard error $\text{se}(\hat{\theta})$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $\text{se}(\hat{\theta})$, that is, $\hat{\theta} \sim N\left(\theta, \text{se}(\hat{\theta})^2\right)$. The probability of the true score at or above the cut score θ_c is estimated as

$$P(\theta \geq \theta_c) = P\left(\frac{\theta - \hat{\theta}}{\text{se}(\hat{\theta})} \geq \frac{\theta_c - \hat{\theta}}{\text{se}(\hat{\theta})}\right) = P\left(\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} < \frac{\hat{\theta} - \theta_c}{\text{se}(\hat{\theta})}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{\text{se}(\hat{\theta})}\right),$$

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

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

4.6.4 CLASSIFICATION ACCURACY

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

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

If a student's estimated 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.

The probability of a student with true ability θ being classified at or above the cut score θ_c, given the student's item scores $\boldsymbol{x} = (x_1, \cdots, x_N)$, can be estimated as

$$P(\theta \geq \theta_c | \boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x}) d\theta},$$

where the likelihood function is

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

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:

$$P(\theta < \theta_c | \boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x}) d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x}) d\theta}$$

Mathematically, we have

$$N_{11} = \sum\nolimits_{i \in N_1} P(\theta_i \geq \theta_c | \boldsymbol{x}),$$

$$N_{01} = \sum\nolimits_{i \in N_1} P(\theta_i < \theta_c | \boldsymbol{x}),$$

$$N_{10} = \sum\nolimits_{i \in N_0} P(\theta_i \geq \theta_c | \boldsymbol{x}), \text{ and}$$

$$N_{00} = \sum\nolimits_{i \in N_0} P(\theta_i < \theta_c | \boldsymbol{x}),$$

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:

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

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

To estimate the consistency, we assume the students are tested twice independently; hence, the probability of the student being classified as at or above the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

Similarly, the probability of consistency for at or above the cut score is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c|\boldsymbol{x}) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of consistency for below the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of inconsistency is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta \int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}, \text{ and}$$

$$P(\theta_1 < \theta_c, \theta_2 \geq \theta_c|\boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta \int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}.$$

The consistent index is computed as $\frac{N_{11}+N_{00}}{N}$, where

$$N_{11} = \sum_{i \in N} P(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c|\boldsymbol{x}),$$

$$N_{01} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} \geq \theta_c|\boldsymbol{x}),$$

$$N_{10} = \sum\nolimits_{i \in N} P(\theta_i \geq \theta_c, \theta_{i,2} < \theta_c | \boldsymbol{x}),$$

$$N_{00} = \sum\nolimits_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} < \theta_c | \boldsymbol{x}), \text{ and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

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**
			ELA			
3	0.92	0.91	0.94	0.88	0.88	0.92
4	0.92	0.91	0.94	0.89	0.88	0.92
5	0.94	0.92	0.93	0.92	0.89	0.90
6	0.92	0.92	0.96	0.88	0.88	0.95
7	0.92	0.91	0.96	0.88	0.88	0.94
8	0.93	0.93	0.95	0.90	0.90	0.93
9	0.92	0.92	0.95	0.89	0.89	0.93
10	0.92	0.92	0.95	0.89	0.89	0.93
11	0.92	0.92	0.95	0.88	0.89	0.93

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
			Mathematics			
3	0.95	0.92	0.94	0.93	0.89	0.91
4	0.94	0.93	0.95	0.92	0.90	0.92
5	0.94	0.93	0.95	0.91	0.90	0.93
6	0.93	0.94	0.95	0.91	0.91	0.94
7	0.93	0.94	0.95	0.91	0.91	0.93
8	0.93	0.94	0.96	0.90	0.92	0.95
Algebra I	0.92	0.93	0.95	0.89	0.90	0.94
Geometry	0.90	0.93	0.97	0.86	0.91	0.96
Algebra II	0.90	0.93	0.96	0.87	0.90	0.95

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
G3E	Overall	0.92	0.91	0.94	0.88	0.88	0.92
	Female	0.91	0.91	0.94	0.88	0.88	0.92
	Male	0.92	0.92	0.95	0.89	0.88	0.93
	African American	0.91	0.92	0.96	0.87	0.88	0.95
	Hispanic/Latino	0.91	0.92	0.96	0.87	0.88	0.94
	Asian	0.93	0.92	0.91	0.90	0.88	0.88
	White	0.92	0.91	0.92	0.89	0.87	0.89
	Hawaiian/Pacific	0.91	0.91	0.95	0.87	0.87	0.93
	American Indian	0.91	0.94	0.98	0.88	0.91	0.97
	Multiple Ethnicities	0.92	0.91	0.93	0.89	0.87	0.90
	LEP	0.93	0.96	0.99	0.91	0.95	0.99
	SPED	0.94	0.96	0.98	0.91	0.94	0.97
	FRL	0.91	0.92	0.96	0.87	0.88	0.95
	Accommodations	0.94	0.96	0.99	0.91	0.94	0.99
G4E	Overall	0.92	0.91	0.94	0.89	0.88	0.92
	Female	0.92	0.91	0.94	0.89	0.88	0.92
	Male	0.92	0.92	0.95	0.89	0.88	0.93
	African American	0.91	0.91	0.97	0.87	0.88	0.95
	Hispanic/ Latino	0.91	0.91	0.96	0.88	0.88	0.95
	Asian	0.95	0.93	0.91	0.93	0.90	0.87
	White	0.94	0.91	0.92	0.91	0.88	0.88
	Hawaiian/Pacific	0.92	0.92	0.95	0.89	0.89	0.93
	American Indian	0.91	0.92	0.98	0.87	0.89	0.97
	Multiple Ethnicities	0.93	0.91	0.93	0.90	0.88	0.90
	LEP	0.92	0.96	1.00	0.89	0.94	0.99

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	SPED	0.93	0.95	0.98	0.91	0.93	0.98
	FRL	0.91	0.91	0.96	0.87	0.88	0.95
	Accommodations	0.93	0.96	0.99	0.90	0.94	0.99
G5E	Overall	0.94	0.92	0.93	0.92	0.89	0.90
	Female	0.94	0.92	0.92	0.92	0.89	0.89
	Male	0.94	0.93	0.93	0.91	0.90	0.91
	African American	0.93	0.93	0.95	0.90	0.90	0.93
	Hispanic/ Latino	0.93	0.92	0.95	0.90	0.89	0.93
	Asian	0.96	0.94	0.89	0.95	0.91	0.85
	White	0.96	0.93	0.90	0.94	0.90	0.86
	Hawaiian/Pacific	0.95	0.92	0.93	0.93	0.89	0.90
	American Indian	0.92	0.93	0.97	0.89	0.90	0.96
	Multiple Ethnicities	0.95	0.92	0.92	0.93	0.89	0.88
	LEP	0.92	0.96	0.99	0.89	0.95	0.99
	SPED	0.94	0.96	0.98	0.92	0.95	0.98
	FRL	0.93	0.92	0.95	0.90	0.89	0.93
	Accommodations	0.93	0.96	0.99	0.90	0.95	0.99
G6E	Overall	0.92	0.92	0.96	0.88	0.88	0.95
	Female	0.92	0.91	0.96	0.88	0.88	0.94
	Male	0.92	0.92	0.97	0.88	0.89	0.96
	African American	0.91	0.93	0.98	0.88	0.90	0.97
	Hispanic/ Latino	0.90	0.92	0.98	0.87	0.89	0.97
	Asian	0.95	0.91	0.92	0.92	0.88	0.90
	White	0.93	0.91	0.94	0.91	0.87	0.92
	Hawaiian/Pacific	0.92	0.90	0.97	0.88	0.87	0.96
	American Indian	0.90	0.94	0.99	0.86	0.91	0.99
	Multiple Ethnicities	0.93	0.91	0.95	0.90	0.88	0.94
	LEP	0.92	0.97	1.00	0.88	0.96	1.00
	SPED	0.93	0.97	0.99	0.91	0.96	0.99
	FRL	0.90	0.92	0.98	0.87	0.89	0.97
	Accommodations	0.93	0.97	1.00	0.90	0.96	1.00
G7E	Overall	0.92	0.91	0.96	0.88	0.88	0.94
	Female	0.92	0.91	0.95	0.88	0.87	0.94
	Male	0.92	0.92	0.96	0.89	0.89	0.95
	African American	0.91	0.92	0.98	0.88	0.89	0.97
	Hispanic/Latino	0.91	0.92	0.97	0.87	0.89	0.96
	Asian	0.95	0.91	0.92	0.92	0.88	0.89
	White	0.93	0.90	0.94	0.90	0.87	0.91
	Hawaiian/Pacific	0.91	0.90	0.95	0.88	0.87	0.94
	American Indian	0.91	0.94	0.99	0.87	0.91	0.98
	Multiple Ethnicities	0.92	0.90	0.94	0.89	0.87	0.92
	LEP	0.94	0.98	1.00	0.92	0.97	1.00

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

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

Note: Hawaiian/Pacific = Native Hawaiian/Pacific Islander; American Indian = American Indian or Alaskan; LEP = Limited English Proficiency; SPED = Special Education; FRL = Free or Reduced-Price 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
G3M	Overall	0.95	0.92	0.94	0.93	0.89	0.91
	Female	0.94	0.92	0.94	0.92	0.89	0.92
	Male	0.95	0.93	0.93	0.93	0.90	0.91
	African American	0.94	0.92	0.96	0.91	0.90	0.95
	Hispanic/Latino	0.94	0.92	0.95	0.91	0.89	0.94
	Asian	0.97	0.94	0.89	0.96	0.92	0.85
	White	0.96	0.92	0.91	0.95	0.89	0.87
	Hawaiian/Pacific	0.95	0.91	0.94	0.92	0.89	0.91
	American Indian	0.92	0.94	0.98	0.89	0.91	0.97
	Multiple Ethnicities	0.96	0.92	0.92	0.94	0.89	0.89
	LEP	0.92	0.95	0.99	0.89	0.94	0.98
	SPED	0.94	0.95	0.98	0.92	0.94	0.97
	FRL	0.94	0.92	0.96	0.91	0.89	0.94
	Accommodations	0.93	0.96	0.99	0.91	0.94	0.98
G4M	Overall	0.94	0.93	0.95	0.92	0.90	0.92
	Female	0.94	0.92	0.95	0.92	0.89	0.93
	Male	0.95	0.93	0.95	0.92	0.90	0.92

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	African American	0.93	0.93	0.97	0.90	0.91	0.96
	Hispanic/Latino	0.93	0.93	0.96	0.91	0.90	0.95
	Asian	0.97	0.94	0.90	0.96	0.91	0.86
	White	0.96	0.93	0.92	0.94	0.89	0.89
	Hawaiian/Pacific	0.94	0.92	0.94	0.91	0.90	0.92
	American Indian	0.93	0.94	0.98	0.89	0.91	0.97
	Multiple Ethnicities	0.95	0.92	0.94	0.93	0.89	0.91
	LEP	0.92	0.96	0.99	0.89	0.94	0.99
	SPED	0.94	0.96	0.98	0.91	0.94	0.98
	FRL	0.93	0.93	0.96	0.91	0.90	0.95
	Accommodations	0.94	0.96	0.99	0.91	0.94	0.98
G5M	Overall	0.94	0.93	0.95	0.91	0.90	0.93
	Female	0.94	0.92	0.95	0.91	0.90	0.93
	Male	0.94	0.93	0.95	0.91	0.91	0.93
	African American	0.92	0.94	0.97	0.89	0.92	0.97
	Hispanic/Latino	0.93	0.93	0.96	0.90	0.90	0.95
	Asian	0.97	0.94	0.90	0.96	0.92	0.87
	White	0.95	0.92	0.93	0.93	0.90	0.90
	Hawaiian/Pacific	0.95	0.91	0.95	0.93	0.88	0.93
	American Indian	0.92	0.94	0.98	0.88	0.92	0.97
	Multiple Ethnicities	0.94	0.92	0.94	0.92	0.90	0.92
	LEP	0.91	0.96	0.99	0.88	0.95	0.99
	SPED	0.93	0.97	0.99	0.91	0.95	0.98
	FRL	0.92	0.93	0.97	0.90	0.91	0.95
	Accommodations	0.93	0.97	0.99	0.90	0.95	0.99
G6M	Overall	0.93	0.94	0.95	0.91	0.91	0.94
	Female	0.93	0.93	0.95	0.90	0.91	0.94
	Male	0.94	0.94	0.95	0.91	0.92	0.94
	African American	0.93	0.95	0.97	0.90	0.93	0.96
	Hispanic/Latino	0.93	0.94	0.97	0.90	0.92	0.95
	Asian	0.96	0.94	0.92	0.95	0.92	0.89
	White	0.94	0.93	0.93	0.92	0.90	0.91
	Hawaiian/Pacific	0.92	0.93	0.96	0.89	0.91	0.95
	American Indian	0.92	0.95	0.98	0.89	0.93	0.97
	Multiple Ethnicities	0.94	0.93	0.94	0.91	0.91	0.92
	LEP	0.93	0.97	0.99	0.91	0.96	0.99
	SPED	0.95	0.97	0.99	0.93	0.97	0.99
	FRL	0.93	0.94	0.97	0.90	0.92	0.96
	Accommodations	0.95	0.98	0.99	0.92	0.97	0.99
G7M	Overall	0.93	0.94	0.95	0.91	0.91	0.93
	Female	0.93	0.93	0.95	0.90	0.91	0.93
	Male	0.94	0.94	0.95	0.91	0.92	0.93

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	African American	0.93	0.95	0.97	0.91	0.93	0.96
	Hispanic/Latino	0.93	0.94	0.97	0.90	0.92	0.95
	Asian	0.96	0.93	0.93	0.94	0.91	0.89
	White	0.94	0.93	0.93	0.91	0.90	0.90
	Hawaiian/Pacific	0.94	0.95	0.94	0.91	0.92	0.92
	American Indian	0.93	0.95	0.98	0.91	0.94	0.97
	Multiple Ethnicities	0.94	0.93	0.94	0.91	0.90	0.92
	LEP	0.96	0.98	0.99	0.94	0.97	0.99
	SPED	0.96	0.98	0.99	0.94	0.97	0.99
	FRL	0.93	0.94	0.97	0.90	0.92	0.95
	Accommodations	0.96	0.98	0.99	0.94	0.98	0.99
G8M	Overall	0.93	0.94	0.96	0.90	0.92	0.95
	Female	0.93	0.94	0.96	0.90	0.91	0.95
	Male	0.93	0.95	0.96	0.91	0.92	0.95
	African American	0.92	0.95	0.98	0.89	0.93	0.97
	Hispanic/Latino	0.93	0.95	0.97	0.90	0.93	0.96
	Asian	0.95	0.94	0.93	0.93	0.91	0.90
	White	0.93	0.93	0.94	0.91	0.90	0.92
	Hawaiian/Pacific	0.94	0.92	0.95	0.91	0.89	0.93
	American Indian	0.93	0.96	0.98	0.90	0.95	0.98
	Multiple Ethnicities	0.93	0.94	0.95	0.90	0.91	0.94
	LEP	0.95	0.98	0.99	0.93	0.97	0.99
	SPED	0.96	0.98	0.99	0.94	0.98	0.99
	FRL	0.93	0.95	0.97	0.89	0.93	0.96
	Accommodations	0.96	0.99	1.00	0.94	0.98	0.99
Algebra I	Overall	0.92	0.93	0.95	0.89	0.90	0.94
	Female	0.92	0.92	0.95	0.88	0.89	0.94
	Male	0.92	0.93	0.96	0.89	0.91	0.94
	African American	0.91	0.93	0.98	0.87	0.91	0.97
	Hispanic/Latino	0.91	0.93	0.97	0.87	0.90	0.96
	Asian	0.96	0.94	0.91	0.94	0.92	0.88
	White	0.93	0.92	0.94	0.91	0.90	0.91
	Hawaiian/Pacific	0.91	0.92	0.96	0.88	0.89	0.94
	American Indian	0.90	0.94	0.98	0.86	0.92	0.98
	Multiple Ethnicities	0.93	0.92	0.95	0.90	0.89	0.93
	LEP	0.92	0.96	0.99	0.88	0.94	0.99
	SPED	0.93	0.97	0.99	0.90	0.96	0.99
	FRL	0.91	0.93	0.97	0.88	0.90	0.96
	Accommodations	0.93	0.97	0.99	0.90	0.96	0.99
Geometry	Overall	0.90	0.93	0.97	0.86	0.91	0.96
	Female	0.90	0.93	0.97	0.85	0.90	0.96
	Male	0.90	0.93	0.97	0.86	0.91	0.96

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

4.6.7 RELIABILITY FOR SUBGROUPS IN THE POPULATION

Exhibits 4.6.7.1 and 4.6.7.2 show the 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[26] , FRL, and accommodations). As the exhibits indicate, 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	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
3	Reliability	0.90	0.89	0.90	0.88	0.89	0.89	0.88	0.86	0.89	0.89	0.87	0.79	0.88	0.82
	Variance	979	957	993	844	1041	898	852	675	937	946	789	490	830	586
4	Reliability	0.90	0.90	0.90	0.89	0.89	0.90	0.89	0.87	0.89	0.89	0.88	0.80	0.89	0.84
	Variance	1048	1022	1054	896	1081	1061	894	743	987	1025	855	487	871	630
5	Reliability	0.91	0.91	0.92	0.91	0.89	0.91	0.91	0.90	0.90	0.91	0.89	0.84	0.91	0.87
	Variance	1416	1345	1449	1306	1357	1338	1269	1055	1280	1365	1106	699	1248	846
6	Reliability	0.90	0.89	0.90	0.89	0.89	0.89	0.89	0.86	0.89	0.90	0.84	0.78	0.88	0.80
	Variance	1042	999	1053	912	1136	893	874	694	1015	1070	671	454	863	535
7	Reliability	0.90	0.89	0.90	0.89	0.89	0.88	0.89	0.87	0.89	0.89	0.85	0.80	0.89	0.82
	Variance	1189	1116	1217	1050	1255	1004	1031	841	1133	1156	801	587	1003	650
8	Reliability	0.92	0.91	0.92	0.91	0.91	0.92	0.91	0.89	0.91	0.92	0.87	0.84	0.91	0.83
	Variance	1289	1217	1288	1124	1336	1244	1140	916	1192	1243	805	647	1115	639
9	Reliability	0.90	0.90	0.90	0.89	0.90	0.91	0.89	0.86	0.90	0.90	0.83	0.83	0.89	0.85
	Variance	987	947	984	817	1171	1059	825	640	985	969	564	584	833	658
10	Reliability	0.91	0.90	0.91	0.90	0.90	0.90	0.89	0.87	0.90	0.90	0.86	0.86	0.90	0.81
	Variance	1019	956	1048	998	1031	1013	913	762	951	997	761	791	935	604
11	Reliability	0.90	0.89	0.90	0.89	0.90	0.89	0.88	0.85	0.90	0.90	0.84	0.86	0.89	0.83
	Variance	1072	993	1120	944	1125	959	883	697	1095	1126	731	794	923	698

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

[26] 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: Mathematics

Grade	Statistic	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	0.92	0.93	0.87	0.92	0.92	0.92	0.90	0.91	0.93	0.91	0.92	0.92
	Variance	1985	1857	2107	1935	1643	1876	1795	1643	1795	1860	2192	1440	1826	1748
4	Reliability	0.93	0.93	0.93	0.93	0.89	0.93	0.93	0.92	0.92	0.93	0.93	0.91	0.93	0.92
	Variance	2070	1928	2205	1979	1712	1982	1882	1700	1856	2035	2160	1517	1893	1853
5	Reliability	0.92	0.92	0.93	0.92	0.90	0.92	0.92	0.90	0.92	0.92	0.90	0.88	0.92	0.88
	Variance	1817	1694	1932	1647	1763	1789	1601	1400	1709	1809	1590	1175	1596	1337
6	Reliability	0.93	0.93	0.93	0.92	0.92	0.93	0.92	0.90	0.93	0.93	0.88	0.86	0.92	0.87
	Variance	1942	1768	2108	1657	2056	1787	1669	1381	1839	1964	1437	1095	1649	1236
7	Reliability	0.93	0.92	0.93	0.91	0.91	0.93	0.92	0.90	0.92	0.93	0.87	0.83	0.92	0.83
	Variance	1866	1745	1983	1547	1963	1918	1555	1294	1758	1818	1147	876	1540	919
8	Reliability	0.92	0.92	0.92	0.90	0.93	0.92	0.91	0.88	0.92	0.92	0.82	0.83	0.91	0.79
	Variance	1591	1470	1706	1278	2058	1432	1375	1066	1576	1609	864	875	1348	724
Algebra I	Reliability	0.91	0.90	0.92	0.88	0.91	0.90	0.89	0.86	0.91	0.91	0.82	0.82	0.90	0.80
	Variance	1378	1257	1487	1057	1427	1278	1146	918	1387	1388	807	787	1229	739
Geo	Reliability	0.90	0.90	0.91	0.84	0.93	0.90	0.87	0.83	0.91	0.92	0.76	0.79	0.87	0.67
	Variance	1401	1324	1476	948	1894	1294	1073	875	1495	1649	731	766	1080	551
Algebra II	Reliability	0.91	0.90	0.91	0.87	0.93	0.91	0.88	0.83	0.92	0.91	0.81	0.85	0.87	0.73
	Variance	1511	1385	1641	1154	1927	1456	1191	906	1615	1561	978	1107	1177	687

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

4.6.8 SUBSCALE RELIABILITY

Reliability estimates associated with the subscales for the 2019 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	0.74	0.74	0.78
Grade 4	0.75	0.75	0.78
Grade 5	0.78	0.79	0.76
Grade 6	0.77	0.73	0.75
Grade 7	0.79	0.73	0.73
Grade 8	0.79	0.78	0.80
Grade 9	0.78	0.73	0.79

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 10	0.79	0.76	0.77
Grade 11	0.79	0.71	0.74

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	0.67	0.75	0.85
Grade 4	0.78	0.68	0.87
Grade 5	0.77	0.75	0.84

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	0.81	0.77	0.72	0.61
Grade 7	0.75	0.70	0.73	0.76

Exhibit 4.6.8.4 Subscale Reliabilities: Mathematics Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability & the Number System
Grade 8	0.80	0.67	0.56	0.73

Exhibit 4.6.8.5 Subscale Reliabilities: Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.81	0.79	0.65
Algebra II	0.78	0.70	0.75

Exhibit 4.6.8.6 Subscale Reliabilities: Geometry

	Circles, Geometric Measurement & Dimension, and Modeling	Congruence	Geometric Properties with Equations	Similarity, Right Triangles & Trigonometry
Geometry	0.42	0.70	0.57	0.72

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.[27] 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	0.70		0.95	
	Writing & Language	0.65	0.65	0.86	0.86
4	Literature	0.74		0.98	
	Writing & Language	0.68	0.68	0.88	0.89
5	Literature	0.78		0.99	
	Writing & Language	0.68	0.68	0.89	0.88
6	Literature	0.73		0.97	
	Writing & Language	0.66	0.64	0.86	0.86
7	Literature	0.73		0.96	
	Writing & Language	0.67	0.65	0.89	0.89
8	Literature	0.76		0.96	
	Writing & Language	0.71	0.70	0.89	0.88
9	Literature	0.73		0.96	
	Writing & Language	0.65	0.65	0.83	0.86
10	Literature	0.74		0.96	
	Writing & Language	0.66	0.64	0.85	0.84
11	Literature	0.72		0.97	
	Writing & Language	0.67	0.65	0.88	0.89

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

Exhibit 4.7.2 Subscale Intercorrelations: Mathematics Grades 3–5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.74		1.00	
	OAT_NBT	0.75	0.81	0.94	1.00
4	MDG	0.71		0.98	
	OAT_NBT	0.77	0.77	1.00	1.00
5	MDG	0.74		0.97	
	OAT_NBT	0.80	0.76	1.00	0.96

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

Exhibit 4.7.3 Subscale Intercorrelations: Mathematics Grade 6 & 7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.81			1.00		
	RP	0.78	0.79		1.00	1.00	
	GSP	0.72	0.73	0.68	1.00	1.00	1.00
7	NS	0.78			1.00		
	RP	0.80	0.77		1.00	1.00	
	GSP	0.76	0.75	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)	0.76			1.00		
	Geometry(G)	0.71	0.64		1.00	1.00	
	SPNS	0.79	0.71	0.66	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		0.99	
	Statistics	0.73	0.72	1.00	1.00
Algebra II	Functions	0.77		1.00	
	Statistics	0.76	0.73	1.00	1.00

Exhibit 4.7.6 Subscale Intercorrelations and Reliability Estimates: Geometry

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	GP	CGM_GPE	C	GP
Geometry	C	0.68			1.00		
	GP	0.69	0.71		1.00	1.00	
	SRTT	0.70	0.73	0.73	1.00	1.00	1.00

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

4.8 HANDSCORING AGREEMENT RATE

For grades in which statistical models were constructed for machine scoring of essay responses, Measurement, Inc. (MI) handscored over 4,100 responses per prompt, with each response double scored and any discrepant scores routed for a final resolution score. At each grade, students responded to one of two randomly selected writing tasks. Exhibit 4.8.1 shows the summary of the rater agreement for the writing prompts administered on the AzMERIT spring 2019 online tests. The rater agreement reports show percentages of exact agreement (Equal), adjacent scores (Adj. Low or Adj. High), and nonadjacent scores (Non-Adj Low or Non-Adj High). The tables also identify mismatched scores when there is a difference involving nonscorable condition codes (Mismatch NS), or a nonscorable/scorable mix (MM NS/Score). Exhibit 4.8.1 provides a summary of those results, showing the mean exact agreement rate for dimension scores across grades. Generally exact agreement rates ranged from 65%–70%, with little variability across the essay prompts.

Exhibit 4.8.1 ELA Writing Prompt Rater Agreement Report: Spring 2019 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	10,303	1,774	0.9	19.5	57.5	19.5	0.9	0.0	1.7
	Evidence/Elaboration	10,305	1,774	1.0	18.1	60.1	18.1	1.0	0.0	1.7
	Conventions	10,565	1,774	0.2	15.5	67.0	15.5	0.2	0.0	1.7
4	Purpose/Organization	10,646	1,898	0.5	16.0	66.9	16.0	0.5	0.0	0.0
	Evidence/Elaboration	10,647	1,898	0.6	16.1	66.7	16.1	0.6	0.0	0.0
	Conventions	10,998	1,898	1.2	17.9	61.9	17.9	1.2	0.0	0.0
5	Purpose/Organization	10,856	1,950	0.2	16.4	66.9	16.4	0.2	0.0	0.0
	Evidence/Elaboration	10,861	1,950	0.5	16.4	66.4	16.4	0.5	0.0	0.0
	Conventions	11,159	1,950	0.5	15.3	68.5	15.3	0.5	0.0	0.0
6	Purpose/Organization	11,205	1,996	1.4	20.3	55.7	20.3	1.4	0.0	0.9
	Evidence/Elaboration	11,210	1,996	1.7	20.1	55.5	20.1	1.7	0.0	0.9
	Conventions	11,434	1,996	0.7	12.2	73.3	12.2	0.7	0.0	0.9
7	Purpose/Organization	10,052	1,792	1.6	19.3	58.3	19.3	1.6	0.0	0.0
	Evidence/Elaboration	10,063	1,792	2.2	20.0	55.6	20.0	2.2	0.0	0.0
	Conventions	10,336	1,792	0.2	17.3	65.1	17.3	0.2	0.0	0.0
8	Purpose/Organization	9,286	1,664	1.7	21.0	54.6	21.0	1.7	0.0	0.0
	Evidence/Elaboration	9,286	1,664	1.7	21.5	53.5	21.5	1.7	0.0	0.0
	Conventions	9,492	1,664	1.0	13.2	71.8	13.2	1.0	0.0	0.0
9	Purpose/Organization	6,262	1,102	0.5	16.4	65.3	16.4	0.5	0.0	0.9
	Evidence/Elaboration	6,264	1,102	0.6	16.0	65.9	16.0	0.6	0.0	0.9
	Conventions	6,395	1,102	0.5	12.1	74.0	12.1	0.5	0.0	0.9
10	Purpose/Organization	5,206	932	1.0	19.4	59.0	19.4	1.0	0.0	0.2
	Evidence/Elaboration	5,210	932	1.4	18.1	60.7	18.1	1.4	0.0	0.2
	Conventions	5,337	932	0.5	14.5	69.7	14.5	0.5	0.0	0.2
11	Purpose/Organization	4,610	818	0.6	20.0	58.7	20.0	0.6	0.0	0.0
	Evidence/Elaboration	4,613	818	1.0	19.3	59.4	19.3	1.0	0.0	0.0
	Conventions	4,716	818	1.0	12.6	72.9	12.6	1.0	0.0	0.0

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.[28]

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 2019 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),

[28] 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[29]

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.[30] 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 specific 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.

[29] Standard 4.7: The procedures used to develop, review, and try out items and to select items from the item pool should be documented.
[30] 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.[31] Items used to develop the spring 2019 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, spring 2017, and spring 2018, and subsequent operational test administration in spring 2017, spring 2018, and spring 2019. 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

[31] 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.

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 it meets the following criteria:

- 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.
- 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, spring 2018, and spring 2019 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:[32]

- 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 to 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.

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 2019 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.[33] 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 subsgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to

[32] 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.

[33] 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.

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 G 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 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 .05.

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 was between .23 and .27. Items with biserials less than .23 were automatically rejected.

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 0. 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-Price Lunch/Free or Reduced-Price 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 ($\Delta_{\text{hat}\ MH}$) 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* χ^2 is significant and $\lvert\Delta_{\text{hat } MH}\rvert \geq 1.5$
	B	*MH* χ^2 is significant and $\lvert\Delta_{\text{hat } MH}\rvert < 1.5$
	A	*MH* χ^2 is not significant
Polytomous Items	C	*MH* χ^2 is significant and $\lvert SMD\rvert / \lvert SD\rvert \geq .25$
	B	*MH* χ^2 is significant and $\lvert SMD\rvert / \lvert SD\rvert < .25$
	A	*MH* χ^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.[34]

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.

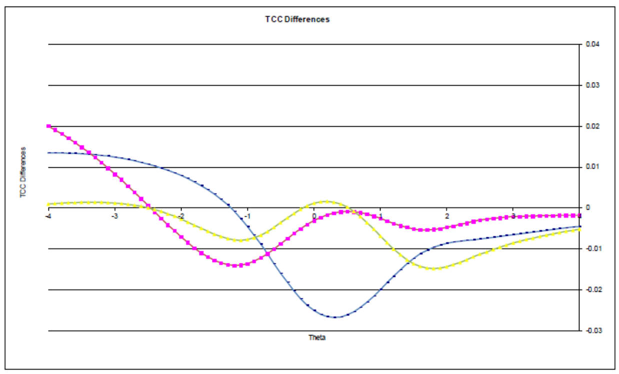
Immediately upon generation of a test form, Form Builder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, Form Builder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. 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,

[34] 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.

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 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 bookmaps (test maps) for use in scoring, forms development, and other processes.

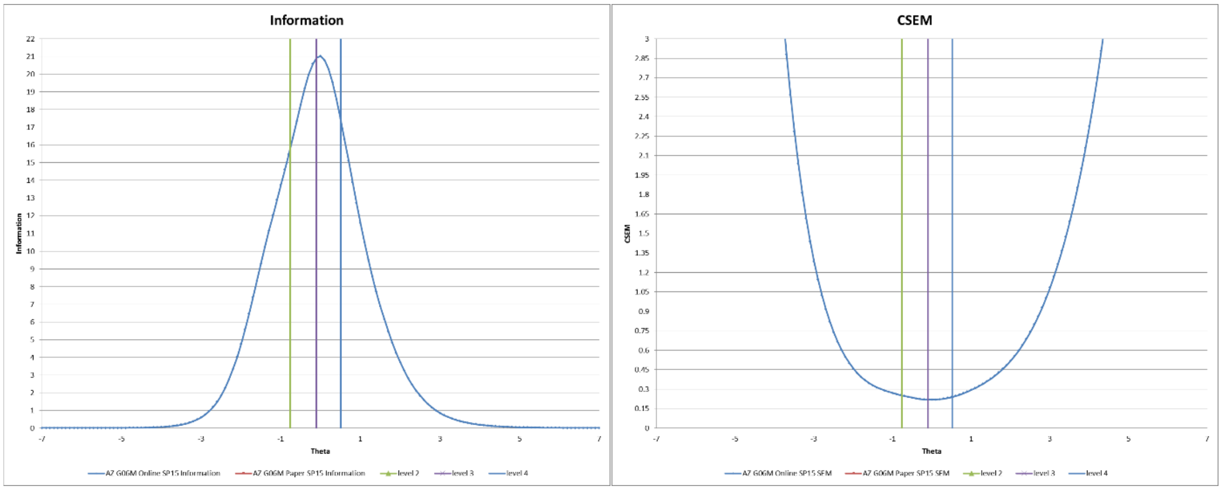
Exhibit 5.5.1.1 Test Characteristics Curve Differences



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 2019 operational test forms are presented in Appendix H.

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 SEMs were evaluated to ensure equivalence between online and paper-pencil test forms.

Exhibit 5.5.1.2 Test Information and Standard Errors Relative to Performance Standards



5.5.2 TEST INFORMATION FUNCTION

Test information function is particularly important and useful in operational testing because it provides information about the precision with which each person's ability measure is estimated. Larger amounts of test information are associated with greater measurement precision. For a set of items that appears on an operational test form, test information can be computed from the item difficulty estimates of these items as a function of student ability. Unlike classical test theory, in which measurement precision is assumed to be the same across all scores, precision in Rasch measurement is conditioned on each score along the ability continuum. The conditional standard error of measurement (CSEM) is calculated as the reciprocal of the square root of the test information function, and thus the CSEM is lowest when information is highest. In a fixed-length test format, ability levels around both ends of the continuum are measured with less precision because there are usually fewer items targeting the levels around both extremes, while ability levels around the middle of the continuum are measured with greater precision because generally more items are developed for these levels.

Test information function (TIF) may be presented as follows:

$$T(\theta) = \sum_{i=1}^{k} p_i(\theta) \times (1 - p_i(\theta)),$$

where $T(\theta)$ is the test information across *k* operational items at a given ability θ, and $p_i(\theta)$ refers to the probability of correct response to item *i* conditioned on the ability θ.

To better depict measurement error at various points along the scale, which is congruent with the *Standards for Educational and Psychological Testing,* the graphs and the values of test information function (TIF) for the spring 2018 online forms and the spring 2019 online forms are presented in Appendix I. Additionally, the graph and the values of the ratio for information function between the spring 2018 online forms and the spring 2019 online forms are presented in Appendix I.

5.5.3 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.[35] 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 as 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 (DTCs), School Test Coordinators (STCs), 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 DTC 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 Administrator 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.[36] DTCs and STCs were responsible for ensuring that all test administration personnel (for both PBT and CBT) were properly trained prior to the start of testing using the various resources.

---

[35] 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.

[36] 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

Manuals and guides on test administrations are available on the AzMERIT Portal.[37] 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 information on the following topics:

- Steps to take prior to accessing the system and logging in
- Navigating the TA Interface
- 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.[38] 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

---

administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

[37] 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.

[38] 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.

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.

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.*[39] 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.[40]

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.

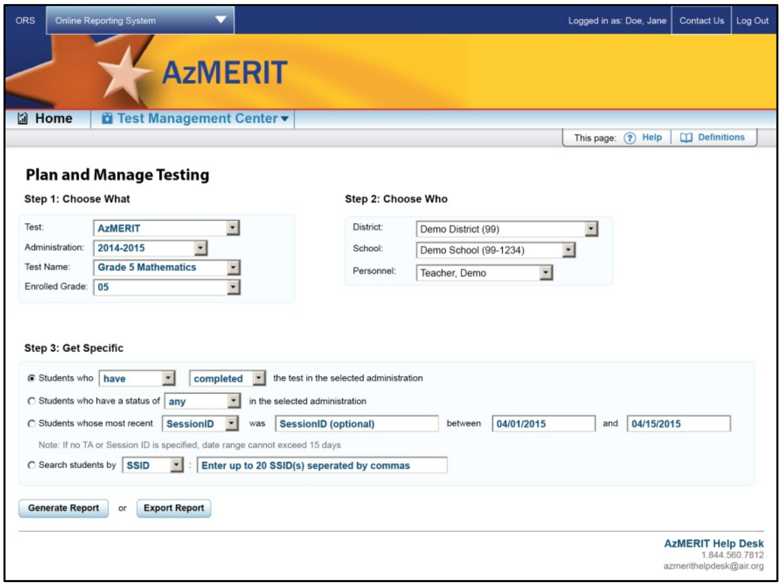
The School Test Coordinator and TAs worked together to determine the most appropriate testing option(s), testing environment, and the average time needed to complete each test. The appropriate protocols were established to maintain a quiet testing environment throughout the testing session. TAs also 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.

[39] 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.

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

6.2.1 MANAGING TESTING

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



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.[41] 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

[41] Standard 3.4: Test takers should receive comparable treatment during the test administration and scoring process.
[41] 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.
[41] Standard 6.4: The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

- 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
- 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 competed 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
Area Boundaries	Click anywhere on the selected-response text or button for multiple-choice options
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen *Test Instructions and Help*.
Highlighter	Highlight text in a passage or item.
Line Reader	This allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	This allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In mathematics, comments are attached to a specific test item and available throughout the session.
Pause and Restart	This allows the session to be paused at any time and restarted and taken over a one-day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	This allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.

Universal Test Tool	Description
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended-response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

6.3.3 SUBJECT-AREA TOOLS FOR COMPUTER-BASED AND PAPER-BASED TESTING

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	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. PBT: Schools must make published, paper dictionaries and thesauruses available to students. Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned off.
Writing Guide	Writing	CBT: Students have access to 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 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. PBT: Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.

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.[42] 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.[43]

ACCOMMODATIONS FOR STUDENTS WITH AN INJURY

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

Exhibit 6.3.4.1 Accommodations for Students with an Injury

Accommodation	Description
Adult Transcription	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. 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.

[42] 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.
[43] 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.

Accommodation	Description
Assistive Technology	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. This accommodation also requires Adult Transcription (see above for rules on Adult Transcription).
Rest/Breaks	Students may take breaks during testing sessions to rest.

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

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

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

Exhibit 6.3.4.3 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT mathematics test.
Adult Transcription	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. 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.
ASL and Closed Caption	In CBTs, this is available for the listening items on the Reading ELA test.
Assistive Technology	This is the use of assistive technology for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Braille Test Booklet	Provide a paper braille test booklet. This accommodation requires Adult Transcription (see above for rules on Adult Transcription).
Large Print Test Booklet	CBT: Either increase default zoom settings 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.
Paper-Pencil Test Booklet	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.
Read Aloud Test Content	CBT: Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the mathematics test. PBT: Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test. Reading aloud the content of the Reading portion of the ELA test.
Rest/Breaks	Provide students with breaks during testing sessions to rest.
Sign Test Content	Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the mathematics test. Signing the content of the Reading portion of the ELA test.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

6.4 SYSTEM SECURITY

6.4.1 SECURE SYSTEM DESIGN

AIR has developed a custom single sign-on application that is made available on Arizona's secure portal. This application is used to support access to AIR's system in accordance with the Arizona's user ID and password policy. Authorized users can log in to Arizona's single sign-on using their current user IDs and passwords and can be redirected to AIR's portal, where they have access to AIR's secure applications, such as 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.[44] 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.

---

[44] 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.*[45]

[45] Standard 6.7: Test users are responsible for 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’s TIDE. The ADE made the final decision on whether to approve the requested test invalidation. District Test Coordinators could track the status and final decisions of requested test invalidations in TIDE.

6.6 DATA FORENSICS PROGRAM

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

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

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 $var(e_i) = \sigma^2(1 - h_{ii})$. 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

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

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).

6.6.4 RESPONSE CHANGE AND RESPONSE SIMILARITY

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

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

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_{ti} 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 $\hat{\sigma}^2_{jk}$, 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
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 2018–2019 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

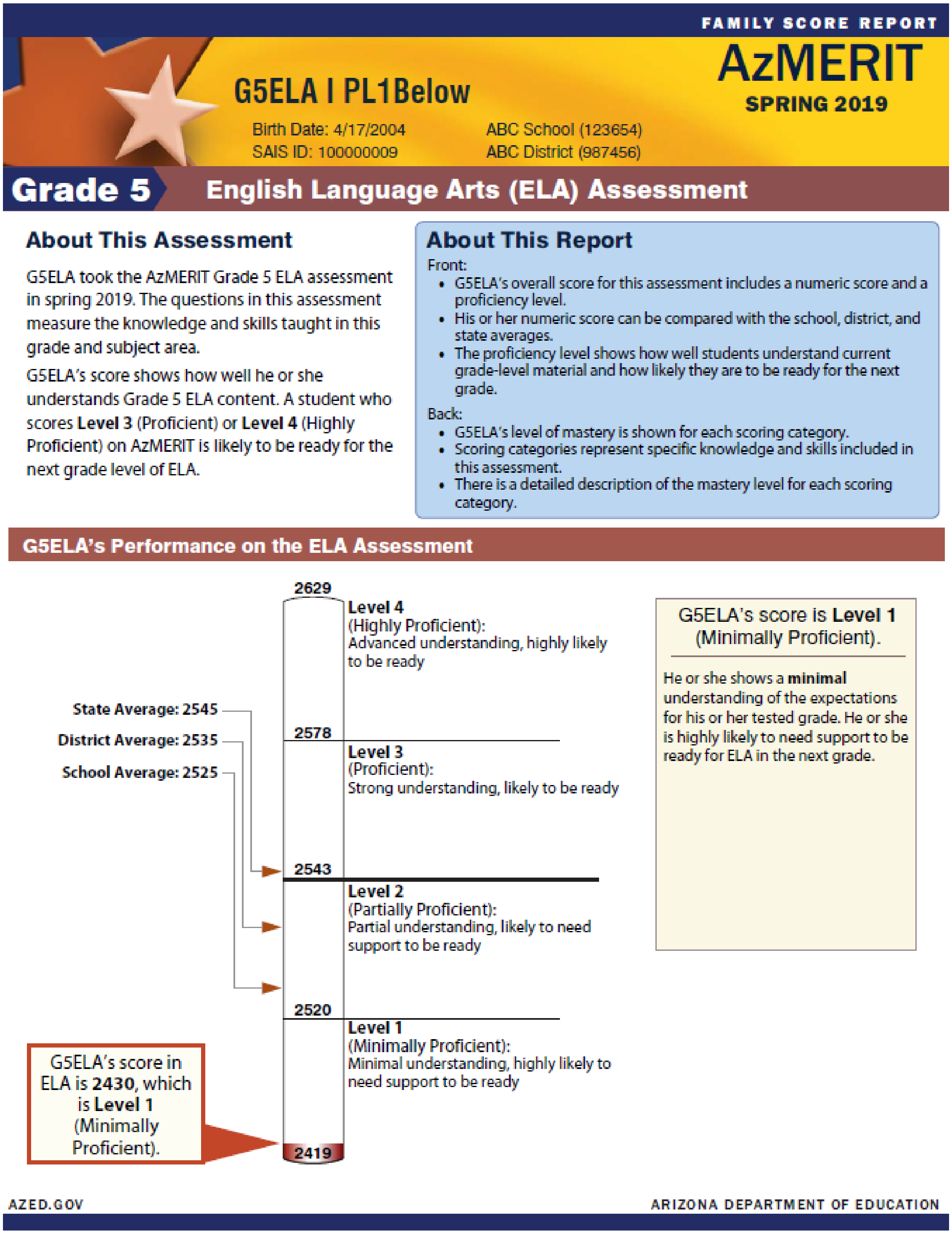
The state provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning and classroom instruction. All reporting systems for the AzMERIT, both paper-pencil and online, are designed with stakeholders in mind–such as teachers, 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.[46] 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.

[46] 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

AzMERIT

SPRING 2019

G5ELA I PL1Below

Birth Date: 4/17/2004
SAIS ID: 100000009

ABC School (123654)
ABC District (987456)

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

About This Assessment

G5ELA took the AzMERIT Grade 5 ELA assessment in spring 2019. The questions in this assessment measure the knowledge and skills taught in this grade and subject area.

G5ELA's score shows how well he or 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:

- G5ELA's overall score for this assessment includes a numeric score and a proficiency level.
- His or 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:

- G5ELA'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.

G5ELA's Performance on the ELA Assessment

G5ELA's score is **Level 1** (Minimally Proficient).

He or she shows a **minimal** understanding of the expectations for his or her tested grade. He or she is highly likely to need support to be ready for ELA in the next grade.

AZED.GOV

ARIZONA DEPARTMENT OF EDUCATION

Spring 2019 987456-9

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:
    - 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; and
- 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.

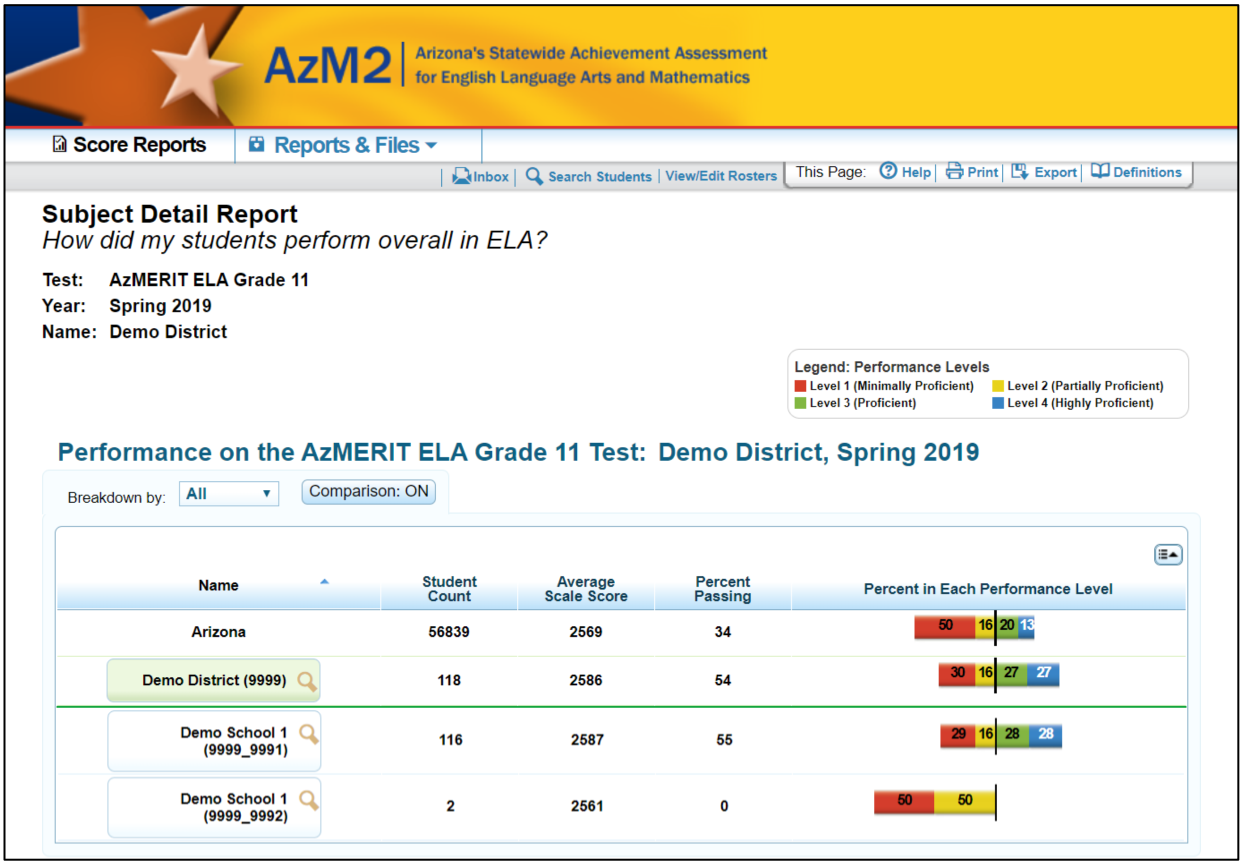
On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district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

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

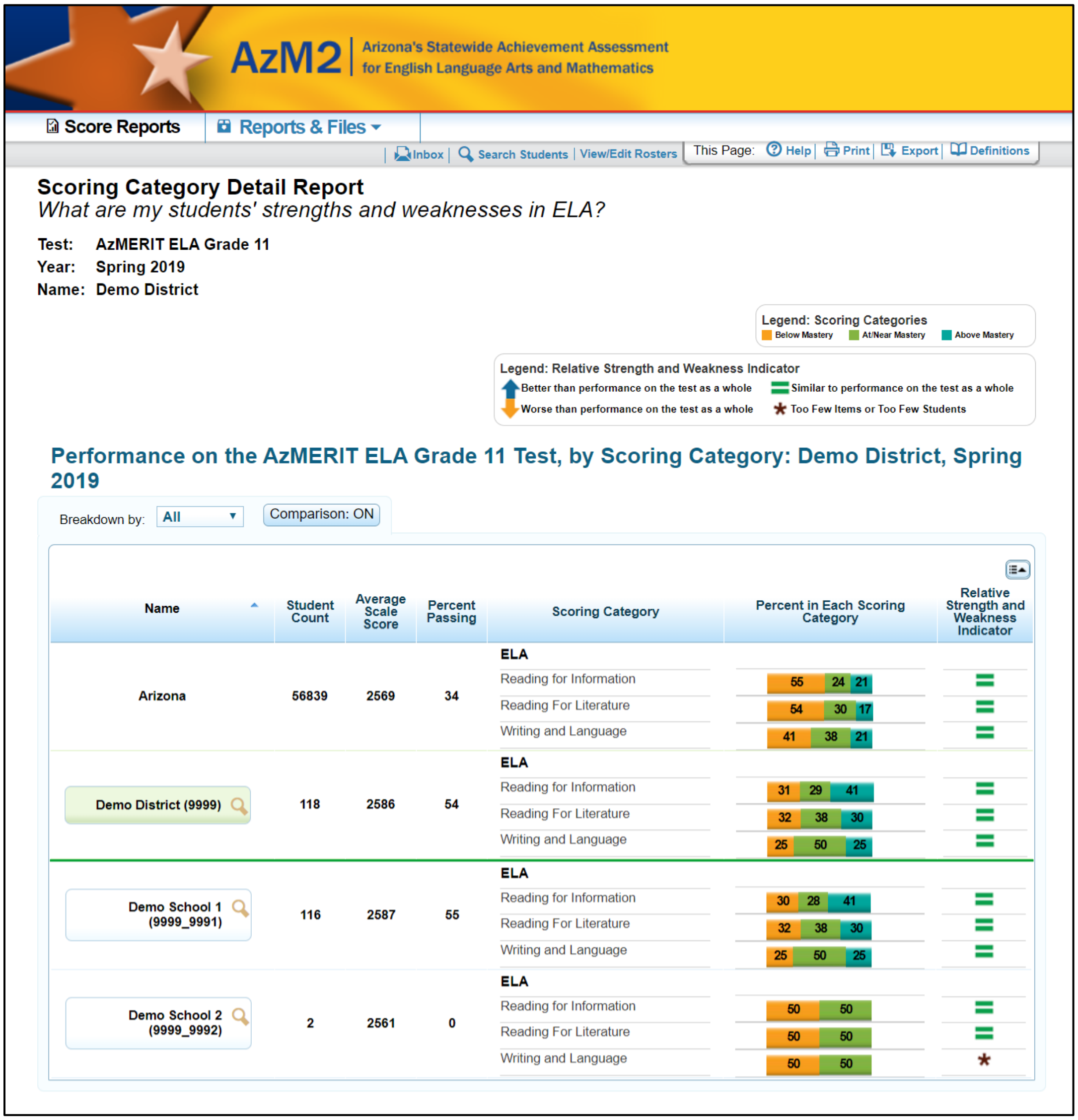
SUBJECT DETAIL REPORTS



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

The Subject Detail Report page shows the following data:

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



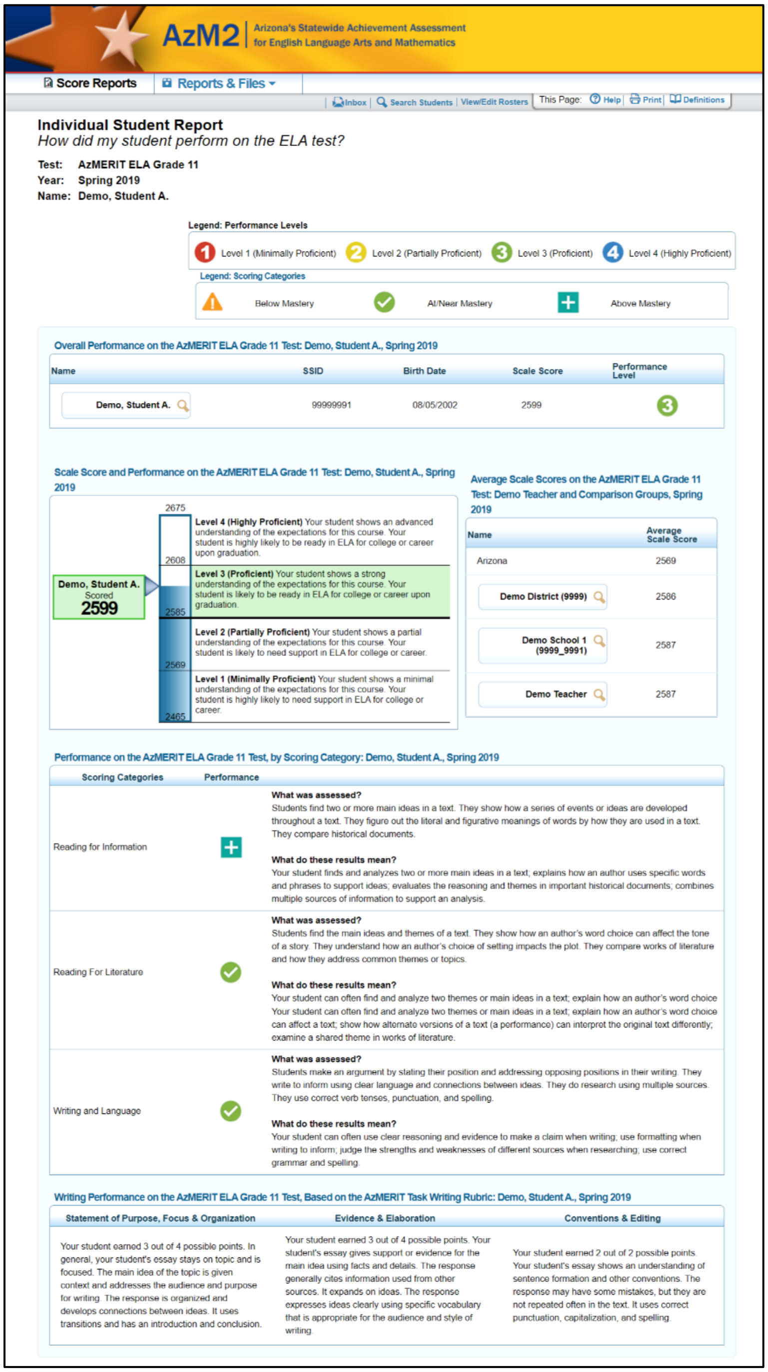
Aggregated scoring category detail reports follow the layout of the subject detail reports, displaying the performance data for the state, districts, schools, teachers, and classes. In addition, these reports include a relative strength and weakness indicator for each category.

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 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


AzM2 Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Score Reports | Reports & Files

Inbox | Search Students | View/Edit Rosters | This Page: Help | Print | Definitions

Individual Student Report

How did my student perform on the ELA test?

Test: AzMERIT ELA Grade 11
Year: Spring 2019
Name: Demo, Student A.

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

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

Overall Performance on the AzMERIT ELA Grade 11 Test: Demo, Student A., Spring 2019

Name	SSID	Birth Date	Scale Score	Performance Level
Demo, Student A.	99999991	08/05/2002	2599	3

Scale Score and Performance on the AzMERIT ELA Grade 11 Test: Demo, Student A., Spring 2019

Demo, Student A. Scored 2599

2675
Level 4 (Highly Proficient) Your student shows an advanced understanding of the expectations for this course. Your student is highly likely to be ready in ELA for college or career upon graduation.
2608
Level 3 (Proficient) Your student shows a strong understanding of the expectations for this course. Your student is likely to be ready in ELA for college or career upon graduation.
2585
Level 2 (Partially Proficient) Your student shows a partial understanding of the expectations for this course. Your student is likely to need support in ELA for college or career.
2569
Level 1 (Minimally Proficient) Your student shows a minimal understanding of the expectations for this course. Your student is highly likely to need support in ELA for college or career.
2465

Average Scale Scores on the AzMERIT ELA Grade 11 Test: Demo Teacher and Comparison Groups, Spring 2019

Name	Average Scale Score
Arizona	2569
Demo District (9999)	2586
Demo School 1 (9999_9991)	2587
Demo Teacher	2587

Performance on the AzMERIT ELA Grade 11 Test, by Scoring Category: Demo, Student A., Spring 2019

Scoring Categories	Performance	
Reading for Information	Above Mastery	**What was assessed?** Students find two or more main ideas in a text. They show how a series of events or ideas are developed throughout a text. They figure out the literal and figurative meanings of words by how they are used in a text. They compare historical documents. **What do these results mean?** Your student finds and analyzes two or more main ideas in a text; explains how an author uses specific words and phrases to support ideas; evaluates the reasoning and themes in important historical documents; combines multiple sources of information to support an analysis.
Reading For Literature	At/Near Mastery	**What was assessed?** Students find the main ideas and themes of a text. They show how an author's word choice can affect the tone of a story. They understand how an author's choice of setting impacts the plot. They compare works of literature and how they address common themes or topics. **What do these results mean?** Your student can often find and analyze two themes or main ideas in a text; explain how an author's word choice Your student can often find and analyze two themes or main ideas in a text; explain how an author's word choice can affect a text; show how alternate versions of a text (a performance) can interpret the original text differently; examine a shared theme in works of literature.
Writing and Language	At/Near Mastery	**What was assessed?** Students make an argument by stating their position and addressing opposing positions in their writing. They write to inform using clear language and connections between ideas. They do research using multiple sources. They use correct verb tenses, punctuation, and spelling. **What do these results mean?** Your student can often use clear reasoning and evidence to make a claim when writing; use formatting when writing to inform; judge the strengths and weaknesses of different sources when researching; use correct grammar and spelling.

Writing Performance on the AzMERIT ELA Grade 11 Test, Based on the AzMERIT Task Writing Rubric: Demo, Student A., Spring 2019

Statement of Purpose, Focus & Organization	Evidence & Elaboration	Conventions & Editing
Your student earned 3 out of 4 possible points. In general, your student's essay stays on topic and is focused. The main idea of the topic is given context and addresses the audience and purpose for writing. The response is organized and develops connections between ideas. It uses transitions and has an introduction and conclusion.	Your student earned 3 out of 4 possible points. Your student's essay gives support or evidence for the main idea using facts and details. The response generally cites information used from other sources. It expands on ideas. The response expresses ideas clearly using specific vocabulary that is appropriate for the audience and style of writing.	Your student earned 2 out of 2 possible points. Your student's essay shows an understanding of sentence formation and other conventions. The response may have some mistakes, but they are not repeated often in the text. It uses correct punctuation, capitalization, and spelling.

Individual Student Reports (ISRs), 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.[47] 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 2017–2018 school year indicates the same level of achievement as an average scale score of 2450 for grade 3 students in the 2018–2019 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.

---

[47] 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
ELA				
Grade 3	2395–2496	2497–2508	2509–2540	2541–2605
Grade 4	2400–2509	2510–2522	2523–2558	2559–2610
Grade 5	2419–2519	2520–2542	2543–2577	2578–2629
Grade 6	2431–2531	2532–2552	2553–2596	2597–2641
Grade 7	2438–2542	2543–2560	2561–2599	2600–2648
Grade 8	2448–2550	2551–2571	2572–2603	2604–2658
Grade 9	2454–2554	2555–2576	2577–2605	2606–2664
Grade 10	2458–2566	2567–2580	2581–2605	2606–2668
Grade 11	2465–2568	2569–2584	2585–2607	2608–2675
Mathematics				
Grade 3	3395–3494	3495–3530	3531–3572	3573–3605
Grade 4	3435–3529	3530–3561	3562–3605	3606–3645
Grade 5	3478–3562	3563–3594	3595–3634	3635–3688
Grade 6	3512–3601	3602–3628	3629–3662	3663–3722
Grade 7	3529–3628	3629–3651	3652–3679	3680–3739
Grade 8	3566–3649	3650–3672	3673–3704	3705–3776
Algebra I	3577–3660	3661–3680	3681–3719	3720–3787
Geometry	3609–3672	3673–3696	3697–3742	3743–3819
Algebra II	3629–3689	3690–3710	3711–3750	3751–3839

ELA and 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).[48] 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 American College Testing (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 grades 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 grades 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

[48] 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.

to consider the location of these 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.

8.2 RECOMMENDED PERFORMANCE STANDARDS

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

Exhibit 8.2.1 Final Recommended Performance Standards for AzMERIT

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% at or Above	Theta	% at or Above	Theta	% at or Above
ELA						
3	-0.09	56	0.29	41	1.36	10
4	0.14	57	0.6	39	1.8	5
5	-0.13	63	0.63	30	1.8	3
6	-0.12	61	0.58	34	2.03	4
7	-0.02	59	0.61	33	1.9	4
8	-0.06	60	0.64	33	1.72	6
9	-0.12	53	0.59	27	1.57	6
10	0.11	51	0.58	30	1.42	8
11	-0.02	46	0.52	26	1.27	8
Mathematics						
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
ELA				
3	44	15	31	10
4	43	19	33	5
5	37	33	27	3
6	39	27	30	4
7	41	26	29	4
8	40	27	26	6
9	47	26	21	6
10	49	21	22	8
11	54	20	17	8
Mathematics				
3	27	31	27	15
4	29	29	32	10
5	29	31	27	13
6	38	30	21	11
7	48	22	18	13
8	43	24	20	13
Algebra I	45	23	23	9
Geometry	47	24	24	6
Algebra II	47	24	23	6

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

Exhibit 8.2.3 Percentages of Students Meeting AzMERIT and Benchmark Proficient Standards

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

9. SCALING AND EQUATING

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

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

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

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (1PL; also known as the Rasch model), is used to calibrate 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 θ can be written as

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

with the constraint that $\sum_{k=0}^{0}(\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i. Item parameters for the assessments were calibrated following the spring administration in 2015 and vertical scales were established for reporting both 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 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.[49] 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.

9.1.1 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 toward 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 2019, 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.

9.1.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

[49] 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 θ given responses to N items, $\boldsymbol{x} = \{x_i\}$, can be expressed as:

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

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

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 θ is found via the following iterative routine:

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

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).[50] 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

---

[50] 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 upward. 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
G3→G5	2.75	2.67	0.81	0.78
G3→G6	3.90	3.73	1.19	1.15
G3→G7	4.48	4.28	1.44	1.39
G3→G8	5.69	5.39	1.76	1.70
G3 M → Algebra I & G3 ELA → G9 ELA	6.07	5.76	1.97	1.88
G3 M → Geometry & G3 ELA → G10 ELA	7.15	6.86	2.12	1.98
G3 M → Algebra II & G3 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

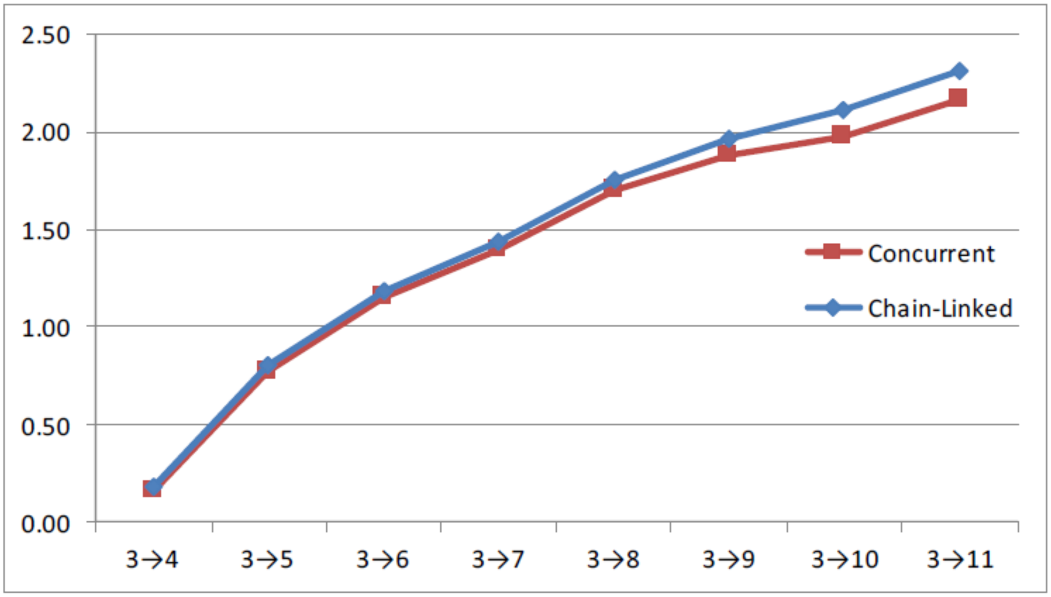
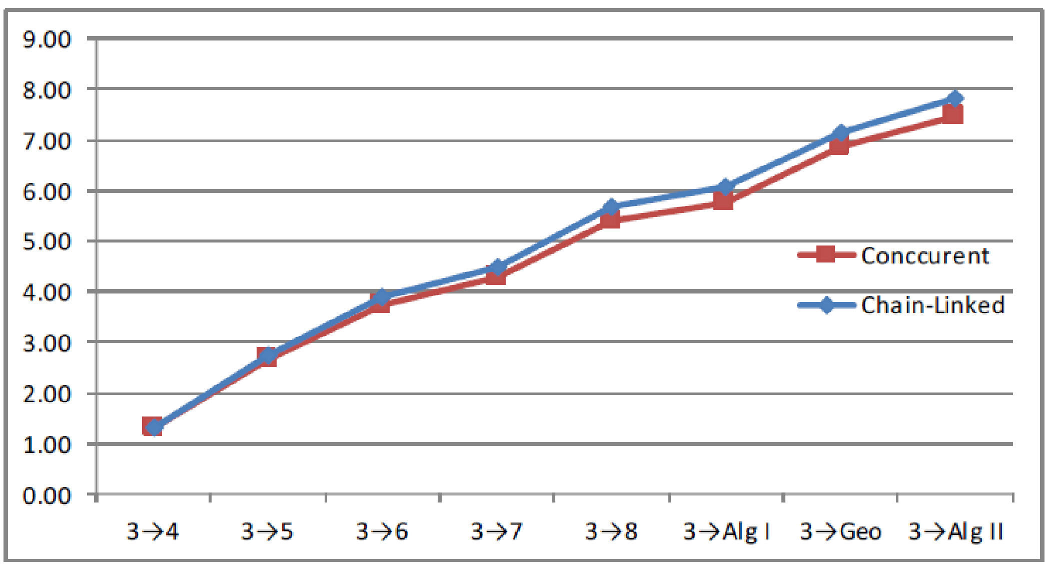


Exhibit 9.2.2.5 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: Mathematics



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 2019 VERTICAL LINKING STUDY

It has been four 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 2019, the vertical linking study was repeated to evaluate results of the 2015 vertical linking study.

Both chain linking and concurrent calibration approaches were used to produce the 2019 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. As expected, the 2019 linking constants produced by chain-linking and concurrent calibration converged. The 2019 vertical linking constants resulting from chain linking and concurrent calibration in ELA and mathematics assessments are presented in Exhibits 9.2.2.6 and 9.2.2.7.

Exhibit 9.2.2.6 Vertical Linking Constants Resulting from Chain-Linking and Concurrent Calibration: ELA

ELA	Chain-Linked	Concurrent
G3E	0	0
G4E	0.48	0.48
G5E	1.04	1.05
G6E	1.43	1.45
G7E	1.67	1.69
G8E	2.03	2.06
G9E	2.23	2.26
G10E	2.48	2.49
G11E	2.61	2.63

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

Mathematics	Chain-Linked	Concurrent
G3M	0	0
G4M	1.55	1.45
G5M	2.98	2.80
G6M	4.17	3.93
G7M	4.74	4.48
G8M	5.55	5.26
Alg I	6.17	5.82
Geometry	6.67	6.24
Alg II	7.09	6.70

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 2019 is presented graphically in Exhibit 9.2.2.8 and Exhibit 9.2.2.9 for ELA and mathematics, respectively.

Exhibit 9.2.2.8 Comparison of 2015 and 2019 Vertical Linking Constants Estimated from Chain-Linking Calibrations: ELA

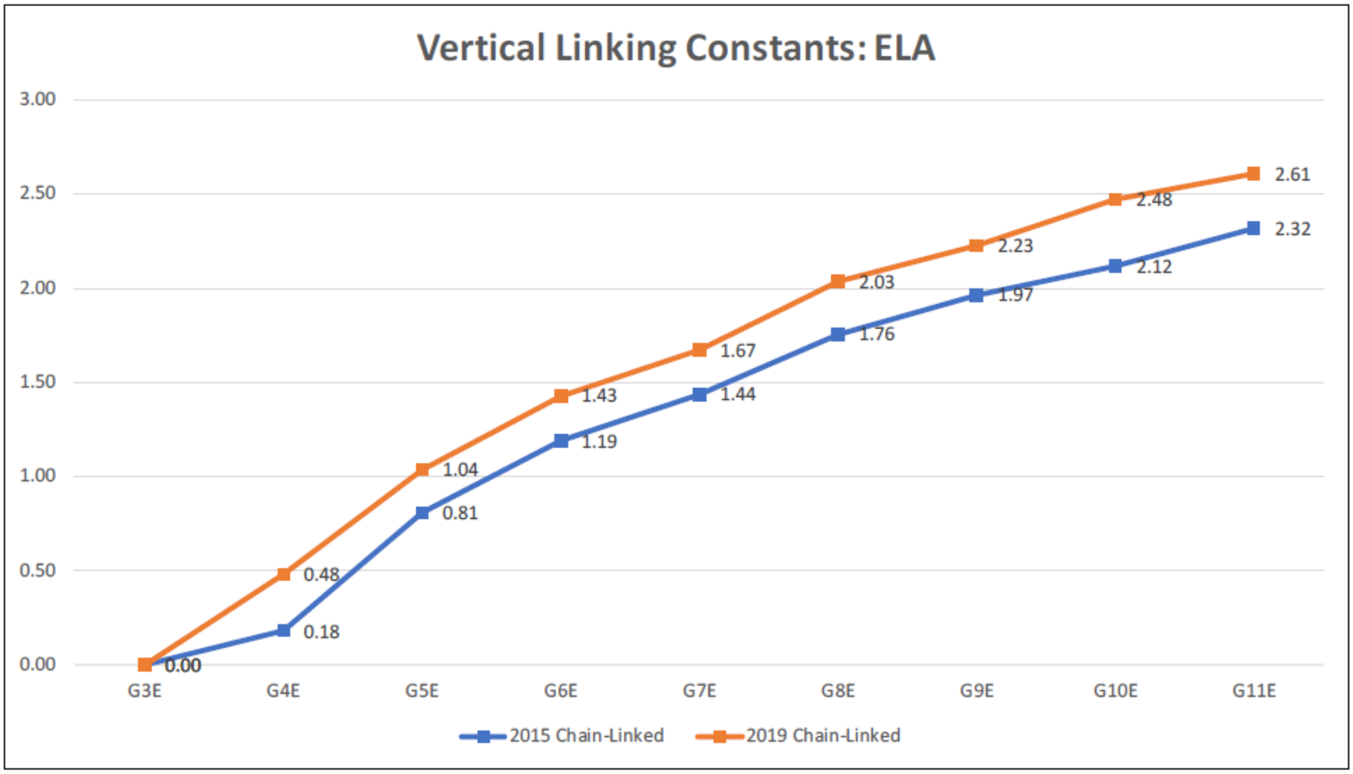
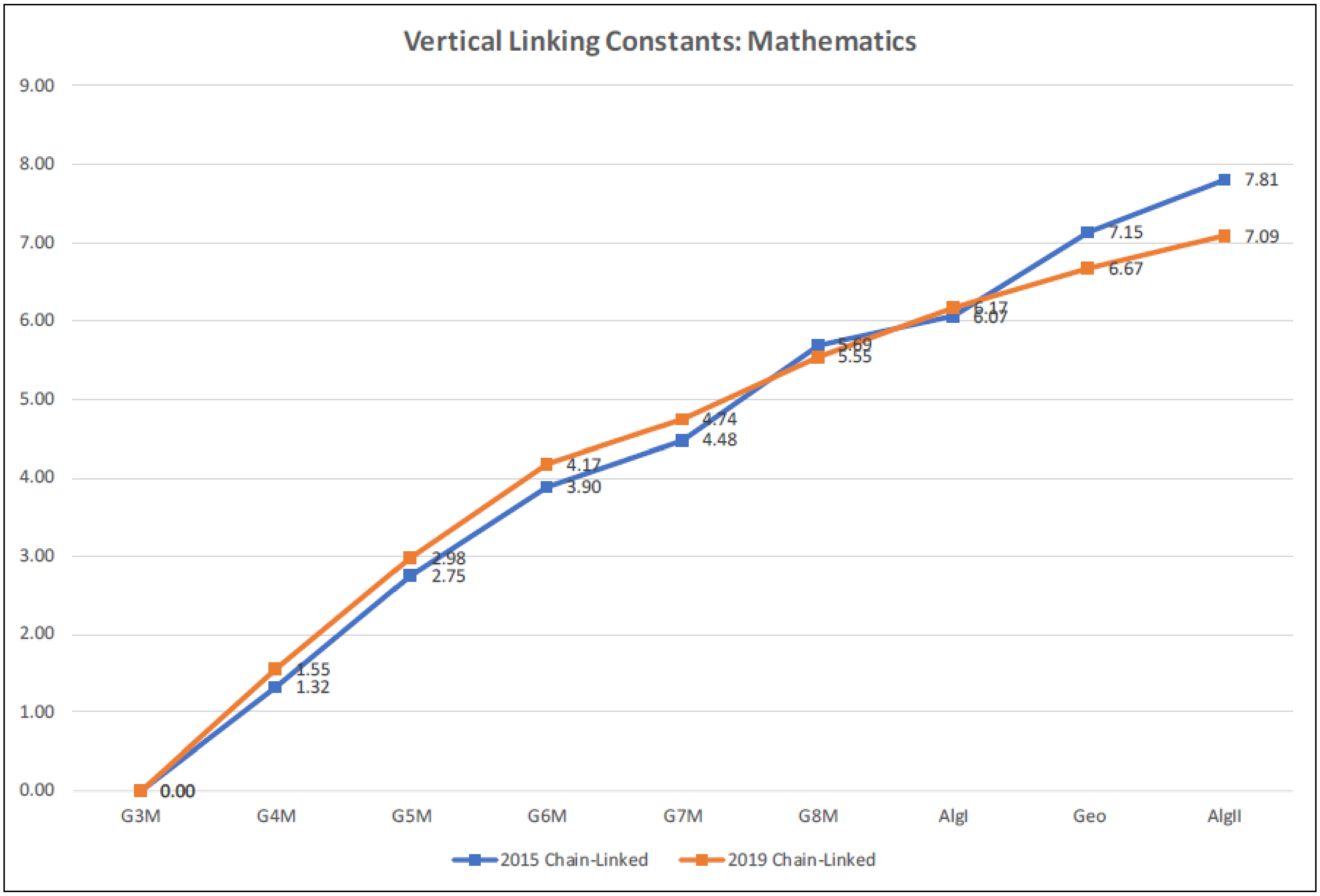


Exhibit 9.2.2.9 Comparison of 2015 and 2019 Vertical Linking Constants Estimated from Chain-Linking Calibrations: Mathematics



Additionally, Exhibits 9.2.2.10 and 9.2.2.11 show the comparison of the chain-linking results obtained in 2015 and 2019 along with the standard error of the linking constants for ELA and mathematics, respectively. Similarity between the 2015 and 2019 vertical linking results is observed with respect to the difference between linking constants by grade. For ELA, although the vertical linking constants by grade in 2019 are uniformly higher than those in 2015, the difference between the 2015 and 2019 ELA linking constant for each grade is not larger than 0.4 logit. For mathematics, the vertical linking constants for grades 8, Geometry, and Algebra II in 2019 are smaller than those in 2015, while the vertical linking constants for the other grades in 2019 are larger than those in 2015. The difference between the 2015 and 2019 mathematics linking constant for each grade is not larger than 0.5 logit, except for Algebra II, which is at 0.72 logit.

Exhibit 9.2.2.10 Vertical Linking Constants from 2015 and 2019: ELA

ELA	2015 Chain-Linked	2019 Chain-Linked	SE of 2019 Chain Linking Constant
G3E	0	0	NA
G4E	0.18	0.48	0.05
G5E	0.81	1.04	0.07
G6E	1.19	1.43	0.08
G7E	1.44	1.67	0.11
G8E	1.76	2.03	0.11
G9E	1.97	2.23	0.11

G10E	2.12	2.48	0.11
G11E	2.32	2.61	0.12

Exhibit 9.2.2.11 Vertical Linking Constants from 2015 and 2019: Mathematics

Mathematics	2015 Chain-Linked	2019 Chain-Linked	SE of 2019 Chain Linking Constant
G3M	0	0	NA
G4M	1.32	1.55	0.04
G5M	2.75	2.98	0.05
G6M	3.9	4.17	0.06
G7M	4.48	4.74	0.06
G8M	5.69	5.55	0.09
Alg I	6.07	6.17	0.09
Geometry	7.15	6.67	0.1
Alg II	7.81	7.09	0.1

The vertical linking results are also similar between 2015 and 2019 in terms of the overall growth patterns across grades, as shown in Exhibits 9.2.2.12 and 9.2.2.13. 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 logit 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.

Exhibit 9.2.2.12 Vertical Growth between Grades for 2019: ELA

ELA	# of Common Vertical Linking Items	Growth between Grades	SE of Growth
G3E_G4E	34	0.48	0.05
G4E_G5E	41	0.56	0.05
G5E_G6E	35	0.39	0.04
G6E_G7E	33	0.24	0.07
G7E_G8E	37	0.36	0.03
G8E_G9E	38	0.19	0.02
G9E_G10E	36	0.25	0.02
G10E_G11E	36	0.13	0.04

Exhibit 9.2.2.13 Vertical Growth between Grades for 2019: Mathematics

Mathematics	# of Common Vertical Linking Items	Growth between Grades	SE of Growth
G3M_G4M	43	1.55	0.04
G4M_G5M	43	1.43	0.03
G5M_G6M	41	1.19	0.04

G6M_G7M	26	0.57	0.02
G7M_G8M	43	0.81	0.06
G8M_AlgI	43	0.62	0.03
G8M_Geo	42	1.12	0.03
AlgI_AlgII	42	0.92	0.02

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.

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,[51]

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

where $a = 30, d = 2500$ for ELA tests, and $a = 30, d = 3500$ for mathematics tests. $\theta_V = \theta + c$, where θ is the on-grade ability estimate and c is a vertical linking constant listed below for each of the tests, as described in the previous section. For reporting, the on-grade ability estimate is truncated at $\pm$ 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
		ELA		
Grade 3	2395–2496	2497–2508	2509–2540	2541–2605
Grade 4	2400–2509	2510–2522	2523–2558	2559–2610
Grade 5	2419–2519	2520–2542	2543–2577	2578–2629
Grade 6	2431–2531	2532–2552	2553–2596	2597–2641
Grade 7	2438–2542	2543–2560	2561–2599	2600–2648
Grade 8	2448–2550	2551–2571	2572–2603	2604–2658
Grade 9	2454–2554	2555–2576	2577–2605	2606–2664
Grade 10	2458–2566	2567–2580	2581–2605	2606–2668
Grade 11	2465–2568	2569–2584	2585–2607	2608–2675
		Mathematics		
Grade 3	3395–3494	3495–3530	3531–3572	3573–3605
Grade 4	3435–3529	3530–3561	3562–3605	3606–3645
Grade 5	3478–3562	3563–3594	3595–3634	3635–3688
Grade 6	3512–3601	3602–3628	3629–3662	3663–3722
Grade 7	3529–3628	3629–3651	3652–3679	3680–3739

[51] 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
		Mathematics		
Grade 8	3566–3649	3650–3672	3673–3704	3705–3776
Algebra I	3577–3660	3661–3680	3681–3719	3720–3787
Geometry	3609–3672	3673–3696	3697–3742	3743–3819
Algebra II	3629–3689	3690–3710	3711–3750	3751–3839

9.4 LINKING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to reporting test scores for the spring 2015 and spring 2016 administrations of AzMERIT, AIR and ADE performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration.[52]

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

[52] 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 grades 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
G3M	2015	−0.71	−0.77	0.06	1.80
	2016—Mathematics MC Match	−0.84	−0.57	−0.27	−8.10
G4M	2015	−0.40	−0.48	0.08	2.40
	2016—2014 AIMS Match	−0.43	−0.25	−0.17	−5.10
	2016—2015 AzMERIT Match	−0.57	−0.43	−0.14	−4.20
	2016—Mathematics MC Match	−0.41	−0.24	−0.17	−5.10
G5M	2015	−0.09	−0.09	−0.01	−0.30
	2016—2014 AIMS Match	−0.06	−0.02	−0.04	−1.20
	2016—2015 AzMERIT Match	−0.16	−0.12	−0.03	−0.90
	2016—Mathematics MC Match	−0.07	−0.06	0.00	0.00
G6M	2015	0.07	0.01	0.07	2.10
	2016—2014 AIMS Match	−0.01	0.04	−0.05	−1.50
	2016—2015 AzMERIT Match	−0.09	−0.06	−0.03	−0.90
G7M	2015	0.15	0.07	0.08	2.40
	2016—2014 AIMS Match	0.18	0.07	0.11	3.30
	2016—2015 AzMERIT Match	0.11	−0.03	0.14	4.20
G8M	2015	0.43	0.32	0.11	3.30
	2016—2014 AIMS Match	0.56	0.55	0.00	0.00
	2016—2015 AzMERIT Match	0.47	0.47	0.01	0.30
Alg I	2015	0.29	0.23	0.05	1.50
	2016—2014 AIMS Match	0.64	0.51	0.13	3.90
	2016—2015 AzMERIT Match	0.72	0.57	0.15	4.50
Geo	2015	1.12	0.99	0.13	3.90
	2016—2014 AIMS Match	1.34	1.15	0.20	6.00
	2016—2015 AzMERIT Match	1.19	1.03	0.16	4.80
Alg II	2015	1.45	1.36	0.09	2.70
	2016—2014 AIMS Match	1.45	1.17	0.28	8.40
	2016—2015 AzMERIT Match	1.06	0.91	0.15	4.50

For grade 3 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.[53] 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

[53] 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 AMERICAN COLLEGE TESTING COLLEGE-READY CUT ON AZMERIT

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the American College Testing (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.

<u>ELA:</u> 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

> Ŷ = 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-Price 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.

<u>Mathematics:</u> 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

> Ŷ = 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-Price 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 Square Error (RMSE) was calculated as the square root of the average squared errors found between the actual ACT score point and the model fitted values. Furthermore, we repeated this sampling and model fitting process 100 times to see how the RMSE varied across random samples. For ELA, the average RMSE was 5.03 and the standard deviation of the RMSE was 0.02 across the 100 replications. For mathematics, the average RMSE was 2.79 and the standard deviation was 0.02. The standard deviation of the RMSE was very small indicating that the sample selected for the modeling has no significant impact on the model fitting.

In addition, the equipercentile equating method was used to verify the linking between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT cut score 22 is 2585.72 for ELA and 3727.46 for mathematics. These cut scores are consistent with those identified using regression models.

10. CONSTRUCTED-RESPONSE SCORING

The 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 handscoring requirements. AIR provides the automated electronic scoring and MI provides all handscoring for the AzMERIT tests. This section describes the process for configuring and validating machine rubrics and the process for handscoring, 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 2019 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 grades 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.[54]

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

[54] 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 handscored, and agreement between human- and machine-assigned scores is examined. The cross-validation process ensures that the rubric generalizes across all responses and that the statistical model identified during training does not capitalize on peculiarities in the training set.

Exhibit 10.1.2.1 presents agreement indicators for the two initial human raters, and between the resolved human and statistical rubric score, for the two writing prompts randomly assigned in each grade in the spring 2019 administration.[55] Please see the 2016 AzMERIT Technical Report, available at www.azed.gov, for the values for the complete list of prompts**.** Indicators include percentage exact agreement, Pearson's correlation, a quadratic 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 2019, Exhibit 10.1.2.2 presents the correlations among the dimension scores.

[55] 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 2019 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N of Human	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted K	SMD	% Exact	Pearson r	Weighted κ	SMD
3	13022	Conventions	2	2092	1.49	1.62	0.68	0.63	0.65	0.70	0.52	0.02	0.71	0.76	0.60	0.20
		Elaboration	4		2.06	2.02	0.72	0.60	0.61	0.57	0.47	0.00	0.63	0.68	0.51	0.06
		Organization	4		2.14	2.08	0.74	0.60	0.67	0.61	0.53	0.03	0.64	0.68	0.52	0.09
3	13025	Conventions	2	2093	1.46	1.52	0.71	0.66	0.59	0.68	0.49	0.01	0.66	0.73	0.58	0.09
		Elaboration	4		2.03	2.01	0.75	0.66	0.61	0.59	0.48	0.01	0.71	0.73	0.61	0.02
		Organization	4		2.05	1.99	0.80	0.74	0.64	0.59	0.51	0.01	0.68	0.65	0.56	0.07
4	13120	Conventions	2	2091	1.20	1.15	0.68	0.64	0.63	0.66	0.53	0.04	0.64	0.69	0.54	0.07
		Elaboration	4		1.31	1.26	0.49	0.46	0.52	0.76	0.48	0.02	0.57	0.82	0.56	0.09
		Organization	4		1.46	1.45	0.55	0.52	0.61	0.76	0.57	0.04	0.59	0.79	0.59	0.02
4	13119	Conventions	2	2094	1.29	1.32	0.64	0.63	0.60	0.68	0.52	0.06	0.67	0.73	0.59	0.04
		Elaboration	4		1.38	1.33	0.53	0.50	0.47	0.71	0.42	0.04	0.59	0.79	0.56	0.10
		Organization	4		1.53	1.51	0.60	0.53	0.59	0.70	0.51	0.03	0.65	0.77	0.60	0.03
5	13247	Conventions	2	2097	1.45	1.48	0.66	0.62	0.69	0.74	0.60	0.04	0.71	0.76	0.62	0.04
		Elaboration	4		1.78	1.81	0.62	0.59	0.56	0.65	0.47	0.05	0.65	0.74	0.57	0.06
		Organization	4		1.94	1.92	0.65	0.61	0.65	0.69	0.54	0.02	0.69	0.77	0.61	0.03
5	13246	Conventions	2	2093	1.46	1.49	0.61	0.62	0.63	0.73	0.56	0.10	0.71	0.78	0.65	0.06
		Elaboration	4		1.61	1.59	0.55	0.51	0.55	0.69	0.48	0.07	0.61	0.78	0.58	0.03
		Organization	4		1.83	1.81	0.66	0.56	0.61	0.67	0.51	0.00	0.62	0.71	0.53	0.03
6	13307	Conventions	2	2095	1.46	1.49	0.66	0.64	0.64	0.68	0.53	0.03	0.69	0.74	0.60	0.05
		Elaboration	4		1.60	1.57	0.67	0.61	0.62	0.66	0.52	0.00	0.67	0.74	0.59	0.05
		Organization	4		1.84	1.79	0.73	0.63	0.63	0.64	0.52	0.02	0.68	0.70	0.57	0.06
6	13306	Conventions	2	2097	1.59	1.63	0.59	0.59	0.56	0.70	0.47	0.07	0.63	0.76	0.55	0.08
		Elaboration	4		1.70	1.64	0.64	0.56	0.55	0.65	0.46	0.01	0.62	0.73	0.55	0.09
		Organization	4		1.91	1.88	0.71	0.63	0.61	0.66	0.51	0.05	0.63	0.69	0.53	0.05
7	13401	Conventions	2	2084	1.67	1.71	0.50	0.52	0.63	0.81	0.59	0.05	0.68	0.83	0.63	0.07
		Elaboration	4		1.84	1.86	0.54	0.50	0.58	0.72	0.50	0.01	0.67	0.82	0.62	0.03
		Organization	4		2.01	2.00	0.55	0.42	0.63	0.74	0.53	0.04	0.66	0.83	0.58	0.01
7	13406	Conventions	2	2090	1.45	1.51	0.62	0.60	0.58	0.70	0.50	0.03	0.72	0.78	0.65	0.10
		Elaboration	4		1.76	1.77	0.54	0.52	0.60	0.74	0.54	0.03	0.61	0.79	0.57	0.03
		Organization	4		1.92	1.92	0.52	0.45	0.54	0.70	0.45	0.04	0.66	0.84	0.61	0.00

Grade	ITS ID	Dimensions	Score Point	N of Human	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted K	SMD	% Exact	Pearson r	Weighted κ	SMD
8	13454	Conventions	2	2677	1.55	1.59	0.63	0.61	0.69	0.79	0.63	0.03	0.72	0.80	0.65	0.06
		Elaboration	4		1.93	1.96	0.71	0.68	0.75	0.78	0.69	0.02	0.73	0.75	0.63	0.05
		Organization	4		2.06	2.04	0.76	0.72	0.76	0.75	0.68	0.01	0.76	0.76	0.67	0.02
8	13439	Conventions	2	2719	1.62	1.70	0.58	0.53	0.62	0.78	0.55	0.02	0.73	0.84	0.66	0.12
		Elaboration	4		2.11	2.08	0.71	0.64	0.75	0.74	0.65	0.01	0.72	0.75	0.62	0.05
		Organization	4		2.21	2.20	0.81	0.75	0.72	0.69	0.62	0.05	0.79	0.75	0.69	0.01
9	13556	Conventions	2	1594	1.65	1.72	0.59	0.53	0.62	0.80	0.54	0.00	0.72	0.84	0.65	0.09
		Elaboration	4		1.90	1.91	0.66	0.60	0.79	0.81	0.72	0.06	0.69	0.77	0.61	0.01
		Organization	4		2.00	2.03	0.65	0.63	0.74	0.77	0.66	0.03	0.77	0.83	0.71	0.03
9	13555	Conventions	2	2956	1.58	1.62	0.60	0.55	0.75	0.81	0.67	0.01	0.77	0.84	0.71	0.08
		Elaboration	4		1.88	1.88	0.61	0.55	0.76	0.82	0.70	0.03	0.72	0.82	0.65	0.00
		Organization	4		2.07	2.04	0.67	0.62	0.79	0.81	0.72	0.02	0.78	0.82	0.72	0.06
10	13638	Conventions	2	2580	1.60	1.68	0.57	0.52	0.58	0.71	0.51	0.05	0.59	0.76	0.52	0.15
		Elaboration	4		2.02	2.01	0.69	0.63	0.65	0.69	0.56	0.00	0.71	0.77	0.63	0.02
		Organization	4		2.10	2.12	0.73	0.68	0.69	0.67	0.58	0.00	0.73	0.74	0.64	0.02
10	13637	Conventions	2	1417	1.59	1.65	0.58	0.54	0.58	0.69	0.49	0.06	0.60	0.77	0.53	0.09
		Elaboration	4		1.92	1.90	0.68	0.64	0.70	0.75	0.62	0.02	0.73	0.77	0.65	0.05
		Organization	4		2.06	2.08	0.72	0.64	0.74	0.76	0.66	0.03	0.75	0.78	0.67	0.01
11	13720	Conventions	2	2091	1.59	1.65	0.56	0.53	0.56	0.76	0.52	0.02	0.66	0.79	0.60	0.10
		Elaboration	4		1.96	1.92	0.76	0.72	0.65	0.60	0.52	0.01	0.74	0.72	0.63	0.05
		Organization	4		2.24	2.25	0.73	0.62	0.70	0.67	0.58	0.02	0.73	0.76	0.64	0.01
11	13721	Conventions	2	2090	1.59	1.63	0.57	0.55	0.57	0.73	0.49	0.04	0.66	0.79	0.60	0.08
		Elaboration	4		2.23	2.24	0.74	0.67	0.65	0.61	0.52	0.04	0.76	0.76	0.67	0.01
		Organization	4		2.33	2.33	0.68	0.64	0.62	0.64	0.50	0.03	0.71	0.77	0.65	0.00

Note: Weighted K = Quadratic weighted kappa; SMD = Standardized Mean Difference

Exhibit 10.1.2.2 Summary of Dimension Intercorrelations for Spring 2019 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	Correlations Among Dimensions	
				Conventions	Elaboration
3	13022	Conventions	2		
		Elaboration	4	0.67	
		Organization	4	0.55	0.86
3	13025	Conventions	2		
		Elaboration	4	0.47	
		Organization	4	0.67	0.82
4	13120	Conventions	2		
		Elaboration	4	0.45	
		Organization	4	0.58	0.72
4	13119	Conventions	2		
		Elaboration	4	0.52	
		Organization	4	0.72	0.54
5	13247	Conventions	2		
		Elaboration	4	0.54	
		Organization	4	0.60	0.84
5	13246	Conventions	2		
		Elaboration	4	0.67	
		Organization	4	0.68	0.67
6	13307	Conventions	2		
		Elaboration	4	0.67	
		Organization	4	0.68	0.88
6	13306	Conventions	2		
		Elaboration	4	0.56	
		Organization	4	0.62	0.74
7	13401	Conventions	2		
		Elaboration	4	0.62	
		Organization	4	0.58	0.76
7	13406	Conventions	2		
		Elaboration	4	0.61	
		Organization	4	0.58	0.73
8	13454	Conventions	2		
		Elaboration	4	0.67	
		Organization	4	0.54	0.86
8	13439	Conventions	2		
		Elaboration	4	0.54	
		Organization	4	0.45	0.86
9	13556	Conventions	2		
		Elaboration	4	0.54	
		Organization	4	0.50	0.76
9	13555	Conventions	2		
		Elaboration	4	0.50	
		Organization	4	0.59	0.80
10	13638	Conventions	2		
		Elaboration	4	0.44	
		Organization	4	0.39	0.85
10	13637	Conventions	2		
		Elaboration	4	0.40	
		Organization	4	0.55	0.80
11	13720	Conventions	2		
		Elaboration	4	0.66	
		Organization	4	0.54	0.80

Grade	ITS ID	Dimensions	Score Point	Correlations Among Dimensions	
				Conventions	Elaboration
11	13721	Conventions	2		
		Elaboration	4	0.56	
		Organization	4	0.59	0.82

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 duplicate text is more 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 percentages of the machine-assigned condition codes for spring 2017 administrations and Exhibit 10.1.3.2 presents percentages of the machine-assigned condition codes for spring 2018 administrations. Exhibit 10.1.3.3 presents percentages of the machine-assigned condition codes for spring 2019 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
G3E	13023	9	0	0	1	0	0	0
	13026	13	0	0	1	0	0	0
G4E	13094	26	0	0	1	0	0	0
	13095	9	0	0	1	0	0	0
G5E	13236	7	0	0	0	0	0	0
	13239	10	0	0	0	0	0	0
G6E	13304	9	0	0	0	0	0	0
	13308	6	0	0	0	0	0	0
G7E	13402	12	0	0	0	0	0	0
	13403	3	0	0	0	0	0	0
G8E	13437	7	0	0	0	2	0	2
	13452	4	0	0	0	0	0	0
G9E	13557	4	0	0	0	1	3	3
	13566	4	0	0	0	0	0	0
G10E	13639	4	0	0	0	0	6	6
	13640	1	0	0	0	0	3	0
G11E	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
G3E	**13021**	12	0	0	1	0	0	0
	13024	10	0	0	1	0	0	0
G4E	**13118**	6	0	0	1	0	0	0
	13121	5	0	0	1	0	0	0
G5E	**13237**	9	0	0	0	0	0	0
	13238	4	0	0	0	0	0	0
G6E	**13305**	8	0	0	0	0	0	0
	13309	5	0	0	0	0	0	0
G7E	**13400**	8	0	0	0	0	0	0
	13405	5	0	0	0	0	0	0
G8E	**13438**	4	0	0	0	1	1	1
	13453	4	0	0	0	2	0	2

G9E	**13554**	5	0	0	0	2	2	2
	13565	2	0	0	0	1	0	0
G10E	**13635**	2	0	0	0	0	0	0
	13636	5	0	0	0	0	0	0
G11E	**13723**	1	0	0	0	0	0	0
	13725	3	0	0	1	0	0	0

Exhibit 10.1.3.3 Frequency of Machine-Assigned Condition Codes for Spring 2019 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**
G3E	**13022**	9	0	0	1	0	0	0
	13025	12	0	0	1	0	0	0
G4E	**13119**	7	0	0	1	0	0	0
	13120	5	0	0	1	0	0	0
G5E	**13246**	8	0	0	0	0	0	0
	13247	3	0	0	0	0	0	0
G6E	**13306**	9	0	0	0	0	0	0
	13307	4	0	0	0	0	0	0
G7E	**13401**	4	0	0	0	0	0	0
	13406	9	0	0	0	0	0	0
G8E	**13439**	5	0	0	0	1	1	0
	13454	4	0	0	0	0	0	0
G9E	**13555**	2	0	0	0	0	0	0
	13556	4	0	0	0	1	2	1
G10E	**13637**	1	0	0	0	0	0	0
	13638	2	0	0	0	0	0	1
G11E	**13720**	4	0	0	1	0	0	0
	13721	1	0	0	0	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 2019, all the essays were autoscored, and the essay responses with the low confidence index were routed to MI for human verification. The final essay score was the human verification score when present.

10.2.1 HANDSCORING PROCESS

MI's handscoring efforts are managed via the Virtual Scoring Center (VSC) software, which is composed of two primary subsystems: VSC Capture and VSC Score. Images of student responses to open ended items were sent to VSC Score, which is a web-based environment for scoring constructed-response items by scorers working in an online environment. VSC Score is a secure, centrally administered environment used by site-based scorers. The interface enabled scorers to evaluate constructed-response items and writing assessments from images. VSC Score has the following capabilities:

- Defining scorer roles and qualifications based on training, security requirements, or prior history
- Managing and randomly routing scorers' responses that require second readings in a double-blind manner
- Allowing project leaders to spot-check scorers, monitor reliability, and offer feedback
- Allowing scorers to flag responses for a variety of reasons (unusual approaches, nonscorable issues, etc.)
- Generating status reports at project milestones (such as percentage of items scored)
- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
- Accommodating 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.[56]

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 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

[56] 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.

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.[57]

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 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

[57] 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.

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 blueline 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 contactor. 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
Item Analysis Report	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Forensic Analysis	To monitor testing irregularities	Early detection of testing irregularities

ITEM ANALYSIS REPORT

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as 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 δ_{ij} is aggregated across all n students for item i,

$$\bar{\delta}_i = \frac{1}{n}\sum_{i=}^{n}(\delta_{ij}).$$

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. I-28–I-40), Retrieved from http://www.azed.gov/assessment/files/2014/11/spring-2015-azmerit-standard-setting_091415-full-report.pdf.

Sireci, S. G. & Rios, J. A. (2013). Decisions that make difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice, 19*(2–3), 170–187, DOI: 10.1080/13803611.2013.767621.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter, *Psychometrika, 66*, 331–342. doi:10.1007/BF02294437.

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing, *The Philippine 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.

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

Calculator Guidance

The AzMERIT calculator guidelines are designed to provide appropriate support for students while still measuring a student's mastery of the standards. On tests where calculators are permitted, it is ideal for a student to use the recommended acceptable calculator. If the recommended calculator is not available, students may use a calculator with less functionality. The Desmos Scientific and Graphing calculators have been customized for AzMERIT and are embedded in online tests that allow the use of a calculator.

These guidelines are for the assessment only. They are not intended to limit instruction in the classroom. Technology is a part of the Arizona Mathematics Standards, and students should still be interacting with technology as appropriate for engaging with and learning the standards.

Grades 3-6: **No calculators permitted on AzMERIT.**

Grades 7-8: **Scientific calculator permitted on AzMERIT Math Part 1 only.**
No calculators permitted on AzMERIT Math Part 2.
Scientific calculator should include these functions: standard four functions (addition, subtraction, multiplication, and division), decimal, change sign (+/-), parentheses, square root, and π.
They may NOT include: any problem solving or programming capabilities, place values, and inequalities. *Sample acceptable calculator: TI-30X IIS or similar.*

High School End-of-Course Tests: **Graphing calculators permitted on AzMERIT Math Part 1 and Part 2.**
No calculators with Computer Algebra System (CAS) features are allowed. Calculators may NOT be capable of communication with other calculators through infrared sensors. NO instruction or formula cards, or other information regarding the operation of calculators such as operating manuals are permitted. The memory of any calculator with programming capability must be cleared, reset, or disabled when students enter the testing room. Many calculators have a testing mode that will allow these features to be disabled and will meet the requirements of AzMERIT. Check the calculator documentation for instructions on enabling this mode. If the memory of any calculator is password protected, and cannot be cleared or reset, the calculator may NOT be used. Items for the EOC tests are written with these types of calculators in mind; however students may use a scientific calculator if they choose to do so. *Sample acceptable calculators: TI-84 Plus, Casio FX-9750GII, or similar.*

Additional Guidance:

- Students are not allowed to share calculators during a testing session.
- The AzMERIT online calculators available for the computer-based assessment are available for practice use on the Calculator and Tutorials site at http://azmeritportal.org/tutorials/.
- For EOC tests only, an online version of the scientific and graphing calculator will be available in the Secure Browser for students taking the paper-based version of the test. Students will not need to sign in to select the online calculator.
- No laptop, tablet, or phone-based calculators are allowed to be used during the AzMERIT assessment unless they are used to access the AzMERIT Secure Browser.
- The applicable portion of the computer-based assessment will include the acceptable online version of approved calculator. Providing handheld calculators is not a requirement for schools choosing the computer-based assessment. However, students may use an acceptable handheld calculator in addition to or instead of the online calculator.

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

English Language Arts Assessment Blueprint

Grade 3		
Strands	**Min**	**Max**
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	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)

For more information go to www.azed.gov/AzMERIT

AzMERIT

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

Mathematics Assessment Blueprint

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 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 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 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%

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%

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

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%

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

Percentage of Points by Depth of Knowledge Level			
Grade	DOK Level 1	DOK Level 2	DOK Level 3
3-11	10%-20%	60%-70%	12%-30%

Within a test, approximately 70% of the assessment will be on major content within that grade or course.

Revised by ADE on 8/19/15

For more information go to www.azed.gov/AzMERIT

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

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	80994.729	1638				
Metric	81897.510	1679	Configural	902.781 (41)	< .01	.000
Scalar	86181.627	1720	Metric	4284.117 (41)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	35572.452	1638				
Metric	36296.535	1679	Configural	724.083 (41)	< .01	.000
Scalar	36567.679	1720	Metric	271.144 (41)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	67063.171	1638				
Metric	69607.650	1679	Configural	2544.478 (41)	< .01	.000
Scalar	71127.244	1720	Metric	1519.594 (41)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	32996.702	1638				
Metric	33086.624	1679	Configural	89.922 (41)	< .01	.000
Scalar	33495.278	1720	Metric	408.654 (41)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	34937.031	1638				
Metric	35925.633	1679	Configural	988.601 (41)	< .01	.000
Scalar	36581.042	1720	Metric	655.409 (41)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	33879.417	1638				
Metric	33930.102	1679	Configural	50.685 (41)	0.14	.000
Scalar	34050.517	1720	Metric	120.415 (41)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	80866.453	1638				
Metric	83191.089	1679	Configural	2324.635 (41)	< .01	.000
Scalar	85497.255	1720	Metric	2306.167 (41)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	81074.356	1638				
Metric	82817.189	1679	Configural	1742.833 (41)	< .01	.000
Scalar	83332.052	1720	Metric	514.863 (41)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	79837.226	1638				
Metric	83158.192	1679	Configural	3320.966 (41)	< .01	.000
Scalar	84594.316	1720	Metric	1436.124 (41)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	80948.338	1638				
Metric	82121.967	1679	Configural	1173.629 (41)	< .01	.000
Scalar	82896.727	1720	Metric	774.759 (41)	< .01	.000

Appendix C.1b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	62800.225	1686	< .01	0.953	0.031
Model B-1	27931.666	1686	< .01	0.953	0.032
Model B-2	53667.395	1686	< .01	0.948	0.032
Model B-3	24899.711	1686	< .01	0.954	0.031
Model B-4	28147.438	1686	< .01	0.952	0.032
Model B-5	26523.526	1686	< .01	0.954	0.031
Model C	59284.487	1686	< .01	0.951	0.030
Model D	60655.181	1686	< .01	0.954	0.031
Model E	55473.682	1686	< .01	0.958	0.029
Model F	56219.017	1686	< .01	0.958	0.030

Appendix C.2a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	54806.784	1804				
Metric	55820.859	1847	Configural	1014.075 (43)	< .01	.000
Scalar	61111.392	1890	Metric	5290.533 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	25516.261	1804				
Metric	26323.723	1847	Configural	807.462 (43)	< .01	.000
Scalar	26852.676	1890	Metric	528.953 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	45838.620	1804				
Metric	48957.028	1847	Configural	3118.408 (43)	< .01	.000
Scalar	51481.248	1890	Metric	2524.220 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	23782.522	1804				
Metric	23880.634	1847	Configural	98.112 (43)	< .01	.000
Scalar	24636.110	1890	Metric	755.476 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	24668.606	1804				
Metric	25663.607	1847	Configural	995.000 (43)	< .01	.000
Scalar	27043.568	1890	Metric	1379.962 (43)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	24456.677	1804				
Metric	24529.782	1847	Configural	73.105 (43)	< .01	.000
Scalar	24697.510	1890	Metric	167.728 (43)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	54579.152	1804				
Metric	57435.718	1847	Configural	2856.566 (43)	< .01	.000
Scalar	59989.286	1890	Metric	2553.568 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	54933.759	1804				
Metric	56818.010	1847	Configural	1884.251 (43)	< .01	.000
Scalar	57396.250	1890	Metric	578.240 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	24077.870	1638				
Metric	25716.898	1679	Configural	1639.027 (41)	< .01	.000
Scalar	27948.466	1720	Metric	2231.568 (41)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	54849.209	1804				
Metric	55898.392	1847	Configural	1049.184 (43)	< .01	.000
Scalar	56934.317	1890	Metric	1035.924 (43)	< .01	.001

Appendix C.2b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	72597.769	1856	< .01	0.964	0.031
Model B-1	32369.240	1856	< .01	0.960	0.032
Model B-2	60180.160	1856	< .01	0.960	0.031
Model B-3	28261.170	1856	< .01	0.960	0.031
Model B-4	31996.825	1856	< .01	0.960	0.032
Model B-5	30605.972	1856	< .01	0.959	0.032
Model C	68049.140	1856	< .01	0.962	0.030
Model D	71296.253	1856	< .01	0.963	0.031
Model E	21978.749	1684	< .01	0.984	0.018
Model F	62922.852	1856	< .01	0.970	0.029

Appendix C.3a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	60765.639	1804				
Metric	62184.094	1847	Configural	1418.454 (43)	< .01	.000
Scalar	67363.052	1890	Metric	5178.958 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	25965.785	1804				
Metric	26982.437	1847	Configural	1016.652 (43)	< .01	.000
Scalar	27539.645	1890	Metric	557.208 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	50329.080	1804				
Metric	53870.536	1847	Configural	3541.456 (43)	< .01	.001
Scalar	57000.564	1890	Metric	3130.028 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	24123.387	1804				
Metric	24351.037	1847	Configural	227.650 (43)	< .01	.000
Scalar	24884.526	1890	Metric	533.489 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	14872.639	1638				
Metric	16403.176	1679	Configural	1530.537 (41)	< .01	.001
Scalar	17395.960	1720	Metric	992.784 (41)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	24736.705	1804				
Metric	24837.410	1847	Configural	100.705 (43)	< .01	.000
Scalar	24946.861	1890	Metric	109.451 (43)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	58934.248	1804				
Metric	62831.178	1847	Configural	3896.930 (43)	< .01	.001
Scalar	66656.571	1890	Metric	3825.393 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	60003.870	1804				
Metric	62414.328	1847	Configural	2410.458 (43)	< .01	.001
Scalar	63680.678	1890	Metric	1266.350 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	34190.233	1638				
Metric	37693.207	1679	Configural	3502.973 (41)	< .01	.001
Scalar	39231.806	1720	Metric	1538.599 (41)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	34601.966	1638				
Metric	36115.908	1679	Configural	1513.942 (41)	< .01	.001
Scalar	37076.824	1720	Metric	960.916 (41)	< .01	.000

Appendix C.3b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	84254.495	1853	< .01	0.967	0.033
Model B-1	34563.314	1853	< .01	0.962	0.032
Model B-2	70708.641	1853	< .01	0.961	0.033
Model B-3	29671.425	1853	< .01	0.962	0.031
Model B-4	13416.939	1681	< .01	0.983	0.021
Model B-5	31773.563	1853	< .01	0.962	0.032
Model C	78599.382	1853	< .01	0.960	0.032
Model D	81443.717	1853	< .01	0.966	0.033
Model E	30658.211	1681	< .01	0.985	0.021
Model F	28855.211	1681	< .01	0.986	0.020

Appendix C.4a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	54417.352	1804				
Metric	56208.818	1847	Configural	1791.467 (43)	< .01	.000
Scalar	62040.187	1890	Metric	5831.369 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	15194.031	1720				
Metric	16406.408	1762	Configural	1212.377 (42)	< .01	.001
Scalar	16952.948	1804	Metric	546.540 (42)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	44619.321	1804				
Metric	48584.334	1847	Configural	3965.013 (43)	< .01	.000
Scalar	50451.640	1890	Metric	1867.307 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	21872.746	1804				
Metric	21970.501	1847	Configural	97.756 (43)	< .01	.000
Scalar	22528.661	1890	Metric	558.160 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	14587.291	1720				
Metric	16337.438	1762	Configural	1750.146 (42)	< .01	.001
Scalar	16973.068	1804	Metric	635.631 (42)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	22544.583	1804				
Metric	22674.955	1847	Configural	130.372 (43)	< .01	.001
Scalar	22767.110	1890	Metric	92.155 (43)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	32715.460	1720				
Metric	37130.689	1762	Configural	4415.229 (42)	< .01	.001
Scalar	42288.263	1804	Metric	5157.575 (42)	< .01	.002
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	54170.444	1804				
Metric	56412.828	1847	Configural	2242.384 (43)	< .01	.000
Scalar	57113.053	1890	Metric	700.225 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	25486.184	1638				
Metric	28224.572	1679	Configural	2738.388 (41)	< .01	.001
Scalar	30019.391	1720	Metric	1794.820 (41)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	25845.369	1638				
Metric	27493.846	1679	Configural	1648.477 (41)	< .01	.001
Scalar	28919.405	1720	Metric	1425.559 (41)	< .01	.000

Appendix C.4b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	58422.336	1854	< .01	0.969	0.028
Model B-1	18646.057	1767	< .01	0.973	0.024
Model B-2	51773.783	1854	< .01	0.970	0.028
Model B-3	22082.074	1854	< .01	0.970	0.026
Model B-4	18016.961	1767	< .01	0.973	0.023
Model B-5	23982.436	1854	< .01	0.970	0.027
Model C	41377.867	1767	< .01	0.969	0.024
Model D	58452.377	1854	< .01	0.973	0.028
Model E	21716.322	1682	< .01	0.983	0.017
Model F	20534.547	1682	< .01	0.985	0.017

Appendix C.5a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	65972.861	1804				
Metric	67616.300	1847	Configural	1643.438 (43)	< .01	.000
Scalar	73058.632	1890	Metric	5442.333 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	33082.946	1804				
Metric	34241.458	1847	Configural	1158.512 (43)	< .01	.000
Scalar	34759.616	1890	Metric	518.158 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	55435.943	1804				
Metric	59332.955	1847	Configural	3897.011 (43)	< .01	.001
Scalar	61046.787	1890	Metric	1713.832 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	31993.900	1804				
Metric	32158.307	1847	Configural	164.407 (43)	< .01	.000
Scalar	32576.139	1890	Metric	417.831 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	32094.869	1804				
Metric	33579.795	1847	Configural	1484.926 (43)	< .01	.001
Scalar	34520.225	1890	Metric	940.430 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	31879.036	1804				
Metric	31935.219	1847	Configural	56.182 (43)	0.09	.000
Scalar	32001.567	1890	Metric	66.349 (43)	0.01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	63904.936	1804				
Metric	68116.865	1847	Configural	4211.929 (43)	< .01	.001
Scalar	72718.806	1890	Metric	4601.940 (43)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	66240.347	1804				
Metric	68082.274	1847	Configural	1841.927 (43)	< .01	.000
Scalar	68744.604	1890	Metric	662.331 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	65403.263	1804				
Metric	68349.188	1847	Configural	2945.925 (43)	< .01	.000
Scalar	70045.119	1890	Metric	1695.931 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	65522.411	1804				
Metric	66933.415	1847	Configural	1411.004 (43)	< .01	.000
Scalar	68208.910	1890	Metric	1275.495 (43)	< .01	.000

Appendix C.5b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	53058.918	1852	< .01	0.969	0.026
Model B-1	21152.685	1852	< .01	0.969	0.025
Model B-2	42502.431	1852	< .01	0.966	0.026
Model B-3	18063.996	1852	< .01	0.970	0.023
Model B-4	20751.782	1852	< .01	0.969	0.025
Model B-5	18945.754	1852	< .01	0.970	0.024
Model C	47108.365	1852	< .01	0.967	0.025
Model D	49132.357	1852	< .01	0.971	0.025
Model E	44310.687	1852	< .01	0.973	0.024
Model F	42049.125	1852	< .01	0.975	0.023

Appendix C.6a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	106651.524	1804				
Metric	109057.577	1847	Configural	2406.053 (43)	< .01	.000
Scalar	113039.654	1890	Metric	3982.077 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	46944.940	1804				
Metric	48250.319	1847	Configural	1305.379 (43)	< .01	.000
Scalar	48794.956	1890	Metric	544.637 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	88651.815	1804				
Metric	93185.147	1847	Configural	4533.331 (43)	< .01	.001
Scalar	95605.881	1890	Metric	2420.734 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	43890.829	1804				
Metric	44160.869	1847	Configural	270.041 (43)	< .01	.000
Scalar	44852.358	1890	Metric	691.489 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	46514.730	1804				
Metric	48918.249	1847	Configural	2403.519 (43)	< .01	.001
Scalar	50057.699	1890	Metric	1139.450 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	44575.203	1804				
Metric	44665.232	1847	Configural	90.029 (43)	< .01	.000
Scalar	44782.148	1890	Metric	116.916 (43)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	103476.066	1804				
Metric	108541.455	1847	Configural	5065.389 (43)	< .01	.000
Scalar	115468.728	1890	Metric	6927.273 (43)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	106993.902	1804				
Metric	109597.400	1847	Configural	2603.498 (43)	< .01	.000
Scalar	110837.061	1890	Metric	1239.661 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	106485.869	1804				
Metric	109472.175	1847	Configural	2986.305 (43)	< .01	.000
Scalar	112490.792	1890	Metric	3018.618 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	106556.337	1804				
Metric	108659.002	1847	Configural	2102.664 (43)	< .01	.000
Scalar	110995.261	1890	Metric	2336.260 (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	*df*	P-Value		
Model A	70727.735	1857	< .01	0.966	0.031
Model B-1	27876.052	1857	< .01	0.967	0.028
Model B-2	59506.353	1857	< .01	0.962	0.031
Model B-3	24069.385	1857	< .01	0.968	0.027
Model B-4	28332.975	1857	< .01	0.966	0.029
Model B-5	25057.031	1857	< .01	0.968	0.028
Model C	59568.439	1857	< .01	0.965	0.028
Model D	69031.849	1857	< .01	0.966	0.030
Model E	56294.926	1857	< .01	0.971	0.027
Model F	53354.483	1857	< .01	0.974	0.027

Appendix C.7a. Global Model Fit Indices of Measurement Invariance Tests for Grade 9 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	63898.554	1978				
Metric	65349.847	2023	Configural	1451.293 (45)	< .01	.000
Scalar	70010.551	2068	Metric	4660.704 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	30463.408	1978				
Metric	31301.998	2023	Configural	838.590 (45)	< .01	.000
Scalar	31689.948	2068	Metric	387.950 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	53764.116	1978				
Metric	56516.020	2023	Configural	2751.904 (45)	< .01	.001
Scalar	58626.470	2068	Metric	2110.449 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	28113.531	1978				
Metric	28286.264	2023	Configural	172.733 (45)	< .01	.000
Scalar	28788.481	2068	Metric	502.217 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	30035.776	1978				
Metric	31015.238	2023	Configural	979.462 (45)	< .01	.000
Scalar	32405.924	2068	Metric	1390.686 (45)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	28323.268	1978				
Metric	28379.150	2023	Configural	55.882 (45)	0.13	.000
Scalar	28432.244	2068	Metric	53.095 (45)	0.19	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	64052.520	1978				
Metric	65517.611	2023	Configural	1465.092 (45)	< .01	.000
Scalar	68091.706	2068	Metric	2574.095 (45)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	64251.684	1978				
Metric	65491.686	2023	Configural	1240.002 (45)	< .01	.000
Scalar	66021.714	2068	Metric	530.028 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	63676.238	1978				
Metric	65279.346	2023	Configural	1603.108 (45)	< .01	.000
Scalar	67127.594	2068	Metric	1848.248 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	23673.822	1804				
Metric	23862.245	1847	Configural	188.423 (43)	< .01	.000
Scalar	24370.029	1890	Metric	507.785 (43)	< .01	.000

Appendix C.7b. Global Model Fit Indices of Scalar Invariance Model for Grade 9 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	57018.524	2031	< .01	0.959	0.029
Model B-1	22760.584	2031	< .01	0.965	0.026
Model B-2	46898.090	2031	< .01	0.957	0.029
Model B-3	19985.092	2031	< .01	0.966	0.025
Model B-4	23170.933	2031	< .01	0.965	0.027
Model B-5	19290.359	2031	< .01	0.968	0.025
Model C	47493.152	2031	< .01	0.962	0.026
Model D	52155.403	2031	< .01	0.963	0.028
Model E	46555.586	2031	< .01	0.966	0.026
Model F	18791.209	1851	< .01	0.983	0.017

Appendix C.8a. Global Model Fit Indices of Measurement Invariance Tests for Grade 10 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	55981.170	1978				
Metric	57888.649	2023	Configural	1907.478 (45)	< .01	.000
Scalar	62183.919	2068	Metric	4295.270 (45)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	28657.247	1978				
Metric	29330.296	2023	Configural	673.049 (45)	< .01	.000
Scalar	29896.565	2068	Metric	566.269 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	47449.195	1978				
Metric	49595.841	2023	Configural	2146.646 (45)	< .01	.000
Scalar	52707.049	2068	Metric	3111.208 (45)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	26895.296	1978				
Metric	27090.134	2023	Configural	194.839 (45)	< .01	.000
Scalar	27744.624	2068	Metric	654.490 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	27825.829	1978				
Metric	28793.820	2023	Configural	967.991 (45)	< .01	.000
Scalar	29837.709	2068	Metric	1043.889 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	27045.805	1978				
Metric	27104.820	2023	Configural	59.015 (45)	0.08	.000
Scalar	27180.445	2068	Metric	75.626 (45)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	55982.504	1978				
Metric	57532.853	2023	Configural	1550.349 (45)	< .01	.000
Scalar	59925.038	2068	Metric	2392.185 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	56449.811	1978				
Metric	57848.224	2023	Configural	1398.414 (45)	< .01	.000
Scalar	58393.061	2068	Metric	544.837 (45)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	56563.959	1978				
Metric	57590.376	2023	Configural	1026.417 (45)	< .01	.000
Scalar	58270.433	2068	Metric	680.057 (45)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	26557.552	1804				
Metric	26839.997	1847	Configural	282.445 (43)	< .01	.001
Scalar	27256.338	1890	Metric	416.341 (43)	< .01	.000

Appendix C.8b. Global Model Fit Indices of Scalar Invariance Model for Grade 10 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	63425.114	2031	< .01	0.961	0.032
Model B-1	34263.879	2031	< .01	0.952	0.034
Model B-2	62686.462	2031	< .01	0.947	0.035
Model B-3	31092.182	2031	< .01	0.953	0.033
Model B-4	33309.871	2031	< .01	0.953	0.033
Model B-5	30099.400	2031	< .01	0.954	0.032
Model C	67113.490	2031	< .01	0.951	0.033
Model D	67288.580	2031	< .01	0.962	0.033
Model E	66308.374	2031	< .01	0.966	0.033
Model F	19712.602	1851	< .01	0.982	0.018

Appendix C.9a. Global Model Fit Indices of Measurement Invariance Tests for Grade 11 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	41896.965	1978				
Metric	43039.650	2023	Configural	1142.685 (45)	< .01	.000
Scalar	46537.066	2068	Metric	3497.416 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	22444.011	1978				
Metric	23109.034	2023	Configural	665.024 (45)	< .01	.000
Scalar	23411.632	2068	Metric	302.597 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	35369.923	1978				
Metric	37710.592	2023	Configural	2340.668 (45)	< .01	.000
Scalar	37710.592	2068	Metric	1191.058 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	21492.602	1978				
Metric	21609.160	2023	Configural	116.558 (45)	< .01	.000
Scalar	21975.769	2068	Metric	366.609 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	21689.850	1978				
Metric	22409.489	2023	Configural	719.639 (45)	< .01	.000
Scalar	22894.929	2068	Metric	485.439 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	21146.209	1978				
Metric	21214.809	2023	Configural	68.600 (45)	0.01	.000
Scalar	21280.512	2068	Metric	65.703 (45)	0.02	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	41508.314	1978				
Metric	42684.767	2023	Configural	1176.453 (45)	< .01	.000
Scalar	44644.048	2068	Metric	1959.281 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	41814.958	1978				
Metric	43134.016	2023	Configural	1319.058 (45)	< .01	.000
Scalar	43426.568	2068	Metric	292.552 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	41826.419	1978				
Metric	42484.953	2023	Configural	658.533 (45)	< .01	.000
Scalar	43128.201	2068	Metric	643.249 (45)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	17117.884	1804				
Metric	17321.950	1847	Configural	204.066 (43)	< .01	.000
Scalar	17790.828	1890	Metric	468.877 (43)	< .01	.000

Appendix C.9b. Global Model Fit Indices of Scalar Invariance Model for Grade 11 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	39422.760	2028	< .01	0.976	0.026
Model B-1	15080.122	2028	< .01	0.975	0.022
Model B-2	33129.137	2028	< .01	0.973	0.026
Model B-3	13281.795	2028	< .01	0.977	0.021
Model B-4	14845.100	2028	< .01	0.977	0.022
Model B-5	12822.114	2028	< .01	0.977	0.021
Model C	41343.117	2028	< .01	0.977	0.027
Model D	44110.417	2028	< .01	0.978	0.028
Model E	39372.673	2028	< .01	0.980	0.026
Model F	9935.452	1848	< .01	0.989	0.013

Appendix C.10a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	78160.944	1890				
Metric	79107.506	1934	Configural	946.561 (44)	< .01	.000
Scalar	83624.228	1978	Metric	4516.722 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	30898.987	1890				
Metric	32972.455	1934	Configural	2073.468 (44)	< .01	.001
Scalar	33738.304	1978	Metric	765.849 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	60878.314	1890				
Metric	65361.997	1934	Configural	4483.682 (44)	< .01	.001
Scalar	69684.223	1978	Metric	4322.226 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	28423.824	1890				
Metric	28873.420	1934	Configural	449.595 (44)	< .01	.000
Scalar	29323.699	1978	Metric	450.279 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	29915.817	1890				
Metric	32355.812	1934	Configural	2439.995 (44)	< .01	.001
Scalar	33664.939	1978	Metric	1309.127 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	29052.995	1890				
Metric	29131.303	1934	Configural	78.307 (44)	< .01	.000
Scalar	29250.743	1978	Metric	119.441 (44)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	72273.321	1890				
Metric	80218.160	1934	Configural	7944.839 (44)	< .01	.001
Scalar	83545.691	1978	Metric	3327.531 (44)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	75227.614	1890				
Metric	79059.724	1934	Configural	3832.111 (44)	< .01	.001
Scalar	80830.793	1978	Metric	1771.069 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	75044.479	1890				
Metric	80072.099	1934	Configural	5027.620 (44)	< .01	.001
Scalar	81446.820	1978	Metric	1374.721 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	75600.687	1890				
Metric	79007.396	1934	Configural	3406.709 (44)	< .01	.000
Scalar	79919.023	1978	Metric	911.627 (44)	< .01	.000

Appendix C.10b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	52843.234	1933	< .01	0.980	0.027
Model B-1	19291.075	1933	< .01	0.980	0.024
Model B-2	41858.468	1933	< .01	0.977	0.026
Model B-3	14286.174	1933	< .01	0.982	0.021
Model B-4	19018.308	1933	< .01	0.980	0.024
Model B-5	16545.170	1933	< .01	0.981	0.022
Model C	50589.790	1933	< .01	0.977	0.026
Model D	49357.670	1933	< .01	0.980	0.026
Model E	48125.947	1933	< .01	0.981	0.025
Model F	46584.052	1933	< .01	0.981	0.025

Appendix C.11a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	149808.323	1890				
Metric	151752.644	1934	Configural	1944.321 (44)	< .01	.000
Scalar	156959.206	1978	Metric	5206.562 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	57945.525	1890				
Metric	61313.773	1934	Configural	3368.248 (44)	< .01	.001
Scalar	62063.010	1978	Metric	749.237 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	118962.702	1890				
Metric	125560.872	1934	Configural	6598.170 (44)	< .01	.001
Scalar	129329.519	1978	Metric	3768.647 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	52765.730	1890				
Metric	53108.172	1934	Configural	342.442 (44)	< .01	.000
Scalar	53686.700	1978	Metric	578.528 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	56213.478	1890				
Metric	59472.669	1934	Configural	3259.191 (44)	< .01	.001
Scalar	60528.460	1978	Metric	1055.791 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	54499.132	1890				
Metric	54718.311	1934	Configural	219.179 (44)	< .01	.000
Scalar	54845.849	1978	Metric	127.538 (44)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	141973.938	1890				
Metric	153104.173	1934	Configural	11130.235 (44)	< .01	.001
Scalar	157359.394	1978	Metric	4255.221 (44)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	146616.996	1890				
Metric	151361.688	1934	Configural	4744.692 (44)	< .01	.000
Scalar	153571.408	1978	Metric	2209.720 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	145883.815	1890				
Metric	153474.268	1934	Configural	7590.453 (44)	< .01	.001
Scalar	155345.697	1978	Metric	1871.429 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	147139.987	1890				
Metric	151941.792	1934	Configural	4801.805 (44)	< .01	.000
Scalar	153200.199	1978	Metric	1258.407 (44)	< .01	.000

Appendix C.11b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	137740.611	1935	< .01	0.951	0.043
Model B-1	50312.815	1935	< .01	0.947	0.039
Model B-2	109542.313	1935	< .01	0.944	0.042
Model B-3	39775.128	1935	< .01	0.947	0.036
Model B-4	48144.731	1935	< .01	0.948	0.039
Model B-5	45591.275	1935	< .01	0.945	0.038
Model C	131478.597	1935	< .01	0.943	0.042
Model D	131674.623	1935	< .01	0.949	0.042
Model E	124201.772	1935	< .01	0.953	0.040
Model F	122875.783	1935	< .01	0.953	0.040

Appendix C.12a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	89879.268	1890				
Metric	91078.608	1934	Configural	1199.340 (44)	< .01	.000
Scalar	98750.244	1978	Metric	7671.636 (44)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	35581.097	1890				
Metric	37823.371	1934	Configural	2242.274 (44)	< .01	.001
Scalar	39125.950	1978	Metric	1302.579 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	74482.664	1890				
Metric	79684.296	1934	Configural	5201.631 (44)	< .01	.001
Scalar	82313.900	1978	Metric	2629.605 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	32255.863	1890				
Metric	32462.664	1934	Configural	206.801 (44)	< .01	.000
Scalar	32845.407	1978	Metric	382.742 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	35237.409	1890				
Metric	37425.107	1934	Configural	2187.698 (44)	< .01	.000
Scalar	38213.810	1978	Metric	788.703 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	33531.114	1890				
Metric	33620.976	1934	Configural	89.861 (44)	< .01	.000
Scalar	33749.522	1978	Metric	128.546 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	88524.729	1890				
Metric	94377.294	1934	Configural	5852.566 (44)	< .01	.000
Scalar	99742.330	1978	Metric	5365.036 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	90310.310	1890				
Metric	94473.130	1934	Configural	4162.820 (44)	< .01	.001
Scalar	95588.397	1978	Metric	1115.267 (44)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	90474.146	1890				
Metric	94676.048	1934	Configural	4201.902 (44)	< .01	.001
Scalar	96686.120	1978	Metric	2010.073 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	90652.668	1890				
Metric	93176.559	1934	Configural	2523.890 (44)	< .01	.000
Scalar	95016.785	1978	Metric	1840.227 (44)	< .01	.000

Appendix C.12b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	92764.516	1934	< .01	0.967	0.034
Model B-1	33653.865	1934	< .01	0.967	0.031
Model B-2	75483.234	1934	< .01	0.963	0.034
Model B-3	27438.170	1934	< .01	0.968	0.029
Model B-4	32604.187	1934	< .01	0.968	0.031
Model B-5	30200.541	1934	< .01	0.968	0.030
Model C	87347.169	1934	< .01	0.961	0.033
Model D	88940.408	1934	< .01	0.966	0.033
Model E	83309.589	1934	< .01	0.969	0.032
Model F	79501.943	1934	< .01	0.969	0.032

Appendix C.13a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	71927.787	2068				
Metric	74167.029	2114	Configural	2239.242 (46)	< .01	.000
Scalar	82009.164	2160	Metric	7842.134 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	27927.974	2068				
Metric	30624.050	2114	Configural	2696.075 (46)	< .01	.001
Scalar	31743.728	2160	Metric	1119.678 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	56011.322	2068				
Metric	63124.095	2114	Configural	7112.774 (46)	< .01	.001
Scalar	65292.428	2160	Metric	2168.333 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	25925.639	2068				
Metric	26241.414	2114	Configural	315.775 (46)	< .01	.000
Scalar	26757.121	2160	Metric	515.707 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	27675.588	2068				
Metric	31008.184	2114	Configural	3332.596 (46)	< .01	.001
Scalar	31867.075	2160	Metric	858.891 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26477.927	2068				
Metric	26660.830	2114	Configural	182.903 (46)	< .01	.000
Scalar	26814.117	2160	Metric	153.287 (46)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	66572.143	2068				
Metric	73503.667	2114	Configural	6931.524 (46)	< .01	.001
Scalar	81808.386	2160	Metric	8304.719 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	70688.562	2068				
Metric	74953.847	2114	Configural	4265.285 (46)	< .01	.000
Scalar	75860.444	2160	Metric	906.597 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	70205.018	2068				
Metric	76194.787	2114	Configural	5989.768 (46)	< .01	.001
Scalar	77877.935	2160	Metric	1683.148 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	70763.174	2068				
Metric	73859.888	2114	Configural	3096.714 (46)	< .01	.000
Scalar	75963.643	2160	Metric	2103.755 (46)	< .01	.000

Appendix C.13b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	56307.636	2114	< .01	0.982	0.025
Model B-1	20042.357	2114	< .01	0.984	0.022
Model B-2	41369.643	2114	< .01	0.981	0.024
Model B-3	16920.800	2114	< .01	0.984	0.021
Model B-4	19285.367	2114	< .01	0.984	0.022
Model B-5	18134.305	2114	< .01	0.984	0.022
Model C	47727.317	2114	< .01	0.980	0.023
Model D	49687.525	2114	< .01	0.983	0.024
Model E	44701.519	2114	< .01	0.984	0.022
Model F	42915.489	2114	< .01	0.985	0.022

Appendix C.14a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	62973.112	2068				
Metric	64770.084	2114	Configural	1796.972 (46)	< .01	.000
Scalar	72251.290	2160	Metric	7481.206 (46)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	26725.587	2068				
Metric	28954.565	2114	Configural	2228.978 (46)	< .01	.001
Scalar	30492.049	2160	Metric	1537.484 (46)	< .01	.001
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	50623.349	2068				
Metric	56982.253	2114	Configural	6358.905 (46)	< .01	.001
Scalar	60317.785	2160	Metric	3335.532 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	24858.369	2068				
Metric	25185.206	2114	Configural	326.838 (46)	< .01	.000
Scalar	26108.017	2160	Metric	922.811 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	25879.611	2068				
Metric	28055.961	2114	Configural	2176.350 (46)	< .01	.001
Scalar	29842.601	2160	Metric	1786.640 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	25329.393	2068				
Metric	25480.506	2114	Configural	151.113 (46)	< .01	.000
Scalar	25601.169	2160	Metric	120.663 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	59009.521	2068				
Metric	64244.336	2114	Configural	5234.815 (46)	< .01	.001
Scalar	71013.794	2160	Metric	6769.458 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	62085.001	2068				
Metric	65141.586	2114	Configural	3056.585 (46)	< .01	.000
Scalar	65910.254	2160	Metric	768.668 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	61365.906	2068				
Metric	64678.185	2114	Configural	3312.279 (46)	< .01	.000
Scalar	68208.055	2160	Metric	3529.869 (46)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	61850.270	2068				
Metric	63517.284	2114	Configural	1667.014 (46)	< .01	.000
Scalar	66040.176	2160	Metric	2522.892 (46)	< .01	.000

Appendix C.14b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	53364.741	2113	< .01	0.983	0.025
Model B-1	19766.315	2113	< .01	0.984	0.022
Model B-2	42011.783	2113	< .01	0.981	0.024
Model B-3	16596.761	2113	< .01	0.985	0.021
Model B-4	19694.600	2113	< .01	0.984	0.022
Model B-5	17272.505	2113	< .01	0.985	0.021
Model C	43993.814	2113	< .01	0.983	0.022
Model D	46376.171	2113	< .01	0.985	0.023
Model E	41300.242	2113	< .01	0.986	0.022
Model F	37695.212	2113	< .01	0.988	0.021

Appendix C.15a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	68293.125	2068				
Metric	69707.385	2114	Configural	1414.260 (46)	< .01	.000
Scalar	74987.504	2160	Metric	5280.119 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	32360.019	2068				
Metric	33414.996	2114	Configural	1054.977 (46)	< .01	.000
Scalar	34483.303	2160	Metric	1068.307 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	56446.867	2068				
Metric	59470.343	2114	Configural	3023.476 (46)	< .01	.000
Scalar	61581.184	2160	Metric	2110.842 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	31121.158	2068				
Metric	31525.348	2114	Configural	404.190 (46)	< .01	.001
Scalar	31982.615	2160	Metric	457.267 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	31736.319	2068				
Metric	33128.483	2114	Configural	1392.164 (46)	< .01	.000
Scalar	34645.775	2160	Metric	1517.292 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	31232.633	2068				
Metric	31306.857	2114	Configural	74.224 (46)	.01	.000
Scalar	31398.483	2160	Metric	91.626 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	65845.837	2068				
Metric	69385.939	2114	Configural	3540.102 (46)	< .01	.001
Scalar	74445.545	2160	Metric	5059.606 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	68177.572	2068				
Metric	69813.507	2114	Configural	1635.935 (46)	< .01	.000
Scalar	70402.468	2160	Metric	588.961 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	67702.089	2068				
Metric	69034.659	2114	Configural	1332.570 (46)	< .01	.000
Scalar	71447.266	2160	Metric	2412.608 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	67705.713	2068				
Metric	68912.779	2114	Configural	1207.066 (46)	< .01	.000
Scalar	70822.393	2160	Metric	1909.614 (46)	< .01	.000

Appendix C.15b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	76737.656	2115	< .01	0.969	0.031
Model B-1	32503.392	2115	< .01	0.967	0.031
Model B-2	61214.696	2115	< .01	0.966	0.031
Model B-3	28962.963	2115	< .01	0.967	0.030
Model B-4	31296.509	2115	< .01	0.968	0.030
Model B-5	30111.091	2115	< .01	0.968	0.030
Model C	62323.554	2115	< .01	0.968	0.028
Model D	70749.515	2115	< .01	0.969	0.030
Model E	62820.482	2115	< .01	0.971	0.028
Model F	54499.692	2115	< .01	0.976	0.026

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

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	57261.907	2068				
Metric	58948.980	2114	Configural	1687.074 (46)	< .01	.000
Scalar	65288.745	2160	Metric	6339.765 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	26670.105	2068				
Metric	28327.319	2114	Configural	1657.215 (46)	< .01	.001
Scalar	29380.482	2160	Metric	1053.162 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	46683.841	2068				
Metric	51014.945	2114	Configural	4331.104 (46)	< .01	.001
Scalar	53711.429	2160	Metric	2696.484 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	25278.971	2068				
Metric	25533.947	2114	Configural	254.976 (46)	< .01	.000
Scalar	26130.445	2160	Metric	596.498 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	26302.113	2068				
Metric	27698.861	2114	Configural	1396.748 (46)	< .01	.001
Scalar	28947.904	2160	Metric	1249.042 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	24976.011	2068				
Metric	25055.539	2114	Configural	79.527 (46)	< .01	.000
Scalar	25150.084	2160	Metric	94.546 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	55690.713	2068				
Metric	58077.270	2114	Configural	2386.557 (46)	< .01	.000
Scalar	62391.081	2160	Metric	4313.812 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	57626.310	2068				
Metric	59084.191	2114	Configural	1457.881 (46)	< .01	.000
Scalar	59557.445	2160	Metric	473.254 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	56775.848	2068				
Metric	58483.187	2114	Configural	1707.339 (46)	< .01	.001
Scalar	60720.261	2160	Metric	2237.074 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	57891.193	2068				
Metric	58294.308	2114	Configural	403.116 (46)	< .01	.001
Scalar	59197.430	2160	Metric	903.121 (46)	< .01	.000

Appendix C.16b. Global Model Fit Indices of Scalar Invariance Model for Algebra I

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	55198.552	2114	< .01	0.976	0.027
Model B-1	22135.114	2114	< .01	0.978	0.024
Model B-2	43364.508	2114	< .01	0.974	0.026
Model B-3	19928.104	2114	< .01	0.979	0.024
Model B-4	21136.739	2114	< .01	0.979	0.024
Model B-5	19449.580	2114	< .01	0.980	0.023
Model C	43462.256	2114	< .01	0.977	0.024
Model D	49658.435	2114	< .01	0.977	0.025
Model E	41805.442	2114	< .01	0.981	0.023
Model F	33721.420	2114	< .01	0.982	0.021

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

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	54699.570	2068				
Metric	55250.423	2114	Configural	550.853 (46)	< .01	.001
Scalar	58070.887	2160	Metric	2820.464 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	28365.255	2068				
Metric	29110.470	2114	Configural	745.215 (46)	< .01	.000
Scalar	29985.563	2160	Metric	875.093 (46)	< .01	.001
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	45580.499	2068				
Metric	48411.947	2114	Configural	2831.448 (46)	< .01	.001
Scalar	50054.053	2160	Metric	1642.105 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	27711.264	2068				
Metric	27980.026	2114	Configural	268.762 (46)	< .01	.000
Scalar	28363.675	2160	Metric	383.649 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	28181.361	2068				
Metric	29208.438	2114	Configural	1027.077 (46)	< .01	.001
Scalar	30010.573	2160	Metric	802.134 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	27313.112	2068				
Metric	27374.041	2114	Configural	60.929 (46)	.07	.001
Scalar	27490.094	2160	Metric	116.053 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	53253.030	2068				
Metric	54478.592	2114	Configural	1225.562 (46)	< .01	.000
Scalar	58375.966	2160	Metric	3897.375 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	53992.488	2068				
Metric	55606.554	2114	Configural	1614.066 (46)	< .01	.001
Scalar	56004.159	2160	Metric	397.605 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	54133.699	2068				
Metric	55210.489	2114	Configural	1076.790 (46)	< .01	.000
Scalar	56103.674	2160	Metric	893.185 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	54495.762	2068				
Metric	54883.984	2114	Configural	388.222 (46)	< .01	.001
Scalar	55567.969	2160	Metric	683.985 (46)	< .01	.000

Appendix C.17b. Global Model Fit Indices of Scalar Invariance Model for Geometry

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	47585.006	2113	< .01	0.972	0.027
Model B-1	21844.952	2113	< .01	0.975	0.026
Model B-2	40609.805	2113	< .01	0.969	0.027
Model B-3	22656.273	2113	< .01	0.972	0.027
Model B-4	21394.680	2113	< .01	0.975	0.026
Model B-5	21304.556	2113	< .01	0.974	0.027
Model C	37682.658	2113	< .01	0.973	0.024
Model D	44370.674	2113	< .01	0.973	0.026
Model E	35683.277	2113	< .01	0.979	0.023
Model F	22710.357	2113	< .01	0.985	0.018

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

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	34408.276	2068				
Metric	35099.034	2114	Configural	690.759 (46)	< .01	.000
Scalar	37766.637	2160	Metric	2667.603 (46)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	18980.975	2068				
Metric	19330.667	2114	Configural	349.692 (46)	< .01	.000
Scalar	19900.001	2160	Metric	569.334 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	28801.371	2068				
Metric	30550.412	2114	Configural	1749.041 (46)	< .01	.000
Scalar	32535.110	2160	Metric	1984.699 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	18609.019	2068				
Metric	18818.749	2114	Configural	209.730 (46)	< .01	.000
Scalar	19279.885	2160	Metric	461.136 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	18451.304	2068				
Metric	19100.677	2114	Configural	649.372 (46)	< .01	.000
Scalar	19929.627	2160	Metric	828.951 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	18147.871	2068				
Metric	18220.087	2114	Configural	72.215 (46)	< .01	.000
Scalar	18324.483	2160	Metric	104.396 (46)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	34441.356	2068				
Metric	34855.298	2114	Configural	413.941 (46)	< .01	.000
Scalar	35981.930	2160	Metric	1126.632 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	34187.244	2068				
Metric	35020.447	2114	Configural	833.203 (46)	< .01	.000
Scalar	35943.713	2160	Metric	923.266 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	34276.503	2068				
Metric	35019.694	2114	Configural	743.191 (46)	< .01	.000
Scalar	35722.677	2160	Metric	702.983 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	34831.485	2068				
Metric	35021.623	2114	Configural	190.137 (46)	< .01	.000
Scalar	35246.939	2160	Metric	225.316 (46)	< .01	.000

Appendix C.18b. Global Model Fit Indices of Scalar Invariance Model for Algebra II

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	31151.771	2114	< .01	0.980	0.023
Model B-1	14437.573	2114	< .01	0.983	0.022
Model B-2	26165.329	2114	< .01	0.979	0.023
Model B-3	14786.681	2114	< .01	0.982	0.023
Model B-4	13184.534	2114	< .01	0.985	0.021
Model B-5	13172.752	2114	< .01	0.984	0.021
Model C	24721.244	2114	< .01	0.981	0.021
Model D	28207.867	2114	< .01	0.981	0.022
Model E	23495.477	2114	< .01	0.985	0.020
Model F	13035.227	2114	< .01	0.991	0.014

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

Parameter	Spring 2018 G3E to Spring 2019 G4E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2528.03	0.13	<.0001	0.00
Female *vs.* Male (β_{01})	0.62	0.12	<.0001	0.01
Special Education Status *vs*. Non-SPED (β_{02})	-5.48	0.25	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-10.39	0.42	<.0001	-0.09
Low income *vs.* Non-Low Income(β_{04})	-2.58	0.13	<.0001	-0.04
Asian *vs.* White (β_{05})	4.20	0.44	<.0001	0.02
Hispanic *vs.* White (β_{06})	-4.02	0.15	<.0001	-0.06
African American *vs.* White (β_{07})	-4.69	0.31	<.0001	-0.03
Hawaiian/Pacific Islander *vs.* White (β_{08})	-1.99	1.04	0.0554	0.00
American Indian *vs.* White (β_{09})	-7.41	0.35	<.0001	-0.05
Multiple *vs.* White (β_{010})	-0.67	0.36	0.0604	0.00
Slope (β_{10})	0.75	0.00	<.0001	0.77
Female *vs.* Male (β_{11})	-0.01	0.00	0.0451	-0.01
Special Education Status *vs.* Non-SPED (β_{12})	0.04	0.01	<.0001	0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.13	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.01
Asian *vs.* White (β_{15})	-0.01	0.01	0.6450	0.00
Hispanic *vs.* White (β_{16})	0.00	0.00	0.5453	0.00
African American *vs.* White (β_{17})	0.00	0.01	0.8454	0.00
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.03	0.03	0.4249	0.00
American Indian *vs.* White (β_{19})	-0.01	0.01	0.2016	0.00
Multiple *vs.* White (β_{110})	0.02	0.01	0.0405	0.00

Parameter	Spring 2018 G4E to Spring 2019 G5E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2545.77	0.15	<.0001	0.00
Female *vs.* Male (β_{01})	2.50	0.14	<.0001	0.03
Special Education Status *vs* . Non-SPED (β_{02})	-9.09	0.29	<.0001	-0.08
Limited English Proficiency vs. Non-LEP (β_{03})	-10.59	0.46	<.0001	-0.08
Low income *vs.* Non-Low Income(β_{04})	-1.77	0.15	<.0001	-0.02
Asian *vs.* White (β_{05})	2.96	0.49	<.0001	0.01
Hispanic *vs.* White (β_{06})	-2.50	0.17	<.0001	-0.03
African American *vs.* White (β_{07})	-4.34	0.34	<.0001	-0.03
Hawaiian/Pacific Islander *vs.* White (β_{08})	-3.03	1.16	0.0090	0.00
American Indian *vs.* White (β_{09})	-7.02	0.40	<.0001	-0.04
Multiple *vs.* White (β_{010})	-1.27	0.40	0.0015	-0.01
Slope (β_{10})	0.84	0.00	<.0001	0.74
Female *vs.* Male (β_{11})	0.00	0.00	0.4023	0.00
Special Education Status *vs.* Non-SPED (β_{12})	0.07	0.01	<.0001	0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.12	0.01	<.0001	-0.04
Low income vs. Non-Low Income (β_{14})	0.02	0.00	<.0001	0.01
Asian *vs.* White (β_{15})	-0.07	0.01	<.0001	-0.01
Hispanic *vs.* White (β_{16})	0.03	0.01	<.0001	0.02
African American *vs.* White (β_{17})	0.05	0.01	<.0001	0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.04	0.04	0.2405	0.00
American Indian *vs.* White (β_{19})	0.05	0.01	<.0001	0.01
Multiple *vs.* White (β_{110})	0.02	0.01	0.0724	0.00

Parameter	Spring 2018 G5E to Spring 2019 G6E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2546.89	0.13	<.0001	0.00
Female *vs.* Male (β_{01})	2.07	0.12	<.0001	0.03
Special Education Status *vs* . Non-SPED (β_{02})	-5.56	0.28	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-7.15	0.42	<.0001	-0.06
Low income *vs.* Non-Low Income(β_{04})	-1.15	0.13	<.0001	-0.02
Asian *vs.* White (β_{05})	5.61	0.44	<.0001	0.03
Hispanic *vs.* White (β_{06})	-2.16	0.14	<.0001	-0.03
African American *vs.* White (β_{07})	-3.07	0.30	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	-0.72	0.96	0.4570	0.00
American Indian *vs.* White (β_{09})	-4.79	0.35	<.0001	-0.03
Multiple *vs.* White (β_{010})	-1.20	0.35	0.0006	-0.01
Slope (β_{10})	0.77	0.00	<.0001	0.83
Female *vs.* Male (β_{11})	0.00	0.00	0.7859	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.03	0.01	<.0001	-0.01
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.12	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.00	0.01	0.7013	0.00
Hispanic *vs.* White (β_{16})	-0.02	0.00	<.0001	-0.02
African American *vs.* White (β_{17})	-0.02	0.01	0.0158	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.04	0.03	0.2336	0.00
American Indian *vs.* White (β_{19})	-0.07	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	0.00	0.01	0.6394	0.00

Parameter	Spring 2018 G6E to Spring 2019 G7E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2554.37	0.14	<.0001	0.00
Female *vs.* Male (β_{01})	2.24	0.13	<.0001	0.03
Special Education Status *vs* . Non-SPED (β_{02})	-8.06	0.32	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.61	0.50	<.0001	-0.06
Low income *vs.* Non-Low Income(β_{04})	-1.99	0.14	<.0001	-0.03
Asian *vs.* White (β_{05})	4.94	0.49	<.0001	0.02
Hispanic *vs.* White (β_{06})	-1.72	0.16	<.0001	-0.02
African American *vs.* White (β_{07})	-2.89	0.32	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	0.33	1.03	0.7500	0.00
American Indian *vs.* White (β_{09})	-5.27	0.38	<.0001	-0.03
Multiple *vs.* White (β_{010})	0.17	0.40	0.6756	0.00
Slope (β_{10})	0.85	0.00	<.0001	0.80
Female *vs.* Male (β_{11})	-0.03	0.00	<.0001	-0.02
Special Education Status *vs.* Non-SPED (β_{12})	-0.01	0.01	0.4031	0.00
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.11	0.01	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.02	0.01	0.0495	0.00
Hispanic *vs.* White (β_{16})	0.01	0.00	0.1485	0.00
African American *vs.* White (β_{17})	0.02	0.01	0.0994	0.00
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.04	0.03	0.2496	0.00
American Indian *vs.* White (β_{19})	-0.01	0.01	0.4070	0.00
Multiple *vs.* White (β_{110})	0.01	0.01	0.6736	0.00

Parameter	Spring 2018 G7E to Spring 2019 G8E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2561.65	0.14	<.0001	0.00
Female *vs.* Male (β_{01})	3.08	0.13	<.0001	0.04
Special Education Status *vs*. Non-SPED (β_{02})	-8.18	0.33	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-8.68	0.53	<.0001	-0.05
Low income *vs.* Non-Low Income(β_{04})	-1.86	0.14	<.0001	-0.03
Asian *vs.* White (β_{05})	3.30	0.48	<.0001	0.02
Hispanic *vs.* White (β_{06})	-2.65	0.15	<.0001	-0.04
African American *vs.* White (β_{07})	-3.15	0.32	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	-0.78	1.11	0.4799	0.00
American Indian *vs.* White (β_{09})	-6.48	0.38	<.0001	-0.04
Multiple *vs.* White (β_{010})	-0.20	0.41	0.6185	0.00
Slope (β_{10})	0.85	0.00	<.0001	0.81
Female *vs.* Male (β_{11})	0.01	0.00	0.1034	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.06	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.13	0.01	<.0001	-0.04
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.01
Asian *vs.* White (β_{15})	0.02	0.01	0.0856	0.00
Hispanic *vs.* White (β_{16})	0.02	0.00	0.0005	0.01
African American *vs.* White (β_{17})	0.00	0.01	0.8485	0.00
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.01	0.03	0.7172	0.00
American Indian *vs.* White (β_{19})	-0.03	0.01	0.0038	-0.01
Multiple *vs.* White (β_{110})	0.01	0.01	0.6373	0.00

Parameter	Spring 2018 G8E to Spring 2019 G9E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2568.01	0.15	<.0001	0.00
Female *vs.* Male (β_{01})	2.49	0.14	<.0001	0.04
Special Education Status *vs* . Non-SPED (β_{02})	-7.57	0.40	<.0001	-0.07
Limited English Proficiency vs. Non-LEP (β_{03})	-5.06	0.49	<.0001	-0.04
Low income *vs.* Non-Low Income(β_{04})	-1.80	0.17	<.0001	-0.03
Asian *vs.* White (β_{05})	6.21	0.50	<.0001	0.03
Hispanic *vs.* White (β_{06})	-3.22	0.17	<.0001	-0.05
African American *vs.* White (β_{07})	-3.67	0.35	<.0001	-0.03
Hawaiian/Pacific Islander *vs.* White (β_{08})	-1.15	1.09	0.2909	0.00
American Indian *vs.* White (β_{09})	-2.84	0.43	<.0001	-0.02
Multiple *vs.* White (β_{010})	-0.09	0.48	0.8459	0.00
Slope (β_{10})	0.82	0.00	<.0001	0.82
Female *vs.* Male (β_{11})	0.02	0.00	0.0010	0.01
Special Education Status *vs.* Non-SPED (β_{12})	-0.13	0.01	<.0001	-0.05
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.06	0.01	<.0001	-0.02
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.02
Asian *vs.* White (β_{15})	0.04	0.01	0.0042	0.01
Hispanic *vs.* White (β_{16})	-0.05	0.01	<.0001	-0.03
African American *vs.* White (β_{17})	-0.08	0.01	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.01	0.04	0.7799	0.00
American Indian *vs.* White (β_{19})	-0.08	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	-0.02	0.02	0.1082	0.00

Parameter	Spring 2018 G9E to Spring 2019 G10E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2568.82	0.15	<.0001	0.00
Female *vs.* Male (β_{01})	0.68	0.15	<.0001	0.01
Special Education Status *vs*. Non-SPED (β_{02})	-5.22	0.44	<.0001	-0.04
Limited English Proficiency vs. Non-LEP (β_{03})	-3.36	0.54	<.0001	-0.02
Low income *vs.* Non-Low Income(β_{04})	-1.95	0.18	<.0001	-0.03
Asian *vs.* White (β_{05})	4.08	0.53	<.0001	0.02
Hispanic *vs.* White (β_{06})	-2.67	0.17	<.0001	-0.04
African American *vs.* White (β_{07})	-1.96	0.37	<.0001	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{08})	-1.22	1.25	0.3275	0.00
American Indian *vs.* White (β_{09})	-4.82	0.44	<.0001	-0.03
Multiple *vs.* White (β_{010})	-1.24	0.53	0.0182	-0.01
Slope (β_{10})	0.81	0.00	<.0001	0.79
Female *vs.* Male (β_{11})	0.00	0.00	0.4976	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.04	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.03	0.01	0.0161	-0.01
Low income vs. Non-Low Income (β_{14})	0.00	0.01	0.6698	0.00
Asian *vs.* White (β_{15})	-0.02	0.01	0.2177	0.00
Hispanic *vs.* White (β_{16})	0.01	0.01	0.0225	0.01
African American *vs.* White (β_{17})	0.03	0.01	0.0249	0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.05	0.04	0.2351	0.00
American Indian *vs.* White (β_{19})	-0.01	0.01	0.6833	0.00
Multiple *vs.* White (β_{110})	0.05	0.02	0.0098	0.01

Parameter	Spring 2018 G10E to Spring 2019 G11E			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	2571.99	0.16	<.0001	0.00
Female *vs.* Male (β_{01})	2.76	0.17	<.0001	0.04
Special Education Status *vs*. Non-SPED (β_{02})	-7.90	0.48	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-1.05	0.65	0.1069	-0.01
Low income *vs.* Non-Low Income(β_{04})	-1.97	0.20	<.0001	-0.03
Asian *vs.* White (β_{05})	2.79	0.57	<.0001	0.01
Hispanic *vs.* White (β_{06})	-2.63	0.19	<.0001	-0.04
African American *vs.* White (β_{07})	-2.98	0.41	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	-2.40	1.42	0.0902	0.00
American Indian *vs.* White (β_{09})	-4.60	0.49	<.0001	-0.03
Multiple *vs.* White (β_{010})	-1.73	0.58	0.0027	-0.01
Slope (β_{10})	0.82	0.00	<.0001	0.82
Female *vs.* Male (β_{11})	-0.02	0.01	0.0007	-0.01
Special Education Status *vs.* Non-SPED (β_{12})	-0.05	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	0.01	0.02	0.6693	0.00
Low income vs. Non-Low Income (β_{14})	-0.02	0.01	0.0005	-0.01
Asian *vs.* White (β_{15})	0.05	0.01	0.0006	0.01
Hispanic *vs.* White (β_{16})	-0.03	0.01	<.0001	-0.02
African American *vs.* White (β_{17})	-0.01	0.01	0.3308	0.00
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.02	0.04	0.6406	0.00
American Indian *vs.* White (β_{19})	-0.09	0.02	<.0001	-0.02
Multiple *vs.* White (β_{110})	0.00	0.02	0.9935	0.00

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

Parameter	Spring 2018 G3M to Spring 2019 G4M			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3562.63	0.20	<.0001	0.00
Female *vs.* Male (β_{01})	-0.31	0.18	0.0816	0.00
Special Education Status *vs*. Non-SPED (β_{02})	-7.33	0.34	<.0001	-0.05
Limited English Proficiency vs. Non-LEP (β_{03})	-8.10	0.47	<.0001	-0.05
Low income *vs.* Non-Low Income(β_{04})	-3.20	0.19	<.0001	-0.04
Asian *vs.* White (β_{05})	5.85	0.71	<.0001	0.02
Hispanic *vs.* White (β_{06})	-4.13	0.21	<.0001	-0.05
African American *vs.* White (β_{07})	-7.29	0.44	<.0001	-0.04
Hawaiian/Pacific Islander *vs.* White (β_{08})	-0.75	1.48	0.6143	0.00
American Indian *vs.* White (β_{09})	-8.97	0.49	<.0001	-0.04
Multiple *vs.* White (β_{010})	-0.87	0.51	0.0881	0.00
Slope (β_{10})	0.77	0.00	<.0001	0.80
Female *vs.* Male (β_{11})	-0.01	0.00	0.0037	-0.01
Special Education Status *vs.* Non-SPED (β_{12})	0.01	0.01	0.0085	0.01
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.08	0.01	<.0001	-0.03
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.01	0.01	0.4735	0.00
Hispanic *vs.* White (β_{16})	-0.01	0.00	0.0058	-0.01
African American *vs.* White (β_{17})	-0.01	0.01	0.2392	0.00
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.01	0.03	0.6776	0.00
American Indian *vs.* White (β_{19})	-0.04	0.01	0.0003	-0.01
Multiple *vs.* White (β_{110})	0.00	0.01	0.7944	0.00

Parameter	Spring 2018 G4M to Spring 2019 G5M			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3590.26	0.18	<.0001	0.00
Female *vs.* Male (β_{01})	0.91	0.16	<.0001	0.01
Special Education Status *vs*. Non-SPED (β_{02})	-7.13	0.33	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-5.89	0.44	<.0001	-0.04
Low income *vs.* Non-Low Income(β_{04})	-2.14	0.18	<.0001	-0.02
Asian *vs.* White (β_{05})	7.79	0.64	<.0001	0.03
Hispanic *vs.* White (β_{06})	-2.10	0.20	<.0001	-0.02
African American *vs.* White (β_{07})	-5.31	0.41	<.0001	-0.03
Hawaiian/Pacific Islander *vs.* White (β_{08})	0.41	1.38	0.7675	0.00
American Indian *vs.* White (β_{09})	-5.27	0.47	<.0001	-0.03
Multiple *vs.* White (β_{010})	-1.17	0.48	0.0137	0.00
Slope (β_{10})	0.82	0.00	<.0001	0.85
Female *vs.* Male (β_{11})	0.00	0.00	0.3302	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.03	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.12	0.01	<.0001	-0.05
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.01	0.01	0.4332	0.00
Hispanic *vs.* White (β_{16})	-0.05	0.00	<.0001	-0.03
African American *vs.* White (β_{17})	-0.07	0.01	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.00	0.03	0.9595	0.00
American Indian *vs.* White (β_{19})	-0.09	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	0.00	0.01	0.8081	0.00

Parameter	Spring 2018 G5M to Spring 2019 G6M			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3620.44	0.17	<.0001	0.00
Female *vs.* Male (β_{01})	0.40	0.16	0.0114	0.00
Special Education Status *vs*. Non-SPED (β_{02})	-7.27	0.35	<.0001	-0.05
Limited English Proficiency vs. Non-LEP (β_{03})	-7.93	0.45	<.0001	-0.05
Low income *vs.* Non-Low Income(β_{04})	-1.06	0.17	<.0001	-0.01
Asian *vs.* White (β_{05})	4.19	0.64	<.0001	0.02
Hispanic *vs.* White (β_{06})	-4.07	0.19	<.0001	-0.05
African American *vs.* White (β_{07})	-5.76	0.41	<.0001	-0.03
Hawaiian/Pacific Islander *vs.* White (β_{08})	-4.29	1.28	0.0008	-0.01
American Indian *vs.* White (β_{09})	-6.92	0.45	<.0001	-0.03
Multiple *vs.* White (β_{010})	-2.64	0.46	<.0001	-0.01
Slope (β_{10})	0.80	0.00	<.0001	0.85
Female *vs.* Male (β_{11})	0.00	0.00	0.8386	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.04	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.10	0.01	<.0001	-0.04
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.04	0.01	0.0006	0.01
Hispanic *vs.* White (β_{16})	-0.03	0.00	<.0001	-0.02
African American *vs.* White (β_{17})	-0.02	0.01	0.0067	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.02	0.03	0.3797	0.00
American Indian *vs.* White (β_{19})	-0.08	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	0.00	0.01	0.9665	0.00

Parameter	Spring 2018 G6M to Spring 2019 G7M			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3640.33	0.16	<.0001	0.00
Female *vs.* Male (β_{01})	-0.54	0.15	0.0003	-0.01
Special Education Status *vs* . Non-SPED (β_{02})	-8.37	0.34	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-10.56	0.47	<.0001	-0.06
Low income *vs.* Non-Low Income(β_{04})	-1.77	0.16	<.0001	-0.02
Asian *vs.* White (β_{05})	4.43	0.60	<.0001	0.02
Hispanic *vs.* White (β_{06})	-4.14	0.18	<.0001	-0.05
African American *vs.* White (β_{07})	-4.82	0.38	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	-2.01	1.17	0.0848	0.00
American Indian *vs.* White (β_{09})	-7.18	0.42	<.0001	-0.03
Multiple *vs.* White (β_{010})	-0.80	0.44	0.0708	0.00
Slope (β_{10})	0.83	0.00	<.0001	0.89
Female *vs.* Male (β_{11})	-0.01	0.00	0.0002	-0.01
Special Education Status *vs.* Non-SPED (β_{12})	-0.13	0.01	<.0001	-0.06
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.17	0.01	<.0001	-0.06
Low income vs. Non-Low Income (β_{14})	-0.04	0.00	<.0001	-0.02
Asian *vs.* White (β_{15})	0.03	0.01	0.0087	0.01
Hispanic *vs.* White (β_{16})	-0.03	0.00	<.0001	-0.02
African American *vs.* White (β_{17})	-0.04	0.01	<.0001	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.02	0.03	0.3758	0.00
American Indian *vs.* White (β_{19})	-0.08	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	0.01	0.01	0.1627	0.00

Parameter	Spring 2018 G7M to Spring 2019 G8M			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3655.55	0.17	<.0001	0.00
Female *vs.* Male (β_{01})	1.45	0.16	<.0001	0.02
Special Education Status *vs*. Non-SPED (β_{02})	-6.40	0.37	<.0001	-0.05
Limited English Proficiency vs. Non-LEP (β_{03})	-3.39	0.53	<.0001	-0.02
Low income *vs.* Non-Low Income(β_{04})	-0.91	0.17	<.0001	-0.01
Asian *vs.* White (β_{05})	4.79	0.67	<.0001	0.02
Hispanic *vs.* White (β_{06})	-0.69	0.19	0.0003	-0.01
African American *vs.* White (β_{07})	-0.43	0.39	0.2749	0.00
Hawaiian/Pacific Islander *vs.* White (β_{08})	1.89	1.34	0.1572	0.00
American Indian *vs.* White (β_{09})	-2.43	0.43	<.0001	-0.01
Multiple *vs.* White (β_{010})	-0.39	0.50	0.4368	0.00
Slope (β_{10})	0.86	0.00	<.0001	0.89
Female *vs.* Male (β_{11})	-0.03	0.00	<.0001	-0.02
Special Education Status *vs.* Non-SPED (β_{12})	-0.15	0.01	<.0001	-0.06
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.12	0.01	<.0001	-0.04
Low income vs. Non-Low Income (β_{14})	-0.02	0.00	<.0001	-0.01
Asian *vs.* White (β_{15})	0.07	0.01	<.0001	0.01
Hispanic *vs.* White (β_{16})	-0.03	0.00	<.0001	-0.02
African American *vs.* White (β_{17})	-0.05	0.01	<.0001	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.05	0.03	0.1449	0.00
American Indian *vs.* White (β_{19})	-0.09	0.01	<.0001	-0.02
Multiple *vs.* White (β_{110})	-0.01	0.01	0.6388	0.00

Parameter	Spring 2018 G8M to Spring 2019 AlgI			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3669.03	0.20	<.0001	0.00
Female *vs.* Male (β_{01})	3.28	0.19	<.0001	0.05
Special Education Status *vs*. Non-SPED (β_{02})	-6.48	0.45	<.0001	-0.06
Limited English Proficiency vs. Non-LEP (β_{03})	-6.76	0.53	<.0001	-0.05
Low income *vs.* Non-Low Income(β_{04})	-1.70	0.21	<.0001	-0.02
Asian *vs.* White (β_{05})	5.71	0.79	<.0001	0.02
Hispanic *vs.* White (β_{06})	-1.55	0.22	<.0001	-0.02
African American *vs.* White (β_{07})	-1.55	0.44	0.0004	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{08})	1.75	1.41	0.2151	0.00
American Indian *vs.* White (β_{09})	-2.93	0.51	<.0001	-0.02
Multiple *vs.* White (β_{010})	0.06	0.63	0.9254	0.00
Slope (β_{10})	0.74	0.01	<.0001	0.80
Female *vs.* Male (β_{11})	-0.01	0.01	0.0967	-0.01
Special Education Status *vs.* Non-SPED (β_{12})	-0.05	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.07	0.01	<.0001	-0.02
Low income vs. Non-Low Income (β_{14})	0.00	0.01	0.4925	0.00
Asian *vs.* White (β_{15})	-0.01	0.02	0.5454	0.00
Hispanic *vs.* White (β_{16})	-0.04	0.01	<.0001	-0.03
African American *vs.* White (β_{17})	-0.06	0.01	<.0001	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.03	0.04	0.4327	0.00
American Indian *vs.* White (β_{19})	-0.04	0.01	0.0027	-0.01
Multiple *vs.* White (β_{110})	-0.01	0.02	0.4027	0.00

Parameter	Spring 2018 AlgI to Spring 2019 Geo			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3690.77	0.19	<.0001	0.00
Female *vs.* Male (β_{01})	-0.68	0.19	0.0003	-0.01
Special Education Status *vs*. Non-SPED (β_{02})	-6.10	0.53	<.0001	-0.04
Limited English Proficiency vs. Non-LEP (β_{03})	-2.60	0.63	<.0001	-0.01
Low income *vs.* Non-Low Income(β_{04})	-0.98	0.23	<.0001	-0.01
Asian *vs.* White (β_{05})	2.33	0.66	0.0005	0.01
Hispanic *vs.* White (β_{06})	-2.54	0.22	<.0001	-0.03
African American *vs.* White (β_{07})	-6.25	0.48	<.0001	-0.04
Hawaiian/Pacific Islander *vs.* White (β_{08})	0.07	1.58	0.9633	0.00
American Indian *vs.* White (β_{09})	-1.83	0.56	0.0010	-0.01
Multiple *vs.* White (β_{010})	0.09	0.66	0.8929	0.00
Slope (β_{10})	0.85	0.00	<.0001	0.84
Female *vs.* Male (β_{11})	0.01	0.01	0.1683	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.11	0.01	<.0001	-0.04
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.09	0.02	<.0001	-0.02
Low income vs. Non-Low Income (β_{14})	-0.01	0.01	0.0546	-0.01
Asian *vs.* White (β_{15})	0.07	0.01	<.0001	0.02
Hispanic *vs.* White (β_{16})	-0.07	0.01	<.0001	-0.04
African American *vs.* White (β_{17})	-0.10	0.01	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.05	0.05	0.2691	0.00
American Indian *vs.* White (β_{19})	-0.06	0.02	<.0001	-0.01
Multiple *vs.* White (β_{110})	0.02	0.02	0.2328	0.00

Parameter	Spring 2018 Geo to Spring 2019 AlgII			
	Unstandardized Estimate	SE	P value	Standardized Estimate
Intercept (β_{00})	3705.58	0.22	<.0001	0.00
Female *vs.* Male (β_{01})	2.33	0.22	<.0001	0.03
Special Education Status *vs* . Non-SPED (β_{02})	-6.23	0.61	<.0001	-0.04
Limited English Proficiency vs. Non-LEP (β_{03})	0.19	0.72	0.7959	0.00
Low income *vs.* Non-Low Income(β_{04})	-1.66	0.27	<.0001	-0.02
Asian *vs.* White (β_{05})	3.61	0.78	<.0001	0.02
Hispanic *vs.* White (β_{06})	-2.44	0.25	<.0001	-0.03
African American *vs.* White (β_{07})	-2.93	0.56	<.0001	-0.02
Hawaiian/Pacific Islander *vs.* White (β_{08})	1.66	1.87	0.3730	0.00
American Indian *vs.* White (β_{09})	-6.58	0.64	<.0001	-0.03
Multiple *vs.* White (β_{010})	-0.11	0.77	0.8817	0.00
Slope (β_{10})	0.84	0.01	<.0001	0.84
Female *vs.* Male (β_{11})	-0.01	0.01	0.3174	0.00
Special Education Status *vs.* Non-SPED (β_{12})	-0.08	0.01	<.0001	-0.02
Limited English Proficiency *vs.* Non-LEP (β_{13})	0.02	0.02	0.1868	0.00
Low income vs. Non-Low Income (β_{14})	-0.04	0.01	<.0001	-0.02
Asian *vs.* White (β_{15})	0.03	0.02	0.0749	0.01
Hispanic *vs.* White (β_{16})	-0.08	0.01	<.0001	-0.04
African American *vs.* White (β_{17})	-0.06	0.02	<.0001	-0.01
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.06	0.05	0.2548	0.00
American Indian *vs.* White (β_{19})	-0.16	0.02	<.0001	-0.03
Multiple *vs.* White (β_{110})	-0.01	0.02	0.5767	0.00

Appendix E.1—Spring 19 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	13022_C	extendedTextInteraction	-1.75863	-0.83246		-1.29555
2	13022_E	extendedTextInteraction	-1.50579	1.76305	4.40933	1.55553
3	13022_O	extendedTextInteraction	-1.67709	1.56754	4.24577	1.37874
4	13025_C	extendedTextInteraction	-1.58499	-0.47921		-1.0321
5	13025_E	extendedTextInteraction	-1.28491	2.22063	3.90277	1.61283
6	13025_O	extendedTextInteraction	-0.6838	1.77292	4.29041	1.793177
7	17964	choiceInteraction	-1.13667			-1.13667
8	17961	choiceInteraction	1.48294			1.48294
9	17965	choiceInteraction	-0.15009			-0.15009
10	17958	choiceInteraction , choiceInteraction	-0.04249	0.4911		0.224305
11	17968	choiceInteraction	0.85138			0.85138
12	17959	matchInteraction	-0.32436			-0.32436
13	12691	choiceInteraction	0.18858			0.18858
14	12701	choiceInteraction	0.7635			0.7635
15	12746	choiceInteraction	-0.03473			-0.03473
16	12208	choiceInteraction	0.05676			0.05676
17	12216	choiceInteraction	0.21638			0.21638
18	8708	inlineChoiceInteraction	-1.85419			-1.85419
19	8709	inlineChoiceInteraction , inlineChoiceInteraction	-1.02511	-0.31661		-0.67086
20	8710	inlineChoiceInteraction	-1.27986			-1.27986
21	8711	inlineChoiceInteraction	-0.1057			-0.1057
22	12990	choiceInteraction	-0.08639			-0.08639
23	12996	choiceInteraction	0.56311			0.56311
24	12995	choiceInteraction , choiceInteraction	2.32576			2.32576
25	12999	choiceInteraction	0.07539			0.07539
26	12992	customInteraction	0.34963			0.34963
27	17539	choiceInteraction	0.52384			0.52384
28	11867	choiceInteraction	-0.53376			-0.53376
29	12521	choiceInteraction	0.16092			0.16092
30	12417	choiceInteraction	0.83368			0.83368
31	11854	choiceInteraction , choiceInteraction	0.89082			0.89082
32	17883	choiceInteraction	0.23819			0.23819
33	17860	choiceInteraction	-0.10701			-0.10701
34	17878	choiceInteraction	0.15777			0.15777
35	17866	choiceInteraction	0.3867			0.3867
36	17901	matchInteraction				
37	17859	choiceInteraction	-0.60585			-0.60585
38	17861	choiceInteraction	0.36146			0.36146
39	10630	choiceInteraction	-0.90745			-0.90745
40	9414	choiceInteraction	-0.03304			-0.03304

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10628	choiceInteraction	-1.10023			-1.10023
42	9422	choiceInteraction	-0.87555			-0.87555
43	10632	choiceInteraction	0.48179			0.48179
44	10634	choiceInteraction	1.17725			1.17725
45	18115	inlineChoiceInteraction	0.71041			0.71041
46	18118	inlineChoiceInteraction , inlineChoiceInteraction				
47	18134	inlineChoiceInteraction , inlineChoiceInteraction	-1.38166	-0.17171		-0.77669

Appendix E.2—Spring 19 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	13120_C	extendedTextInteraction	-1.25847	1.5145		0.128015
2	13120_E	extendedTextInteraction	1.98955	5.05842	4.4995	3.849157
3	13120_O	extendedTextInteraction	0.90148	4.27097	6.19001	3.787487
4	13119_C	extendedTextInteraction	-1.65829	1.05358		-0.30236
5	13119_E	extendedTextInteraction	1.54979	4.4619	6.46226	4.157983
6	13119_O	extendedTextInteraction	0.64594	4.34742	5.35763	3.45033
7	16003	choiceInteraction	-0.59814			-0.59814
8	16006	choiceInteraction	-0.32739			-0.32739
9	16005	choiceInteraction	-0.59429			-0.59429
10	16008	choiceInteraction	-0.55711			-0.55711
11	16002	choiceInteraction	0.73393			0.73393
12	16009	customInteraction	0.69813			0.69813
13	11840	choiceInteraction	0.09196			0.09196
14	12567	choiceInteraction	-0.08581			-0.08581
15	11837	choiceInteraction	-0.34149			-0.34149
16	11841	choiceInteraction	0.87491			0.87491
17	11844	choiceInteraction , choiceInteraction	1.01626	0.92279		0.969525
18	11967	customInteraction	0.32005			0.32005
19	11847	choiceInteraction	0.29982			0.29982
20	16093	inlineChoiceInteraction	0.34829			0.34829
21	16094	inlineChoiceInteraction , inlineChoiceInteraction	-1.4101	0.48824		-0.46093
22	16095	inlineChoiceInteraction , inlineChoiceInteraction	-1.32027	1.33667		0.0082
23	18527	choiceInteraction	-1.01472			-1.01472
24	18518	choiceInteraction	-0.00942			-0.00942
25	18524	choiceInteraction	1.02876			1.02876
26	18522	choiceInteraction	-0.24254			-0.24254
27	18519	hottextInteraction	-0.33131			-0.33131
28	18525	choiceInteraction	0.73941			0.73941
29	11915	choiceInteraction	-0.27048			-0.27048
30	11930	choiceInteraction	0.5267			0.5267
31	11910	choiceInteraction	0.76403			0.76403
32	11957	customInteraction	0.72112	1.17084		0.94598
33	11949	choiceInteraction	0.02743			0.02743
34	17656	matchInteraction	1.8244			1.8244
35	12317	choiceInteraction	0.80561			0.80561
36	12666	choiceInteraction	0.42527			0.42527
37	12653	choiceInteraction	-0.53559			-0.53559

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
38	12647	choiceInteraction	0.22325			0.22325
39	18542	choiceInteraction	0.22801			0.22801
40	18546	choiceInteraction	-0.75396			-0.75396
41	18541	choiceInteraction	0.28732			0.28732
42	18547	choiceInteraction	0.71132			0.71132
43	18548	choiceInteraction	0.36471			0.36471
44	16080	inlineChoiceInteraction	-0.62177			-0.62177
45	16081	inlineChoiceInteraction , inlineChoiceInteraction	-1.07187	0.37369		-0.34909
46	16084	inlineChoiceInteraction	-0.33913			-0.33913
47	16085	inlineChoiceInteraction	-0.28574			-0.28574

Appendix E.3—Spring 19 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	13247_C	extendedTextInteraction	-2.16137	0.43573		-0.86282
2	13247_E	extendedTextInteraction	-0.45605	3.2947	4.52191	2.45352
3	13247_O	extendedTextInteraction	-1.00143	2.7942	4.59779	2.130187
4	13246_C	extendedTextInteraction	-2.03231	0.35683		-0.83774
5	13246_E	extendedTextInteraction	0.33484	4.37175	6.34064	3.68241
6	13246_O	extendedTextInteraction	-0.55282	3.33662	5.17957	2.654457
7	9303	choiceInteraction	0.14403			0.14403
8	9305	choiceInteraction	-1.06814			-1.06814
9	9300	choiceInteraction	0.08946			0.08946
10	9304	choiceInteraction	0.5267			0.5267
11	9301	choiceInteraction	0.26042			0.26042
12	9302	choiceInteraction	0.397			0.397
13	18059	choiceInteraction	0.83005			0.83005
14	18054	choiceInteraction , choiceInteraction	0.43536			0.43536
15	18049	matchInteraction	-0.36268			-0.36268
16	18050	choiceInteraction	-0.59245			-0.59245
17	18044	choiceInteraction , choiceInteraction	1.31788			1.31788
18	18053	choiceInteraction	1.05518			1.05518
19	18058	choiceInteraction	-0.54115			-0.54115
20	18155	choiceInteraction	0.65584			0.65584
21	18168	choiceInteraction	0.98299			0.98299
22	10659	inlineChoiceInteraction	-0.82784			-0.82784
23	10661	inlineChoiceInteraction , inlineChoiceInteraction	-2.20002	-0.66798		-1.434
24	10662	inlineChoiceInteraction	-1.01469			-1.01469
25	18593	choiceInteraction	-0.42973			-0.42973
26	18597	choiceInteraction	0.26411			0.26411
27	18594	choiceInteraction	0.55591			0.55591
28	18592	choiceInteraction	0.48222			0.48222
29	18590	choiceInteraction , choiceInteraction	0.80012			0.80012
30	14861	choiceInteraction	-0.46617			-0.46617
31	14862	choiceInteraction	-0.51209			-0.51209
32	14866	choiceInteraction	-0.38581			-0.38581
33	14864	choiceInteraction	-1.06794			-1.06794
34	9306	choiceInteraction	-0.33688			-0.33688
35	9308	choiceInteraction	-0.32336			-0.32336
36	9299	choiceInteraction	-0.1376			-0.1376

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
37	9312	customInteraction	2.72402			2.72402
38	18046	choiceInteraction	-0.01825			-0.01825
39	18038	choiceInteraction	-0.7338			-0.7338
40	18040	choiceInteraction	-0.65858			-0.65858
41	18042	choiceInteraction	-0.10591			-0.10591
42	18045	choiceInteraction	0.24955			0.24955
43	18164	choiceInteraction	0.72121			0.72121
44	18051	matchInteraction	1.80364			1.80364
45	9286	inlineChoiceInteraction	-0.23947			-0.23947
46	9287	inlineChoiceInteraction	1.18837			1.18837
47	9288	inlineChoiceInteraction , inlineChoiceInteraction	-1.29281	0.66758		-0.31262

Appendix E.4—Spring 19 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	13307_C	extendedTextInteraction	-2.10835	-0.10872		-1.10854
2	13307_E	extendedTextInteraction	0.4746	3.03111	4.52238	2.67603
3	13307_O	extendedTextInteraction	-0.54732	2.76978	4.2732	2.16522
4	13306_C	extendedTextInteraction	-2.06099	-0.58055		-1.32077
5	13306_E	extendedTextInteraction	0.07056	3.38151	6.19445	3.215507
6	13306_O	extendedTextInteraction	-0.59192	2.51057	5.24377	2.387473
7	18196	choiceInteraction	-0.90573			-0.90573
8	18195	choiceInteraction	0.71971			0.71971
9	18201	choiceInteraction	-0.1235			-0.1235
10	18189	choiceInteraction	-0.47927			-0.47927
11	18202	choiceInteraction	-0.29306			-0.29306
12	13259	choiceInteraction	-0.13029			-0.13029
13	13271	choiceInteraction	-0.32339			-0.32339
14	13287	choiceInteraction , choiceInteraction	0.0232			0.0232
15	13274	choiceInteraction	-0.56873			-0.56873
16	13261	choiceInteraction	-0.02637			-0.02637
17	13260	choiceInteraction	-0.13377			-0.13377
18	13264	choiceInteraction	-1.054			-1.054
19	13263	choiceInteraction	1.61746			1.61746
20	9107	inlineChoiceInteraction	-1.86994			-1.86994
21	9108	inlineChoiceInteraction , inlineChoiceInteraction	-1.32061	1.24437		-0.03812
22	9109	inlineChoiceInteraction , inlineChoiceInteraction	-1.49476	0.62001		-0.43738
23	16031	choiceInteraction	-0.01495			-0.01495
24	16029	choiceInteraction	-0.00757			-0.00757
25	16027	choiceInteraction	1.02018			1.02018
26	16028	choiceInteraction	0.24012			0.24012
27	16033	choiceInteraction	0.33379			0.33379
28	16030	choiceInteraction	0.70326			0.70326
29	16138	choiceInteraction , choiceInteraction	0.62183			0.62183
30	18608	choiceInteraction	-0.12196			-0.12196
31	18615	choiceInteraction	-0.57883			-0.57883
32	18617	choiceInteraction	0.04547			0.04547
33	17483	choiceInteraction	0.85374			0.85374
34	18616	choiceInteraction	0.13575			0.13575
35	18619	choiceInteraction	0.29569			0.29569
36	18660	choiceInteraction	0.27295			0.27295
37	18659	choiceInteraction	0.01174			0.01174
38	18656	choiceInteraction	-0.91868			-0.91868

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	18654	choiceInteraction	0.41557			0.41557
40	18655	choiceInteraction , choiceInteraction	1.21991			1.21991
41	9872	customInteraction	1.87785			1.87785
42	10280	choiceInteraction	0.14517			0.14517
43	9867	choiceInteraction	0.88477			0.88477
44	9865	choiceInteraction	0.65339			0.65339
45	9866	choiceInteraction	0.07074			0.07074
46	13248	inlineChoiceInteraction	-1.12962			-1.12962
47	13250	inlineChoiceInteraction , inlineChoiceInteraction	-1.67091	-0.39014		-1.03053

Appendix E.5—Spring 19 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	13401_C	extendedTextInteraction	-2.93019	-0.87734		-1.90377
2	13401_E	extendedTextInteraction	-1.08712	3.58298	4.84348	2.446447
3	13401_O	extendedTextInteraction	-2.01399	2.99796	5.04182	2.008597
4	13406_C	extendedTextInteraction	-2.53963	0.10313		-1.21825
5	13406_E	extendedTextInteraction	-0.83081	4.35574	5.68822	3.07105
6	13406_O	extendedTextInteraction	-1.24126	3.45562	5.9479	2.720753
7	16197	choiceInteraction	-0.12794			-0.12794
8	16199	customInteraction	-1.38888			-1.38888
9	16198	choiceInteraction	0.00146			0.00146
10	16115	choiceInteraction	0.70087			0.70087
11	16200	choiceInteraction	-0.36837			-0.36837
12	16155	choiceInteraction , choiceInteraction	0.77837			0.77837
13	16201	choiceInteraction	-0.22346			-0.22346
14	16118	choiceInteraction	-0.99956			-0.99956
15	17520	choiceInteraction	-0.53139			-0.53139
16	18718	choiceInteraction	-0.04659			-0.04659
17	18716	choiceInteraction	0.37883			0.37883
18	18720	matchInteraction	0.16355			0.16355
19	18721	choiceInteraction	1.40891			1.40891
20	16120	inlineChoiceInteraction	-1.16575			-1.16575
21	16121	inlineChoiceInteraction	-0.77536			-0.77536
22	16122	inlineChoiceInteraction	-1.17838			-1.17838
23	14807	choiceInteraction	0.5733			0.5733
24	14805	choiceInteraction	-0.66921			-0.66921
25	14809	choiceInteraction	1.76118			1.76118
26	14804	choiceInteraction	1.14494			1.14494
27	9743	choiceInteraction	0.07478			0.07478
28	9741	choiceInteraction	1.30368			1.30368
29	9847	choiceInteraction	0.68817			0.68817
30	9740	choiceInteraction	-1.52845			-1.52845
31	9747	choiceInteraction	1.6429			1.6429
32	9845	choiceInteraction	1.05787			1.05787
33	9610	choiceInteraction	0.90991			0.90991
34	9611	customInteraction	0.44176			0.44176
35	9711	choiceInteraction	0.29644			0.29644
36	9713	choiceInteraction	0.2301			0.2301
37	10695	choiceInteraction	-0.51732			-0.51732
38	10613	choiceInteraction	1.17647			1.17647
39	9750	choiceInteraction	-0.17847			-0.17847

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
40	9709	choiceInteraction	-0.24287			-0.24287
41	18682	choiceInteraction	0.34503			0.34503
42	18684	choiceInteraction	-0.82004			-0.82004
43	18686	choiceInteraction	-1.49275			-1.49275
44	18688	choiceInteraction	0.09794			0.09794
45	16124	inlineChoiceInteraction	0.12234			0.12234
46	16126	inlineChoiceInteraction	-0.04171			-0.04171
47	16127	inlineChoiceInteraction , inlineChoiceInteraction	-1.57246	0.2339		-0.66928

Appendix E.6—Spring 19 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	13454_C	extendedTextInteraction	-2.12549	-0.79592		-1.46071
2	13454_E	extendedTextInteraction	-1.16469	1.93019	3.31328	1.359593
3	13454_O	extendedTextInteraction	-1.30159	1.37223	3.31286	1.127833
4	13439_C	extendedTextInteraction	-2.40518	-1.1545		-1.77984
5	13439_E	extendedTextInteraction	-1.77948	1.72887	2.90301	0.9508
6	13439_O	extendedTextInteraction	-1.71597	1.09631	3.03584	0.805393
7	11813	choiceInteraction	0.19219			0.19219
8	11810	choiceInteraction , choiceInteraction	0.87267			0.87267
9	11814	choiceInteraction	-0.35945			-0.35945
10	11815	choiceInteraction , choiceInteraction	0.62015			0.62015
11	11816	choiceInteraction	-0.07166			-0.07166
12	11811	choiceInteraction	-0.25816			-0.25816
13	11820	customInteraction	-0.7631			-0.7631
14	11812	choiceInteraction	-0.25561			-0.25561
15	12429	choiceInteraction , choiceInteraction	-0.57203	2.62188		1.024925
16	12685	customInteraction	-0.49068			-0.49068
17	12660	choiceInteraction , choiceInteraction	0.15505	-0.11586		0.019595
18	12696	choiceInteraction	-0.8354			-0.8354
19	17735	choiceInteraction	2.06006			2.06006
20	12651	choiceInteraction	0.24816			0.24816
21	12702	choiceInteraction , choiceInteraction	1.64156			1.64156
22	9727	inlineChoiceInteraction	-2.20531			-2.20531
23	9728	inlineChoiceInteraction , inlineChoiceInteraction	-1.65694	-0.20084		-0.92889
24	9729	inlineChoiceInteraction , inlineChoiceInteraction	-1.73003	0.26659		-0.73172
25	18103	choiceInteraction , choiceInteraction	1.0333			1.0333
26	18173	choiceInteraction	0.25292			0.25292
27	18218	choiceInteraction	-0.97689			-0.97689
28	18135	hottextInteraction	0.10977			0.10977
29	18203	choiceInteraction	0.11964			0.11964
30	17776	choiceInteraction	-1.06416			-1.06416
31	12447	choiceInteraction	0.1355			0.1355
32	12454	choiceInteraction	-0.78082			-0.78082
33	12445	choiceInteraction	-0.09919			-0.09919

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
34	12450	choiceInteraction	-0.19005			-0.19005
35	17679	choiceInteraction	0.92372			0.92372
36	18850	choiceInteraction	0.00906			0.00906
37	18849	choiceInteraction	-0.2368			-0.2368
38	18851	choiceInteraction	-0.19757			-0.19757
39	18852	choiceInteraction	0.0936			0.0936
40	18129	choiceInteraction	0.66653			0.66653
41	18131	choiceInteraction	1.69311			1.69311
42	18128	choiceInteraction	0.37957			0.37957
43	18120	choiceInteraction	1.77873			1.77873
44	18119	choiceInteraction , choiceInteraction	1.63812			1.63812
45	16270	inlineChoiceInteraction	-0.42517			-0.42517
46	16272	inlineChoiceInteraction , inlineChoiceInteraction	-1.44518	-0.54267		-0.99393
47	16273	inlineChoiceInteraction , inlineChoiceInteraction	-2.03019	-0.65552		-1.34286

Appendix E.7—Spring 19 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	13556_C	extendedTextInteraction	-2.38602	-1.37075		-1.87839
2	13556_E	extendedTextInteraction	-1.38759	2.43865	3.83392	1.628327
3	13556_O	extendedTextInteraction	-1.76241	1.77841	3.89649	1.304163
4	13555_C	extendedTextInteraction	-2.4999	-1.03399		-1.76695
5	13555_E	extendedTextInteraction	-1.46735	2.51036	4.0225	1.688503
6	13555_O	extendedTextInteraction	-1.7157	1.64362	3.78329	1.23707
7	16445	choiceInteraction	-0.27739			-0.27739
8	16441	customInteraction	-0.56351			-0.56351
9	16446	choiceInteraction , choiceInteraction	1.09465			1.09465
10	16442	choiceInteraction	-0.28203			-0.28203
11	13562	choiceInteraction	0.241			0.241
12	13563	choiceInteraction	0.34102			0.34102
13	13559	choiceInteraction	-1.26254			-1.26254
14	13561	choiceInteraction	0.79566			0.79566
15	13564	choiceInteraction	1.36543			1.36543
16	9047	choiceInteraction	1.04482			1.04482
17	9048	choiceInteraction	-0.38622			-0.38622
18	9052	choiceInteraction , choiceInteraction	0.9943			0.9943
19	9053	choiceInteraction	-0.83869			-0.83869
20	9051	choiceInteraction , choiceInteraction	0.20427			0.20427
21	9734	inlineChoiceInteraction	-0.38567			-0.38567
22	9735	inlineChoiceInteraction , inlineChoiceInteraction	-1.01417	1.57881		0.28232
23	9736	inlineChoiceInteraction	-0.28792			-0.28792
24	9737	inlineChoiceInteraction	-1.41093			-1.41093
25	16488	choiceInteraction	0.26425			0.26425
26	16485	choiceInteraction	0.45266			0.45266
27	16492	choiceInteraction	-0.44001			-0.44001
28	16496	choiceInteraction	0.01893			0.01893
29	16464	choiceInteraction	0.33373			0.33373
30	16493	choiceInteraction	-0.03566			-0.03566
31	13515	choiceInteraction	-0.23416			-0.23416
32	13553	choiceInteraction	0.25005			0.25005
33	13516	choiceInteraction	-0.71545			-0.71545
34	13518	choiceInteraction	0.23285			0.23285
35	13551	choiceInteraction , choiceInteraction	0.69902			0.69902
36	13534	choiceInteraction	-0.67176			-0.67176
37	15034	choiceInteraction	-0.64331			-0.64331
38	15043	choiceInteraction	-0.0511			-0.0511
39	15049	choiceInteraction	0.33294			0.33294
40	15036	choiceInteraction	0.12149			0.12149

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	15047	choiceInteraction	0.60804			0.60804
42	17720	choiceInteraction	-0.43707			-0.43707
43	12817	choiceInteraction	0.74507			0.74507
44	17724	choiceInteraction	0.09931			0.09931
45	12809	choiceInteraction	-0.40091			-0.40091
46	12808	choiceInteraction , choiceInteraction	0.33411	0.2086		0.271355
47	13455	inlineChoiceInteraction	0.34843			0.34843
48	13456	inlineChoiceInteraction , inlineChoiceInteraction	0.10449	1.39644		0.750465
49	13457	inlineChoiceInteraction , inlineChoiceInteraction	-0.84937	0.5127		-0.16834

Appendix E.8—Spring 19 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	13638_C	extendedTextInteraction	-3.08122	-1.11156		-2.09639
2	13638_E	extendedTextInteraction	-1.78924	1.61688	3.39067	1.07277
3	13638_O	extendedTextInteraction	-1.97581	0.93168	3.19227	0.716047
4	13637_C	extendedTextInteraction	-2.97674	-0.96661		-1.97168
5	13637_E	extendedTextInteraction	-1.2983	2.05466	3.36557	1.373977
6	13637_O	extendedTextInteraction	-1.89522	1.40099	3.39837	0.968047
7	12332	choiceInteraction , choiceInteraction	0.64991			0.64991
8	12807	choiceInteraction	-0.77456			-0.77456
9	12328	choiceInteraction	-0.70801			-0.70801
10	12912	customInteraction	-2.31193			-2.31193
11	12327	choiceInteraction	-0.49603			-0.49603
12	12329	choiceInteraction , choiceInteraction	1.45207	1.00639		1.22923
13	12923	choiceInteraction	0.67783			0.67783
14	12928	choiceInteraction	0.94998			0.94998
15	15194	choiceInteraction , choiceInteraction	0.33783			0.33783
16	15196	choiceInteraction	-0.16844			-0.16844
17	15191	choiceInteraction , choiceInteraction	-0.26361			-0.26361
18	15193	choiceInteraction	-0.04387			-0.04387
19	15216	choiceInteraction	-0.02077			-0.02077
20	15192	choiceInteraction	-0.78514			-0.78514
21	15220	choiceInteraction	0.17362			0.17362
22	8760	inlineChoiceInteraction	0.86409			0.86409
23	8761	inlineChoiceInteraction	-0.1046			-0.1046
24	8762	inlineChoiceInteraction	0.57993			0.57993
25	8764	inlineChoiceInteraction , inlineChoiceInteraction	-0.66888	1.12054		0.22583
26	15111	choiceInteraction , choiceInteraction	1.06335			1.06335
27	15103	choiceInteraction	0.28246			0.28246
28	15104	choiceInteraction	-0.30243			-0.30243
29	15138	choiceInteraction	0.00325			0.00325
30	15142	choiceInteraction , choiceInteraction	1.62565			1.62565
31	16231	choiceInteraction	-0.48274			-0.48274
32	16232	choiceInteraction	-1.01634			-1.01634
33	16228	choiceInteraction	-1.27458			-1.27458
34	16234	choiceInteraction	0.59022			0.59022

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
35	16326	choiceInteraction , choiceInteraction	0.88014			0.88014
36	16230	choiceInteraction	-0.30017			-0.30017
37	16233	choiceInteraction	-0.39653			-0.39653
38	16435	choiceInteraction	1.668			1.668
39	16431	customInteraction	0.16419			0.16419
40	16433	choiceInteraction	-0.78444			-0.78444
41	16430	choiceInteraction	1.8919			1.8919
42	17705	choiceInteraction , choiceInteraction	0.54137			0.54137
43	17707	choiceInteraction	-0.62214			-0.62214
44	17706	choiceInteraction	-0.72109			-0.72109
45	12431	choiceInteraction , choiceInteraction	0.26262			0.26262
46	17748	choiceInteraction	-0.1779			-0.1779
47	16331	inlineChoiceInteraction	-0.62445			-0.62445
48	16332	inlineChoiceInteraction , inlineChoiceInteraction	-0.45155	1.99442		0.771435
49	16333	inlineChoiceInteraction , inlineChoiceInteraction	-0.6066	0.74176		0.06758

Appendix E.9—Spring 19 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	13720_C	extendedTextInteraction	-3.06998	-0.79431		-1.93215
2	13720_E	extendedTextInteraction	-1.06366	1.77264	3.26664	1.325207
3	13720_O	extendedTextInteraction	-2.24515	1.13742	3.0911	0.661123
4	13721_C	extendedTextInteraction	-2.98695	-0.98931		-1.98813
5	13721_E	extendedTextInteraction	-2.24616	0.86028	3.32082	0.64498
6	13721_O	extendedTextInteraction	-2.99104	0.50579	3.23517	0.249973
7	12811	choiceInteraction	-1.54233			-1.54233
8	12910	choiceInteraction	-0.462			-0.462
9	12833	choiceInteraction	-0.38774			-0.38774
10	12815	choiceInteraction	-0.94882			-0.94882
11	12832	choiceInteraction , choiceInteraction	2.1149			2.1149
12	12877	choiceInteraction	-0.25171			-0.25171
13	12847	extendedTextInteraction	-0.05338			-0.05338
14	8834	choiceInteraction	-0.09598			-0.09598
15	8856	choiceInteraction , choiceInteraction	0.53849			0.53849
16	8837	choiceInteraction	1.07611			1.07611
17	8855	choiceInteraction , choiceInteraction	1.09807			1.09807
18	8843	choiceInteraction	0.82014			0.82014
19	8841	choiceInteraction	0.06178			0.06178
20	8842	customInteraction	1.03534			1.03534
21	8854	choiceInteraction	1.09846			1.09846
22	8778	inlineChoiceInteraction	0.49041			0.49041
23	8779	inlineChoiceInteraction , inlineChoiceInteraction	-1.89965	-0.47048		-1.18507
24	8780	inlineChoiceInteraction	-0.3298			-0.3298
25	13709	choiceInteraction	-0.74553			-0.74553
26	13707	choiceInteraction	-0.72604			-0.72604
27	13704	customInteraction	1.99108			1.99108
28	13702	choiceInteraction	-0.65174			-0.65174
29	13701	choiceInteraction , choiceInteraction	2.56482			2.56482
30	13673	choiceInteraction	-0.43478			-0.43478
31	13665	choiceInteraction	-0.06295			-0.06295
32	13664	choiceInteraction	0.15898			0.15898
33	13656	choiceInteraction , choiceInteraction	-0.48005			-0.48005
34	13674	choiceInteraction	-0.87138			-0.87138
35	13672	choiceInteraction	-0.00331			-0.00331
36	8791	choiceInteraction	1.31404			1.31404
37	8781	choiceInteraction	0.34351			0.34351
38	8794	customInteraction	0.60324			0.60324
39	8783	choiceInteraction	-0.68116			-0.68116
40	8784	choiceInteraction	-0.25728			-0.25728

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	16303	choiceInteraction	0.57531			0.57531
42	16305	choiceInteraction	-0.28035			-0.28035
43	16307	choiceInteraction	0.06166			0.06166
44	16300	choiceInteraction	-0.66812			-0.66812
45	16308	choiceInteraction	-0.44197			-0.44197
46	16314	choiceInteraction	0.12495			0.12495
47	13644	inlineChoiceInteraction	-1.40824			-1.40824
48	13646	inlineChoiceInteraction	-1.88481			-1.88481
49	13647	inlineChoiceInteraction	-1.47527			-1.47527

Appendix E.10—Spring 19 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	17369	choiceInteraction	-1.30341			-1.30341
2	13740	customInteraction	-0.97445			-0.97445
3	19132	choiceInteraction	-0.35921			-0.35921
4	13969	customInteraction	-0.8999			-0.8999
5	19107	choiceInteraction	-0.80924			-0.80924
6	17358	choiceInteraction	0.14672			0.14672
7	15376	customInteraction	0.94886			0.94886
8	17549	choiceInteraction	0.18854			0.18854
9	15377	customInteraction	0.57968			0.57968
10	13989	choiceInteraction	1.32956			1.32956
11	15383	customInteraction	2.32606			2.32606
12	10454	matchInteraction	2.28432			2.28432
13	10687	choiceInteraction	0.89578			0.89578
14	11120	customInteraction	1.1283			1.1283
15	15566	customInteraction	1.00008			1.00008
16	13970	choiceInteraction	1.04176			1.04176
17	13980	customInteraction	0.25382			0.25382
18	15548	choiceInteraction	0.4375			0.4375
19	13746	customInteraction	-0.1961			-0.1961
20	19163	customInteraction	-0.31568			-0.31568
21	10409	choiceInteraction	-1.87291			-1.87291
22	11647	customInteraction	-0.95156			-0.95156
23	10404	choiceInteraction	-2.42513			-2.42513
24	12921	choiceInteraction	-1.04976			-1.04976
25	9460	customInteraction	-1.67509			-1.67509
26	13965	customInteraction	-1.00408			-1.00408
27	10391	choiceInteraction	-0.4991			-0.4991
28	17348	customInteraction	-0.53365			-0.53365
29	10460	choiceInteraction	-0.44542			-0.44542
30	10465	choiceInteraction	-0.14474			-0.14474
31	10398	customInteraction	0.4498			0.4498
32	9464	customInteraction	1.3174			1.3174
33	17353	choiceInteraction	0.85588			0.85588
34	19108	customInteraction	1.93287			1.93287
35	15389	customInteraction	1.10926			1.10926
36	15371	customInteraction	2.3642			2.3642
37	12569	choiceInteraction	0.72111			0.72111
38	13773	customInteraction	1.11913			1.11913
39	17403	choiceInteraction	0.17901			0.17901
40	12421	customInteraction	1.09421			1.09421

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	15914	choiceInteraction	0.75837			0.75837
42	10439	customInteraction	-1.01284			-1.01284
43	17343	choiceInteraction	-0.76371			-0.76371
44	10679	customInteraction	-0.93916			-0.93916
45	10411	choiceInteraction	-1.61799			-1.61799

Appendix E.11—Spring 19 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	13733	choiceInteraction	-1.76471			-1.76471
2	13760	customInteraction	-1.4998			-1.4998
3	10827	customInteraction	-0.86309			-0.86309
4	13762	matchInteraction	-0.70678			-0.70678
5	12276	customInteraction	-0.57774			-0.57774
6	13320	customInteraction	0.01192			0.01192
7	13756	customInteraction	0.18462			0.18462
8	13780	choiceInteraction	-0.17136			-0.17136
9	10716	customInteraction	1.10308			1.10308
10	17457	customInteraction	1.78262			1.78262
11	10774	customInteraction	1.86572			1.86572
12	13753	customInteraction	1.64173			1.64173
13	15443	customInteraction	0.79332			0.79332
14	15454	customInteraction	1.87834			1.87834
15	14035	customInteraction	0.48837			0.48837
16	15450	customInteraction	0.80359			0.80359
17	17376	choiceInteraction , choiceInteraction	-0.77387	1.8163		0.521215
18	15530	customInteraction	-0.12707			-0.12707
19	10760	customInteraction	-0.20261			-0.20261
20	17406	choiceInteraction	-1.41141			-1.41141
21	13777	customInteraction	-1.17708			-1.17708
22	13769	customInteraction	-1.25463			-1.25463
23	15428	choiceInteraction	-1.58475			-1.58475
24	13993	choiceInteraction	-1.72791			-1.72791
25	13738	customInteraction	-1.68096			-1.68096
26	11675	customInteraction	-0.51088			-0.51088
27	17795	choiceInteraction	-0.88726			-0.88726
28	10744	customInteraction	-0.50014			-0.50014
29	15579	customInteraction	-0.00102			-0.00102
30	9482	customInteraction	-1.86072	1.48656		-0.18708
31	14110	customInteraction	0.82349			0.82349
32	17453	customInteraction	0.72342			0.72342
33	17452	choiceInteraction	1.80546			1.80546
34	11105	customInteraction	1.97493			1.97493
35	15446	choiceInteraction	1.12694			1.12694
36	10756	customInteraction	1.2577			1.2577
37	15562	customInteraction	0.40579			0.40579
38	12271	choiceInteraction	1.68633			1.68633
39	13779	customInteraction	-0.50284			-0.50284
40	11713	customInteraction	-0.56445			-0.56445

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	15417	customInteraction	2.03122			2.03122
42	10750	customInteraction	-0.58904			-0.58904
43	15438	customInteraction	-0.75252			-0.75252
44	10783	choiceInteraction	-0.58052			-0.58052
45	13900	customInteraction	-1.64546			-1.64546

Appendix E.12—Spring 19 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	15497	choiceInteraction	-1.07291			-1.07291
2	15596	customInteraction	-1.54668			-1.54668
3	18899	matchInteraction	-1.66192			-1.66192
4	18898	customInteraction	-0.99542			-0.99542
5	14121	customInteraction	-0.03083			-0.03083
6	11107	choiceInteraction	-0.46606			-0.46606
7	15917	customInteraction	0.46781			0.46781
8	12090	customInteraction	0.5675			0.5675
9	17347	choiceInteraction	1.58258			1.58258
10	15486	customInteraction	0.35879			0.35879
11	10808	customInteraction	0.95699			0.95699
12	19159	customInteraction	0.9571			0.9571
13	15485	customInteraction	0.13277			0.13277
14	14088	customInteraction	0.16327			0.16327
15	11597	choiceInteraction	0.18814			0.18814
16	12223	customInteraction	-0.05863			-0.05863
17	17668	customInteraction	1.41459			1.41459
18	10851	customInteraction	0.92342			0.92342
19	15507	customInteraction	-1.06059	-0.15035		-0.60547
20	10794	customInteraction	-0.07612			-0.07612
21	17374	customInteraction	1.04237			1.04237
22	10811	choiceInteraction	-1.11389			-1.11389
23	14138	choiceInteraction	-1.18708			-1.18708
24	17799	choiceInteraction	-1.37036			-1.37036
25	17411	customInteraction	-1.1349			-1.1349
26	14155	customInteraction	-0.01572			-0.01572
27	10875	choiceInteraction	-0.69325			-0.69325
28	13086	customInteraction	0.28086			0.28086
29	12221	choiceInteraction	-0.3931			-0.3931
30	10839	customInteraction	0.63393			0.63393
31	15558	customInteraction	1.32987			1.32987
32	14156	choiceInteraction	0.78765			0.78765
33	10858	customInteraction	0.63797			0.63797
34	17445	customInteraction	1.60517			1.60517
35	10813	customInteraction	1.41929			1.41929
36	11368	choiceInteraction	0.23111			0.23111
37	10840	customInteraction	0.96138			0.96138
38	9476	customInteraction	1.35515			1.35515
39	10863	choiceInteraction	0.62779			0.62779
40	15506	choiceInteraction	-0.19127			-0.19127

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	14172	customInteraction	0.3286			0.3286
42	15491	customInteraction	-0.12347			-0.12347
43	18916	choiceInteraction	-1.46587			-1.46587
44	14084	customInteraction	-0.88864			-0.88864
45	15918	customInteraction	-0.94523			-0.94523

Appendix E.13—Spring 19 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	10100	choiceInteraction	-2.3786			-2.3786
2	9492	customInteraction	-1.41594			-1.41594
3	10113	choiceInteraction	-1.43848			-1.43848
4	18300	customInteraction	-0.26521			-0.26521
5	13114	choiceInteraction	0.12306			0.12306
6	17447	choiceInteraction , customInteraction	-0.87735	0.45621		-0.21057
7	10093	choiceInteraction	0.03634			0.03634
8	11376	customInteraction	0.75812			0.75812
9	15609	choiceInteraction	0.7296			0.7296
10	10062	customInteraction	0.87598			0.87598
11	19148	customInteraction	1.08204			1.08204
12	17782	customInteraction	1.62913			1.62913
13	9493	customInteraction	0.90776			0.90776
14	17466	customInteraction	1.20167			1.20167
15	18326	choiceInteraction	0.27082			0.27082
16	10070	customInteraction	0.2958			0.2958
17	13117	choiceInteraction	-0.38515			-0.38515
18	11569	customInteraction	-0.10501			-0.10501
19	14224	choiceInteraction	-0.12021			-0.12021
20	10144	customInteraction	-0.75359			-0.75359
21	17652	choiceInteraction	-1.60396			-1.60396
22	15618	customInteraction	-0.92681			-0.92681
23	18936	choiceInteraction	-1.23854			-1.23854
24	18940	choiceInteraction	-1.85532			-1.85532
25	11728	customInteraction	-1.56689			-1.56689
26	10082	choiceInteraction	-1.0193			-1.0193
27	13111	customInteraction	-1.25569			-1.25569
28	18932	choiceInteraction	-0.35917			-0.35917
29	12304	customInteraction	0.52126			0.52126
30	15646	matchInteraction	0.22577			0.22577
31	17655	customInteraction	0.70445			0.70445
32	14426	choiceInteraction	0.85707			0.85707
33	10143	customInteraction	0.87374			0.87374
34	17761	choiceInteraction	1.0454			1.0454
35	14423	customInteraction	2.38799			2.38799
36	15624	customInteraction	1.32882			1.32882
37	13112	customInteraction	1.07232			1.07232
38	10139	choiceInteraction	2.24036			2.24036
39	10078	customInteraction	1.05127			1.05127
40	10111	choiceInteraction	1.60177			1.60177

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	14160	customInteraction	0.77077			0.77077
42	17475	choiceInteraction	-0.18525			-0.18525
43	10050	customInteraction	0.4661			0.4661
44	17853	customInteraction	-0.21343			-0.21343
45	19348	choiceInteraction	-1.38253			-1.38253
46	10064	customInteraction	-0.12482			-0.12482
47	10129	choiceInteraction	-2.30001			-2.30001

Appendix E.14—Spring 19 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	15692	choiceInteraction	-1.97504			-1.97504
2	15666	customInteraction	-1.02157			-1.02157
3	18982	choiceInteraction	-1.37594			-1.37594
4	17346	customInteraction	-0.89568			-0.89568
5	14316	choiceInteraction	-0.84689			-0.84689
6	10298	customInteraction	0.12637			0.12637
7	15657	choiceInteraction	0.06594			0.06594
8	18330	customInteraction	0.25956			0.25956
9	17634	customInteraction	0.5154			0.5154
10	13805	customInteraction	1.00257			1.00257
11	15681	customInteraction	1.22584			1.22584
12	11969	customInteraction	0.89123			0.89123
13	10339	customInteraction	2.28753			2.28753
14	10701	customInteraction	1.03009			1.03009
15	10340	customInteraction	1.98187			1.98187
16	17801	customInteraction	1.12591			1.12591
17	17474	customInteraction	1.86684			1.86684
18	11742	choiceInteraction	-0.71542			-0.71542
19	15682	customInteraction	0.72006			0.72006
20	15697	choiceInteraction	-0.11345			-0.11345
21	17967	customInteraction	0.19747			0.19747
22	18318	choiceInteraction	-1.20998			-1.20998
23	14317	customInteraction	-1.2123			-1.2123
24	12472	choiceInteraction	-1.55344			-1.55344
25	10299	customInteraction	-0.49012			-0.49012
26	10303	choiceInteraction	-1.12478			-1.12478
27	18958	customInteraction	-0.43523			-0.43523
28	10318	choiceInteraction	-0.85168			-0.85168
29	15656	choiceInteraction	-0.9032			-0.9032
30	18984	customInteraction	0.00606			0.00606
31	11580	choiceInteraction	0.232			0.232
32	15669	customInteraction	1.42334			1.42334
33	12423	customInteraction	0.68394			0.68394
34	15670	choiceInteraction	1.29834			1.29834
35	13122	customInteraction	0.33845			0.33845
36	14248	customInteraction	2.28886			2.28886
37	10381	choiceInteraction	0.76292			0.76292
38	17783	customInteraction	0.86552			0.86552
39	15673	choiceInteraction	-0.06771			-0.06771
40	10331	customInteraction	0.38423			0.38423

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10360	choiceInteraction	0.17838			0.17838
42	15667	choiceInteraction	0.10863			0.10863
43	15691	customInteraction	0.62797			0.62797
44	14230	choiceInteraction	-0.5406			-0.5406
45	10370	choiceInteraction	-0.35904			-0.35904
46	14234	choiceInteraction	-1.91832			-1.91832
47	18967	choiceInteraction	-1.31583			-1.31583

Appendix E.15—Spring 19 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	choiceInteraction	-2.01606			-2.01606
2	10567	customInteraction	-0.90128			-0.90128
3	17754	choiceInteraction	-1.4875			-1.4875
4	11304	customInteraction	-0.65743			-0.65743
5	17818	choiceInteraction	-0.17203			-0.17203
6	10564	customInteraction	0.40128			0.40128
7	15732	customInteraction	0.09842			0.09842
8	17657	choiceInteraction	0.32875			0.32875
9	17811	customInteraction	1.24279			1.24279
10	10526	choiceInteraction	0.62216			0.62216
11	11546	customInteraction	2.14921			2.14921
12	10514	customInteraction	1.40567			1.40567
13	8251	customInteraction	0.87585			0.87585
14	13152	customInteraction	0.34924			0.34924
15	14588	choiceInteraction	-0.20867			-0.20867
16	10532	customInteraction	0.16579			0.16579
17	15716	choiceInteraction	0.07104			0.07104
18	15733	customInteraction	0.32537			0.32537
19	15727	choiceInteraction	-0.61674			-0.61674
20	10543	choiceInteraction	-1.05211			-1.05211
21	13814	customInteraction	-0.23429			-0.23429
22	15950	choiceInteraction	-0.81943			-0.81943
23	18989	choiceInteraction	-1.85304			-1.85304
24	10507	choiceInteraction	-1.82784			-1.82784
25	11360	customInteraction	-0.36993			-0.36993
26	18991	choiceInteraction	-2.37179			-2.37179
27	18971	customInteraction	-0.322			-0.322
28	11690	choiceInteraction	-0.34286			-0.34286
29	14275	choiceInteraction	-0.21049			-0.21049
30	17548	choiceInteraction	-0.30328			-0.30328
31	9525	customInteraction	0.25872	0.32204		0.29038
32	10512	customInteraction	0.59204			0.59204
33	14360	customInteraction	0.83407			0.83407
34	17971	choiceInteraction	2.28189			2.28189
35	10546	customInteraction	1.42971			1.42971
36	18033	gapMatchInteraction	-0.25287			-0.25287
37	18325	choiceInteraction	1.33115			1.33115
38	18990	choiceInteraction	-0.63558			-0.63558
39	10527	customInteraction	0.11829			0.11829
40	18977	choiceInteraction	0.44518			0.44518

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	15712	customInteraction	-1.52292	1.44566		-0.03863
42	17844	choiceInteraction	-0.43811			-0.43811
43	14378	customInteraction	-0.29899			-0.29899
44	12005	choiceInteraction	-0.97905			-0.97905
45	18972	choiceInteraction	-1.44192			-1.44192
46	18263	customInteraction	-0.20223			-0.20223
47	10581	choiceInteraction	-1.92813			-1.92813

Appendix E.16—Spring 19 Operational Item Parameter Estimates — Algebra I

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	9707	customInteraction	-2.17345			-2.17345
2	12499	choiceInteraction	-1.37038			-1.37038
3	10934	customInteraction	-0.94776			-0.94776
4	13185	choiceInteraction	1.09113			1.09113
5	10994	choiceInteraction	-0.81846			-0.81846
6	11338	choiceInteraction	-0.27385			-0.27385
7	15927	customInteraction	0.35798			0.35798
8	15783	choiceInteraction	0.15192			0.15192
9	19031	customInteraction	1.33688			1.33688
10	15773	choiceInteraction	0.2895			0.2895
11	18398	customInteraction	2.29375			2.29375
12	19353	choiceInteraction	1.66963			1.66963
13	18384	customInteraction	0.83441			0.83441
14	10896	choiceInteraction	0.61086			0.61086
15	15785	choiceInteraction	0.5957			0.5957
16	18305	choiceInteraction	1.23311			1.23311
17	10882	customInteraction	0.8083			0.8083
18	13976	choiceInteraction	0.19018			0.19018
19	10981	choiceInteraction	-0.28524			-0.28524
20	19363	customInteraction	0.43475			0.43475
21	10905	choiceInteraction	-0.97781			-0.97781
22	19060	choiceInteraction	-1.31912			-1.31912
23	19024	choiceInteraction	-1.22913			-1.22913
24	19225	choiceInteraction	-2.02958			-2.02958
25	10953	choiceInteraction	-0.75521			-0.75521
26	9542	customInteraction	-0.12327	-0.05925		-0.09126
27	11611	customInteraction	0.16231			0.16231
28	10973	choiceInteraction	-1.36424			-1.36424
29	10889	choiceInteraction	-0.16663			-0.16663
30	15764	customInteraction	0.59782			0.59782
31	10942	choiceInteraction	0.09573			0.09573
32	15774	choiceInteraction	1.26697			1.26697
33	19078	customInteraction	1.02254			1.02254
34	10990	choiceInteraction	0.65596			0.65596
35	19344	customInteraction	2.08493			2.08493
36	10965	customInteraction	2.1504			2.1504
37	12699	customInteraction	1.62793			1.62793
38	13972	choiceInteraction	1.33572			1.33572
39	12346	choiceInteraction	0.29896			0.29896
40	12733	customInteraction	0.63644			0.63644

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	10966	choiceInteraction	0.15284			0.15284
42	10988	customInteraction	0.41304			0.41304
43	10977	choiceInteraction	-0.42252			-0.42252
44	9535	customInteraction	-0.04011			-0.04011
45	19021	choiceInteraction	-0.4729			-0.4729
46	19170	choiceInteraction	-1.12731			-1.12731
47	15957	choiceInteraction	-0.91456			-0.91456

Appendix E.17—Spring 19 Operational Item Parameter Estimates — Geometry

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	11083	choiceInteraction	-2.35216			-2.35216
2	11315	choiceInteraction	-1.61279			-1.61279
3	11114	customInteraction	-0.12484			-0.12484
4	15837	customInteraction	-0.29917			-0.29917
5	15805	choiceInteraction	-1.09744			-1.09744
6	10924	customInteraction	-0.20101			-0.20101
7	15815	choiceInteraction	-0.46105			-0.46105
8	12045	customInteraction	0.15431			0.15431
9	12576	choiceInteraction	0.47267			0.47267
10	11923	customInteraction	1.49293			1.49293
11	8246	customInteraction	0.95865			0.95865
12	15175	customInteraction	1.97647			1.97647
13	14663	choiceInteraction	0.24202			0.24202
14	14246	customInteraction	1.5708			1.5708
15	12091	customInteraction	-0.35157			-0.35157
16	15840	customInteraction	0.97272			0.97272
17	11025	customInteraction	-0.16054			-0.16054
18	11613	choiceInteraction	-0.02039			-0.02039
19	11068	choiceInteraction	-0.64576			-0.64576
20	19251	customInteraction	0.61215			0.61215
21	15098	choiceInteraction	-0.9931			-0.9931
22	12350	customInteraction	-0.75689			-0.75689
23	11448	choiceInteraction	-1.87122			-1.87122
24	19040	choiceInteraction	-1.51717			-1.51717
25	15923	customInteraction	0.31054			0.31054
26	11033	choiceInteraction	-0.93043			-0.93043
27	11681	customInteraction	0.13965			0.13965
28	11018	choiceInteraction	-0.61172			-0.61172
29	12341	customInteraction	0.32447			0.32447
30	19219	choiceInteraction	-0.66708			-0.66708
31	14942	choiceInteraction	-0.11675			-0.11675
32	13538	customInteraction	1.10559			1.10559
33	14278	choiceInteraction	-0.24634			-0.24634
34	12047	customInteraction	0.96809			0.96809
35	11523	choiceInteraction	0.9141			0.9141
36	11545	customInteraction	1.96788			1.96788
37	14926	choiceInteraction	1.07479			1.07479
38	15109	customInteraction	1.2575			1.2575
39	12579	choiceInteraction	-0.59389			-0.59389
40	11109	customInteraction	1.05048			1.05048

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	11019	choiceInteraction	-0.44872			-0.44872
42	13500	choiceInteraction	-0.35376			-0.35376
43	9564	customInteraction	0.43952			0.43952
44	15816	choiceInteraction	-0.79983			-0.79983
45	11059	choiceInteraction	-1.01865			-1.01865
46	19241	choiceInteraction	-1.14073			-1.14073
47	11547	customInteraction	0.05275			0.05275

Appendix E.18—Spring 19 Operational Item Parameter Estimates — Algebra II

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13486	choiceInteraction	-2.28427			-2.28427
2	9580	customInteraction	-0.05533			-0.05533
3	11541	choiceInteraction	-1.39444			-1.39444
4	18356	matchInteraction	-0.2285			-0.2285
5	14661	choiceInteraction	-1.05321			-1.05321
6	8253	customInteraction	0.5678			0.5678
7	10187	choiceInteraction	-0.17866			-0.17866
8	15892	customInteraction	0.54523			0.54523
9	12219	choiceInteraction, choiceInteraction	-0.71983			-0.71983
10	18371	customInteraction	1.05647			1.05647
11	10241	choiceInteraction	1.83403			1.83403
12	11330	customInteraction	1.16324			1.16324
13	10240	choiceInteraction	0.00695			0.00695
14	14349	customInteraction	1.02083			1.02083
15	14340	customInteraction	0.11248			0.11248
16	19093	customInteraction	1.01834			1.01834
17	10236	choiceInteraction	-0.89747			-0.89747
18	13204	customInteraction	0.47746			0.47746
19	11603	choiceInteraction	0.4601			0.4601
20	14970	customInteraction	0.1386			0.1386
21	14357	choiceInteraction	-1.3352			-1.3352
22	15880	customInteraction	0.18315			0.18315
23	15885	choiceInteraction	-1.5452			-1.5452
24	10214	choiceInteraction	-1.65318			-1.65318
25	18360	choiceInteraction	-1.41846			-1.41846
26	10233	customInteraction	-0.81277			-0.81277
27	10192	choiceInteraction	-1.21466			-1.21466
28	12096	choiceInteraction	-0.08671			-0.08671
29	14652	customInteraction	-0.27312	0.41036		0.06862
30	10228	matchInteraction	0.55259			0.55259
31	11401	customInteraction	0.64249			0.64249
32	10160	choiceInteraction	-0.65505			-0.65505
33	15873	customInteraction	0.6162			0.6162
34	10193	choiceInteraction	-0.57364			-0.57364
35	19085	customInteraction	1.44725			1.44725
36	9577	customInteraction	1.22957			1.22957
37	10256	customInteraction	1.04876			1.04876
38	9567	customInteraction	0.98222			0.98222
39	18359	choiceInteraction	0.82242			0.82242
40	18378	customInteraction	0.10034			0.10034

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
41	11936	choiceInteraction	-0.51723			-0.51723
42	19243	customInteraction	0.32296			0.32296
43	14350	choiceInteraction	0.76145			0.76145
44	15857	choiceInteraction	-1.09658			-1.09658
45	12725	customInteraction	-0.35279			-0.35279
46	13475	choiceInteraction	-1.25771			-1.25771
47	10227	choiceInteraction	-1.65582			-1.65582

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

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	73,477	77,032	80,273	80,073	79,539	78,657	63,851	58,691	52,827
Female	36,040	37,496	39,340	39,349	39,033	38,840	31,086	29,168	26,468
Male	37,437	39,536	40,933	40,724	40,506	39,817	32,765	29,523	26,359
African American	4,131	4,359	4,401	4,387	4,391	4,326	3,611	3,265	2,906
Asian	1,671	1,751	1,800	1,778	1,807	1,923	1,606	1,595	1,582
Native Hawaiian/Pacific	304	294	289	338	342	293	290	240	197
Hispanic/Latino	33,911	35,901	37,869	37,148	36,504	35,688	27,521	24,519	21,606
American Indian or Alaskan	3,442	3,656	3,796	3,738	3,918	3,869	3,402	2,863	2,471
White	27,228	28,425	29,379	29,947	30,040	30,253	25,766	24,790	22,779
Multiple	2,790	2,646	2,739	2,737	2,537	2,305	1,655	1,419	1,286
Limited English Proficiency	6,339	6,925	7,541	6,829	5,973	4,800	3,915	2,786	1,924
Special Education	9,350	9,965	10,330	9,851	9,197	8,843	6,116	4,942	4,202
Free/Reduced Lunch	31,112	33,034	34,812	32,722	31,662	30,481	17,992	16,385	14,162
Accommodation	3,671	4,024	4,205	3,974	3,385	3,131	1,101	890	665

Appendix F.2 – Number of Participating Students by Demographic Subgroups – ELA Paper + DEI

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	9,302	9,661	9,885	10,161	9,084	8,389	5,496	4,597	4,090
Female	4,632	4,680	4,988	5,030	4,522	4,209	2,635	2,256	2,056
Male	4,670	4,981	4,897	5,131	4,562	4,180	2,861	2,341	2,034
African American	500	512	521	497	522	449	318	266	192
Asian	760	821	814	797	729	662	376	258	225
Native Hawaiian/Pacific	31	27	40	30	25	34	13	8	16
Hispanic/Latino	3,934	3,970	4,264	4,371	3,983	3,651	3,462	2,949	2,583
American Indian or Alaskan	504	562	521	559	354	337	191	131	166
White	3,251	3,450	3,430	3,609	3,238	3,051	1,070	933	862
Multiple	322	319	295	298	233	205	66	52	46
Limited English Proficiency	570	547	699	601	476	360	615	178	88
Special Education	1,007	1,061	1,045	1,078	918	788	568	363	329
Free/Reduced Lunch	3,417	3,568	3,798	3,661	3,204	2,952	1,109	975	840
Accommodation	835	719	727	586	467	393	122	121	49

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

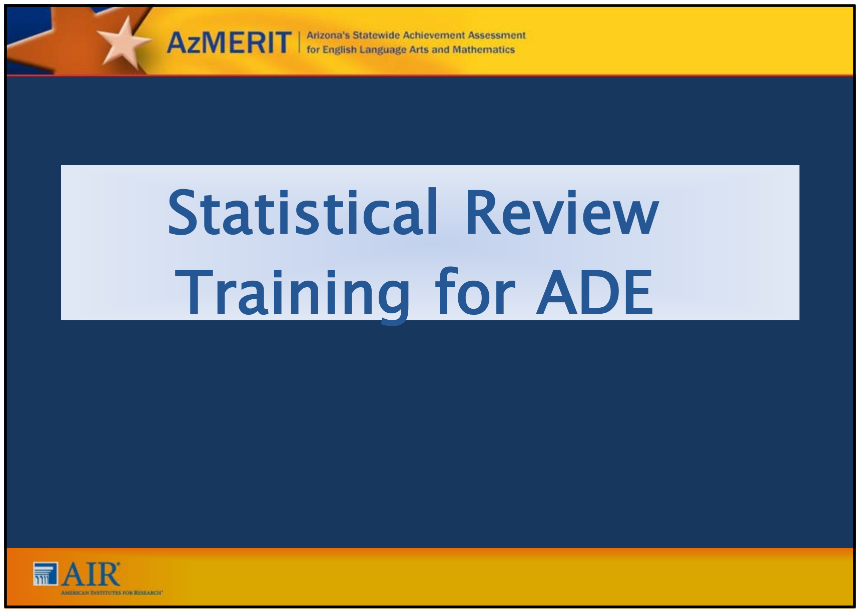
Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	AlgI	Geo	AlgII
All students	73,778	77,198	80,350	80,142	79,779	71,237	70,501	58,130	50,749
Female	36,145	37,563	39,342	39,350	39,114	35,109	34,362	28,895	25,932
Male	37,633	39,635	41,008	40,792	40,665	36,128	36,139	29,235	24,817
African American	4,162	4,383	4,412	4,381	4,417	4,081	3,902	3,165	2,741
Asian	1,674	1,754	1,803	1,784	1,807	1,563	1,807	1,622	1,548
Native Hawaiian/Pacific	305	295	290	340	342	268	331	230	198
Hispanic/Latino	34,062	35,959	37,930	37,167	36,615	32,851	31,107	24,457	20,768
American Indian or Alaskan	3,470	3,675	3,791	3,763	3,941	3,793	3,460	2,809	2,258
White	27,299	28,476	29,382	29,964	30,106	26,600	28,031	24,430	22,024
Multiple	2,806	2,656	2,742	2,743	2,551	2,081	1,863	1,417	1,212
Limited English Proficiency	6,377	6,958	7,558	6,863	6,005	4,581	4,016	2,977	1,894
Special Education	9,463	10,038	10,376	9,875	9,256	8,609	6,502	4,890	3,135
Free/Reduced Lunch	31,208	33,055	34,823	32,689	31,716	28,866	20,461	16,214	13,145
Accommodation	3,555	3,963	4,092	3,842	3,297	3,022	1,109	761	429

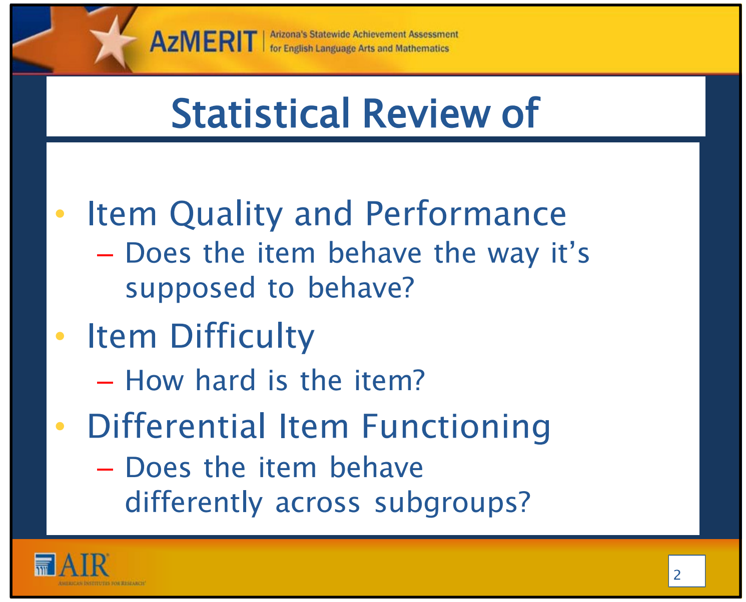
Note: AlgI=Algebra I; Geo=Geometry; AlgII=Algebra II.

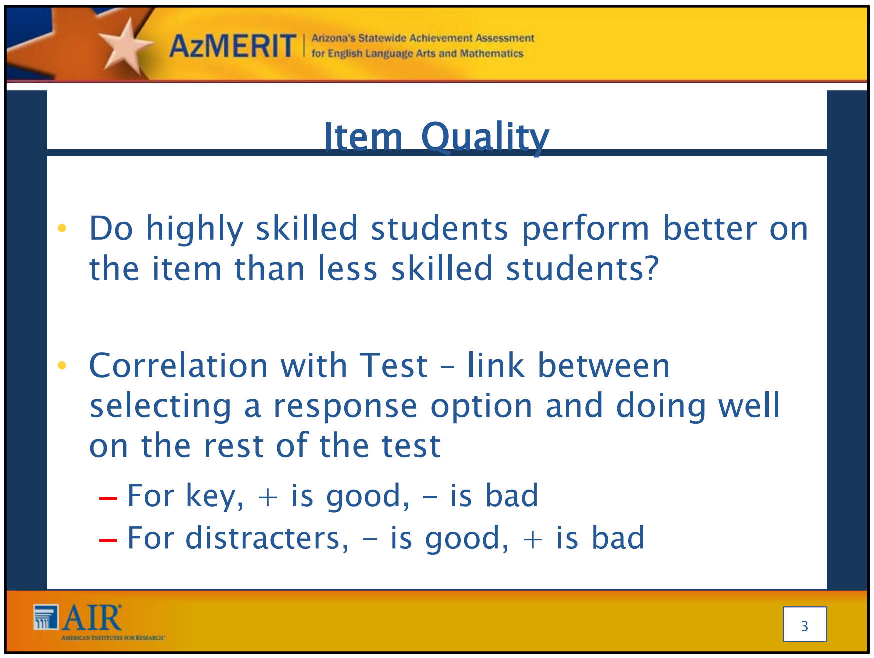
Appendix F.4 – Number of Participating Students by Demographic Subgroups – Mathematics Paper + DEI

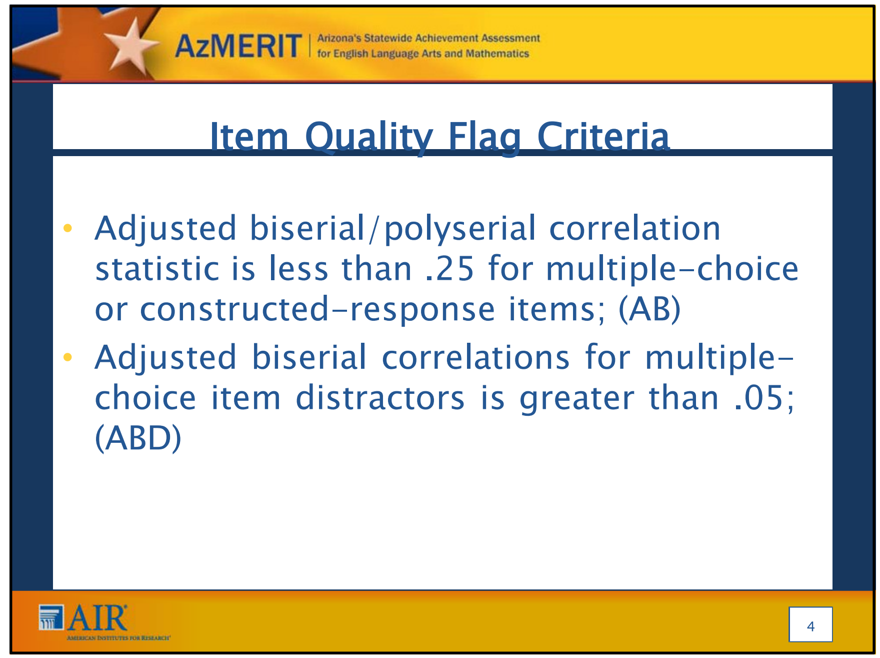
Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	AlgI	Geo	AlgII
All students	9,402	9,721	9,886	10,170	8,972	6,787	6,224	5,197	4,474
Female	4,668	4,712	4,989	5,030	4,475	3,400	3,029	2,485	2,221
Male	4,734	5,009	4,897	5,140	4,497	3,387	3,195	2,712	2,253
African American	507	513	519	497	516	404	355	270	217
Asian	760	820	813	790	663	178	614	386	378
Native Hawaiian/Pacific	31	27	40	30	24	33	17	11	11
Hispanic/Latino	3,967	4,022	4,263	4,378	3,989	3,357	3,473	3,265	2,725
American Indian or Alaskan	509	562	522	561	357	326	188	152	148
White	3,303	3,457	3,435	3,616	3,194	2,346	1,466	1,047	938
Multiple	325	320	294	298	229	143	111	66	57
Limited English Proficiency	575	549	699	601	478	359	560	410	224
Special Education	1,029	1,068	1,049	1,082	917	768	473	388	314
Free/Reduced Lunch	3,445	3,617	3,799	3,669	3,216	2,800	1,236	1,030	827
Accommodation	952	859	747	476	379	353	77	48	32

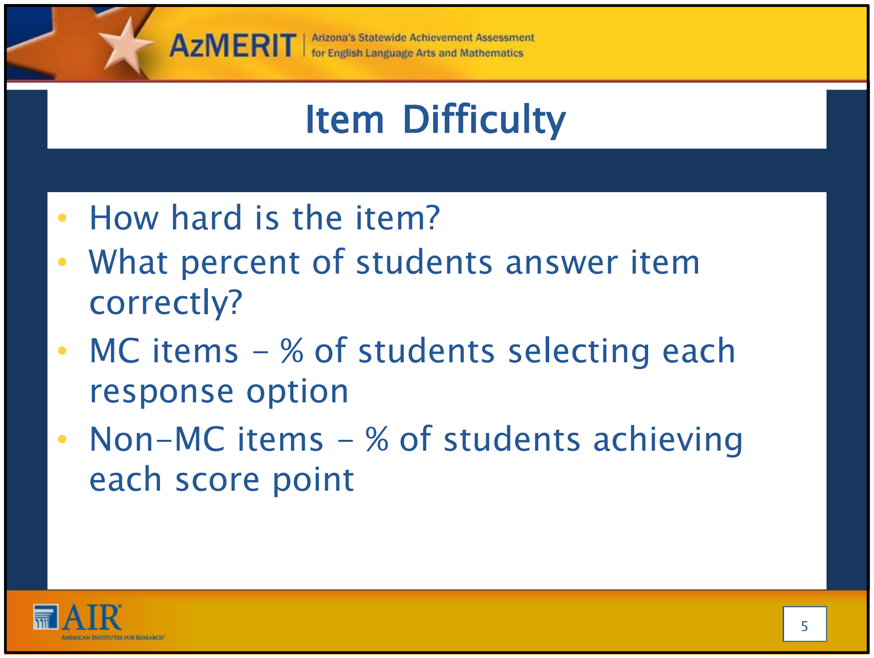
Note: AlgI=Algebra I; Geo=Geometry; AlgII=Algebra II.

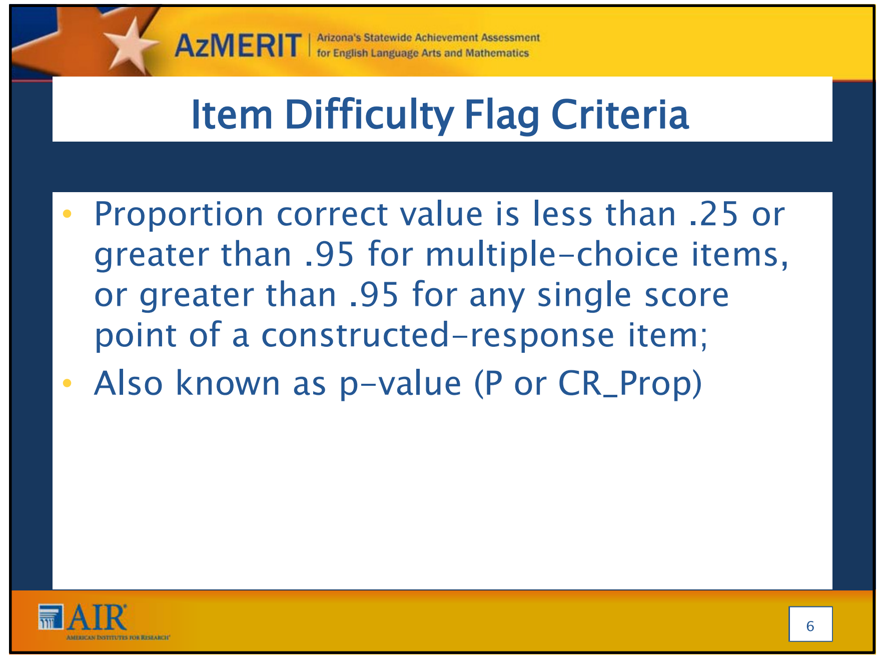

AzMERIT
Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics
Statistical Review
Training for ADE
AIR

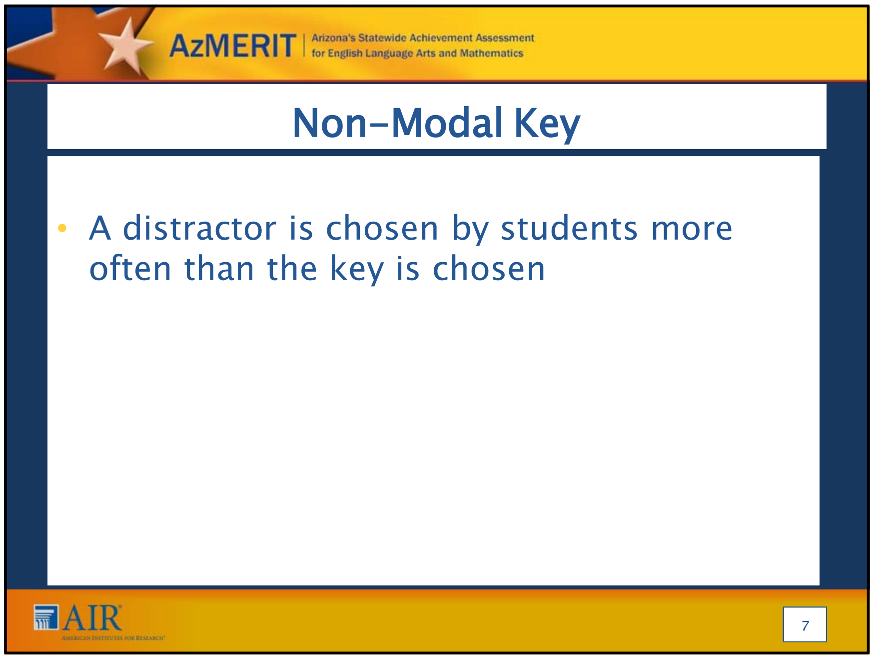

AzMERIT
Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics
Statistical Review of
• Item Quality and Performance
– Does the item behave the way it's supposed to behave?
• Item Difficulty
– How hard is the item?
• Differential Item Functioning
– Does the item behave differently across subgroups?
AIR
2

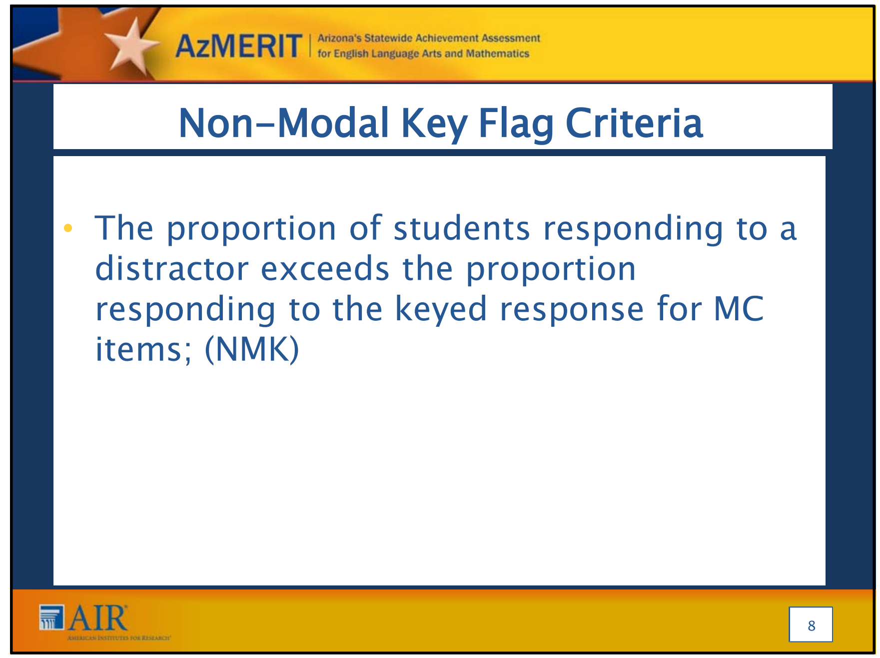

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Quality
• Do highly skilled students perform better on the item than less skilled students?
• Correlation with Test – link between selecting a response option and doing well on the rest of the test
– For key, + is good, – is bad
– For distracters, – is good, + is bad
AIR
3

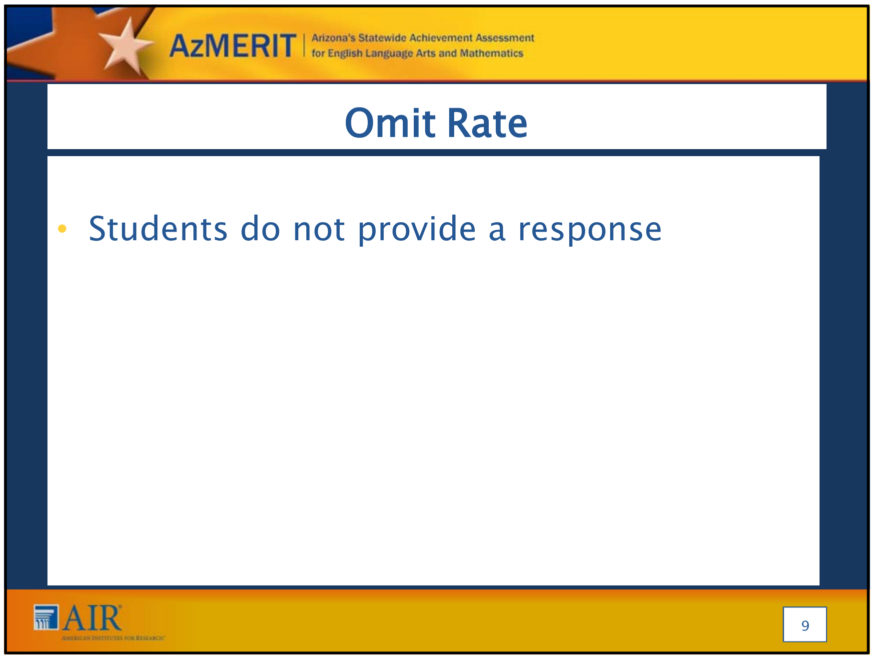

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Quality Flag Criteria
• Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items; (AB)
• Adjusted biserial correlations for multiple-choice item distractors is greater than .05; (ABD)
AIR
4

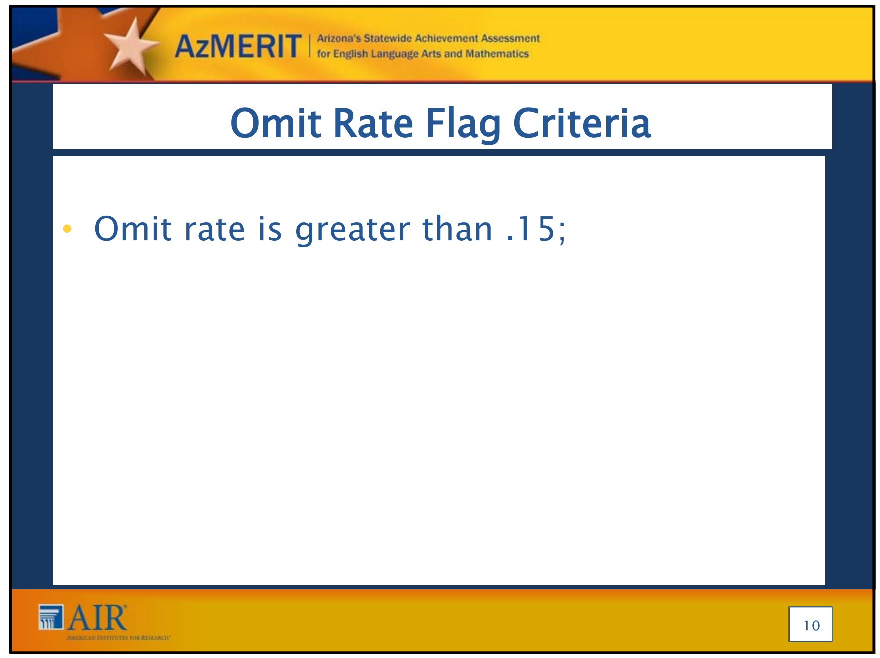

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Difficulty
How hard is the item?
What percent of students answer item correctly?
MC items – % of students selecting each response option
Non-MC items – % of students achieving each score point
AIR
5

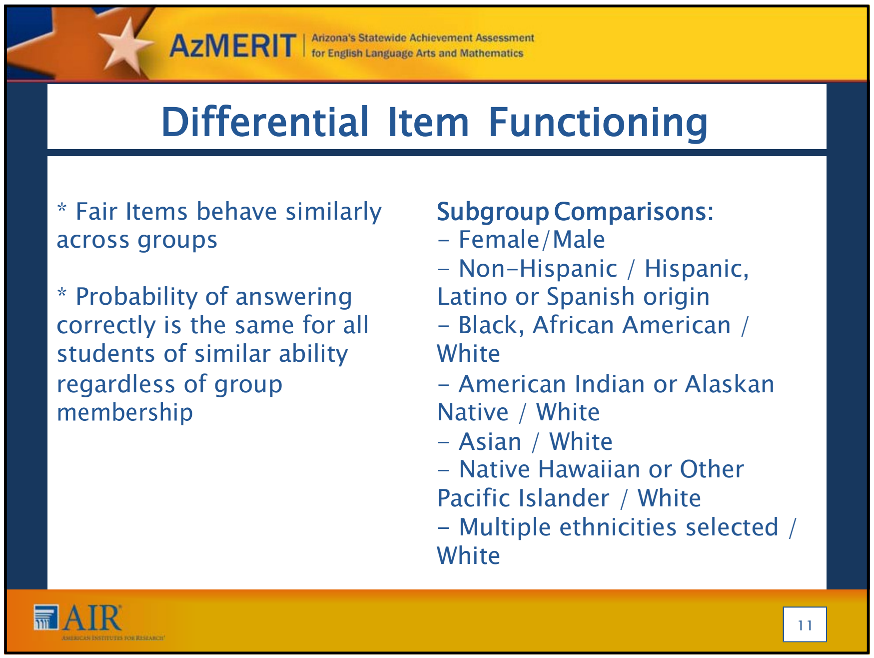

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Difficulty Flag Criteria
Proportion correct value is less than .25 or greater than .95 for multiple-choice items, or greater than .95 for any single score point of a constructed-response item;
Also known as p-value (P or CR_Prop)
AIR
6

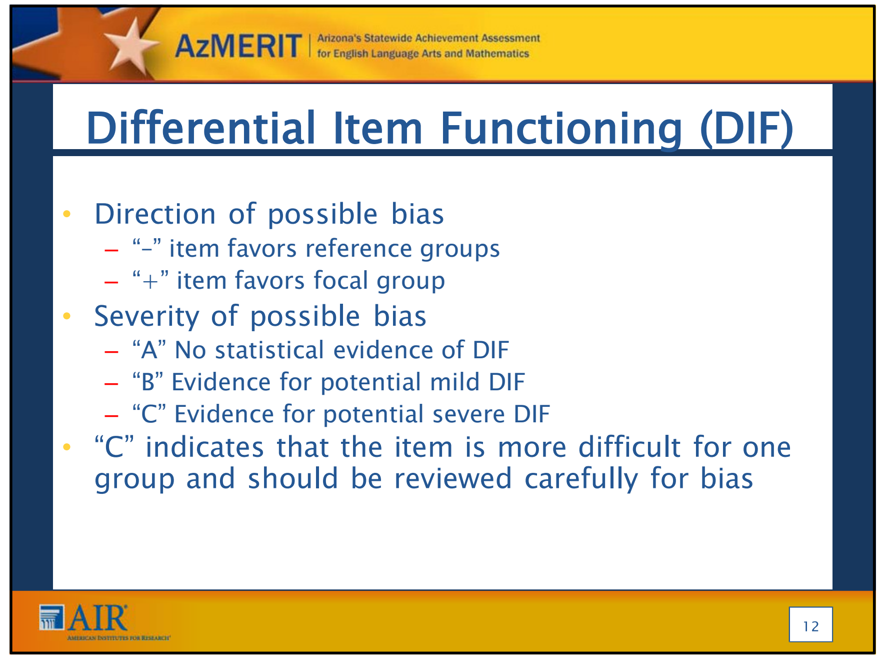

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Non-Modal Key
• A distractor is chosen by students more often than the key is chosen
AIR
7

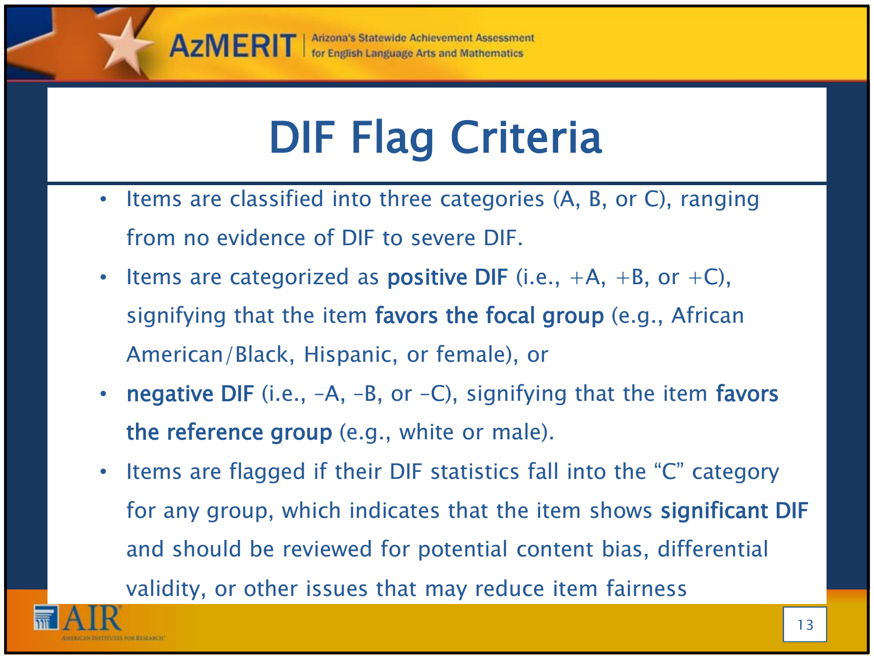

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Non-Modal Key Flag Criteria
• The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items; (NMK)
AIR
8

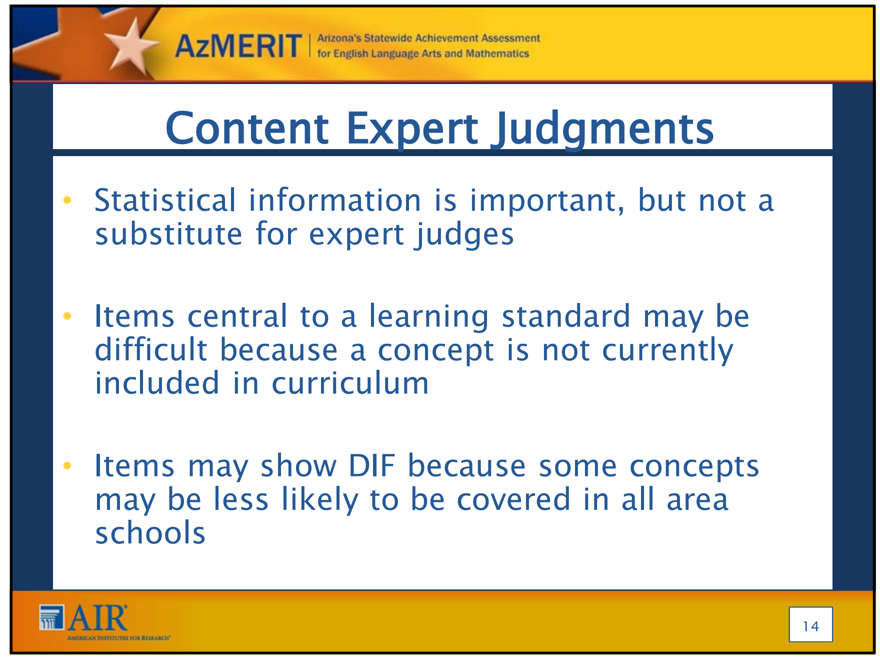

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Omit Rate
Students do not provide a response
AIR
9

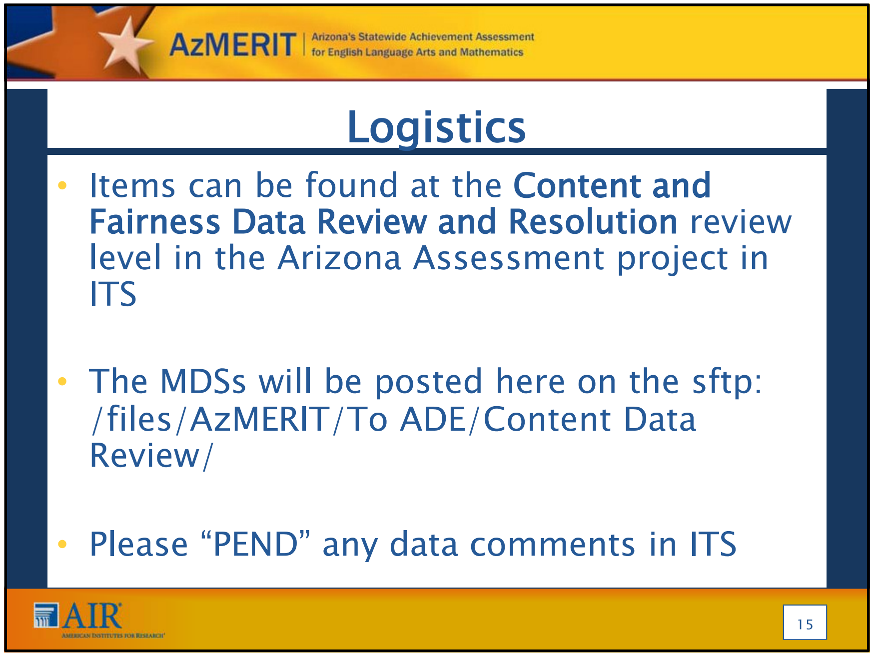

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Omit Rate Flag Criteria
Omit rate is greater than .15;
AIR
10


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
11

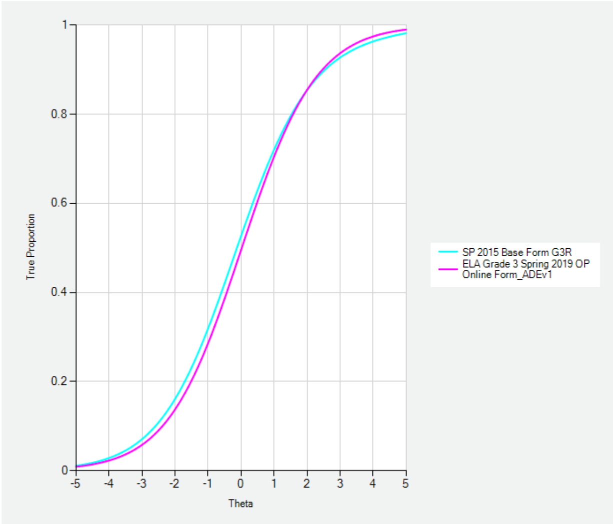

AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Differential Item Functioning (DIF)
Direction of possible bias
– "–" item favors reference groups
– "+" item favors focal group
Severity of possible bias
– "A" No statistical evidence of DIF
– "B" Evidence 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
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
13

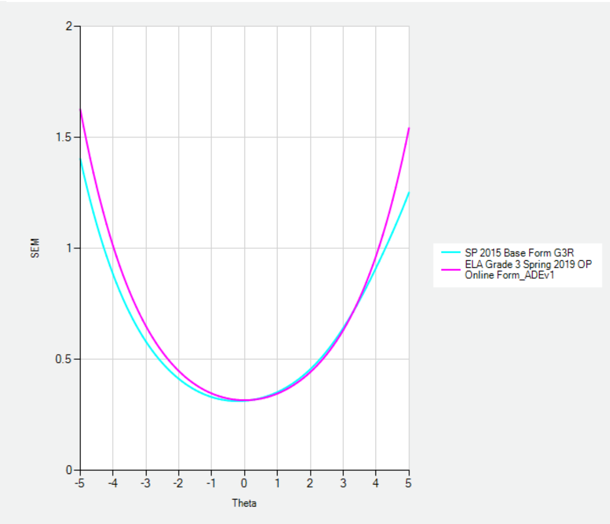

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Content Expert Judgments
Statistical information is important, but not a substitute for expert judges
Items central to a learning standard may be difficult because a concept is not currently included in curriculum
Items may show DIF because some concepts may be less likely to be covered in all area schools
AIR
14


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Logistics
Items can be found at the Content and Fairness Data Review and Resolution review level in the Arizona Assessment project in ITS
The MDSs will be posted here on the sftp: /files/AzMERIT/To ADE/Content Data Review/
Please "PEND" any data comments in ITS
AIR
15

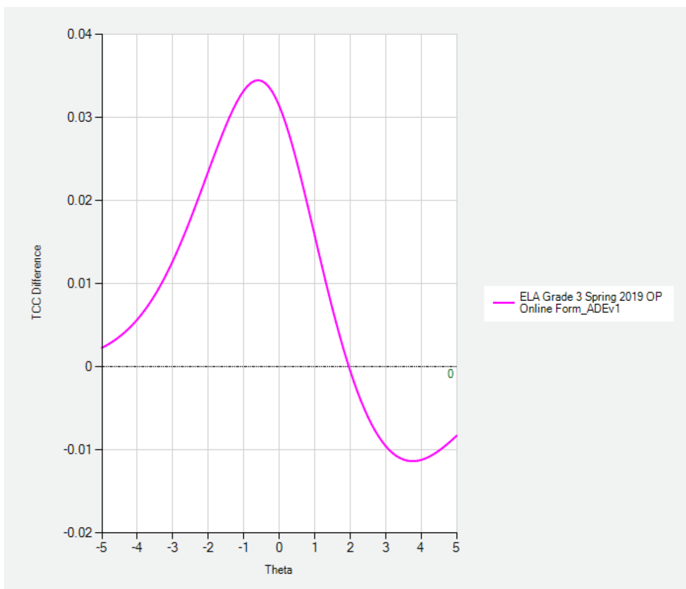
Appendix H.1 – Spring 2019 ELA Grade 3



Test Characteristic Curves

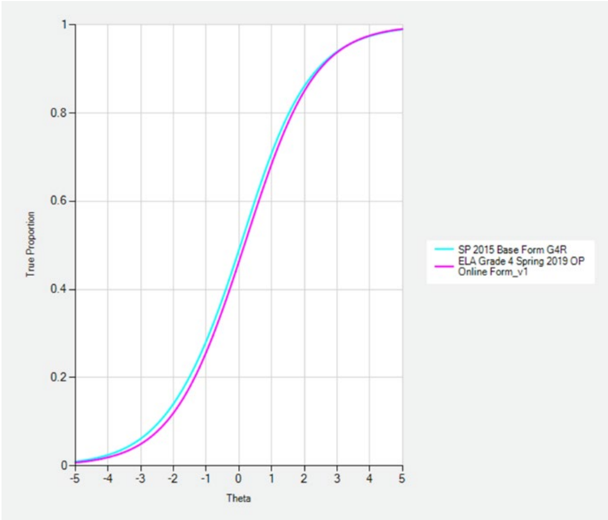


Standard Errors of Measurement

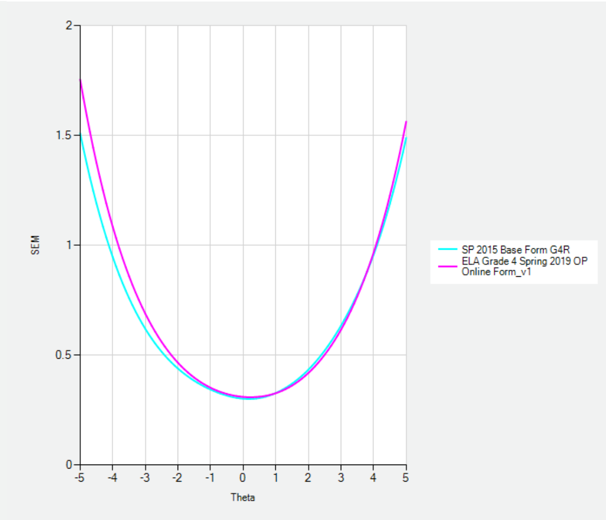


TCC Differences

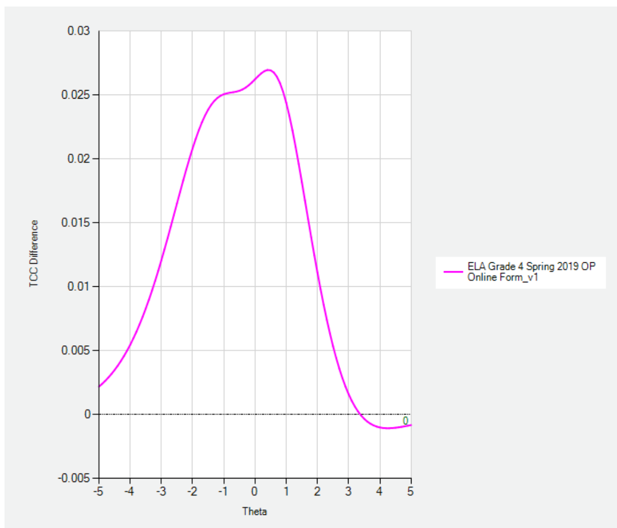
Appendix H.2 – Spring 2019 ELA Grade 4



Test Characteristic Curves

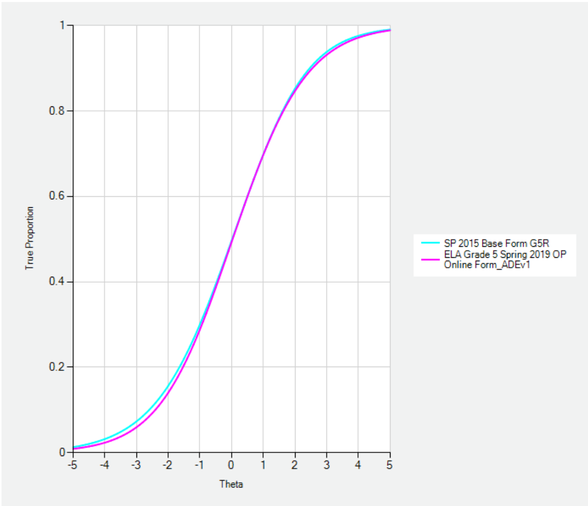


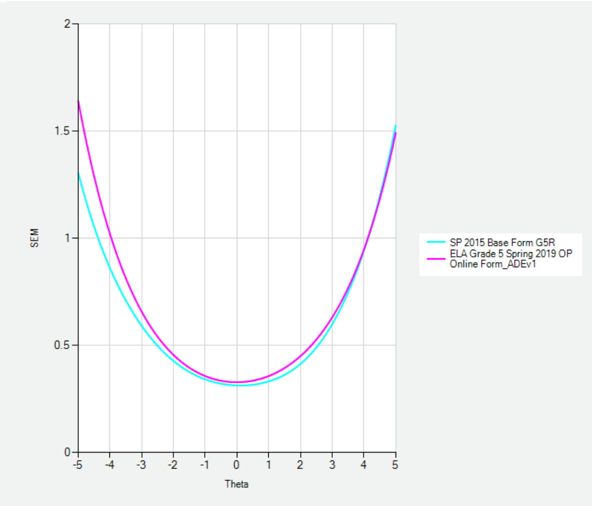
Standard Errors of Measurement



TCC Differences

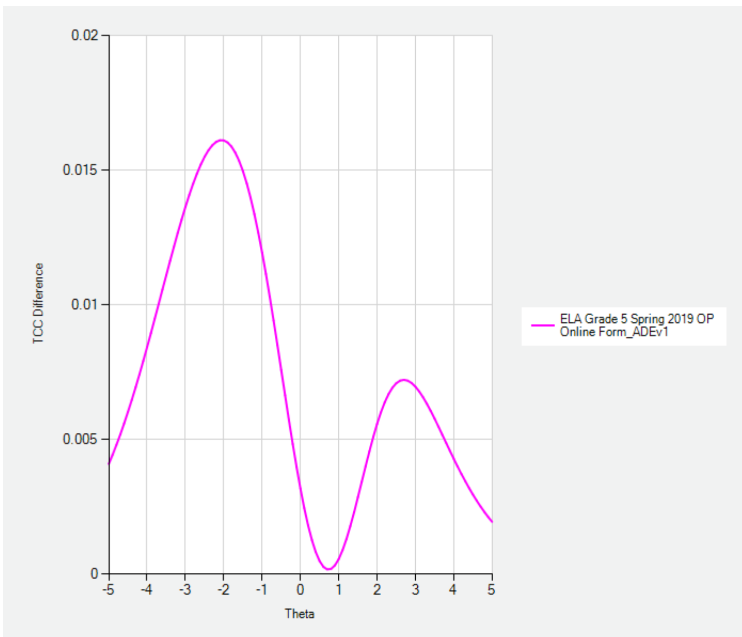
Appendix H.3 – Spring 2019 ELA Grade 5





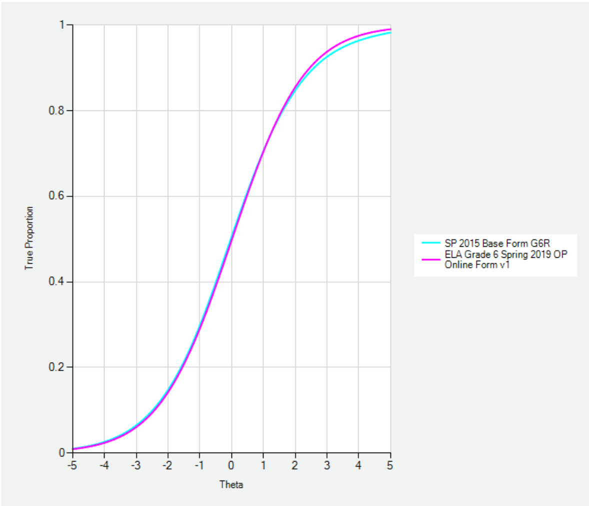
Test Characteristic Curves

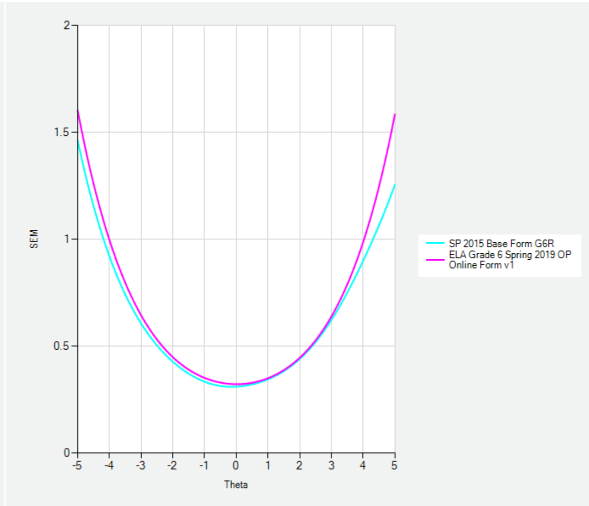
Standard Errors of Measurement



TCC Differences

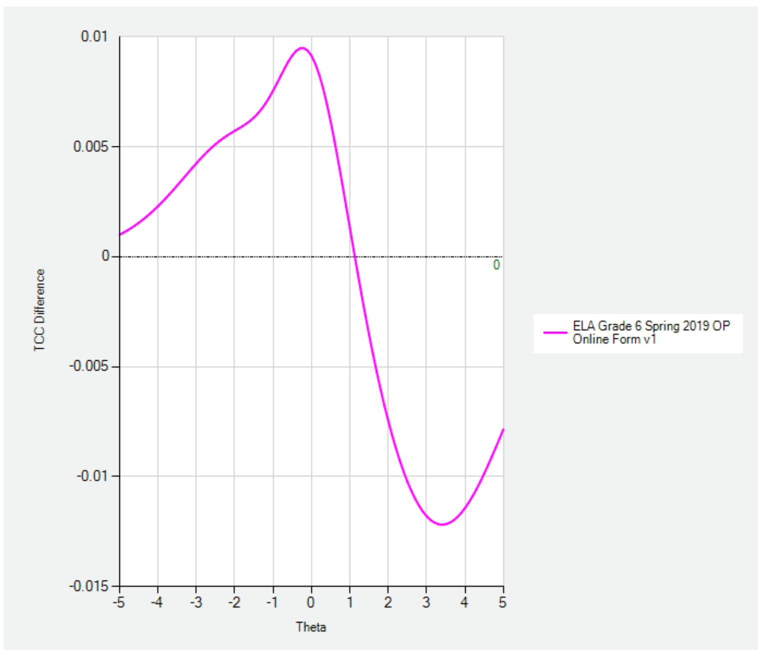
Appendix H.4 – Spring 2019 ELA Grade 6





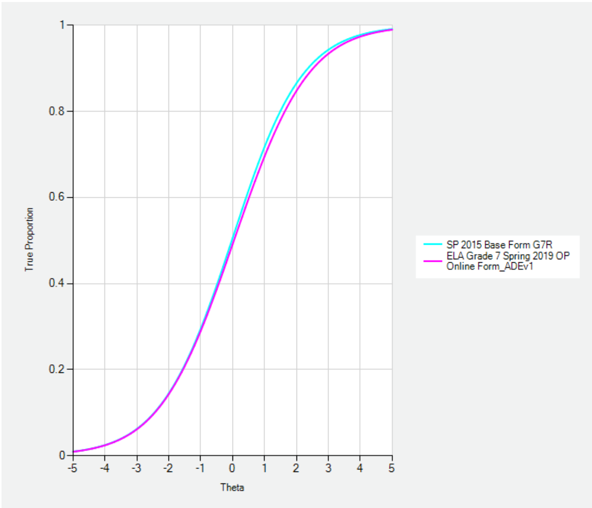
Test Characteristic Curves

Standard Errors of Measurement

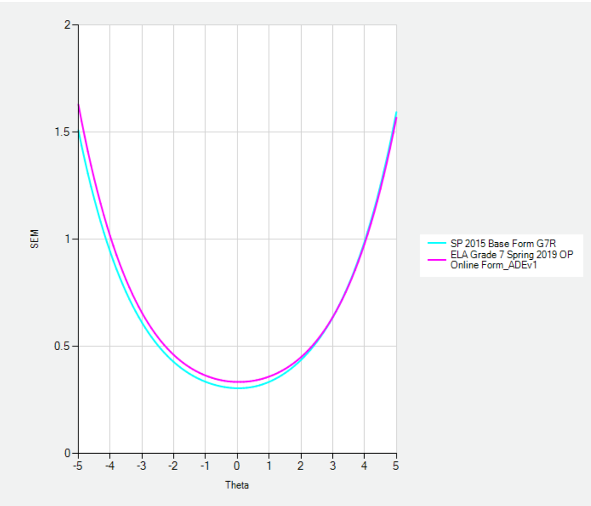


TCC Differences

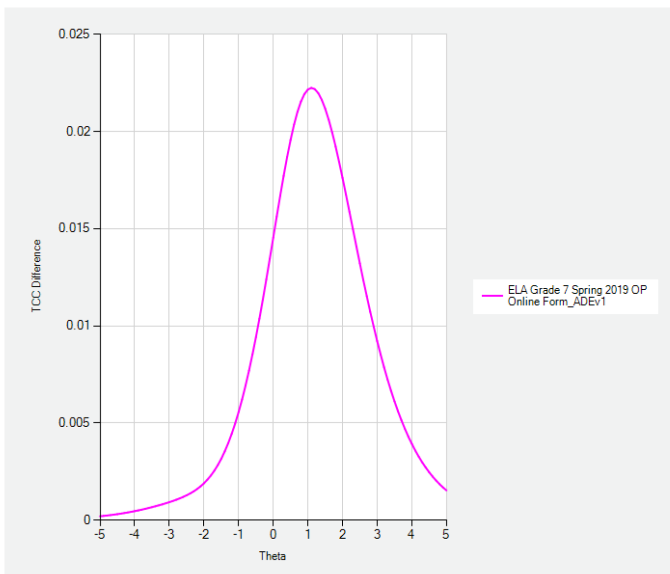
Appendix H.5 – Spring 2019 ELA Grade 7



Test Characteristic Curves

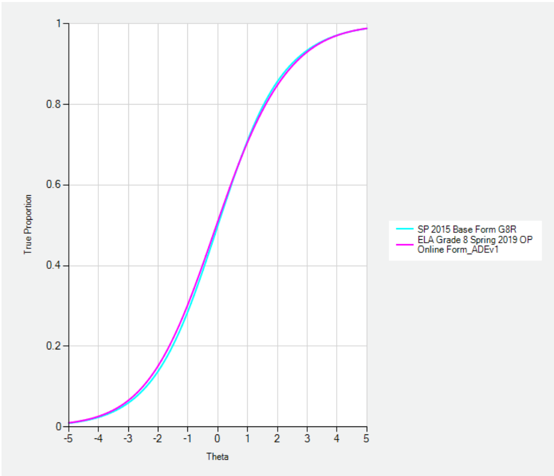


Standard Errors of Measurement

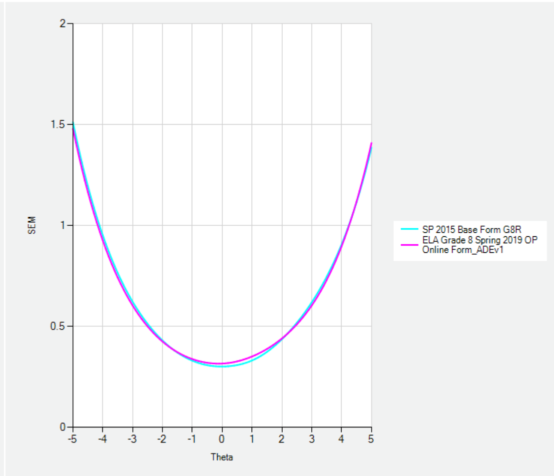


TCC Differences

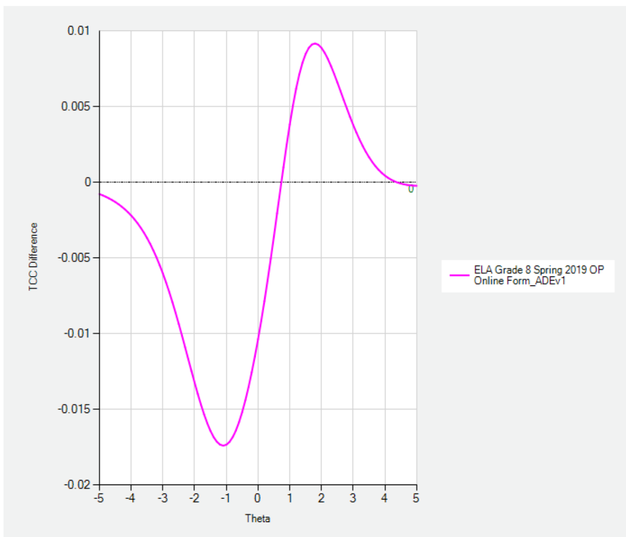
Appendix H.6 – Spring 2019 ELA Grade 8



Test Characteristic Curves

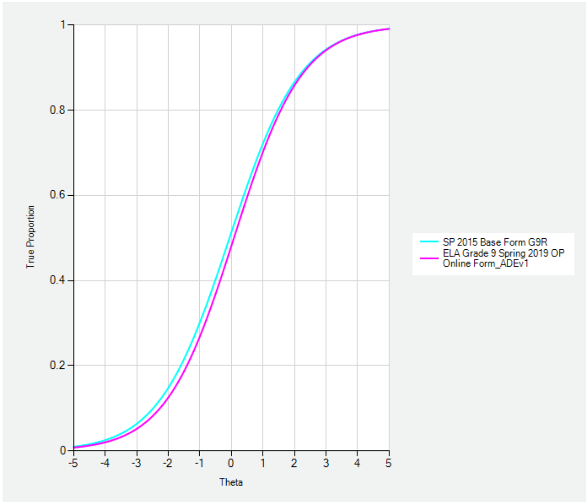


Standard Errors of Measurement

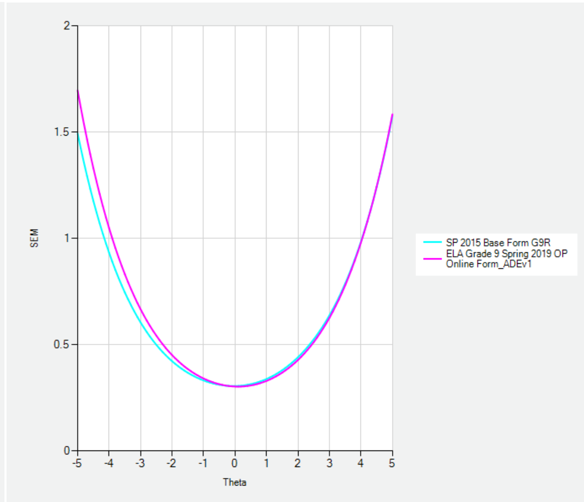


TCC Differences

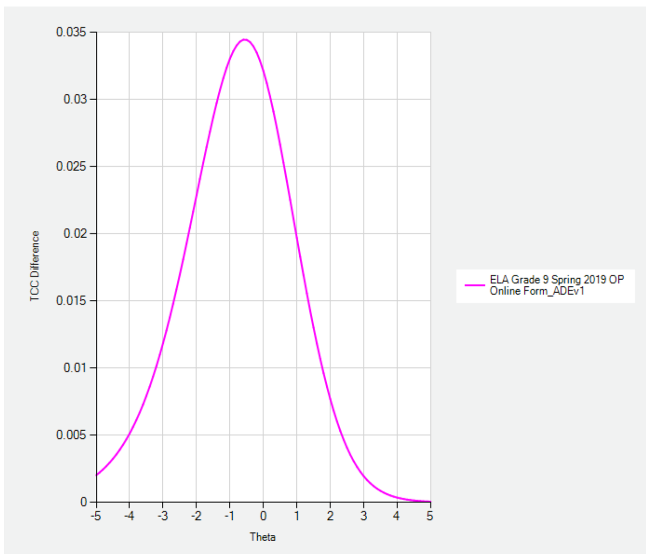
Appendix H.7 – Spring 2019 ELA Grade 9



Test Characteristic Curves

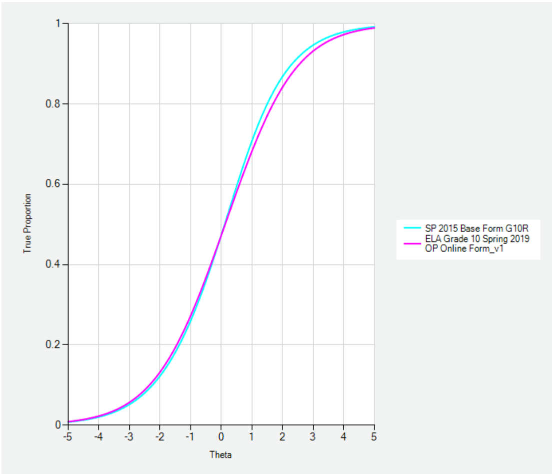


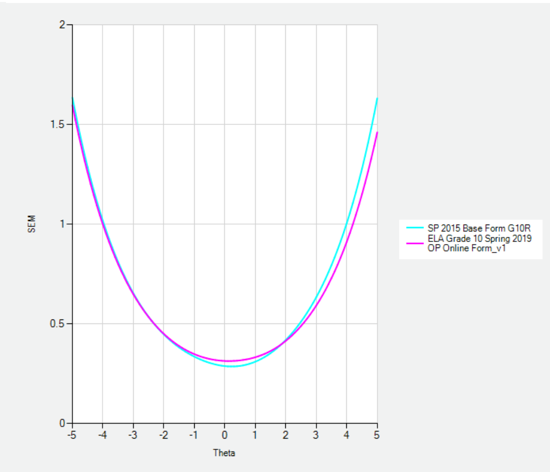
Standard Errors of Measurement



TCC Differences

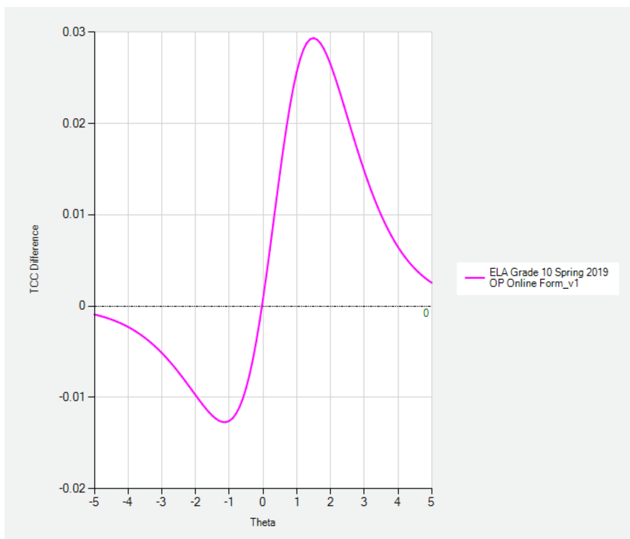
Appendix H.8 – Spring 2019 ELA Grade 10





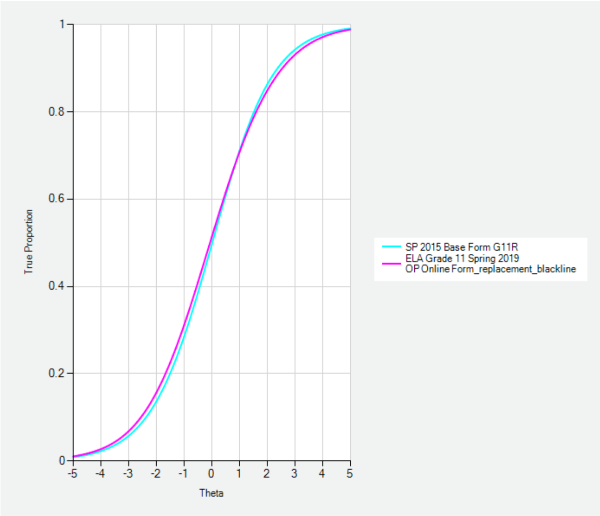
Test Characteristic Curves

Standard Errors of Measurement

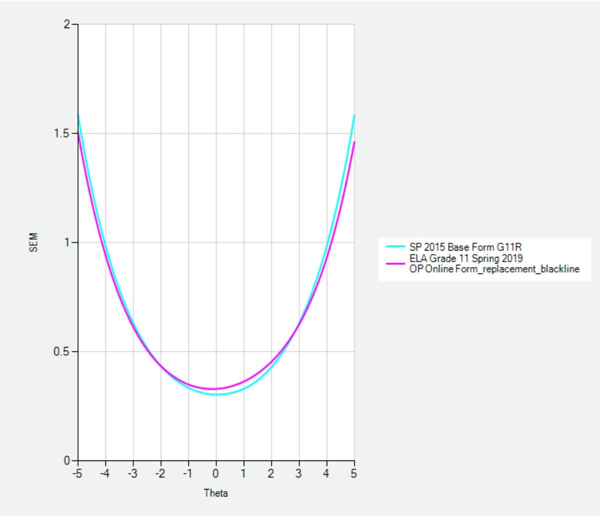


TCC Differences

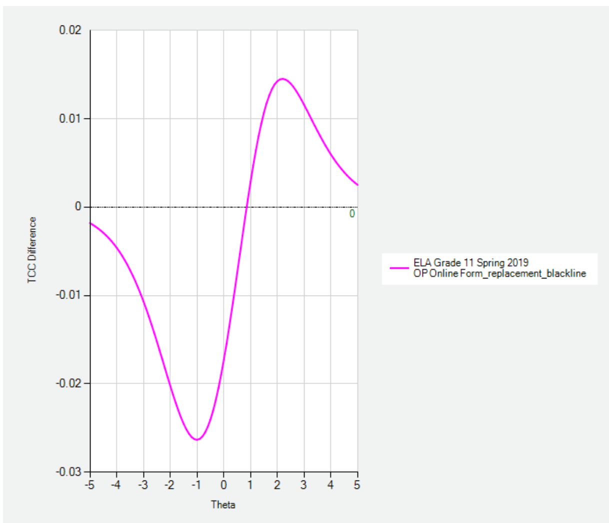
Appendix H.9 – Spring 2019 ELA Grade 11



Test Characteristic Curves

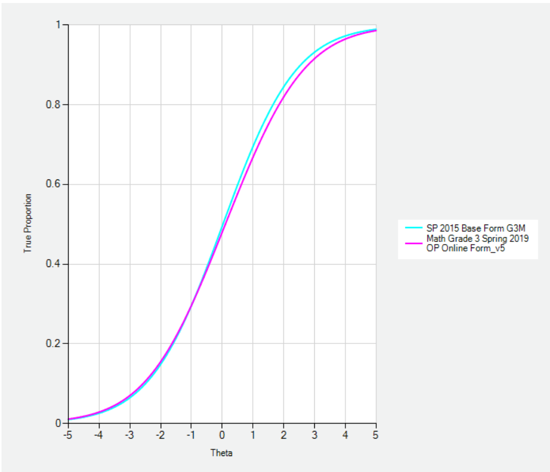


Standard Errors of Measurement

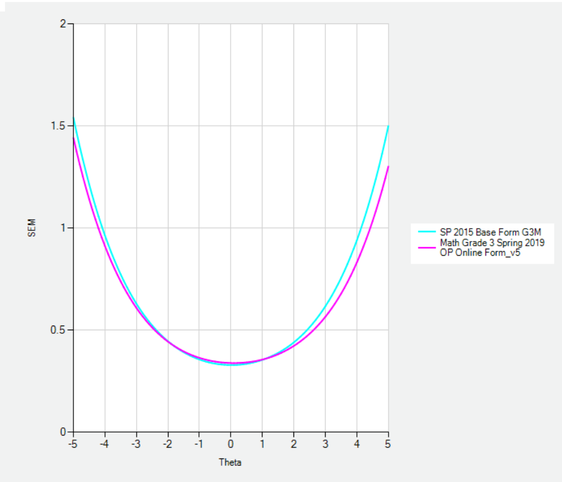


TCC Differences

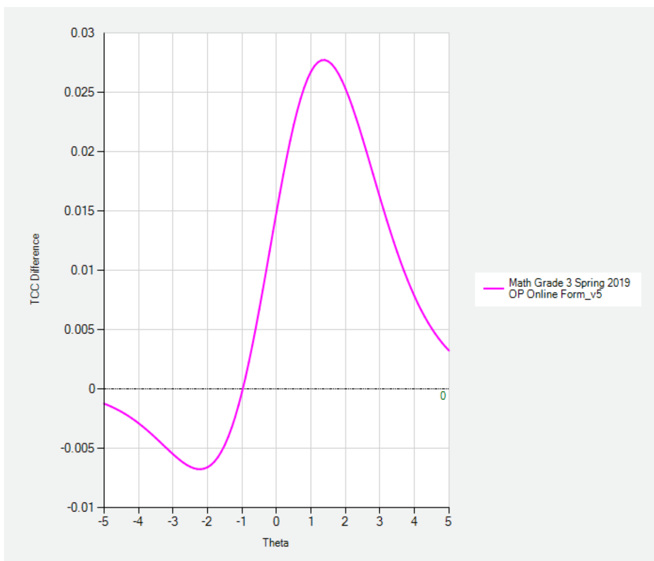
Appendix H.10 – Spring 2019 Math Grade 3



Test Characteristic Curves

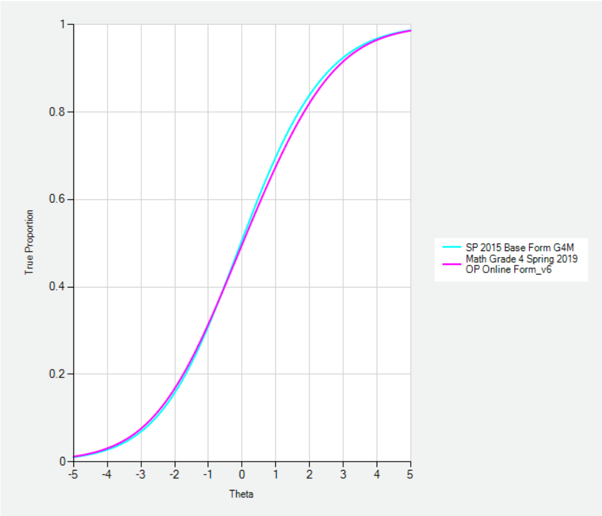


Standard Errors of Measurement

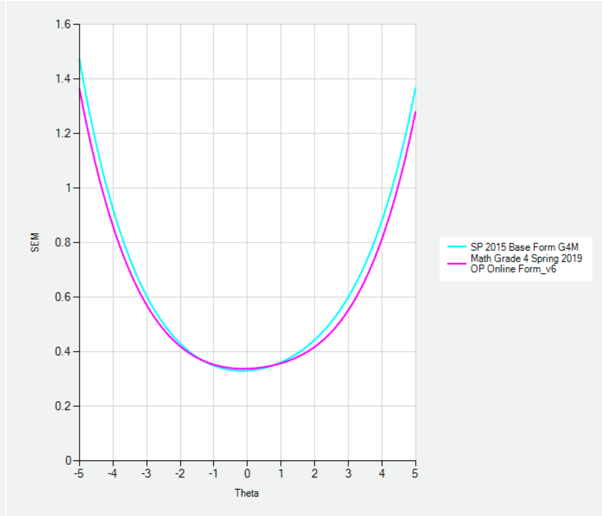


TCC Differences

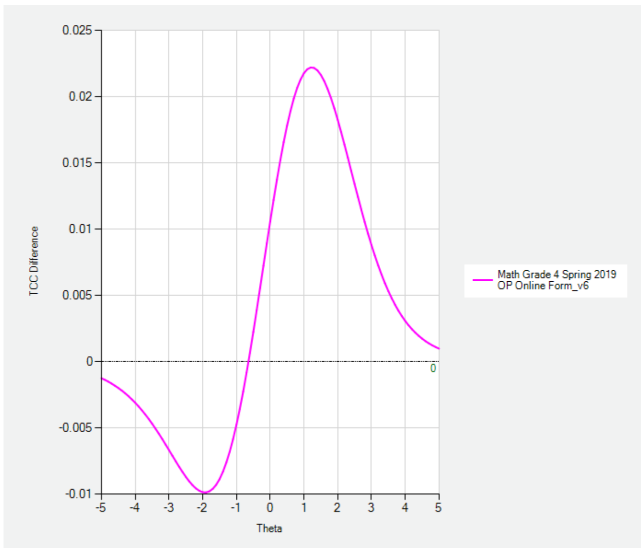
Appendix H.11 – Spring 2019 Math Grade 4



Test Characteristic Curves

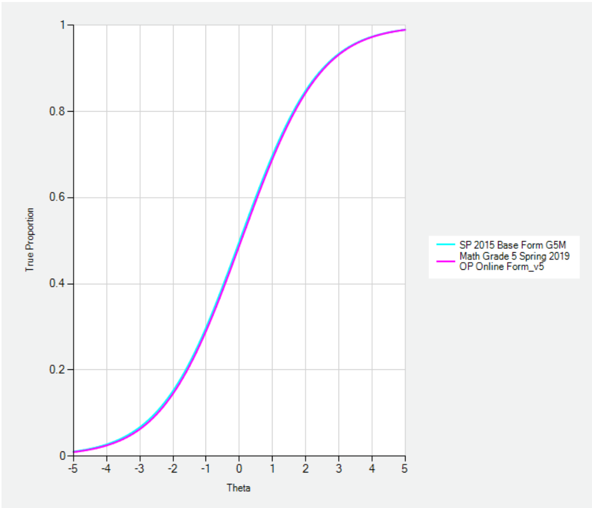


Standard Errors of Measurement

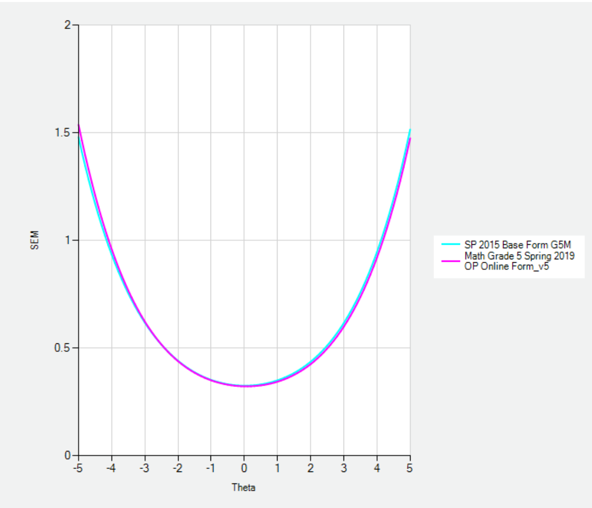


TCC Differences

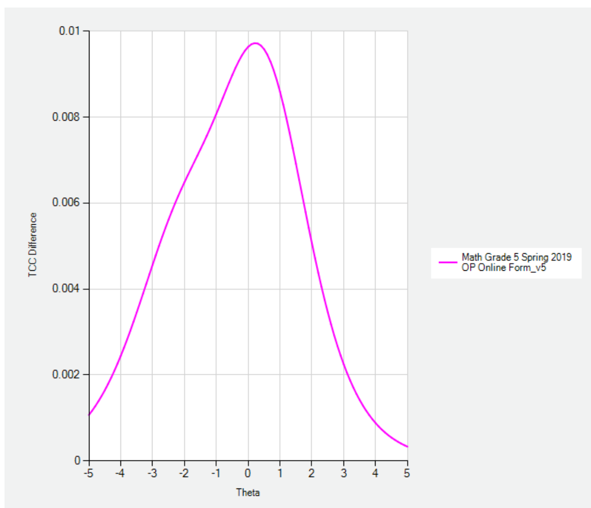
Appendix H.12 – Spring 2019 Math Grade 5



Test Characteristic Curves

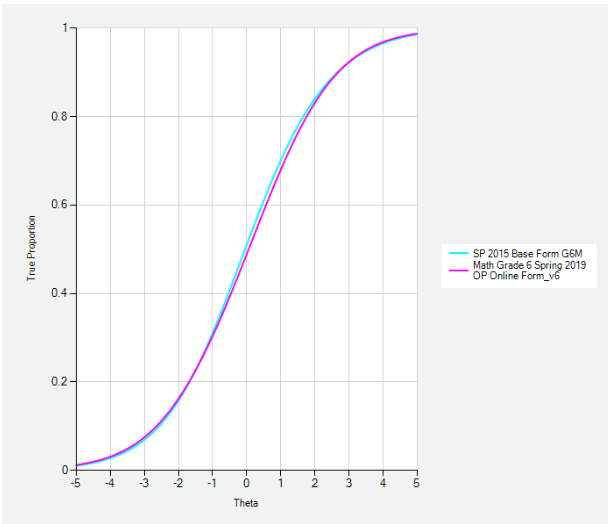


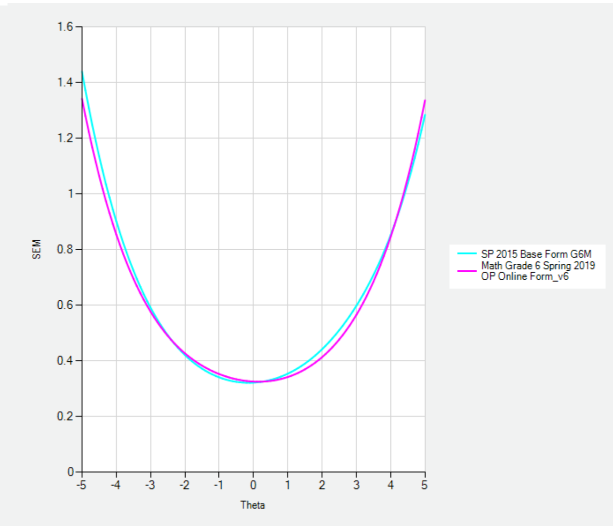
Standard Errors of Measurement



TCC Differences

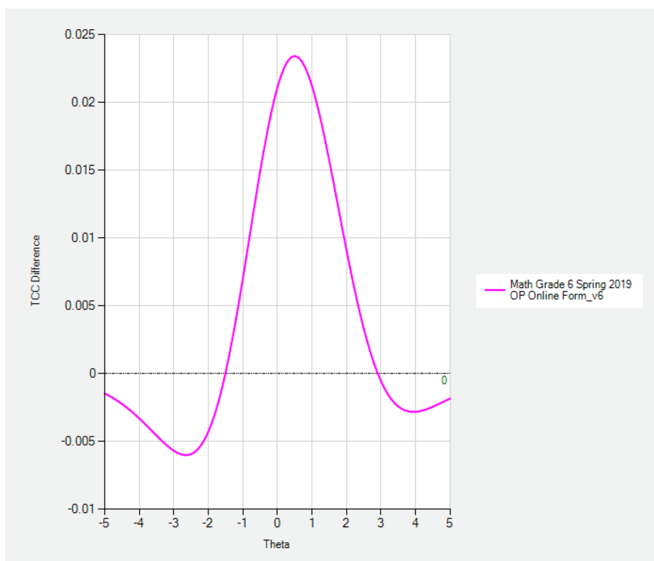
Appendix H.13 – Spring 2019 Math Grade 6





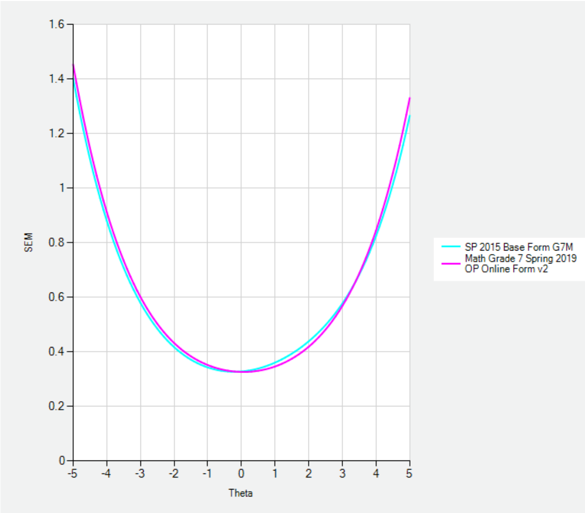
Test Characteristic Curves

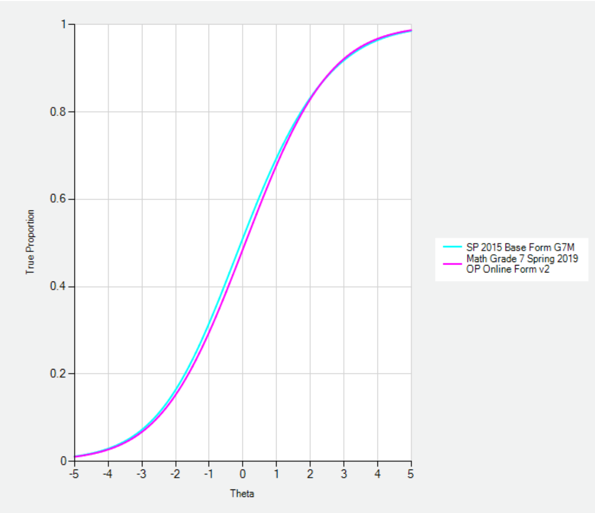
Standard Errors of Measurement



TCC Differences

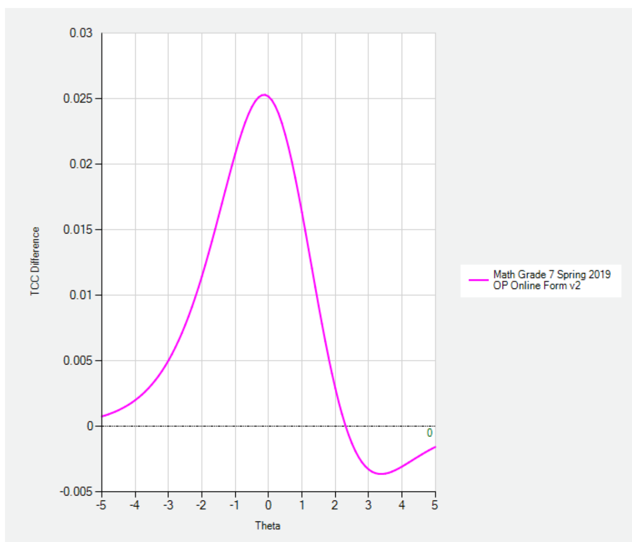
Appendix H.14 – Spring 2019 Math Grade 7





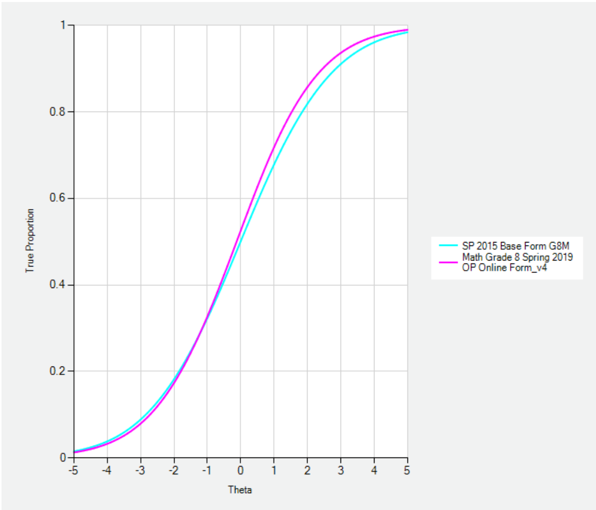
Test Characteristic Curves

Standard Errors of Measurement

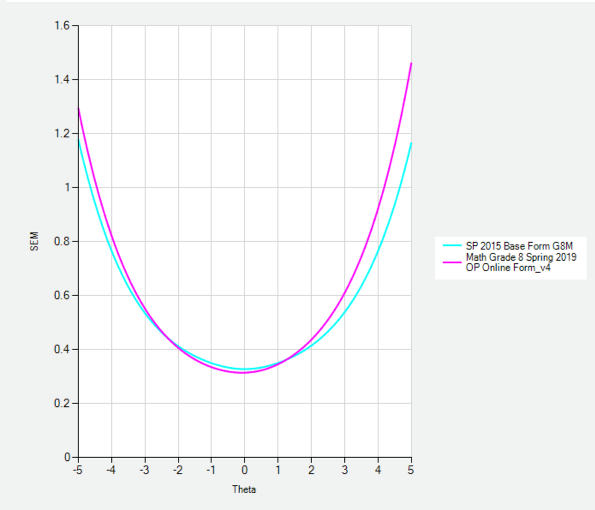


TCC Differences

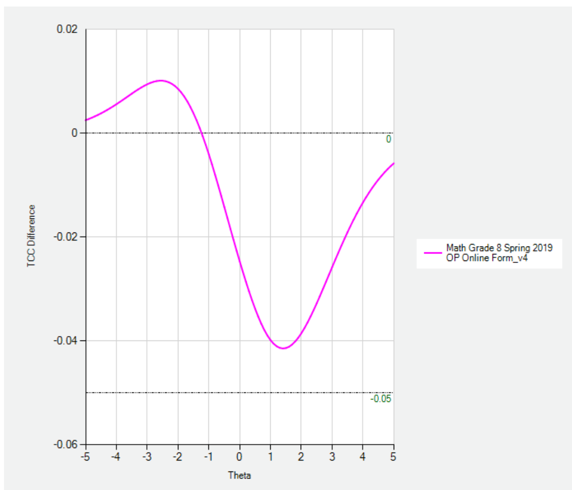
Appendix H.15 – Spring 2019 Math Grade 8



Test Characteristic Curves

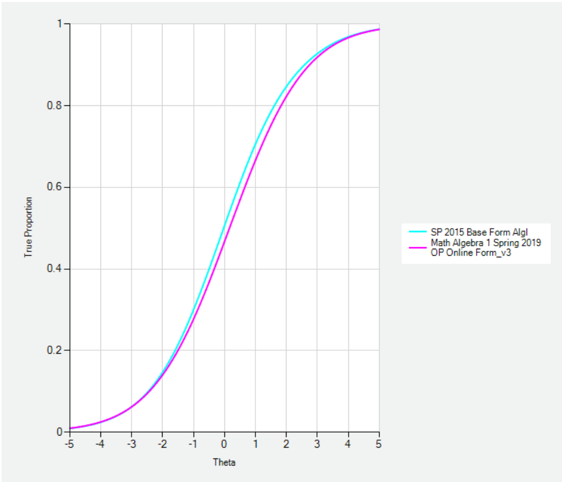


Standard Errors of Measurement

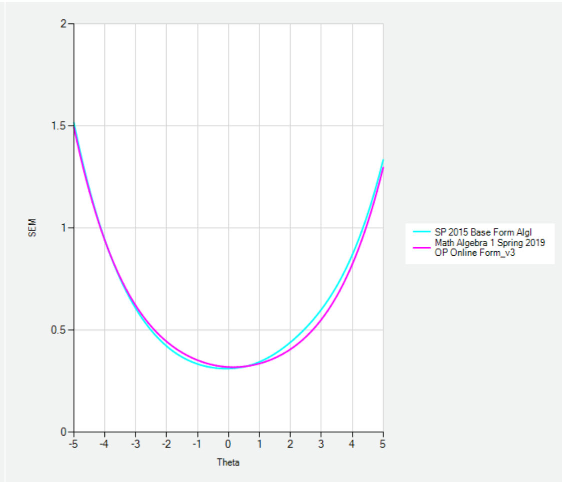


TCC Differences

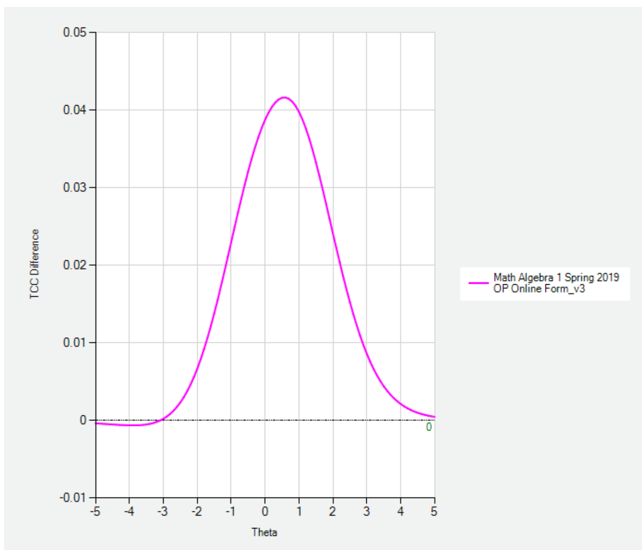
Appendix H.16 – Spring 2019 Math Algebra I



Test Characteristic Curves

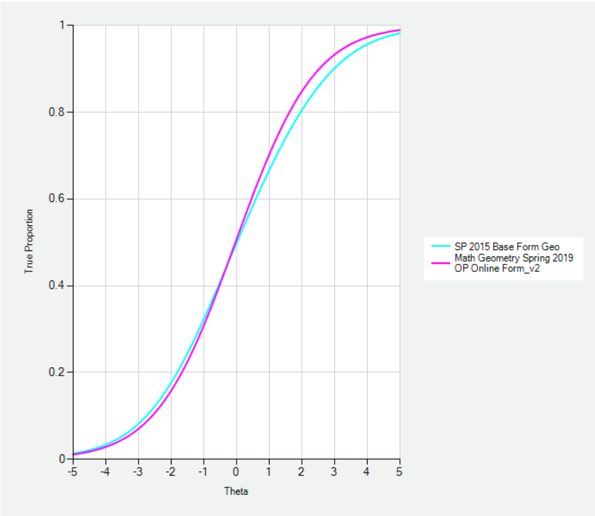


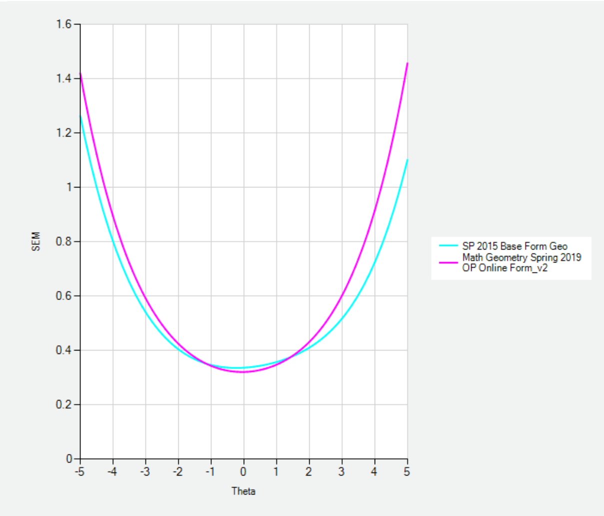
Standard Errors of Measurement



TCC Differences

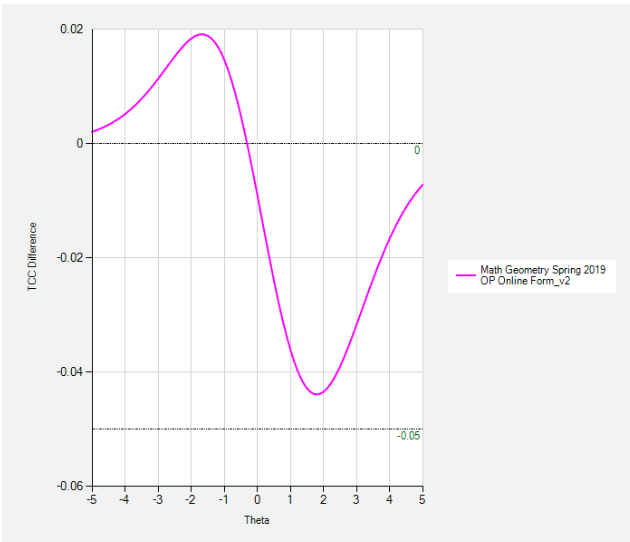
Appendix H.17 – Spring 2019 Math Geometry





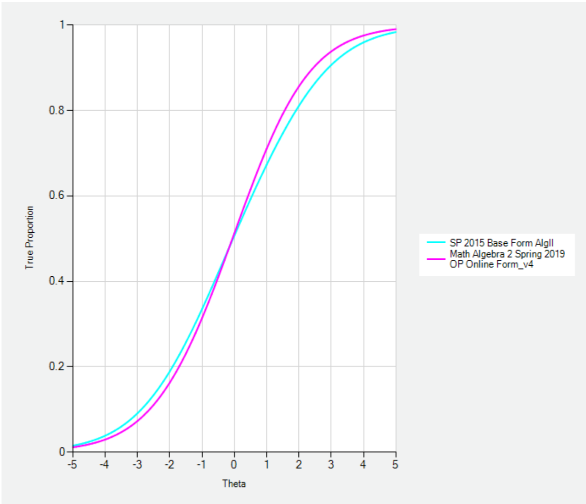
Test Characteristic Curves

Standard Errors of Measurement

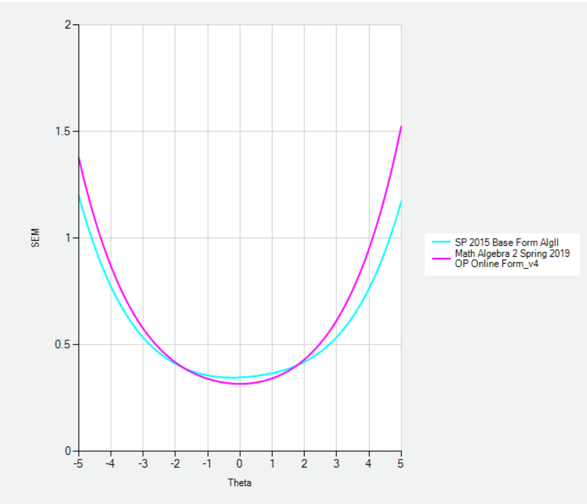


TCC Differences

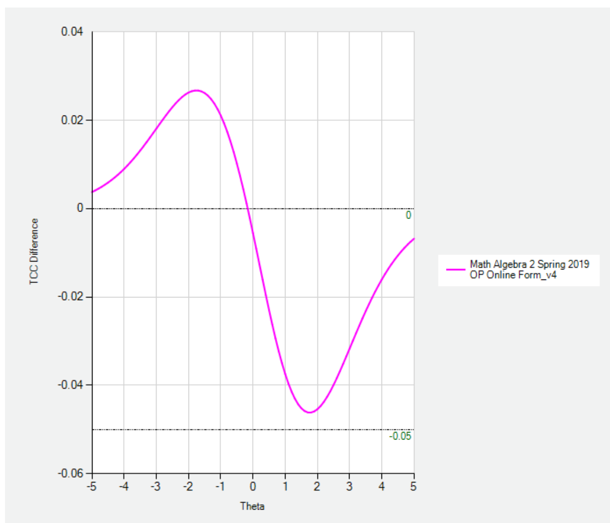
Appendix H.18 – Spring 2019 Math Algebra II



Test Characteristic Curves



Standard Errors of Measurement



TCC Differences

Appendix I.1 – Test Information Function and Ratio of Test Information Function – Grade 3 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.03	1.33
-3.00	0.06	0.05	1.20
-2.50	0.09	0.08	1.13
-2.00	0.12	0.11	1.09
-1.50	0.16	0.15	1.07
-1.00	0.19	0.19	1.00
-0.50	0.21	0.21	1.00
0.00	0.21	0.22	0.95
0.50	0.20	0.21	0.95
1.00	0.17	0.19	0.89
1.50	0.14	0.15	0.93
2.00	0.10	0.11	0.91
2.50	0.07	0.08	0.88
3.00	0.05	0.06	0.83
3.50	0.03	0.04	0.75

Appendix I.2 – Test Information Function and Ratio of Test Information Function – Grade 4 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.03	0.03	1.00
-3.00	0.05	0.05	1.00
-2.50	0.07	0.07	1.00
-2.00	0.10	0.10	1.00
-1.50	0.14	0.14	1.00
-1.00	0.17	0.18	0.94
-0.50	0.20	0.21	0.95
0.00	0.22	0.23	0.96
0.50	0.22	0.23	0.96
1.00	0.20	0.20	1.00
1.50	0.16	0.17	0.94
2.00	0.12	0.12	1.00
2.50	0.09	0.09	1.00
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.3 – Test Information Function and Ratio of Test Information Function – Grade 5 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.03	0.04	0.75
-3.00	0.05	0.05	1.00
-2.50	0.08	0.08	1.00
-2.00	0.11	0.11	1.00
-1.50	0.15	0.15	1.00
-1.00	0.18	0.18	1.00
-0.50	0.21	0.21	1.00
0.00	0.21	0.22	0.95
0.50	0.21	0.21	1.00
1.00	0.18	0.18	1.00
1.50	0.15	0.15	1.00
2.00	0.12	0.11	1.09
2.50	0.09	0.08	1.13
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix J.4 – Test Information Function and Ratio of Test Information Function – Grade 6 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.05	1.20
-2.50	0.08	0.08	1.00
-2.00	0.11	0.11	1.00
-1.50	0.14	0.15	0.93
-1.00	0.17	0.18	0.94
-0.50	0.20	0.21	0.95
0.00	0.21	0.22	0.95
0.50	0.20	0.21	0.95
1.00	0.18	0.19	0.95
1.50	0.15	0.15	1.00
2.00	0.12	0.11	1.09
2.50	0.09	0.08	1.13
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.5 – Test Information Function and Ratio of Test Information Function – Grade 7 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.03	0.04	0.75
-3.00	0.05	0.06	0.83
-2.50	0.08	0.08	1.00
-2.00	0.11	0.11	1.00
-1.50	0.14	0.15	0.93
-1.00	0.18	0.18	1.00
-0.50	0.20	0.20	1.00
0.00	0.21	0.21	1.00
0.50	0.21	0.21	1.00
1.00	0.19	0.18	1.06
1.50	0.16	0.15	1.07
2.00	0.12	0.12	1.00
2.50	0.09	0.09	1.00
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.6 – Test Information Function and Ratio of Test Information Function – Grade 8 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.06	1.00
-2.50	0.08	0.09	0.89
-2.00	0.11	0.12	0.92
-1.50	0.15	0.15	1.00
-1.00	0.18	0.19	0.95
-0.50	0.20	0.21	0.95
0.00	0.21	0.21	1.00
0.50	0.20	0.20	1.00
1.00	0.18	0.17	1.06
1.50	0.15	0.14	1.07
2.00	0.12	0.11	1.09
2.50	0.09	0.08	1.13
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.7 – Test Information Function and Ratio of Test Information Function – Grade 9 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.03	1.33
-3.00	0.06	0.05	1.20
-2.50	0.08	0.07	1.14
-2.00	0.11	0.10	1.10
-1.50	0.15	0.14	1.07
-1.00	0.18	0.18	1.00
-0.50	0.20	0.22	0.91
0.00	0.21	0.23	0.91
0.50	0.20	0.22	0.91
1.00	0.18	0.20	0.90
1.50	0.15	0.16	0.94
2.00	0.12	0.12	1.00
2.50	0.09	0.08	1.13
3.00	0.06	0.05	1.20
3.50	0.04	0.03	1.33

Appendix I.8 – Test Information Function and Ratio of Test Information Function – Grade 10 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.03	0.03	1.00
-3.00	0.05	0.05	1.00
-2.50	0.08	0.07	1.14
-2.00	0.11	0.11	1.00
-1.50	0.15	0.14	1.07
-1.00	0.19	0.18	1.06
-0.50	0.21	0.20	1.05
0.00	0.22	0.22	1.00
0.50	0.21	0.21	1.00
1.00	0.18	0.19	0.95
1.50	0.15	0.16	0.94
2.00	0.11	0.12	0.92
2.50	0.08	0.09	0.89
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.9 – Test Information Function and Ratio of Test Information Function – Grade 11 ELA

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.05	0.04	1.25
-3.00	0.07	0.06	1.17
-2.50	0.10	0.09	1.11
-2.00	0.13	0.12	1.08
-1.50	0.17	0.16	1.06
-1.00	0.19	0.19	1.00
-0.50	0.21	0.21	1.00
0.00	0.20	0.21	0.95
0.50	0.19	0.20	0.95
1.00	0.16	0.17	0.94
1.50	0.13	0.14	0.93
2.00	0.10	0.11	0.91
2.50	0.07	0.08	0.88
3.00	0.05	0.06	0.83
3.50	0.04	0.04	1.00

Appendix I.10 – Test Information Function and Ratio of Test Information Function – Grade 3 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.06	1.00
-2.50	0.08	0.08	1.00
-2.00	0.11	0.11	1.00
-1.50	0.14	0.14	1.00
-1.00	0.17	0.17	1.00
-0.50	0.19	0.19	1.00
0.00	0.19	0.19	1.00
0.50	0.19	0.19	1.00
1.00	0.18	0.18	1.00
1.50	0.16	0.15	1.07
2.00	0.13	0.12	1.08
2.50	0.10	0.10	1.00
3.00	0.07	0.07	1.00
3.50	0.05	0.05	1.00

Appendix I.11 – Test Information Function and Ratio of Test Information Function – Grade 4 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.07	0.86
-2.50	0.09	0.09	1.00
-2.00	0.12	0.12	1.00
-1.50	0.15	0.15	1.00
-1.00	0.18	0.17	1.06
-0.50	0.19	0.18	1.06
0.00	0.20	0.19	1.05
0.50	0.19	0.18	1.06
1.00	0.18	0.17	1.06
1.50	0.15	0.15	1.00
2.00	0.12	0.12	1.00
2.50	0.09	0.09	1.00
3.00	0.06	0.07	0.86
3.50	0.04	0.05	0.80

Appendix I.12 – Test Information Function and Ratio of Test Information Function – Grade 5 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.06	1.00
-2.50	0.09	0.08	1.13
-2.00	0.11	0.11	1.00
-1.50	0.15	0.15	1.00
-1.00	0.17	0.18	0.94
-0.50	0.20	0.20	1.00
0.00	0.20	0.21	0.95
0.50	0.20	0.20	1.00
1.00	0.18	0.18	1.00
1.50	0.15	0.15	1.00
2.00	0.12	0.12	1.00
2.50	0.09	0.09	1.00
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

Appendix I.13 – Test Information Function and Ratio of Test Information Function – Grade 6 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.05	0.04	1.25
-3.00	0.07	0.06	1.17
-2.50	0.09	0.09	1.00
-2.00	0.11	0.11	1.00
-1.50	0.13	0.14	0.93
-1.00	0.16	0.17	0.94
-0.50	0.17	0.19	0.89
0.00	0.18	0.20	0.90
0.50	0.19	0.19	1.00
1.00	0.18	0.18	1.00
1.50	0.15	0.15	1.00
2.00	0.13	0.12	1.08
2.50	0.10	0.09	1.11
3.00	0.07	0.07	1.00
3.50	0.05	0.04	1.25

Appendix I.14 – Test Information Function and Ratio of Test Information Function – Grade 7 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.06	1.00
-2.50	0.09	0.08	1.13
-2.00	0.12	0.11	1.09
-1.50	0.15	0.14	1.07
-1.00	0.17	0.17	1.00
-0.50	0.19	0.19	1.00
0.00	0.20	0.20	1.00
0.50	0.19	0.19	1.00
1.00	0.18	0.18	1.00
1.50	0.15	0.15	1.00
2.00	0.12	0.12	1.00
2.50	0.09	0.09	1.00
3.00	0.06	0.07	0.86
3.50	0.04	0.05	0.80

Appendix I.15 – Test Information Function and Ratio of Test Information Function – Grade 8 Math

Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.05	0.05	1.00
-3.00	0.07	0.07	1.00
-2.50	0.09	0.09	1.00
-2.00	0.12	0.12	1.00
-1.50	0.15	0.15	1.00
-1.00	0.18	0.18	1.00
-0.50	0.19	0.20	0.95
0.00	0.20	0.21	0.95
0.50	0.19	0.20	0.95
1.00	0.17	0.17	1.00
1.50	0.14	0.14	1.00
2.00	0.11	0.11	1.00
2.50	0.09	0.08	1.13
3.00	0.06	0.05	1.20
3.50	0.04	0.04	1.00

Appendix I.16 – Test Information Function and Ratio of Test Information Function – Algebra I

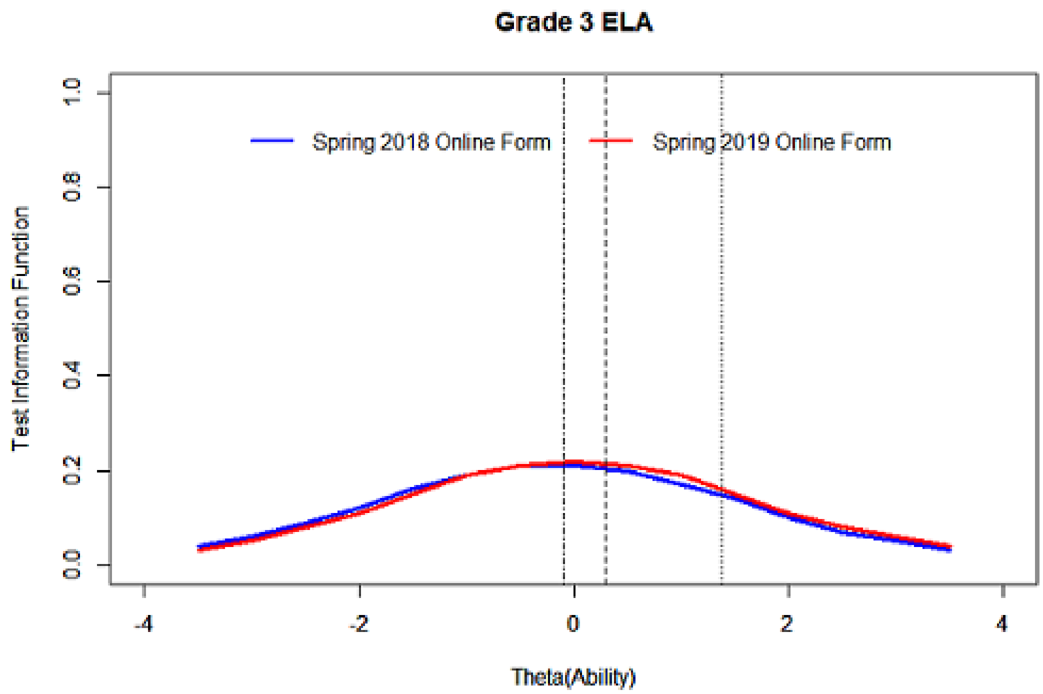
Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.06	0.05	1.20
-2.50	0.08	0.08	1.00
-2.00	0.11	0.11	1.00
-1.50	0.15	0.14	1.07
-1.00	0.18	0.17	1.06
-0.50	0.20	0.19	1.05
0.00	0.21	0.20	1.05
0.50	0.20	0.20	1.00
1.00	0.18	0.18	1.00
1.50	0.15	0.16	0.94
2.00	0.12	0.13	0.92
2.50	0.09	0.10	0.90
3.00	0.06	0.07	0.86
3.50	0.04	0.05	0.80

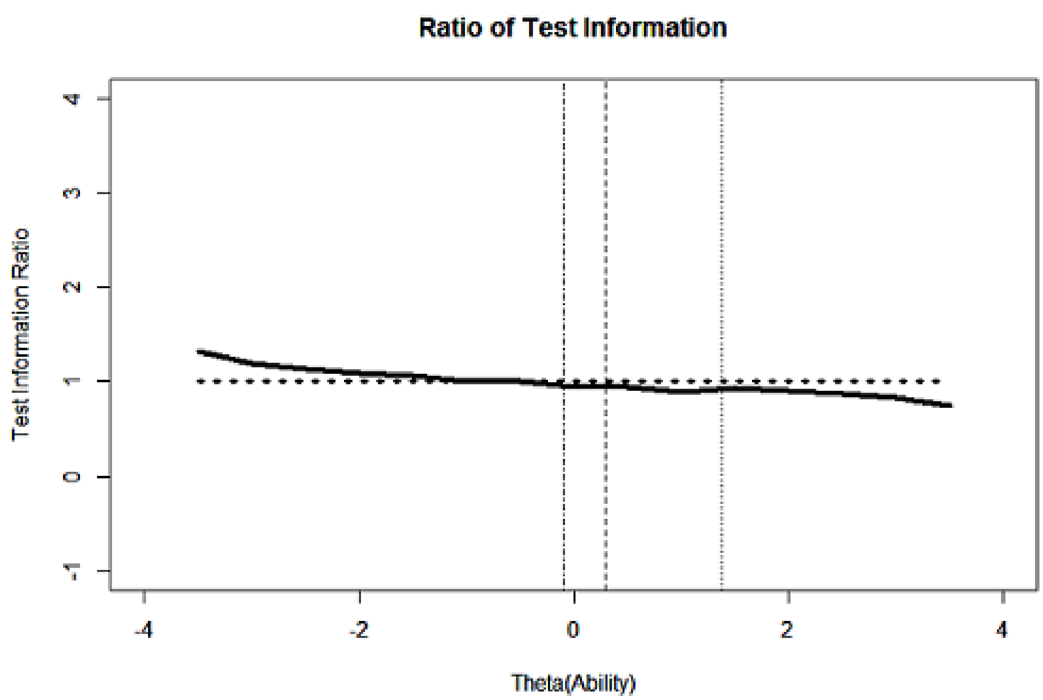
Appendix I.17 – Test Information Function and Ratio of Test Information Function – Geometry

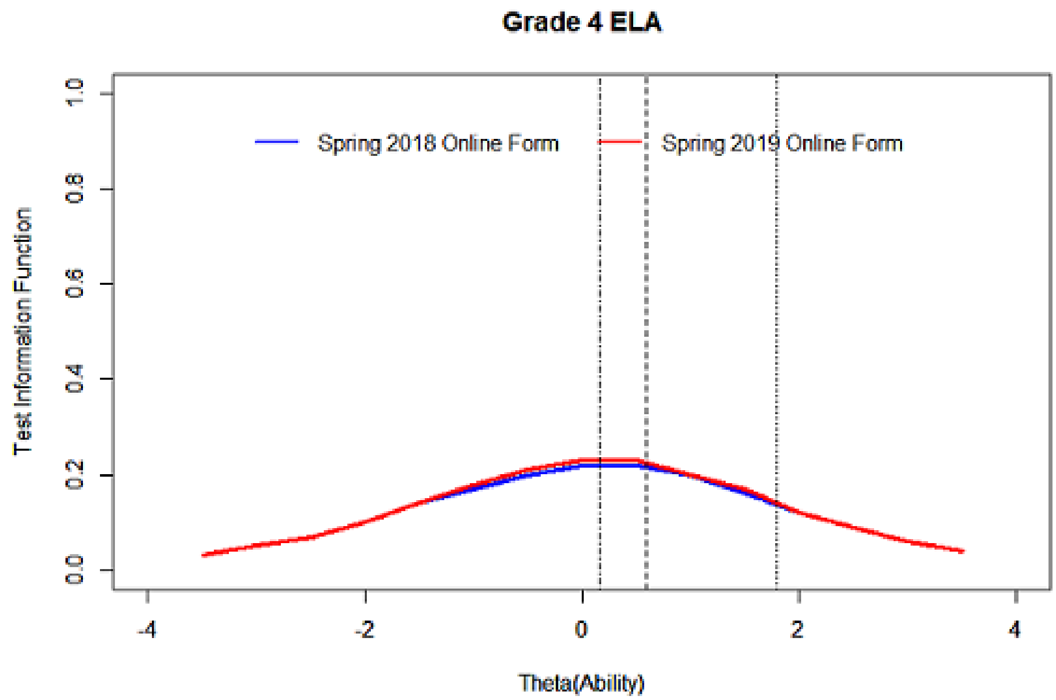
Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.04	0.04	1.00
-3.00	0.07	0.06	1.17
-2.50	0.09	0.09	1.00
-2.00	0.12	0.12	1.00
-1.50	0.15	0.15	1.00
-1.00	0.18	0.18	1.00
-0.50	0.19	0.20	0.95
0.00	0.20	0.21	0.95
0.50	0.19	0.20	0.95
1.00	0.17	0.18	0.94
1.50	0.14	0.15	0.93
2.00	0.12	0.11	1.09
2.50	0.09	0.08	1.13
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

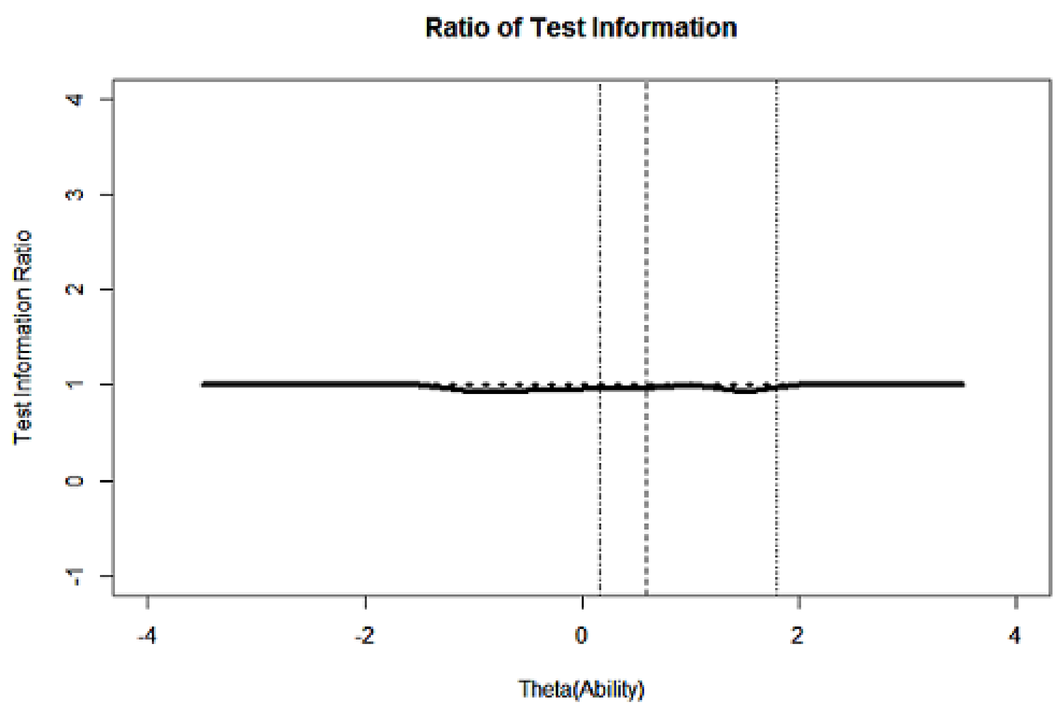
Appendix I.18 – Test Information Function and Ratio of Test Information Function – Algebra II

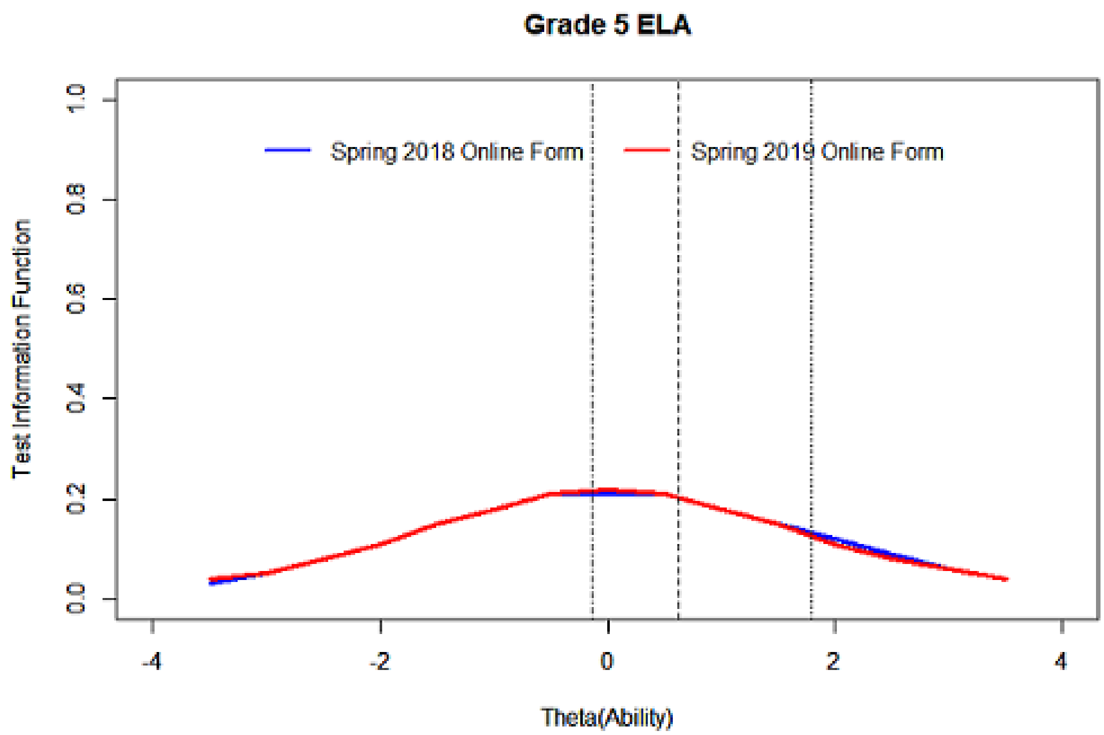
Theta	Spring 2018 Online Form	Spring 2019 Online Form	Ratio (Spring 2018 Online Form / Spring 2019 Online Form)
-3.50	0.05	0.04	1.25
-3.00	0.07	0.06	1.17
-2.50	0.09	0.09	1.00
-2.00	0.12	0.12	1.00
-1.50	0.15	0.15	1.00
-1.00	0.17	0.18	0.94
-0.50	0.18	0.20	0.90
0.00	0.19	0.21	0.90
0.50	0.19	0.20	0.95
1.00	0.18	0.18	1.00
1.50	0.15	0.15	1.00
2.00	0.12	0.11	1.09
2.50	0.09	0.08	1.13
3.00	0.06	0.06	1.00
3.50	0.04	0.04	1.00

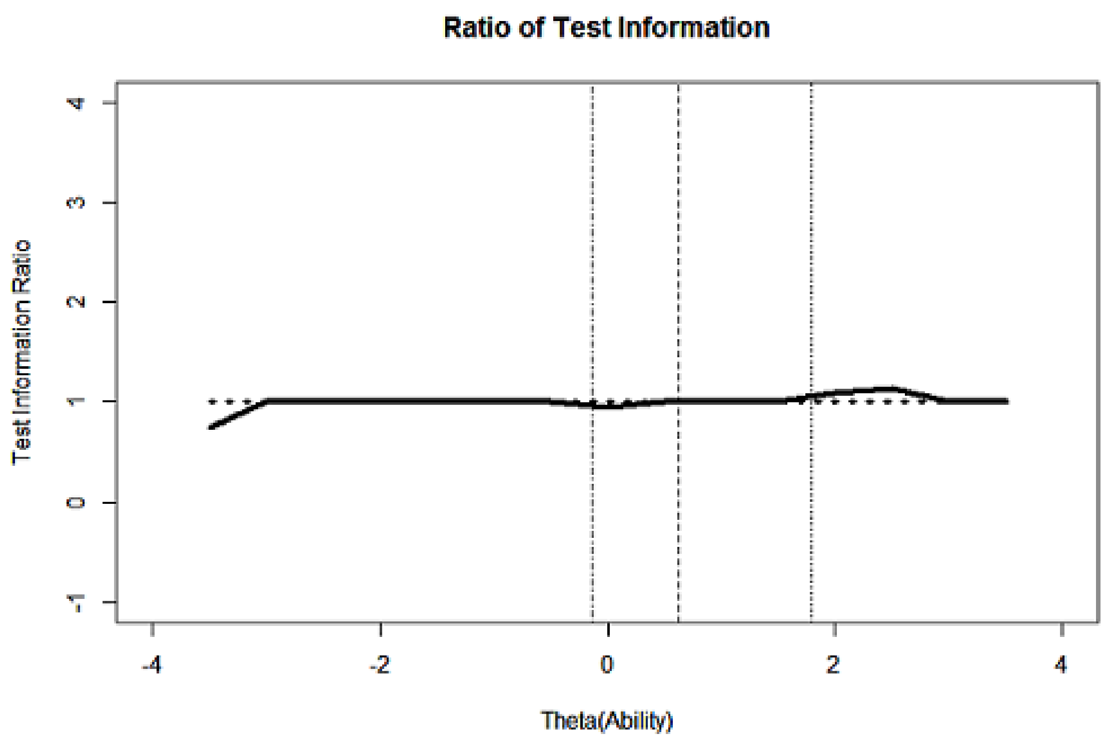

Grade 3 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

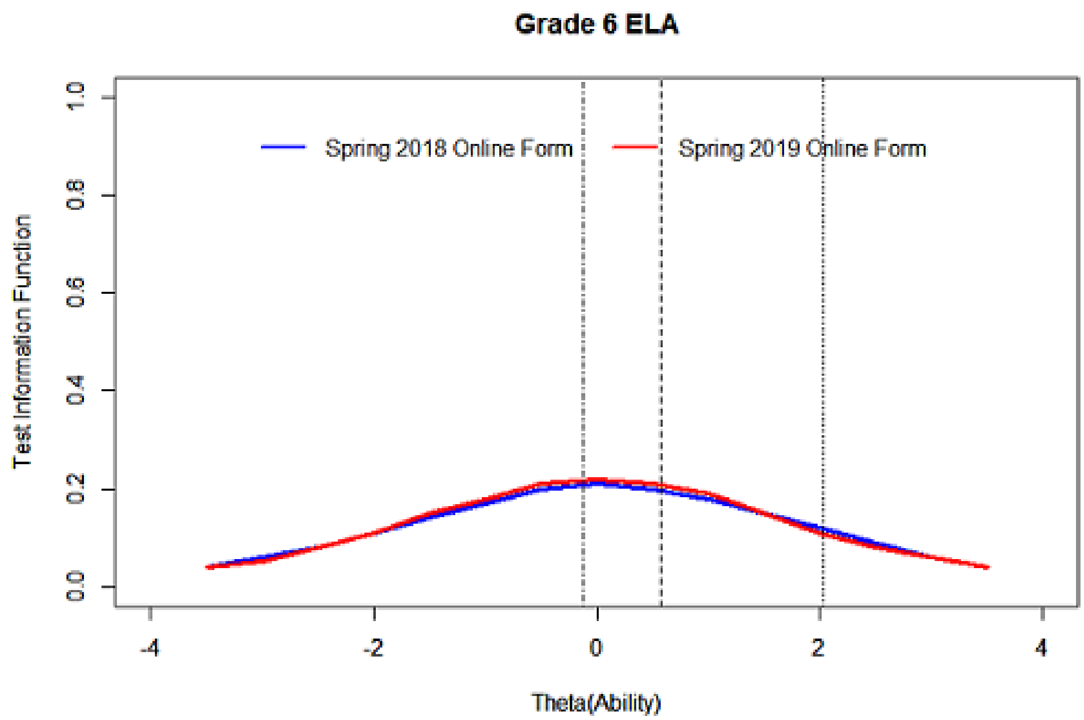

Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

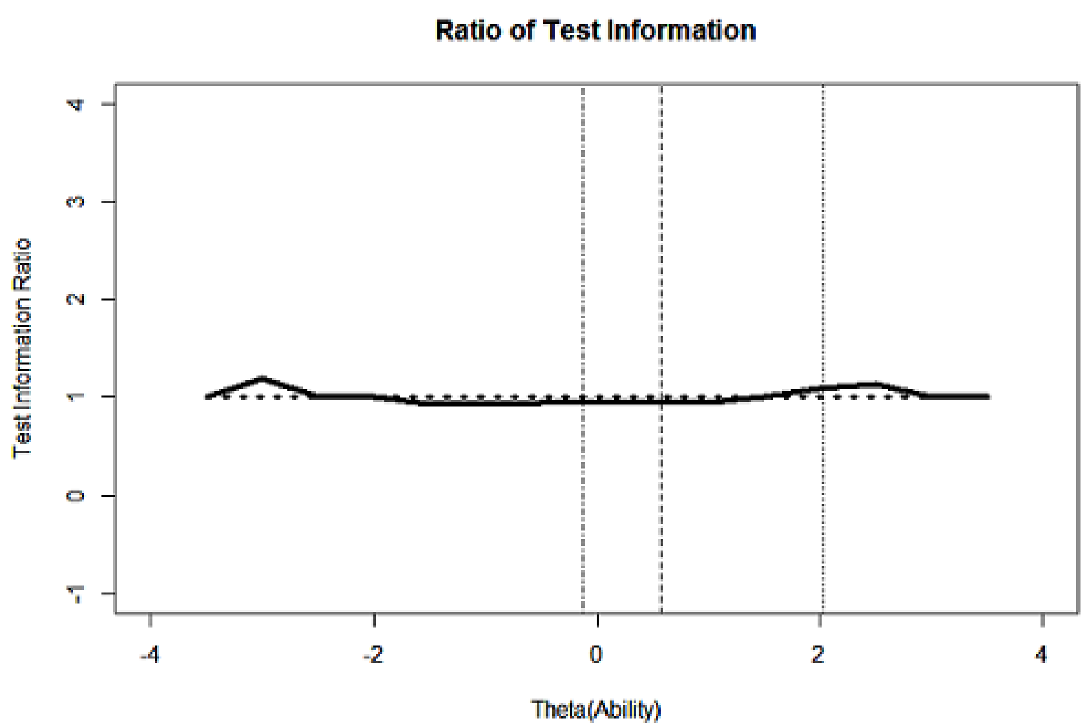

Grade 4 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

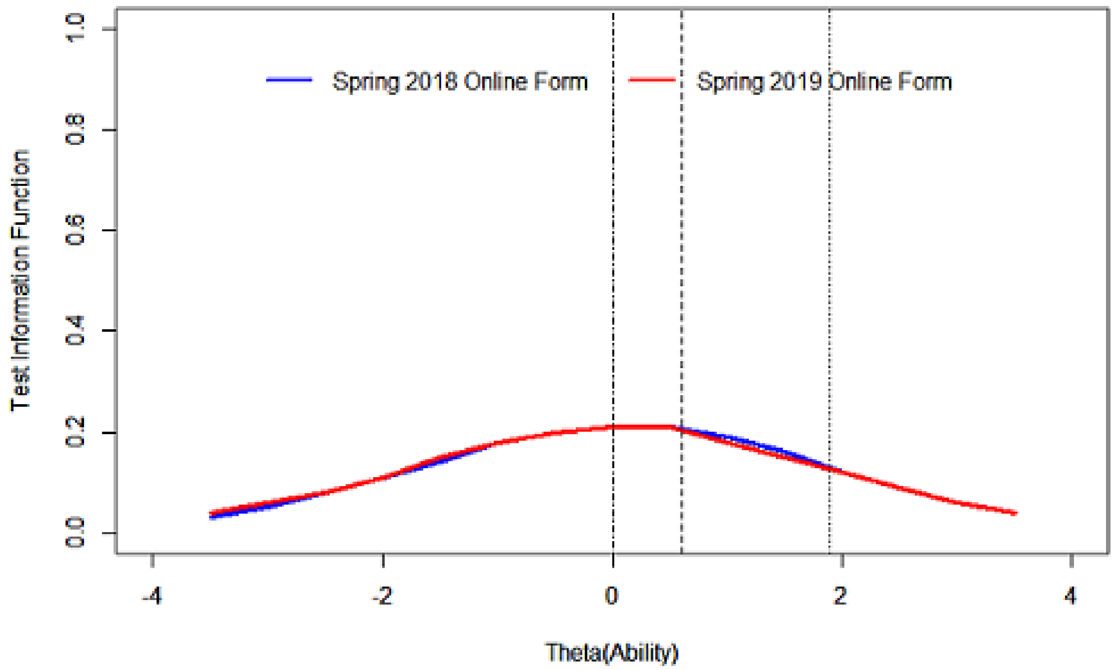

Grade 5 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

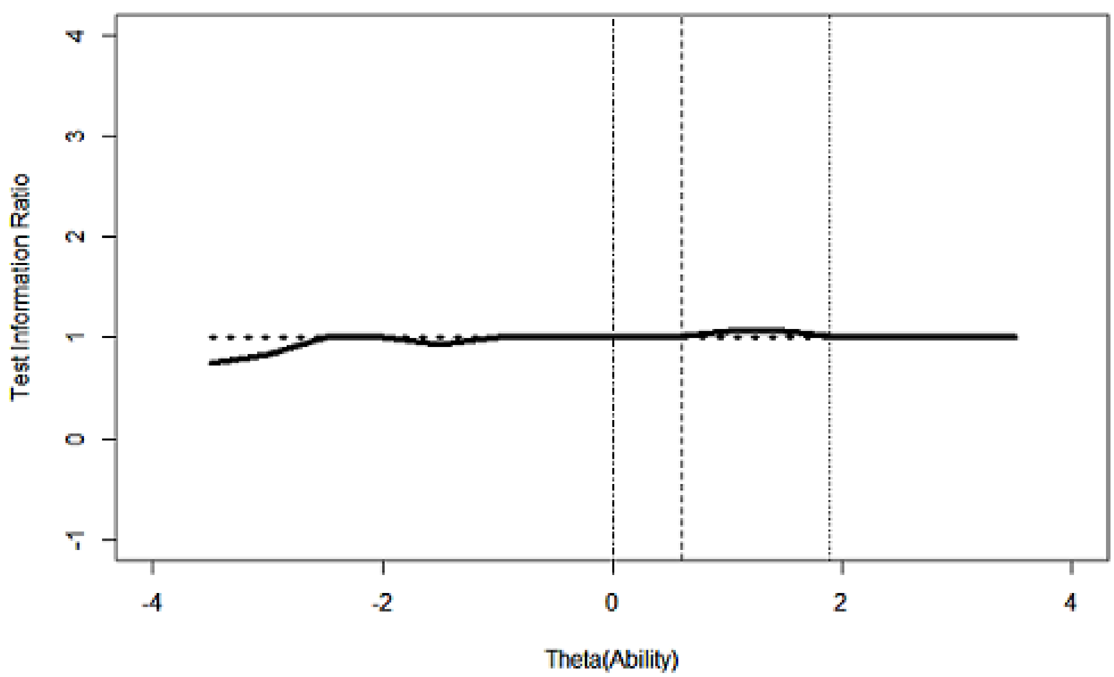

Grade 6 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

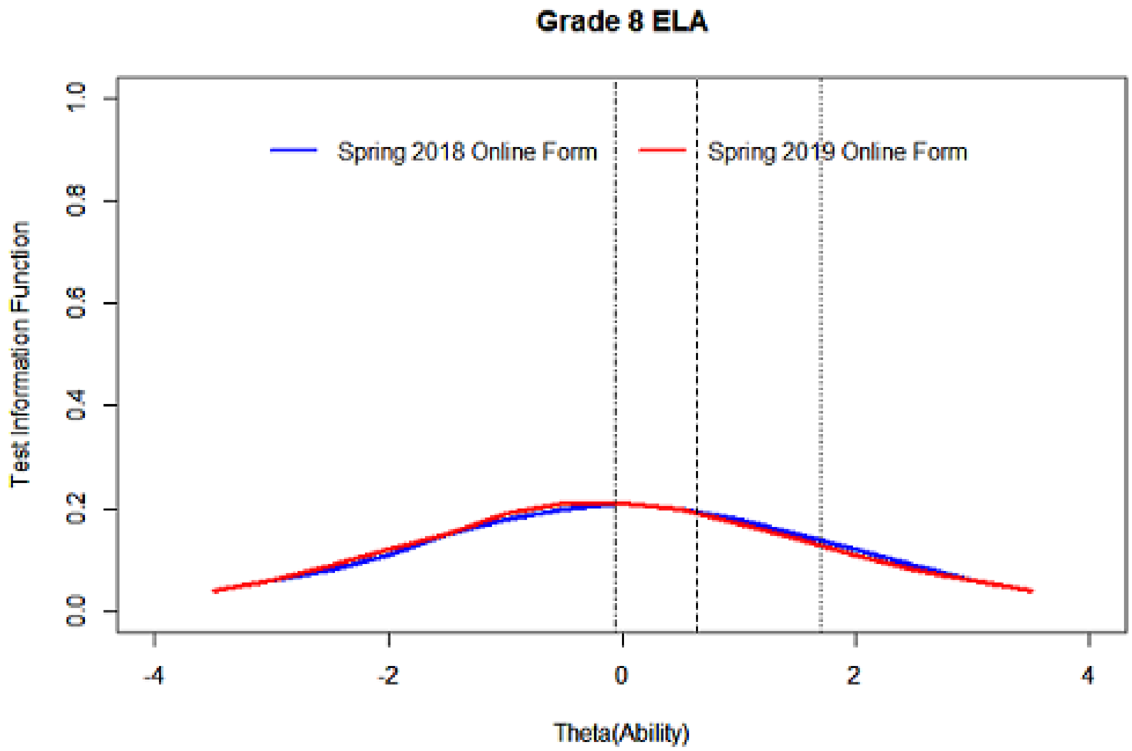

Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

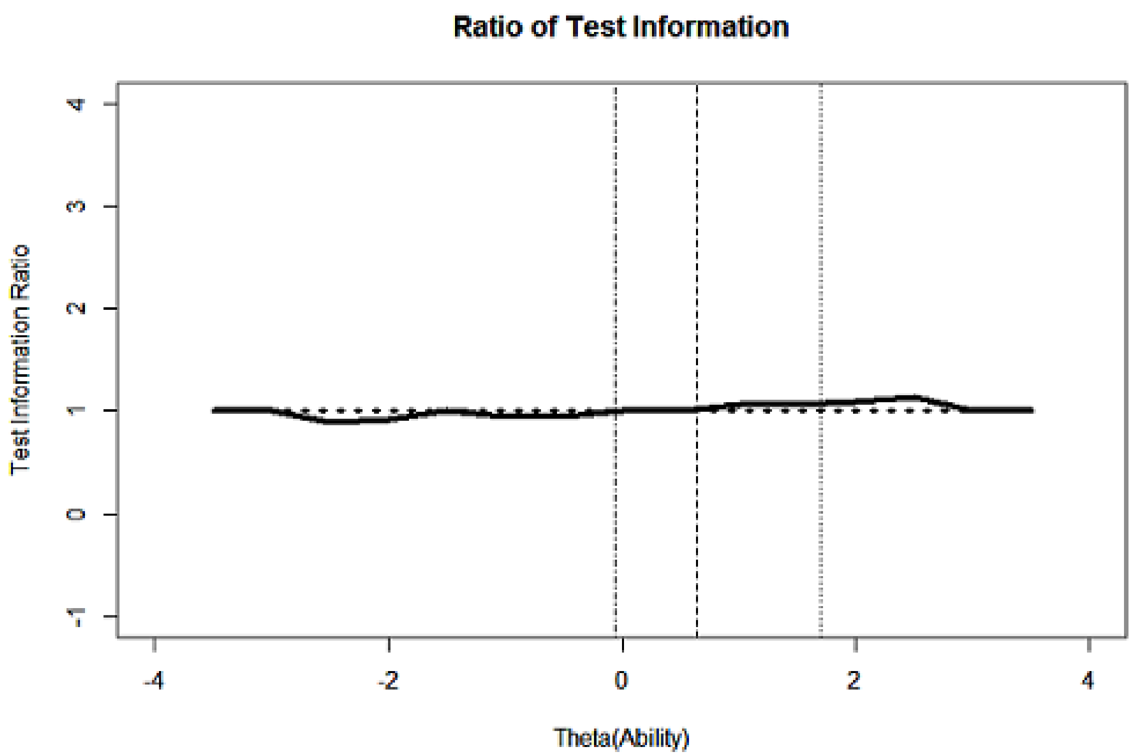
Grade 7 ELA

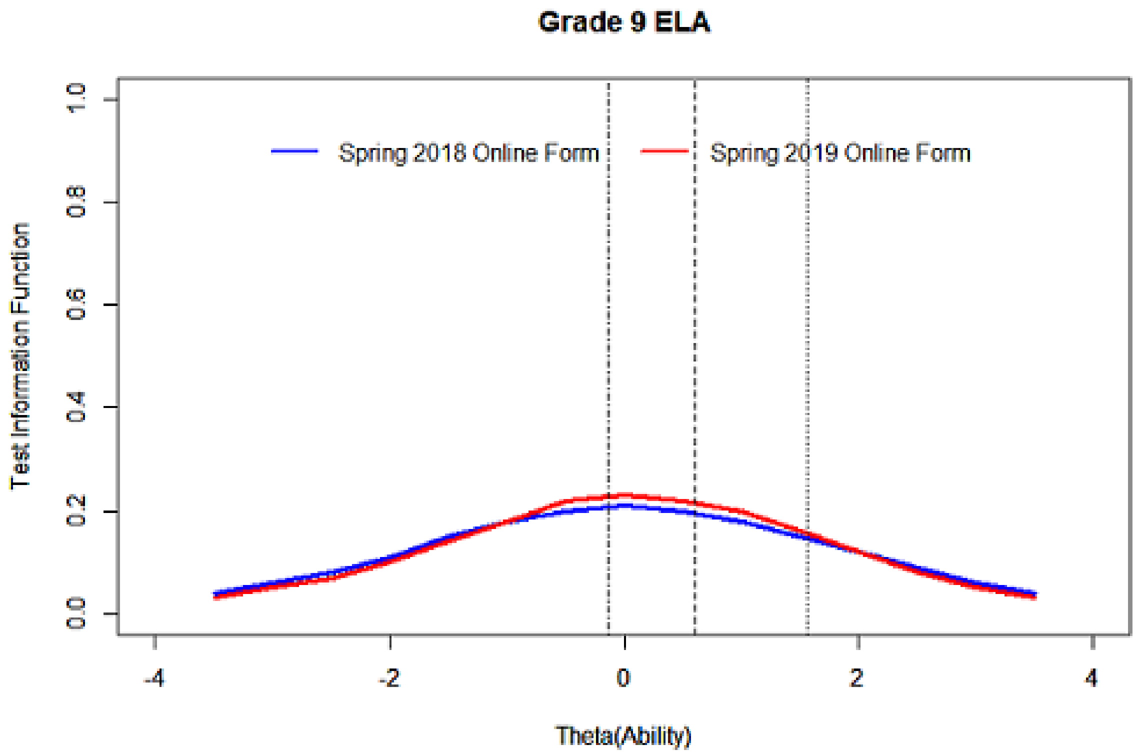


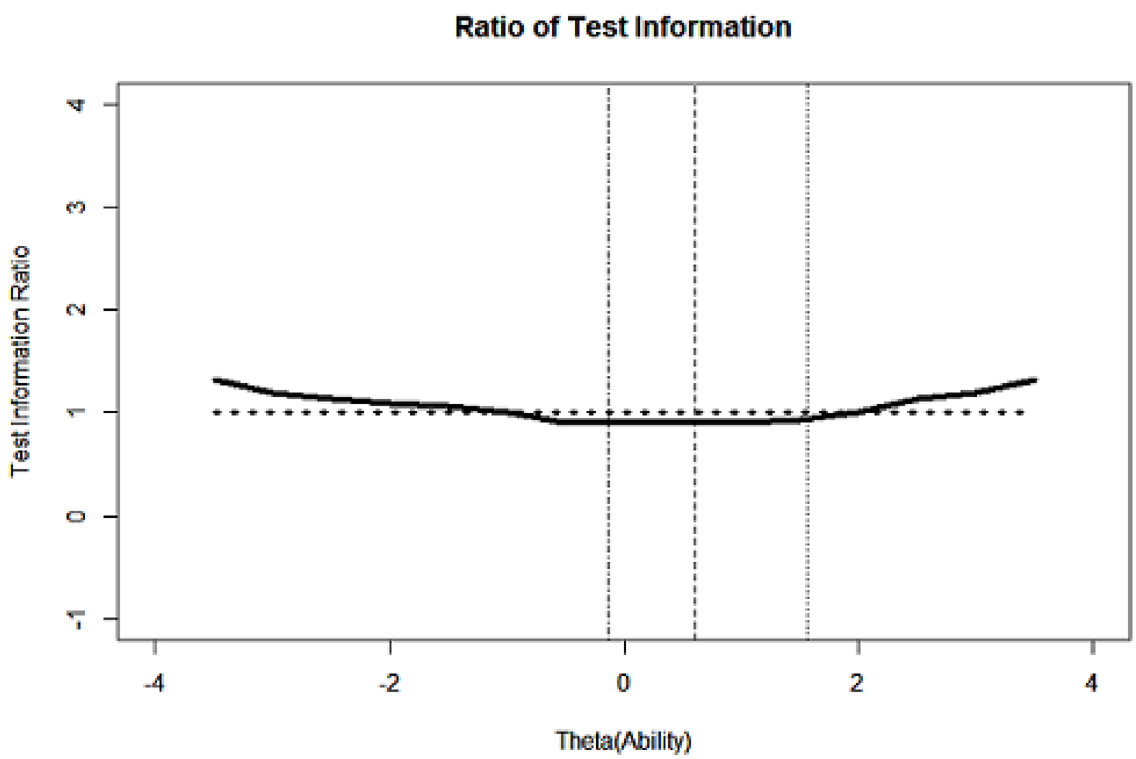
Ratio of Test Information

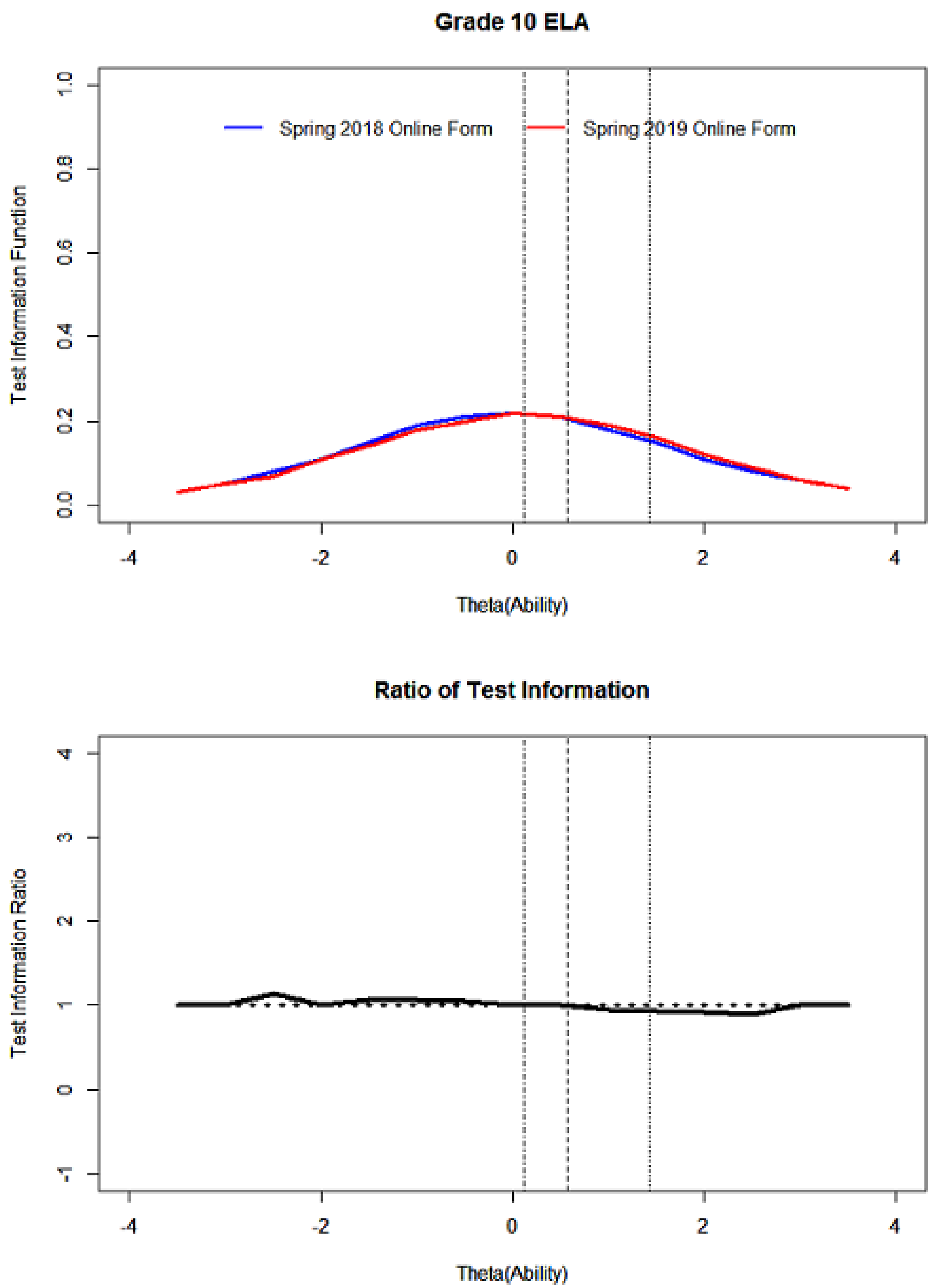


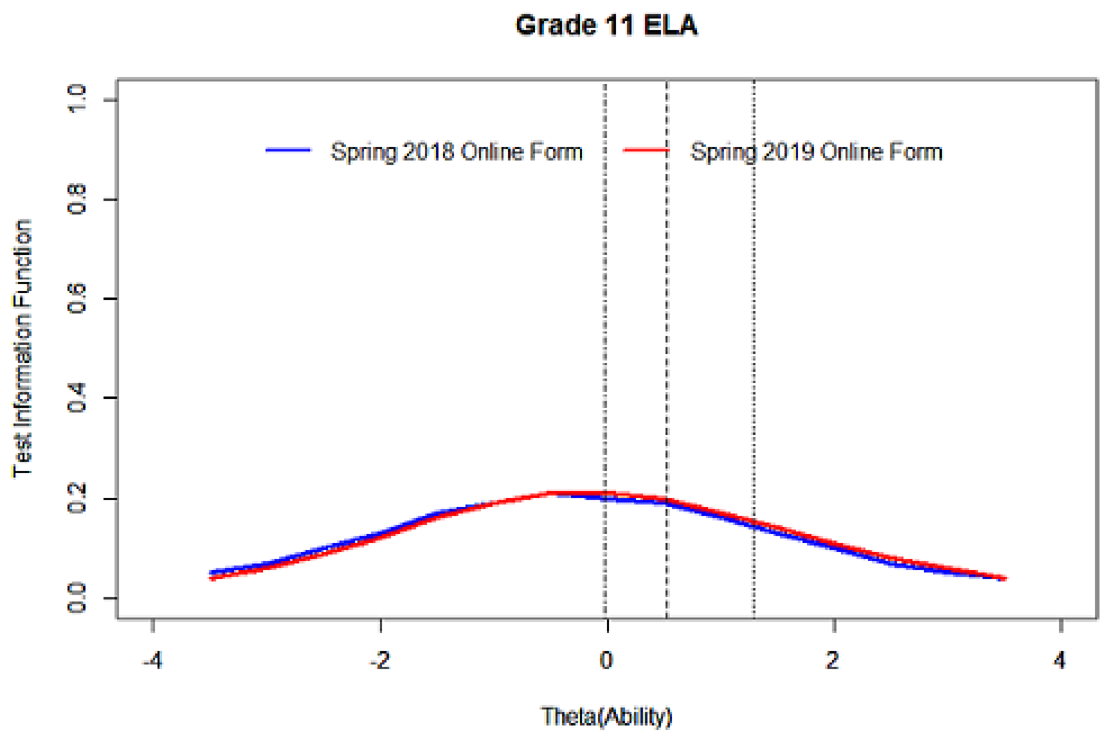

Grade 8 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

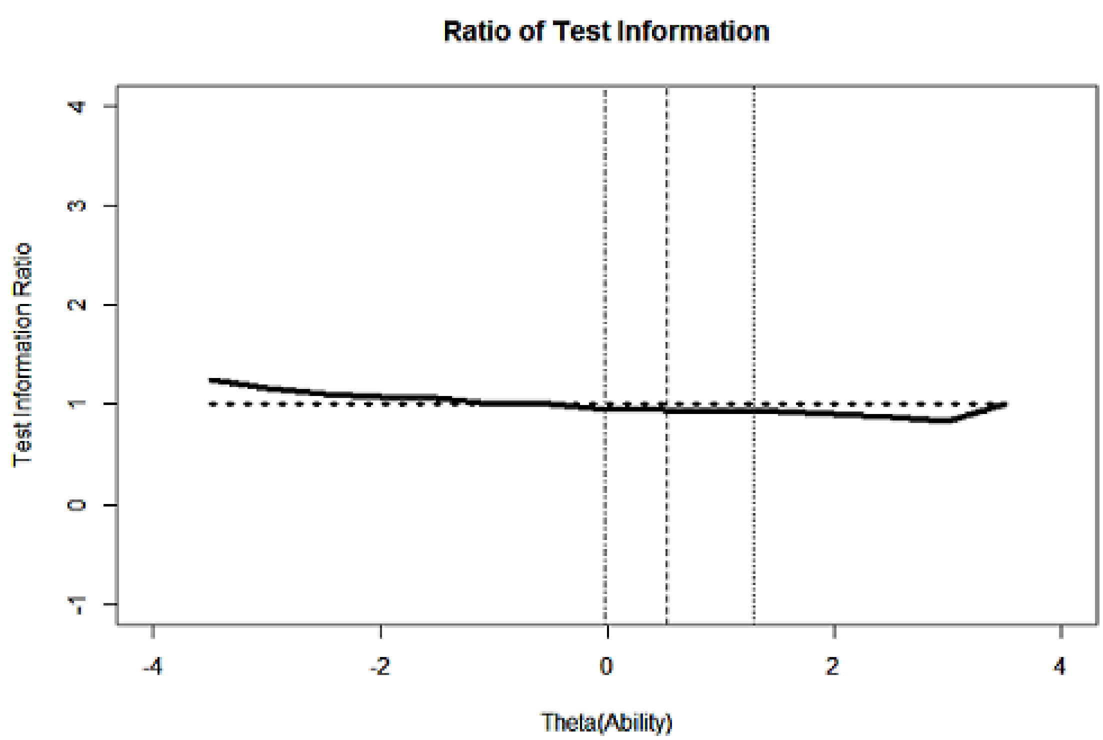

Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

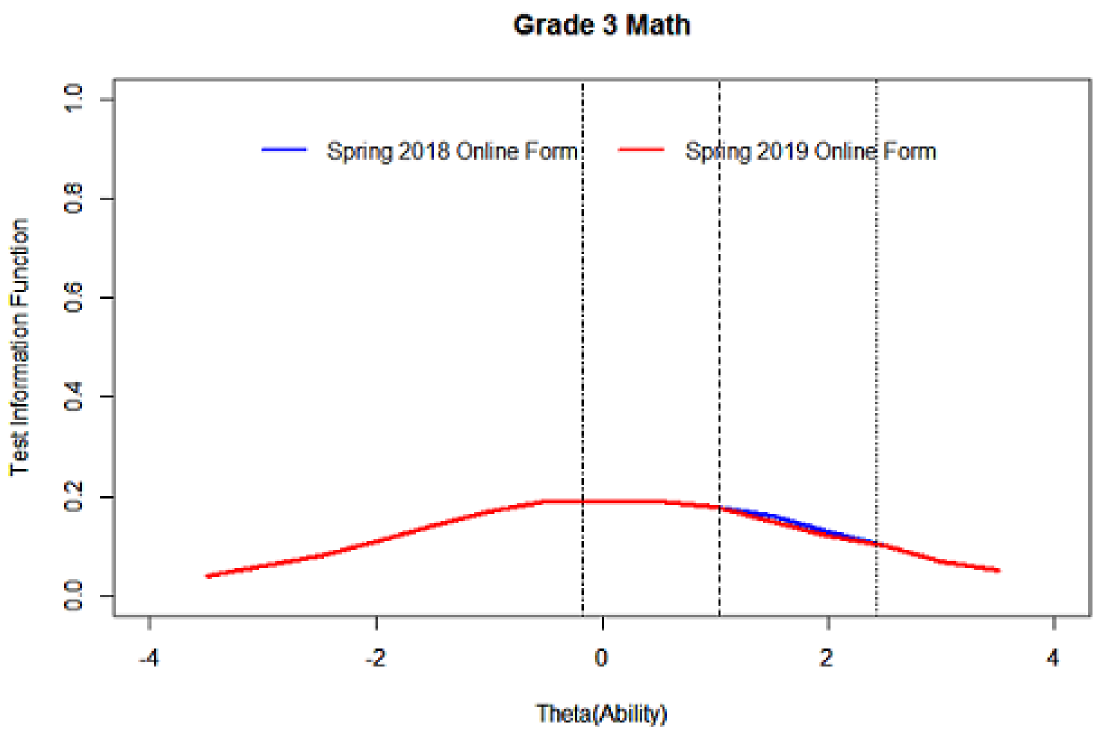

Grade 9 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

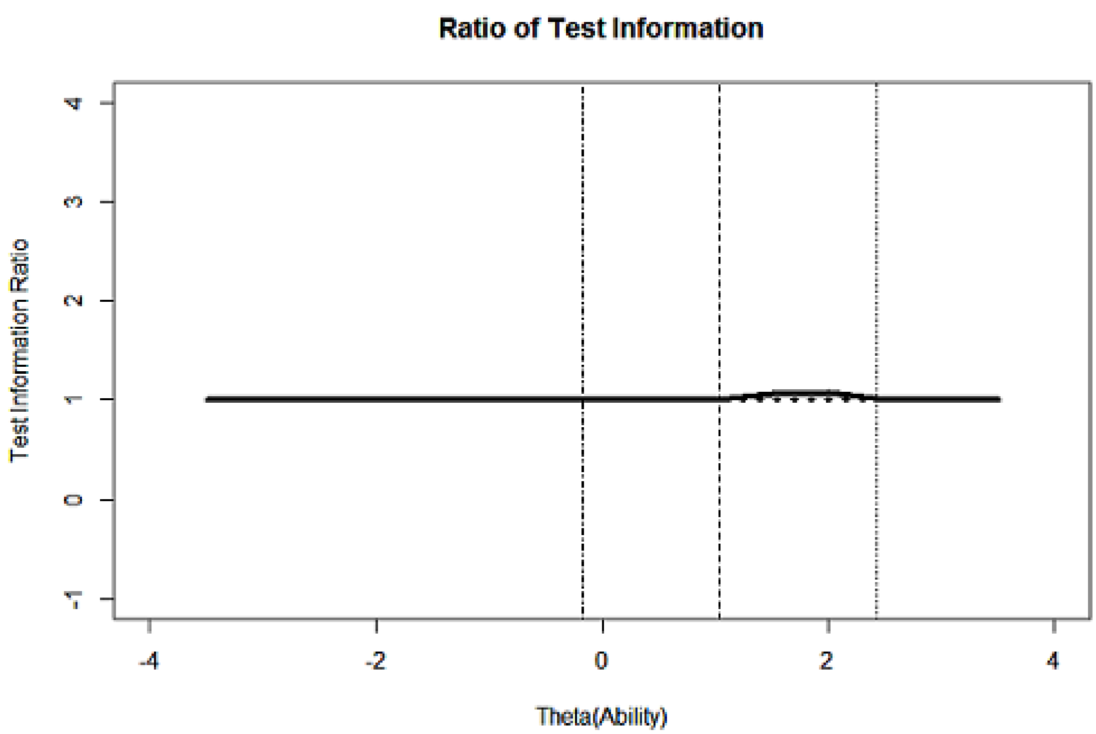

Grade 10 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
Theta(Ability)
Ratio of Test Information
Test Information Ratio
Theta(Ability)


Grade 11 ELA
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

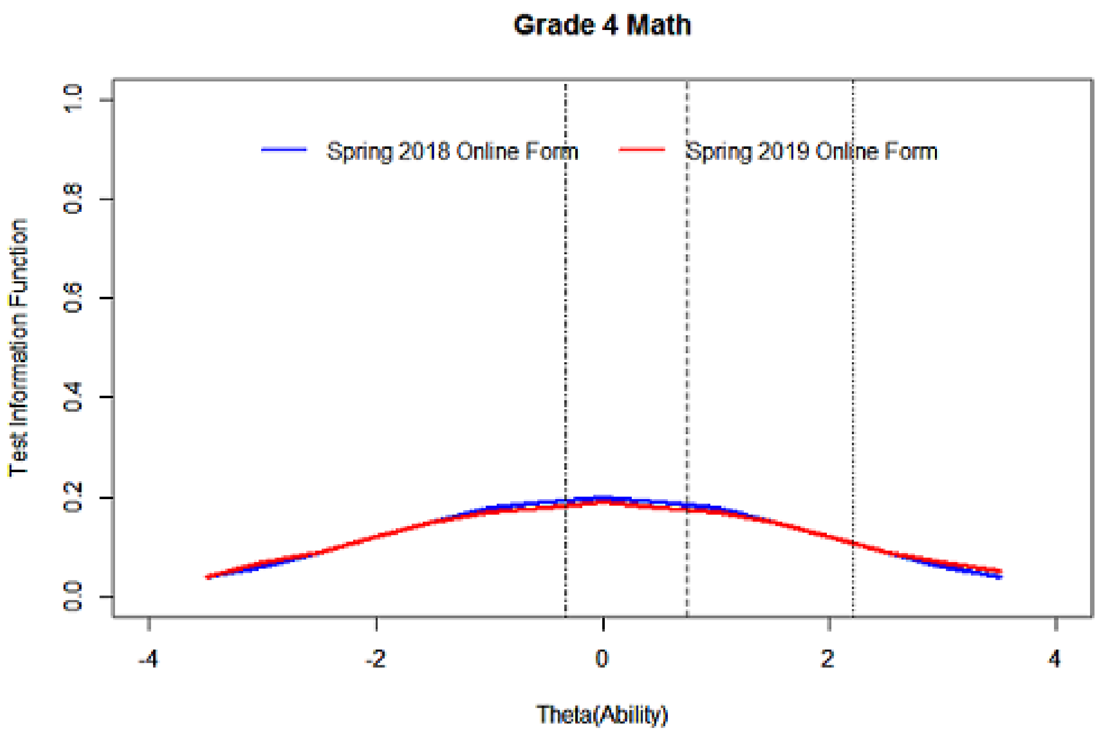

Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

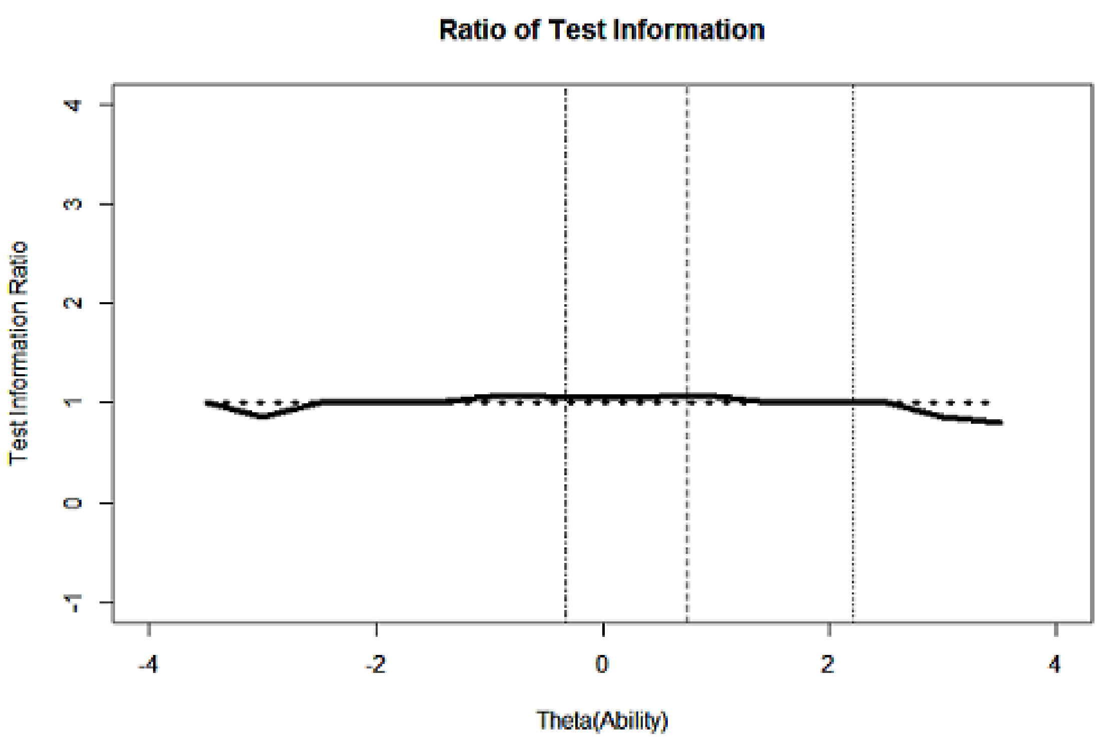


Grade 3 Math
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

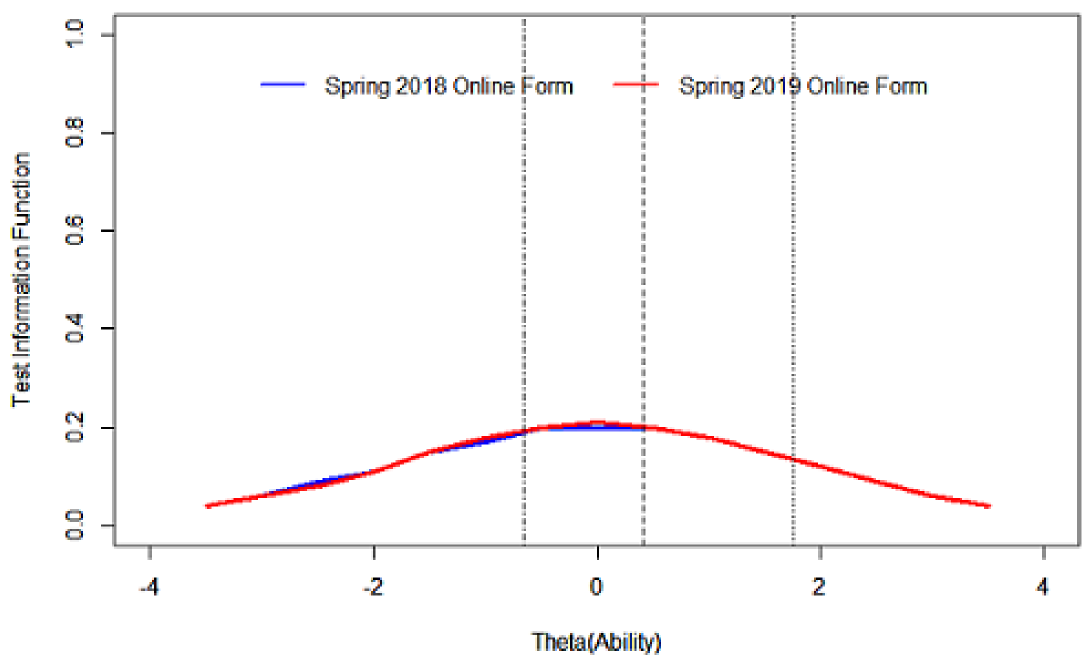


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

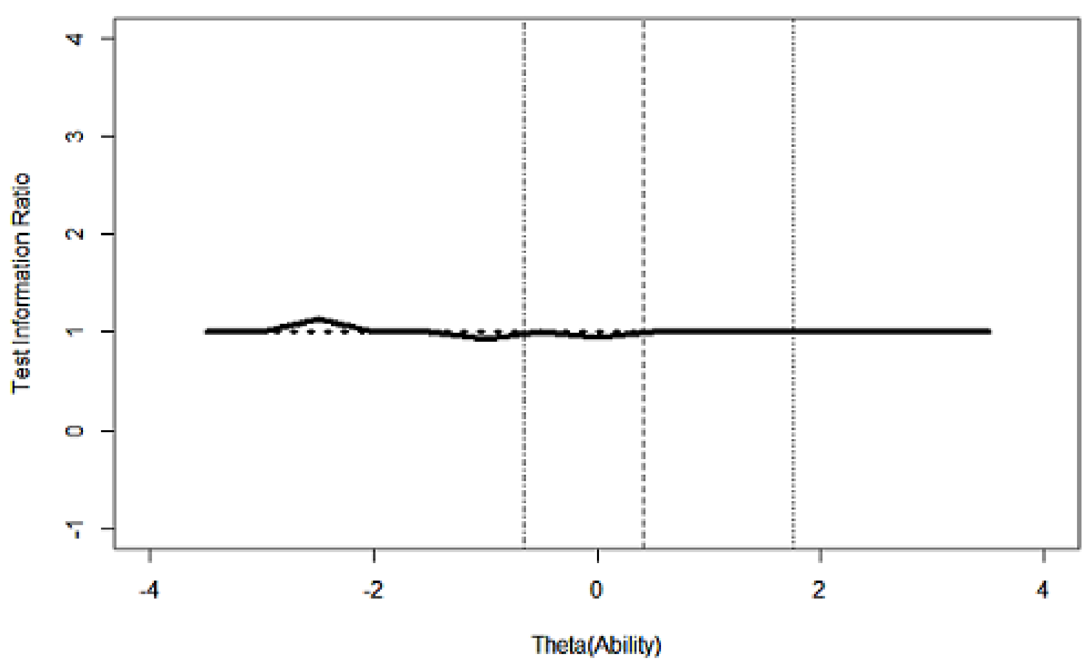

Grade 4 Math
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

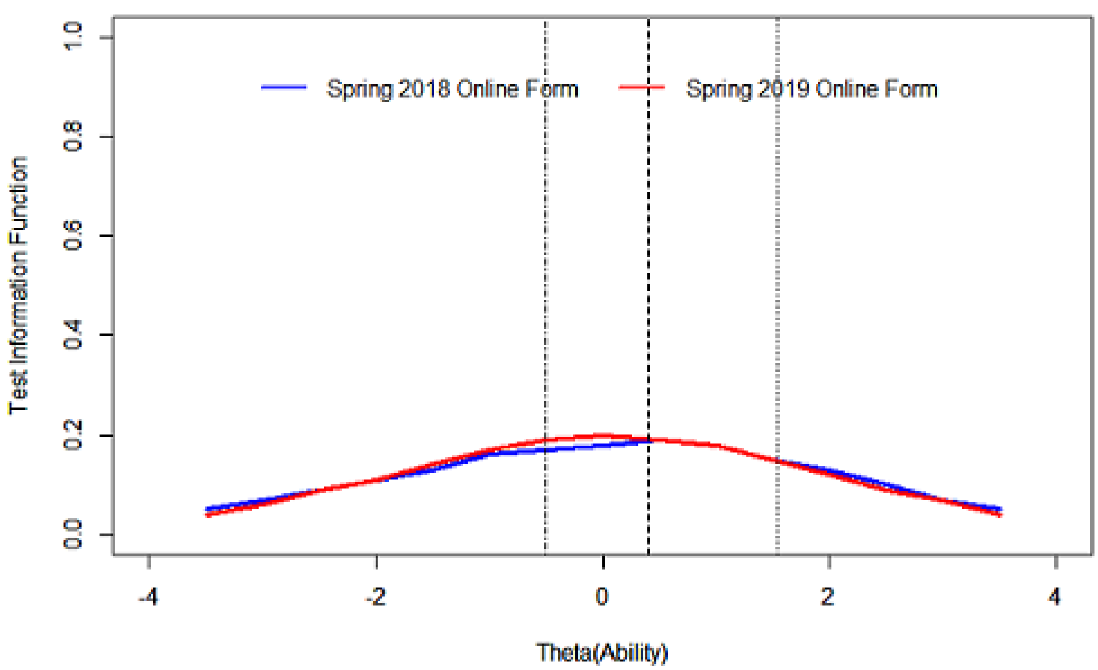
Grade 5 Math



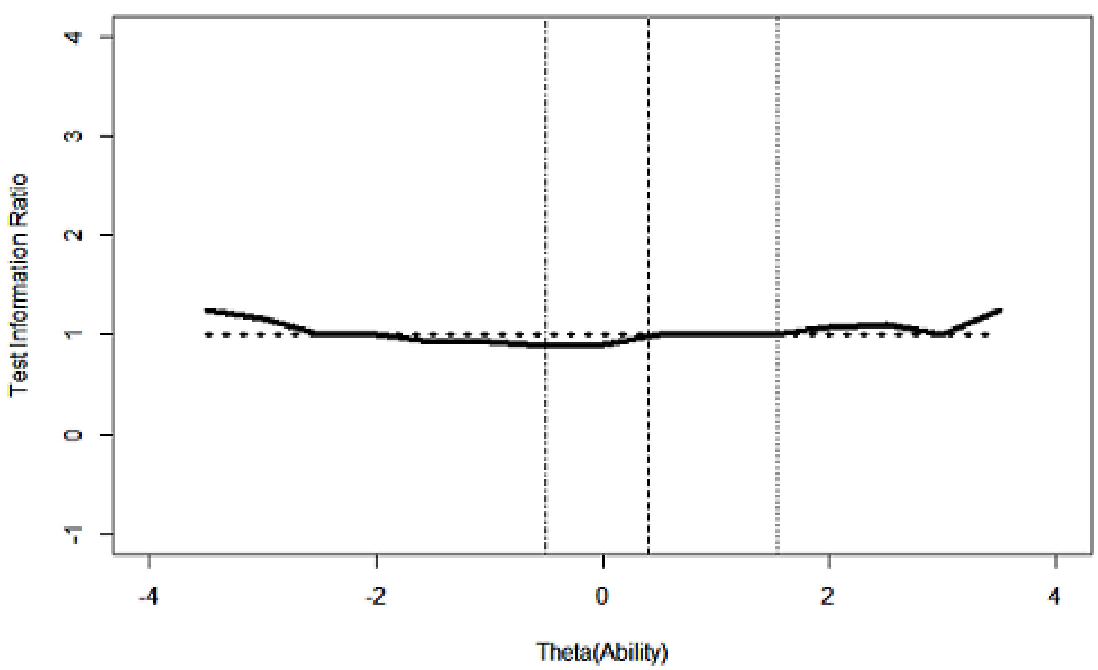
Ratio of Test Information

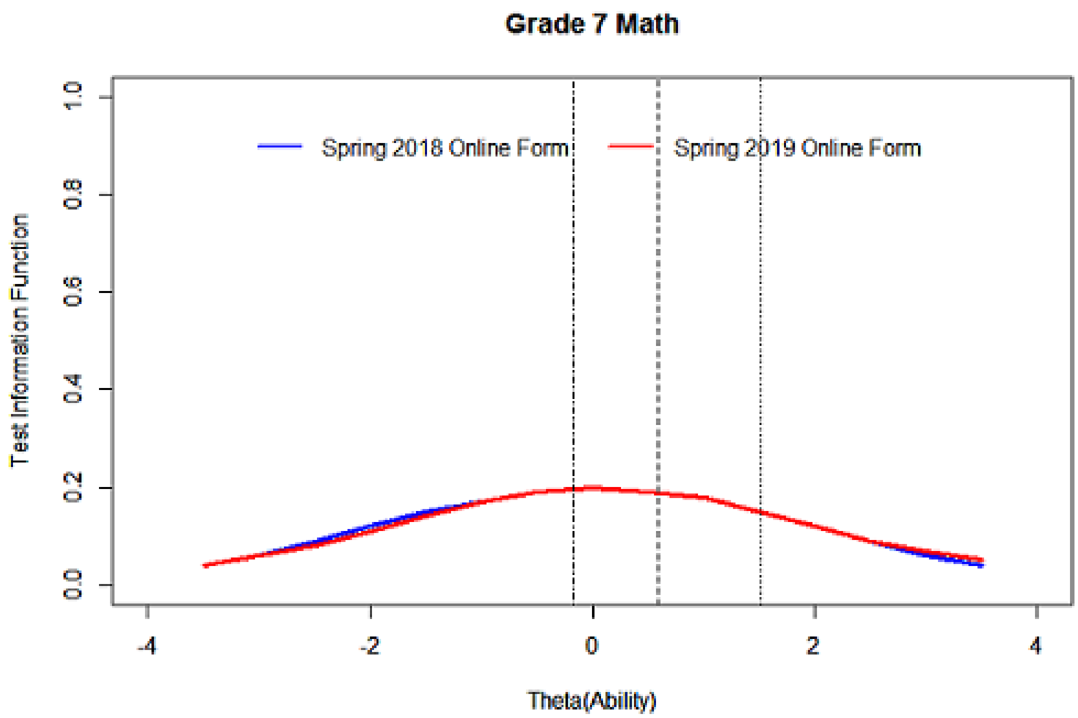


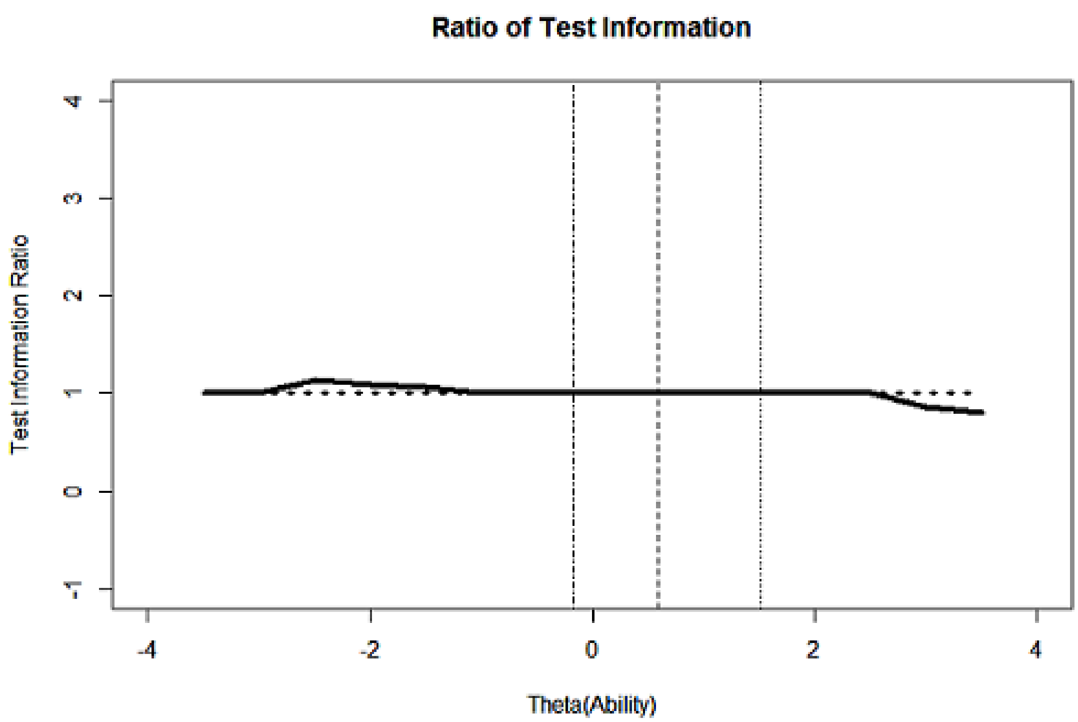
Grade 6 Math

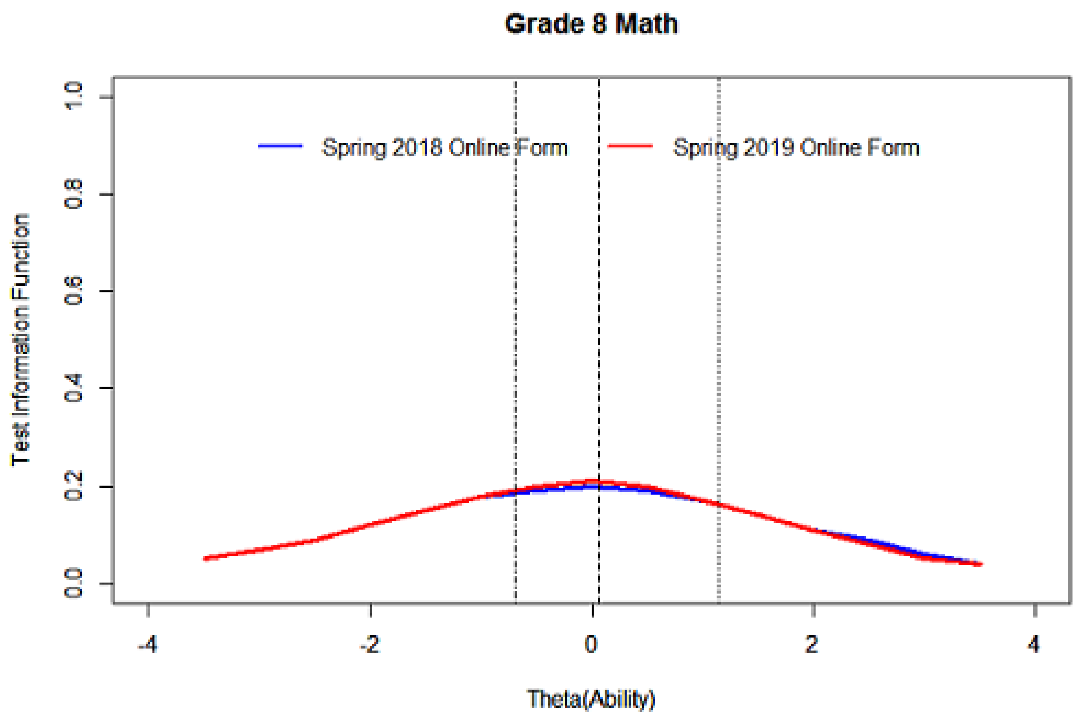


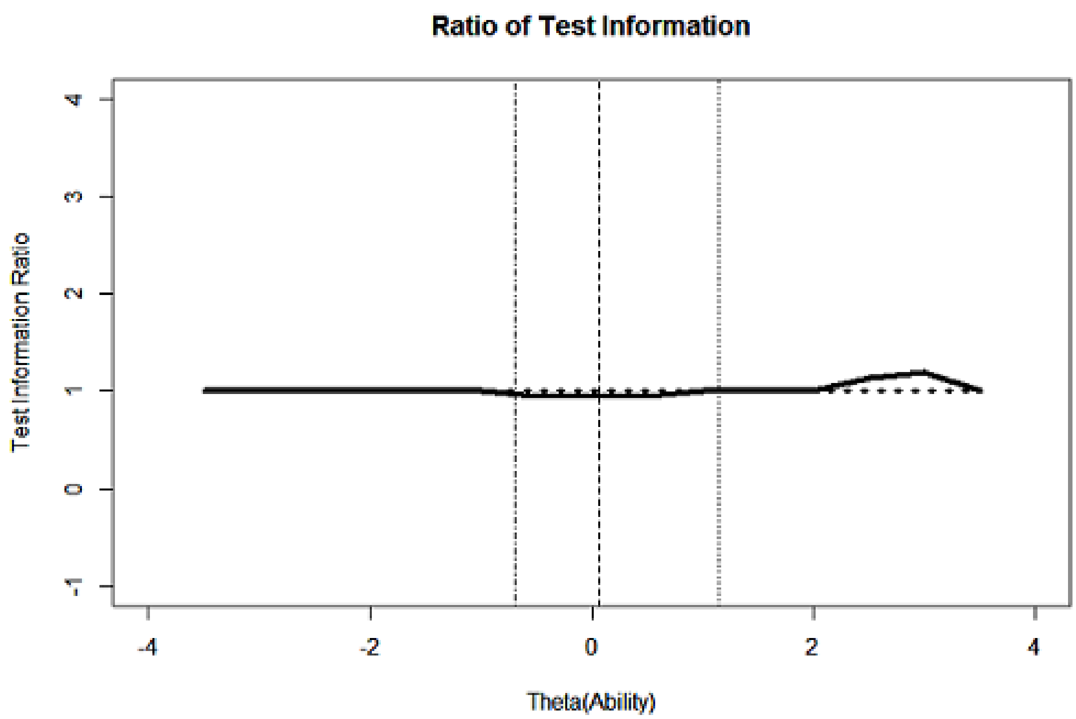
Ratio of Test Information

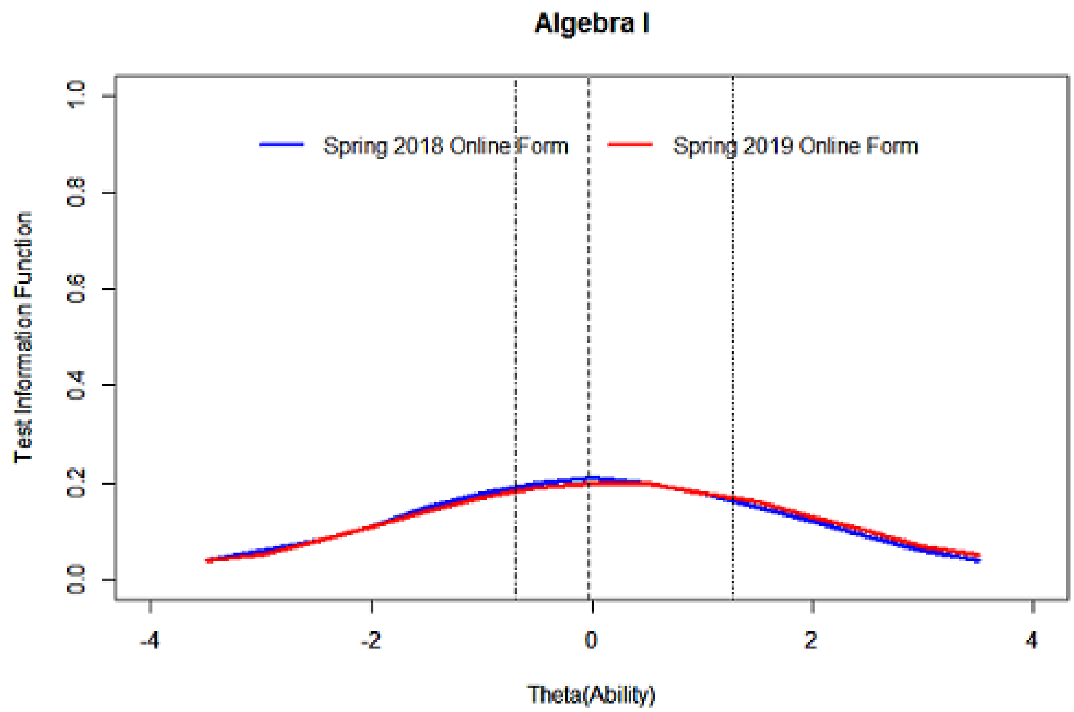


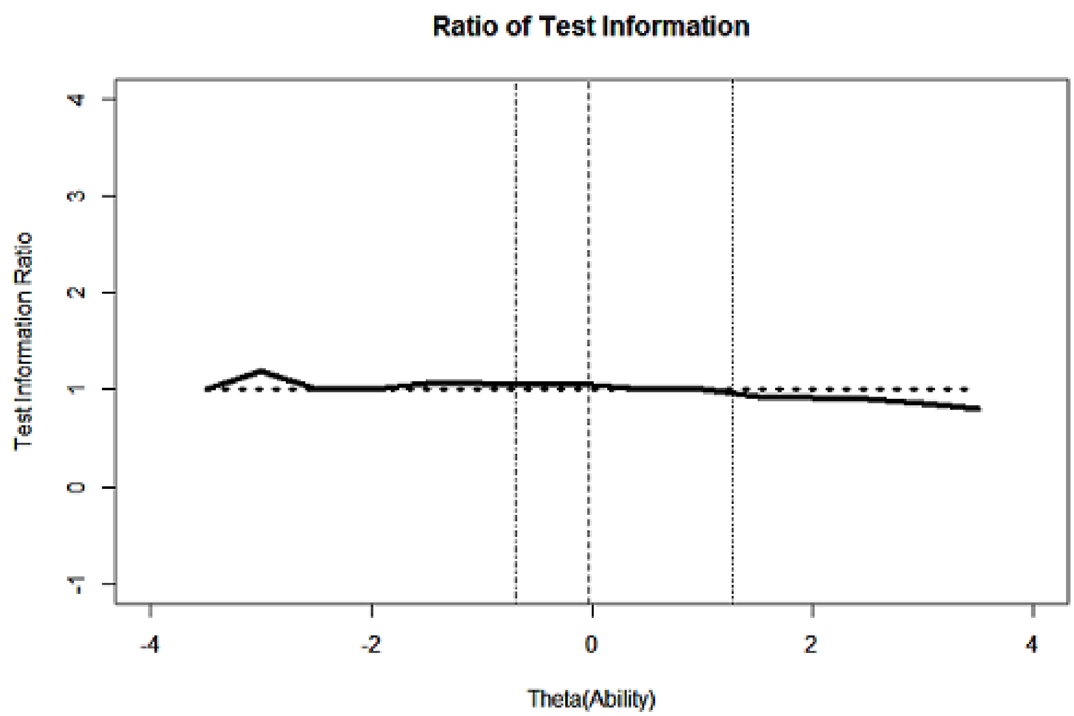

Grade 7 Math
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)

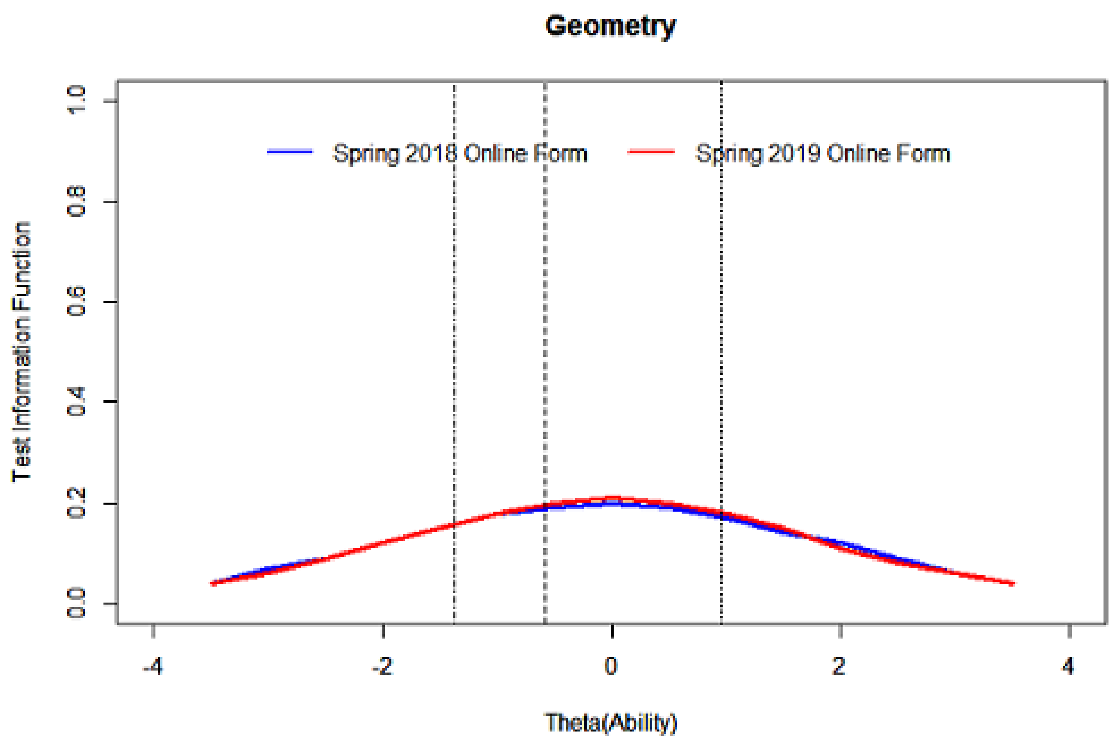

Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

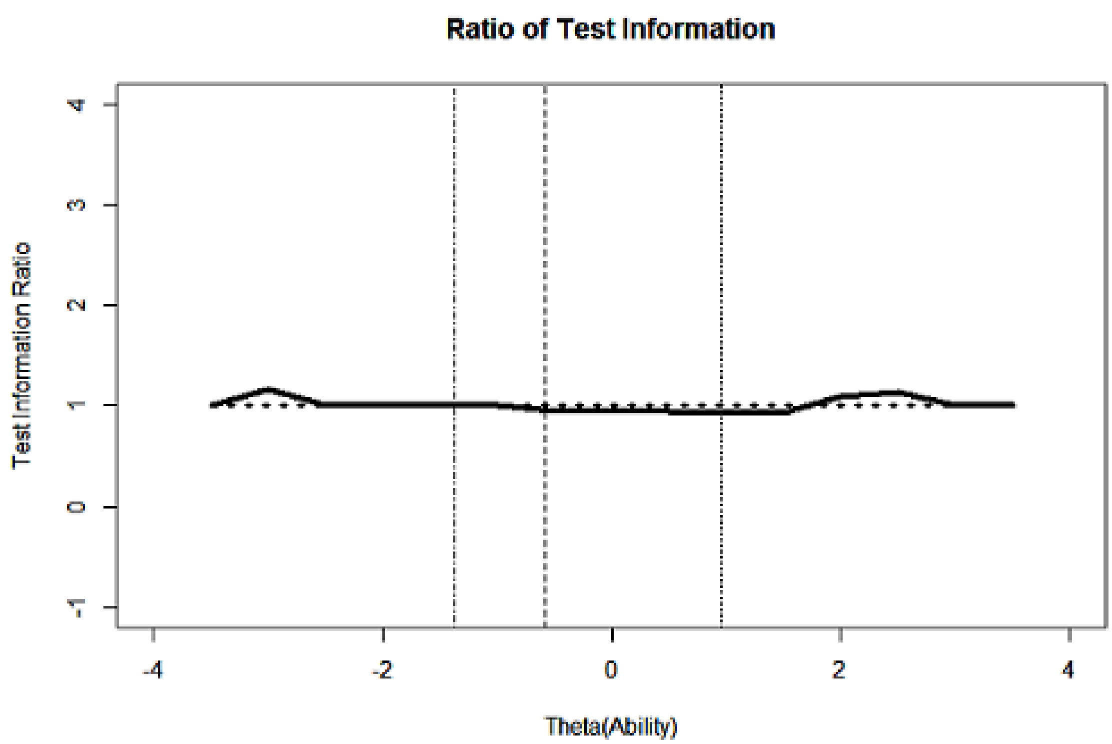

Grade 8 Math
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

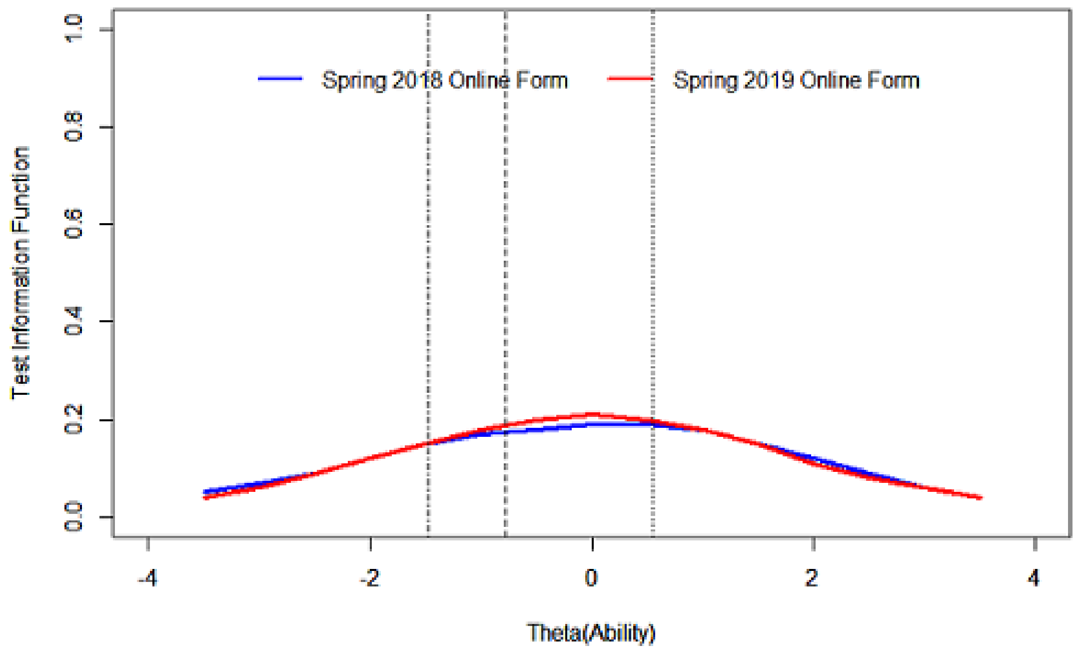

Algebra I
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
Theta(Ability)


Ratio of Test Information
Test Information Ratio
Theta(Ability)


Geometry
Spring 2018 Online Form
Spring 2019 Online Form
Test Information Function
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Theta(Ability)


Ratio of Test Information
Test Information Ratio
-1
0
1
2
3
4
-4
-2
0
2
4
Theta(Ability)

Algebra II



Ratio of Test Information