AIR®

American Institutes for Research®

Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Mathematics

2014-2015 School Year

March 3, 2016

ARIZONA STATEWIDE ASSESSMENTS

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

ELA GRADES 3-11

MATHEMATICS GRADES 3-8, ALGEBRA I, GEOMETRY, AND ALGEBRA II

2014-2015 ANNUAL TECHNICAL REPORT

MARCH 3, 2016

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

TABLE OF CONTENTS

APPENDICES

1. EXECUTIVE SUMMARY: VALIDITY OF AZMERIT TEST SCORE INTERPRETATIONS

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support 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 may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

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

In 2010, Arizona adopted new academic content standards in English language arts (ELA) and math. The Arizona College and Career Ready Standards are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate. In spring 2015, the Arizona Department of Education (ADE) administered Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) to assess proficiency on the new Arizona College and Career Ready Standards for the first time. The AzMERIT measures English language arts (ELA) and mathematics in grades 3-8 and following completion of high school coursework in ELA Grade 9, ELA Grade 10, ELA Grade 11, Algebra I, Geometry, and Algebra II.

Because directly measuring student achievement against of each benchmark in the Arizona College and Career Ready 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 ACCRS. To ensure that each student is assessed on the intended breadth and depth of the ACCRS, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark. Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzMERIT ELA and math test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels. The ELA and mathematics blueprints are also provided as an attachment in Appendix A.

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. 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
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

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

In addition, the AzMERIT test delivery system provides a range of accessibility tools and accommodations for reducing construct irrelevant barriers to accessing test content for virtually all students. The range of accommodations, provided in the online testing environment, far exceed the typical accommodations made available in paper-based test administrations. Paper test forms are available as an accommodation for students testing in online schools should the accommodations provided online not be sufficient to remove barriers to accessing test content. These include both large print and Braille forms, which are also available, for students who need them, in schools administering AzMERIT as a paper-based assessment. Section 5.3 describes available testing tools and accommodations for students testing online and on paper.

1.1 EVIDENCE BASED ON TEST CONTENT

Because the AzMERIT are designed to measure student progress toward achievement of the ACCRS the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the academic standards.

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 state department of education, test developers, and educator and stakeholder committees. Because most of the items used to construct the spring 2015 AzMERIT test forms were drawn from Utah's Student Assessment of Growth and Excellence (SAGE) item banks, item development proceeded from the Utah Core Standards (UCS), and the review process described below was with respect to those standards. However, prior to form development activities for AzMERIT, these items were subjected to an additional round of reviews by content experts and educators in Arizona to ensure the alignment of item content to the ACCRS and the appropriateness of test content for Arizona students.

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. Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical.

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

Because most of the items used operationally on the 2015 AzMERIT were initially developed for Utah's SAGE assessments, we begin with a description of the item review procedures used to ensure item accuracy and alignment with the intended Utah Core Standard. Following a standard item review process, item reviews proceeded initially through a series of internal reviews before items were eligible for review by USOE content experts. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passed through four internal review steps before it was eligible for review by USOE. Those steps include

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

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

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

- For machine-scored constructed-response (MSCR) items, that item responses yield the intended score points based on the rubric, and
- For human scored constructed response items, the scoring rubric clearly explains what characterizes responses at each possible level of achievement.

In addition, rubric-scored items, both machine-scored and human-scored, are validated following field test administration. Machine-scored items go through a rubric validation process wherein samples of student responses are reviewed, along with resulting scores, to ensure that rubrics are enacted as intended. This process is described in Section 11.1. Human-scored items go through a rangefinding process prior to scoring where samples of item responses are used to create scorer training materials and ensure the scoring rubric is appropriate, as described in Section 11.2.

Based on their 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 were sent to USOE for their review. At this stage, items may have been further revised based on any edits or changes requested by USOE, or rejected outright. Items passing through the USOE review level then had to pass through two stakeholder reviews in which committees of Utah educators and stakeholders reviewed 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 SAGE item pools were initially reviewed by

- A Utah content advisory committee, which checked to ensure that each item was
  - aligned to the UCS content standards
  - appropriate for the grade level
  - accurate
  - presented online in a way that is clear and appropriate
- A Utah fairness and sensitivity committee, which checked to ensure that each item and any associated stimulus materials were 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.
- A Utah community panel also reviewed all test items for appropriateness of test content.

Items successfully passing through this committee review process were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is 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 have to pass a three-stage review to be included in the final item pool from which operational forms are created. In the first stage of this review, a team of psychometricians reviews 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.

USOE then reconvened their content review and fairness and sensitivity committees to re-evaluate flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, the content review and fairness and sensitivity committees could recommend that flagged items be rejected or deem the item eligible for inclusion in operational test administrations.

1.1.1 ARIZONA REVIEW

Although the standards frameworks embodied by the ACCRS and UCS share much in common, they are not the same. It is important to emphasize Arizona adopted a standards framework independently from that adopted by Utah, and the AzMERIT test blueprint, which specifies how achievement of the ACCRS is to be measured, was also developed independently of the SAGE blueprints and in fact there are marked differences. Thus, the SAGE item bank was not developed to measure achievement of the ACCRS as implemented in the AzMERIT test design. However, because SAGE was designed as a system of adaptive assessments, the item banks associated with each assessment are relatively large, and most AzMERIT blueprint elements could be met by items in the SAGE banks. There were, however, some AzMERIT specifications that could not be met by items in the SAGE banks, and in those cases items were pulled either from AIR's proprietary item pool or from the AIMS item pools. Very few such items were needed to construct the AzMERIT test forms.

Moreover, the SAGE items were not developed for administration in Arizona and had therefore not been reviewed by Arizona educators and stakeholders. ADE therefore instituted a review process to ensure that each item eligible for inclusion in an AzMERIT test form had been reviewed both by the Department and by Arizona educators.

To perform this review, AIR selected the best potential items from this pool of eligible items and AIR and ADE performed an initial review to check each item for accuracy, fairness, and alignment to Arizona's College and Career Ready Standards. Through this process, the pool of form eligible items was refined. The final set of AzMERIT form eligible items was then presented to committees of Arizona educators for their review. Review committees were charged with confirming each item's accuracy, fairness and alignment to ACCRS. These committees were made up of educators from all over the state. Exhibit 4.5.1 lists the districts that were represented at these meetings.

During these review meetings, Arizona educators had the opportunity to evaluate each item that would be considered eligible for constructing the spring 2015 AzMERIT test forms, and they could move to either approve or reject the inclusion each item in the AzMERIT eligible pool. Only those items approved by the Arizona review committees were eligible for inclusion in AzMERIT test forms.

Arizona educators reviewed a total of 870 operational items at these meetings, 434 in ELA and 436 in math. In ELA, the committees rejected one passage and 8 additional items across grades, all of which were replaced with items and passages that the committees did approve. In math, the committees rejected 24 total items across all grades. These items were also replaced during the meetings, and the replacement items were reviewed and approved by the committees.

1.2 EVIDENCE FOR INTERPRETATION OF PERFORMANCE

Alignment of test content to the Arizona College and Career Ready Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the ACCRS. However, the interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the

ACCRS. AzMERIT test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the ACCRS. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the ACCRS, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standard for the AzMERIT assessments are therefore central to the validity of test score interpretations.

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

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. For each assessment, panelists were provided the approximate location of performance standards for other important assessment systems. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

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

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

Based on the adopted performance standards, Exhibit 1.2.1 shows the estimated percentage of students meeting the AzMERIT proficient standard for each assessment in spring 2015. Exhibit 1.2.1 also shows the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. Exhibit 1.2.1 also presents the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2014 field test administration. As Exhibit 1.2.1 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.

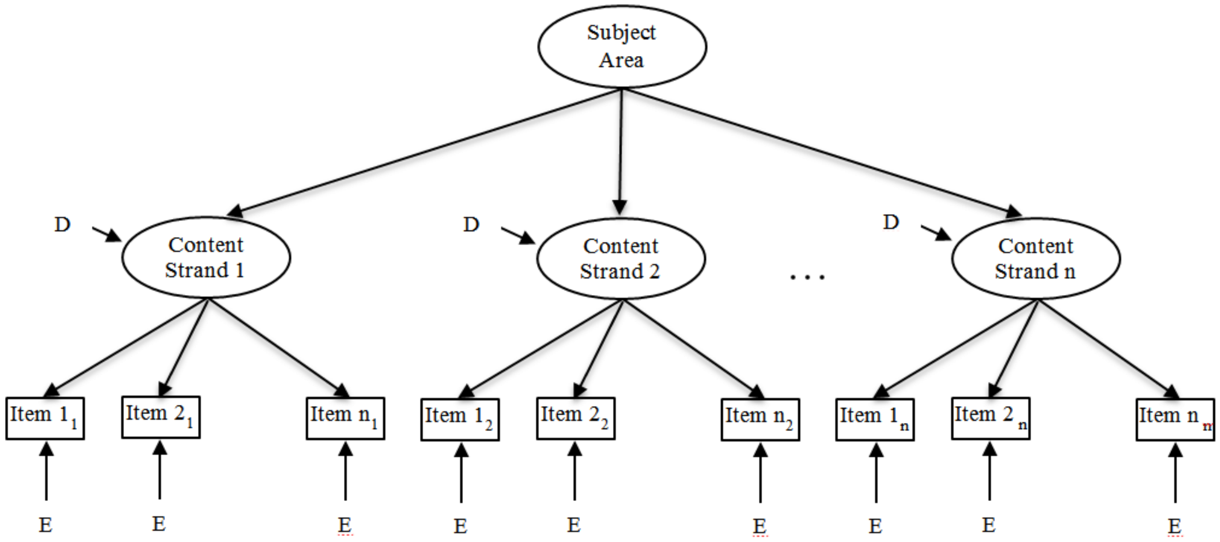
Exhibit 1.2.1 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Test	Percent 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%

1.3 EVIDENCE BASED ON INTERNAL STRUCTURE

Arizona’s 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 1.3.1. 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 1.3.1 Second-Order Structural Model for AzMERIT Assessments



Confirmatory factor analysis was used to evaluate the fit of this structural model to student response data from the AzMERIT test administrations. For each of test forms administered in spring 2015, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ2 statistic, with good model fit indicated by a non-significant χ2 statistic. The χ2 statistic is sensitive to sample size, however, so even well-fitting models will demonstrate highly significant χ2 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), the Root Mean Square of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR) were also used to evaluate model fit.

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

Exhibit 1.3.2 Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Index	Indication of Good Fit
CFI	≥ .95
TLI	≥ .95
RMSEA	≤ .05

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

1.4 ELA RESULTS

1.4.1 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 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 1.4.1.1. All of the statistics indicate the general achievement factor model fit 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 1.4.1.1 Goodness-of-Fit for the AzMERIT ELA First-Order Model

First-Order Models			
Grade	CFI	TLI	RMSEA
3	0.934	0.931	0.047
4	0.949	0.946	0.033
5	0.966	0.964	0.039
6	0.955	0.953	0.043
7	0.974	0.972	0.037
8	0.964	0.963	0.048
9	0.924	0.921	0.039
10	0.948	0.945	0.042
11	0.928	0.925	0.034

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.4.1.2. All of the statistics indicate 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 equal to or greater than .95, with RMSEA values well below the .05 threshold used to indicate good fit.

Exhibit 1.4.1.2 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Second-Order Models			
Grade	CFI	TLI	RMSEA
3	0.958	0.956	0.038
4	0.970	0.969	0.025
5	0.980	0.979	0.030
6	0.973	0.972	0.033
7	0.983	0.982	0.029
8	0.980	0.979	0.036
9	0.962	0.960	0.028
10	0.972	0.970	0.031
11	0.949	0.947	0.029

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.4.1.3. We note that model fit for first-order model of general achievement are also very high and provide 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 general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.4.1.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	13560.7	3	$p < .001$
4	8460.9	3	$p < .001$
5	10944.7	3	$p < .001$
6	12019.8	3	$p < .001$
7	8848.6	3	$p < .001$
8	15590.1	3	$p < .001$
9	8896.6	3	$p < .001$
10	9084.7	3	$p < .001$
11	4412.8	3	$p < .001$

1.4.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.4.2.1. Across all grades, results indicate the second-order models posited by the AzMERIT assessments fit the data well. The CFI and TLI values were all .97 to .99, RMSEA values are all .03 or lower. SRMR values between .02 and .04, well below the values used to indicate good fit.

Exhibit 1.4.2.1 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

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

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.4.2.2. The chi-square difference test shows that across grade levels, the DOK-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.4.2.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	21402.6	4	*p* < .001
4	12053.6	4	*p* < .001
5	17102.9	4	*p* < .001
6	18192.1	4	*p* < .001
7	16351.4	4	*p* < .001
8	25454.7	4	*p* < .001
9	14989.3	4	*p* < .001
10	14920.9	4	*p* < .001
11	8075.1	4	*p* < .001

1.4.3 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 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 1.4.3.1. All of the statistics indicate the general achievement factor model fit 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 1.4.3.1 Goodness-of-Fit for the AzMERIT Mathematics First-Order Model

First-Order Models			
grade	CFI	TLI	RMSEA
3	0.975	0.973	0.027
4	0.976	0.975	0.024
5	0.976	0.975	0.026
6	0.975	0.973	0.023
7	0.982	0.981	0.021
8	0.969	0.967	0.026
Algebra I	0.976	0.975	0.023
Algebra II	0.973	0.971	0.021
Geometry	0.986	0.985	0.018

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 1.4.3.2. The models show very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their.05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.4.3.2 Goodness-of-Fit for the AzMERIT Mathematics Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.979	0.978	0.024
4	0.978	0.977	0.024
5	0.978	0.977	0.025
6	0.976	0.975	0.023
7	0.983	0.982	0.021
8	0.970	0.969	0.026
Algebra I	0.978	0.977	0.022
Algebra II	0.974	0.972	0.020
Geometry	0.987	0.986	0.017

The results of the comparison between the second-order, strand-based model and the first-order, general achievement model are presented in Exhibit 1.4.3.3. Again, model fit for the general achievement first-order model is very high, providing evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether knowledge of the depth of knowledge level of items provides information beyond that provided by the more general model. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.4.3.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	3225.0	3	*p* < .001
4	1326.3	3	*p* < .001
5	1427.0	3	*p* < .001
6	1036.2	4	*p* < .001
7	559.8	4	*p* < .001
8	1039.3	4	*p* < .001
Algebra I	750.9	3	*p* < .001
Algebra II	246.5	3	*p* < .001
Geometry	269.7	4	*p* < .001

1.4.4 MATHEMATICS DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 1.4.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their.05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.4.4.1 Goodness-of-Fit for the AzMERIT Mathematics Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.99	0.99	0.02
Geometry	0.97	0.97	0.02

The results of the comparison between the second-order, DOK-based model and the first-order, general achievement model are presented in Exhibit 1.4.4.2. 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 1.4.4.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	331.4	3	$p < .001$
4	309.5	3	$p < .001$
5	14.9	3	$p < .001$
6	14.5	3	$p < .001$
7	236.6	3	$p < .001$
8	79.2	3	$p < .001$
Algebra I	20.1	3	$p < .001$
Algebra II	26.4	3	$p < .001$
Geometry	20.9	3	$p < .001$

1.5 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.

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

Exhibit 1.5.1 shows the correlations between student test scores on the 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 1.5.1 Correlations between AzMERIT and Locally Administered NWEA Test Scores

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

Exhibit 1.5.2 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 1.5.2 Correlations between AzMERIT and Locally Administered NWEA Test Scores

	AzMERIT ELA/NWEA Math		AzMERIT Math/NWEA Reading	
Grade	**Sample Size**	**Correlation**	**Sample Size**	**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, 2016). These researchers reported the mean correlations between 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 grade levels 3 through 8. These results likewise show good convergence between AzMERIT and other locally administered assessments purporting to measure the same constructs.

1.6 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

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

2. BACKGROUND OF ARIZONA STATEWIDE ASSESSMENTS

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

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

The AzMERIT was first administered in spring 2015, assessing proficiency in ELA in grades 3 through 11, mathematics in grades 3 through 8, and following completion of Algebra I, Geometry, and Algebra II (or similar) coursework. Following the initial administration, the AzMERIT in grades 3 through 8 will be administered in the spring of each academic year; tests assessing high school end-of-course tests will be administered in the fall, spring, and summer of each academic year.

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

2.1 DEVELOPMENT OF ARIZONA COLLEGE AND CAREER READY STANDARDS (ACCRS)

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

2.2 AZMERIT TEST DESIGN

The AzMERIT is a series of fixed form assessments that are intended to be administered online, although the assessment is offered as a dual mode, online and paper, assessment to accommodate schools that are not yet ready to transition to the online testing environment. A common operational base form is administered to all students within a given test grade and subject. Each assessment is comprised of two to three discrete test sessions. The AzMERIT operational item pools include a variety of selected response, machine-scored constructed responses (MSCR), some hand-scored constructed-response items in the paper mathematics forms where MSCR items could not readily be rendered for paper test administration, as well as hand-scored essay responses in ELA assessments.

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

Each ELA assessment included one writing essay prompt that required an extended essay response; these responses were scored by human raters on three distinct scoring dimensions or rubrics: Statement of Purpose/Focus and Organization, Evidence/Elaboration, and Conventions/Editing. In addition, hand-scoring was required for a subset of mathematics items administered on paper, generally equation items, where it was not possible to represent the item on paper in a way that allowed machine-scoring.

3. SUMMARY OF 2015 OPERATIONAL TEST ADMINSTRATION

The following tests were administered in spring 2015:

- ELA (reading and writing) in grades 3 through 11
- Mathematics in grades 3 through 8, and following completion of Algebra I, Geometry, and Algebra II, or similar, coursework

Online administration of the AzMERIT occurred from March 30 through May 8, 2015. The paper version of the AzMERIT was administered between April 13 and April 24, 2015.

Item parameters for the assessments were calibrated following the spring administration, and vertical scales were established for reporting both the ELA and mathematics test scores. A series of linking studies was performed to allow comparison of performance on the AzMERIT with other state and national scales, and a mode comparability study was completed to equate test scores across test administration modes. These studies were completed prior to establishing performance standards in summer 2015, and subsequent scoring and reporting of AzMERIT results.

This section summarizes the operational test results for the spring 2015 administration of the AzMERIT. Detailed descriptions of procedures for item and test development, test administration, scaling, equating, and scoring are presented in subsequent sections.

3.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and mathematics is presented in Exhibits 3.1.1 and 3.1.2 by assessment and subgroup. We note that some students were required to participate in both an end-of-course and a grade level assessment, especially in grade 8 where more advanced students are enrolled in Algebra I courses. The tables in Appendix C show the demographic composition of test takers by mode of test administration.

Exhibit 3.1.1 Number of Students Participating in ELA Assessments, by Test

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
All Students	86432	85295	84937	84547	82737	82929	77220	69822	57574
Female	42605	42054	41338	41608	40782	40884	38134	34507	28343
Male	43827	43241	43599	42939	41955	42045	39086	35315	29231
Unknown	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
African American	4350	4244	4456	4421	4465	4458	3961	3506	2878
Asian	2334	2408	2510	2376	2301	2365	2268	2076	1830
Native Hawaiian/Pacific Islander	282	220	213	215	210	234	169	209	158
Hispanic/Latino	38578	37560	37125	36606	35946	35981	32579	29225	22961
American Indian or Alaskan	4872	4974	4840	4566	4540	4355	3960	3667	3031
White	32947	33241	33394	34116	33162	33426	31554	28749	23776
Multiple Ethnicities	2433	2098	1908	1757	1668	1650	1489	1333	1009

Exhibit 3.1.2 Number of Students Participating in Mathematics Assessments, by Test Name

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
All Students	86880	85760	85206	84814	83509	83306	81993	68136	57259
Female	42766	42245	41431	41705	41095	41007	39995	33785	29044
Male	44114	43514	43775	43109	42414	42298	41997	34350	28214
Unknown	N/A	1	N/A	N/A	N/A	1	1	1	1
African American	4373	4281	4471	4435	4509	4478	4270	3357	2818
Asian	2344	2413	2517	2383	2320	2364	2283	2163	1963
Native Hawaiian/Pacific Islander	284	220	214	216	209	235	201	177	160
Hispanic/Latino	38754	37737	37237	36694	36374	36149	35804	28261	22543
American Indian or Alaskan	4914	5009	4874	4593	4606	4439	4382	3725	2939
White	33101	33412	33466	34217	33321	33507	32176	28195	24627
Multiple Ethnicities	2448	2107	1919	1767	1682	1651	1433	1276	1054

3.2 CLASSICAL ITEM ANALYSIS

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

The item discrimination index indicates the extent to which each item differentiated between those examinees who possessed 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 multiple-choice items is calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we compute the mean total number correct for student scoring within each of the possible score categories. 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.

Exhibit 3.2.1 presents the average item *p*-values/ proportion of total points and adjusted point biserial/polyserial correlations of the spring 2015 online administration of AzMERIT. As indicated in Exhibit 3.2.1, items on the ELA assessments are, on average, easier than items on the mathematics assessments. While mean difficulty of ELA items is relatively consistent across grade level assessments, average difficulty of mathematics items increases across grade level and course assessments. The proportion of students responding to test items in the end of course assessments in mathematics was relatively quite low. Mean biserial correlations for the grade level and end of course assessments are reasonably high and consistent across assessments.

Exhibit 3.2.1 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	0.62	0.29	0.43	0.11
4	0.64	0.22	0.42	0.10
5	0.62	0.29	0.41	0.09
6	0.63	0.25	0.42	0.11
7	0.64	0.29	0.41	0.11
8	0.66	0.34	0.43	0.09
9	0.58	0.24	0.39	0.10
10	0.62	0.26	0.40	0.10
11	0.56	0.24	0.39	0.10
		Mathematics		
3	0.61	0.16	0.47	0.07
4	0.56	0.18	0.45	0.07
5	0.50	0.16	0.47	0.09
6	0.48	0.18	0.44	0.08
7	0.47	0.19	0.45	0.09
8	0.45	0.25	0.45	0.08
Algebra I	0.44	0.16	0.44	0.09
Geometry	0.33	0.19	0.43	0.12
Algebra II	0.28	0.20	0.41	0.10

3.3 ITEM RESPONSE THEORY ANALYSIS

Because the AzMERIT was first administered in 2015, item banks were calibrated following the close of the test window. In addition, a vertical linking study was conducted as part of the spring 2015 administration to produce a vertical scale for scoring and reporting. The procedures for calibration, equating, and scaling of tests is described in the Scaling and Equating section.

The tables in Appendix D provide Rasch and Masters' partial credit model item parameter estimates for the spring 2015 operational test items. Since AzMERIT is an online assessment system, bank item parameters were estimated based only on online responses to test items. Exhibit 3.3.1 presents the mean and standard deviation of the Rasch item parameters by item type for each test for items administered online. Item types include traditional four-option multiple choice (MC) items, technology-enhanced (TE) selected response items which may require students to select one or more options, and machine-scored constructed response (MSCR) items for which students' constructed responses are scored electronically using explicit rubrics. In addition, the average Rasch difficulty is presented for each scoring dimension of the writing prompt administered at each grade. Organization and Evidence/Elaboration dimensions are scored on a 1-4 point rubric, and Conventions/Editing are scored on a 0-2 point rubric. As illustrated in Exhibit 3.3.1, selected-response items are, on average, less difficult than the constructed-response item types. The MSCR item types allow test developers to assess more complex knowledge and skills than can be readily assessed via selected-response items.

Exhibit 3.3.1 Rasch Summary Statistics by Item Type for Items Administered Online

Grade/Course	MC			TE Selected Response			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
ELA												
3	26	-0.18	0.83	9	0.15	1.06	6	0.23	1.21	1.21	1.32	-0.53
4	25	-0.33	0.88	10	-0.07	0.65	6	0.94	1.26	2.02	2.16	-0.79
5	25	-0.12	0.65	10	-0.02	0.93	6	0.05	1.72	1.80	1.81	-0.67
6	29	-0.13	0.77	11	0.06	1.15	1	0.86	NA	1.23	1.49	-0.56
7	22	-0.21	0.71	13	0.19	0.82	6	0.14	0.74	1.29	1.31	-1.21
8	25	0.02	0.69	14	-0.18	1.07	2	0.96	0.20	0.45	0.97	-1.30
9	27	-0.12	0.58	12	0.06	0.80	4	0.06	0.55	1.41	2.04	-1.12
10	29	-0.08	0.53	12	0.20	0.66	2	0.20	1.09	0.25	0.45	-1.14
11	30	-0.17	0.7	10	0.06	0.76	4	1.28	0.94	0.85	1.37	-1.39
Mathematics												
3	39	-0.13	0.88	0	NA	NA	6	0.83	1.02	--	--	--
4	40	-0.22	0.88	0	NA	NA	5	1.78	0.81	--	--	--
5	36	-0.23	0.88	0	NA	NA	9	0.90	0.60	--	--	--
6	40	-0.32	0.75	1	1.68	NA	6	1.87	1.00	--	--	--
7	32	-0.57	0.72	1	1.64	NA	14	1.18	1.03	--	--	--
8	27	-0.80	1.18	1	1.90	NA	19	1.03	1.03	--	--	--
Algebra I	34	-0.48	0.49	0	NA	NA	13	1.25	0.92	--	--	--
Geometry	22	-0.97	0.75	2	1.75	0.89	23	0.78	1.10	--	--	--
Algebra II	20	-1.13	0.94	4	0.96	0.98	23	0.82	1.22	--	--	--

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 flagged if Infit or Outfit values are less than 0.7 or greater than 1.3. Exhibit 3.3.2 summarizes the number of online administered operational test items with Infit and Outfit statistics within the range of .7 to 1.3.

Exhibit 3.3.2 Summary of Item Fit Statistics for Items Administered Online

Grade/ Course	Infit			Outfit		
	Below .7	Between .7-1.3	Above 1.3	Below .7	Between .7-1.3	Above 1.3
ELA						
3	0	44	0	1	38	5
4	0	44	0	2	39	3
5	0	44	0	0	40	4
6	0	44	0	0	42	2
7	0	44	0	0	41	3
8	0	44	0	0	41	3
9	0	46	0	0	46	0
10	0	46	0	0	45	1
11	0	46	0	0	46	0
Mathematics						

Grade/ Course	Infit			Outfit		
	Below .7	Between .7-1.3	Above 1.3	Below .7	Between .7-1.3	Above 1.3
3	0	45	0	1	42	2
4	0	45	0	0	42	3
5	0	45	0	2	39	4
6	0	47	0	2	43	2
7	0	47	0	2	39	6
8	0	46	1	5	38	4
Algebra I	0	46	1	3	42	2
Geometry	0	43	4	9	31	7
Algebra II	0	46	1	11	33	3

3.4 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 Exhibit 3.4.1. Item calibrations for the spring 2015 AzMERIT assessments 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 3.4.1 Test Score Summary Statistics

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
ELA					
3	86432	2500.60	30.18	2395.00	2605.00
4	85295	2513.56	28.82	2400.35	2610.35
5	84937	2527.97	28.06	2419.29	2629.29
6	84547	2540.68	29.93	2431.00	2640.66
7	82737	2547.32	29.02	2438.13	2648.13
8	82929	2558.54	29.33	2448.00	2657.77
9	77220	2558.55	27.05	2453.97	2663.97
10	69822	2567.97	26.97	2458.47	2668.47
11	57574	2569.00	27.94	2464.53	2674.53
Mathematics					
3	86880	3520.93	39.46	3395.00	3605.00
4	85760	3551.93	37.63	3435.07	3644.71
5	85206	3585.07	38.32	3477.64	3687.64
6	84814	3614.88	35.76	3511.89	3721.89
7	83509	3631.61	37.40	3529.36	3739.36
8	83306	3659.91	38.01	3565.81	3775.81
Algebra I	81993	3668.81	33.36	3576.95	3786.95
Geometry	68136	3681.69	34.08	3609.46	3819.46
Algebra II	57259	3693.24	34.35	3629.35	3839.35

The percentage of students in each performance level by grade and content area, as well as the percent of students at or above Proficient are presented in Exhibit 3.4.2.

Exhibit 3.4.2 Percentage of Students in Performance Levels

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
			ELA			
3	86432	44	16	29	10	40
4	85295	42	17	36	5	41
5	84937	37	31	28	4	32
6	84547	39	25	32	4	36
7	82737	42	25	28	5	33
8	82929	40	26	27	7	34
9	77220	45	29	21	5	26
10	69822	47	21	22	9	32
11	57574	51	19	19	10	30
			Mathematics			
3	86880	28	31	28	13	41
4	85760	30	29	31	10	41
5	85206	29	31	28	12	39
6	84814	38	30	22	11	32
7	83509	48	22	18	13	30
8	83306	42	25	20	13	33
Algebra I	81993	45	23	23	9	32
Geometry	68136	43	27	24	6	30
Algebra II	57259	46	24	23	7	30

3.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 3.5.1 presents the percentage of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Hispanic/Latino, American Indian, White, and Multiple Ethnicities. Note that because there are less than 10 students in many categories, performance for Hawaiian/Alaskan Native students is omitted.

Exhibit 3.5.1 Percentage of Students At Each Performance Level by Gender and Ethnicity

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level								
		Overall	Female	Male	African American	Asian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities
						ELA				
3	Highly Proficient	10	12	9	4	23	5	2	18	14
	Proficient	29	32	27	23	42	23	15	38	35
	Part. Proficient	16	17	16	16	14	17	15	16	16

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level								
		Overall	Female	Male	African American	Asian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities
	Min Proficient	44	40	48	57	21	55	68	28	35
4	Highly Proficient	5	6	5	3	18	2	1	9	8
	Proficient	36	38	33	29	50	27	16	48	41
	Part. Proficient	17	17	16	17	13	18	15	17	18
	Min Proficient	42	38	46	51	19	53	68	27	33
5	Highly Proficient	4	5	3	1	12	1	1	7	5
	Proficient	28	31	25	19	44	19	10	40	35
	Part. Proficient	31	32	30	31	26	32	26	31	33
	Min Proficient	37	32	42	48	17	47	64	22	28
6	Highly Proficient	4	5	3	2	12	2	1	7	5
	Proficient	32	36	28	22	49	23	12	44	39
	Part. Proficient	25	26	24	27	20	26	21	25	26
	Min Proficient	39	33	45	48	19	50	66	25	30
7	Highly Proficient	5	6	4	2	15	2	1	8	6
	Proficient	28	32	25	22	45	20	10	40	31
	Part. Proficient	25	26	24	25	20	26	19	25	27
	Min Proficient	42	37	48	51	20	53	71	27	36
8	Highly Proficient	7	9	6	4	19	4	1	12	9
	Proficient	27	30	25	20	40	21	12	35	32
	Part. Proficient	26	27	24	25	21	26	22	26	26
	Min Proficient	40	35	45	52	20	49	65	27	34
9	Highly Proficient	5	6	4	3	17	2	1	8	7
	Proficient	21	25	18	16	37	15	8	30	27
	Part. Proficient	29	31	28	28	23	29	24	31	28
	Min Proficient	45	39	50	54	23	54	68	31	38
10	Highly Proficient	9	11	8	5	27	5	2	15	12
	Proficient	22	25	20	17	31	17	10	29	26
	Part. Proficient	21	23	20	20	19	21	17	22	23
	Min Proficient	47	42	52	59	24	57	72	34	39
11	Highly Proficient	10	12	9	5	29	5	1	16	13
	Proficient	19	21	17	15	26	15	8	25	22
	Part. Proficient	19	20	18	17	17	19	14	21	20
	Min Proficient	51	47	56	63	29	61	77	39	45
Mathematics										
3	Highly Proficient	13	12	13	6	35	7	4	20	16

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level								
		Overall	Female	Male	African American	Asian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities
	Proficient	28	28	28	22	36	24	17	35	32
	Part. Proficient	31	33	30	33	20	34	35	28	30
	Min Proficient	28	27	29	40	8	35	44	17	22
4	Highly Proficient	10	9	10	4	29	5	2	16	12
	Proficient	31	31	31	23	42	25	17	41	35
	Part. Proficient	29	30	28	31	18	32	30	26	30
	Min Proficient	30	30	30	42	10	38	51	17	23
5	Highly Proficient	12	11	13	4	36	6	3	19	14
	Proficient	28	29	26	19	34	23	15	35	31
	Part. Proficient	31	33	30	33	19	34	33	28	31
	Min Proficient	29	28	31	44	11	37	49	18	24
6	Highly Proficient	11	11	11	5	34	5	3	17	12
	Proficient	22	23	20	14	30	17	12	28	23
	Part. Proficient	30	31	29	29	22	31	27	30	33
	Min Proficient	38	35	40	52	14	47	59	25	32
7	Highly Proficient	13	12	14	5	38	7	3	20	14
	Proficient	18	18	18	11	24	13	9	24	19
	Part. Proficient	22	23	21	19	18	21	16	24	22
	Min Proficient	48	48	48	65	20	59	72	32	45
8	Highly Proficient	13	12	14	6	40	7	3	20	15
	Proficient	20	21	19	14	25	16	11	26	22
	Part. Proficient	25	26	24	23	18	25	22	26	24
	Min Proficient	42	40	43	57	16	51	64	28	39
Algebra I	Highly Proficient	9	9	9	4	28	5	2	14	12
	Proficient	23	25	21	15	36	18	13	30	26
	Part. Proficient	23	24	22	23	19	24	20	23	22
	Min Proficient	45	42	48	58	17	54	66	32	40
Geometry	Highly Proficient	6	5	7	2	21	2	1	9	7
	Proficient	24	25	24	15	40	17	12	33	27
	Part. Proficient	27	29	26	27	20	28	24	28	29
	Min Proficient	43	41	44	57	20	53	63	29	37
Algebra II	Highly Proficient	7	6	7	2	23	2	1	10	9
	Proficient	23	24	22	15	38	17	9	30	24
	Part. Proficient	24	26	23	22	21	25	19	26	24
	Min Proficient	46	45	47	60	18	56	70	34	42

4. ITEM DEVELOPMENT & TEST CONSTRUCTION

The AzMERIT assessments are rigorously examined in accordance to the guidelines provided in the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence based on these standards that is necessary to validate assessments for their intended purposes.

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

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 included the state department of education, test developers, and educator and stakeholder committees.

In its base year, most of the items used to construct the AzMERIT test forms were drawn from Utah's Student Assessment of Growth and Excellence (SAGE) item banks, item development proceeded from the Utah Core Standards (UCS), and the review process described below was with respect to those standards. However, prior to form development activities for AzMERIT, these items were subjected to an additional round of reviews by content experts and educators in Arizona to ensure the alignment of item content to the ACCRS and the appropriateness of test content for Arizona students. Following the base year, AzMERIT test forms will utilize items that are field-tested within Arizona and reviewed by ADE and Arizona educators throughout the item development process.

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 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 determined how student achievement of the ACCRS was evaluated, alignment of test blueprints with the content standards was critical. The ELA and mathematics blueprints are also provided as an attachment in Appendix A.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet blueprint. 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, writing genre, and so on. And 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.

In addition to a review-intensive item development process and form construction process that ensures test forms meet complex blueprint specifications, Student Achievement Partners reviewed the AzMERIT English Language Arts and Mathematics tests to determine how well they are aligned to the Common Core State Standards. This review was based on the criteria recently published in Criteria for Procuring and Evaluating High-Quality Assessments (Council of Chief State School Officers, 2014).

4.1 ITEM DEVELOPMENT PROCESS

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

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

The AzMERIT item development process is predicated on a high level of interaction between test developers at AIR and the Arizona Department of Education, 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 (described in this section) and sign-offs before it is locked for fieldtest or operational administration.

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

4.1.1 ITEM WRITING

Test development experts use item specifications to guide the item development process. These item specifications, developed by content experts at AIR, strategically guide the item development process. They are detailed documents that outline content limits and model tasks for a particular standard. Item writers use these specifications while developing items to make the best use of the available item types.

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

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

Detailed item specifications include the following:

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

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

Within each grade or course, all items are aligned according to DOK. Depth of Knowledge refers to the cognitive complexity of the item and the cognitive demands on the student. Based on work done 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 specifications document. 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 (usually 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/argumentative prompts ask students to analyze primary and secondary sources, make sound judgments, and present their opinions and 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/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.

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 students are assessed on their writing skills and not their reading abilities.

The stimulus is followed by a prompt that asks students to write a short essay about the topic. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support their main ideas. For example, the prompt might require students to describe the steps in a process or describe problems that need to be solved.

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 dictate which section they are to be assessed in.

4.1.2 MACHINE-SCORED CONSTRUCTED-RESPONSE ITEM DEVELOPMENT TOOLS

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

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

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

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

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

4.1.3 ITEM TYPES

AzMERIT includes a wide variety of item types that are designed around a broad and growing college 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 items 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.
- Word Builder, in which students respond by entering a single number or word.

- Proposition Response, in which students respond in one English language sentences or more, 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. Most MSCR items remain accessible. If accessibility is sacrificed for some population, test development staff carefully considers the measurement benefit before developing that item.

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

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

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

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

Where possible, MSCR items were rendered for administration on paper test forms, using the gridded response field in the scannable answer documents. Where equation and graphic response items could not be rendered to accept a gridded response on paper forms, responses were hand-scored. This applied to 176 operational mathematics items across tests. For other MSCR items that could not readily be rendered for paper test administration, the item was replaced by another item measuring the same content standards.

All essay response items, whether administered online or on paper, were hand scored in spring 2015.

4.2 ITEM REVIEW

This section describes the multi-step item review process that items travel through from inception, to several rounds of test developer, Department of Education, educator, and stakeholder review, to field testing and final review prior to inclusion on operational test forms. In its base year, AzMERIT test forms largely utilized items that were initially developed for Utah's SAGE assessments; thus, items administered operationally in 2015 went through the review process with Utah DOE and educator reviews, as well final approval by ADE and Arizona educators prior to administration in the state. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning ADE and Arizona educator and parent committees act as reviewers throughout the item development cycle.

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

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

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

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

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

Items passing through the internal review process are sent to the Department for their review. At this stage, items may be further revised based on any edits or changes requested by Department, or rejected outright. Items passing through the Department review level then have to pass through two stakeholder reviews in which committees of educators and stakeholders review each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. For items appearing in the spring 2015 test forms, these reviews were conducted with the Utah State Office of Education and Utah educators and stakeholders. ADE and committees of Arizona educators also reviewed all items considered for inclusion in the construction of AzMERIT test forms. However, due to time constraints in building the spring 2015 test forms, Arizona stakeholders did not review candidate items. Thus, all items considered for inclusion in the AzMERIT 2015 item pools were initially reviewed by:

- A content advisory educator committee, which checked to ensure that each item was

  - aligned to the content standards
  - appropriate for the grade level
  - accurate
  - presented online in a way that is clear and appropriate
- A fairness and sensitivity educator committee, which checks to ensure that each item and any associated stimulus materials are 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 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 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.

Content review and fairness and sensitivity committees were again convened to re-evaluate flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, the content review and fairness and sensitivity committees can recommend that flagged items be rejected or deem the item eligible for inclusion in operational test administrations.

In addition, the spring 2015 AzMERIT forms were augmented with items developed for Arizona's Instrument to Measure Standards, or AIMS assessment, as well as items developed for AIR's AIR Core item bank. AIMS item development followed a similarly rigorous review process, and is documented fully in the AIMS technical reports. AIR Core items followed a development cycle consistent with that described here for Utah's SAGE items.

4.3 FIELD TESTING

To establish a pool of items for constructing future AzMERIT test forms, newly developed test items were embedded in the spring 2015 AzMERIT test forms for field-testing. Embedding field-test items in operational assessments yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations. A number of factors that may influence item difficulty in the context of operational test administrations may be less relevant in stand-alone field-test contexts. For example, in a high-stakes test, such as high school end-of-course (EOC) exams where test performance may impact student grades, students may be motivated to expend greater effort to achieve maximum performance. Conversely, the high-stakes assessments may also be more likely to elicit anxiety in some students, thus impairing their performance on the tests. Even when assessments are low stakes for students, schools often work to convey to students the importance of statewide assessments in ways that are likely not done for independent field tests. While the impact of contextual factors may not be great, embedded field testing ensures that all aspects of the operational testing context influencing item difficulty are incorporated into the resulting item parameter estimates.

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

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

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

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

The embedded field testing algorithm actually consists of two different algorithms – one for identifying which field test items will be administered to which student (the distribution algorithm), and one for selecting the position on the test for each item administered the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field test items designated for administration to each student. This randomization ensure 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, EFT sets were embedded in paper AzMERIT test forms. Four test forms were spiraled within classrooms, reducing clustering effects. The forms were printed, spiraled, and packaged in sets of 10. Wherever the previous package ended, the next one began with the next form.

4.4 ITEM STATISTICS

Following the close of test administration 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. Analysis of field test items includes classical item statistics as well as the IRT item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subsgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by AIR and ADE psychometric and content staff for possible miskey or scoring erors. Items that pass through AIR and ADE statistical review are accepted for future operational use. Appendix E 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.

4.4.1 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 multiple-choice 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 multiple-choice items, the proportion of examinees in the sample selecting the correct answer (p-values) 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 reviews if the p-value was less than .25 or greater than .95.

The item discrimination index indicated the extent to which each item differentiated between those examinees who possessed 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 multiple-choice items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .25.

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

4.4.2 IRT STATS

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

4.4.3 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 a Fairness and Sensitivity Committee. Committee members were asked to reexamine each flagged item, using the statistics as a guide, and 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
- 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

Because of the unreliability of the DIF statistics when calculated on small samples, DIF classifications are suppressed for items where focal or modal groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992, Camilli & Shepard, 1994, Muniz, Hambleton, & Xing, 2001, Sireci & Rios, 2013).

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 that 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 (MH) chi-square 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 for the MC items; the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the CR items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed below. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Exhibit 4.4.3.1.

Exhibit 4.4.3.1 DIF Classification Rules

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

4.5 ARIZONA REVIEW OF BASE YEAR TEST ITEMS

As described, AzMERIT mainly utilized items developed for the Utah SAGE assessments, as well as a small number of operational AIMS items, and AIRCore items for forms administered in spring 2015. While items developed with SAGE, AIMS, and AIRCore followed a rigorous item development process, the item banks were not developed to measure achievement of the Arizona College and Career Ready Standards as implemented in the AzMERIT test design. Thus, an additional review process was instituted for all items selected as candidates for inclusion on AzMERIT to ensure that each item eligible for inclusion in an AzMERIT test form had been reviewed by the Department and by Arizona educators.

To perform this review, AIR selected the best potential items from this pool of eligible items and AIR and ADE performed an initial review to check each item for accuracy, fairness, and alignment to Arizona's College and Career Ready Standards. Through this process, the pool of form eligible items was refined. The final set of AzMERIT form eligible items was then presented to committees of Arizona educators for their review. Review committees were charged with confirming each item's accuracy, fairness and alignment to ACCRS. These committees were made up of educators from all over the state. Exhibit 4.5.1 lists the districts that were represented at these meetings.

During these review meetings, Arizona educators evaluated each item that would be considered eligible for constructing the spring 2015 AzMERIT test forms. Only those items approved by the Arizona review committees were eligible for inclusion in AzMERIT test forms. In addition, the educator committees advised ADE on calculator use policy. For the grade 7 and grade 8 mathematics forms where there are both calculator and non-calculator segments, the committees identified the items for which calculator use would be permitted. The high school mathematics committee advised ADE that students at this level should have access to a calculator throughout the test.

Arizona educators reviewed a total of 870 operational items at these meetings, 434 in ELA and 436 in math. In ELA, the committees rejected one passage and 8 additional items across grades, all of which were replaced with items and passages that the committees did approve. In math, the committees rejected 24 total items across all grades. These items were also replaced during the meetings, and the replacement items were reviewed and approved by the committees.

Exhibit 4.5.1 Arizona Districts Represented at the December 2014 Item Content and Fairness Review

District
Aguila Elementary SD
Apache Junction Unified SD
Avondale Elementary SD
Bicentennial Union High SD
Cartwright SD
Casa Grande Union High SD
Catalina Foothills Unified SD
Chandler Unified SD
Deer Valley Unified SD
Dysart Unified SD
Florence Unified SD
Fowler Elementary SD
Gilbert Public Schools
Glendale Elementary SD
Glendale Union High SD
Higley Unified SD
Isaac SD
J.O. Combs Unified SD
Kingman Unified SD
Kyrene SD
Lake Havasu Unified SD
Leading Edge Academy
Leona Group
Madison Elementary SD
Mayer Unified SD
Mesa Public Schools
Navajo Christian Prep Academy
Nogales Unified SD
Osborn SD
Paradise Schools
Paradise Valley Unified SD
Payson Unified SD
Peoria Unified SD
Phoenix Collegiate Academy
Phoenix Union High SD
Primavera Technical Learning Center
Queen Creek Unified SD
Red Mesa Unified SD
Sahuarita Unified SD
Salt River Pima Maricopa Community Schools
Sanders Unified SD
Scottsdale Unified SD
Sequoia Charter Schools
Superior Unified SD
Tanque Verde Unified SD
Tempe Elementary SD
Tucson Unified SD
Vail SD
Vision Charter School
Washington Elementary SD
Whiteriver Unified SD
Wickenburg Unified SD
Yuma Union High SD

4.6 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 met all of the criteria established by the Department in terms of content, fairness review, and data characteristics. For the 2015 base year AzMERIT forms, item pools were comprised of operational SAGE, AIMS and AIRCore items, each aligned to the ACCRS and reviewed by ADE as described previously.

4.6.1 OPERATIONAL FORM CONSTRUCTION

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

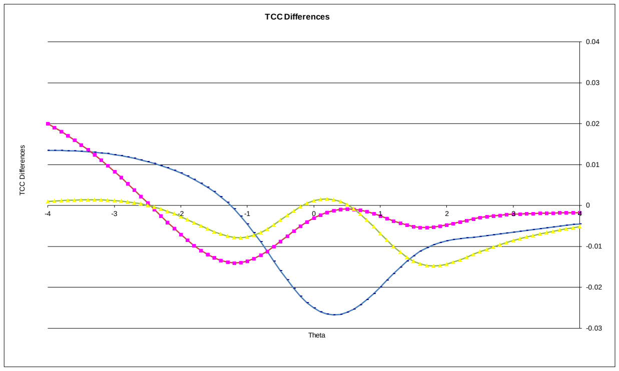
AIR’s test developers used the FormBuilder software to help construct operational forms. FormBuilder interfaces with AIR’s Item Tracking System (ITS) to extract test information and interactively create test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers built a test map. 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, the FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, the FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allow another opportunity to ensure that poorly performing items are not included in operational test forms.

The FormBuilder also plotted the distribution of item difficulties, both classical and IRT indices, to both flag extremely easy or difficult items and to ensure that the distribution of item difficulties was consistent across test forms.

As test developers built forms, FormBuilder generated TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 4.6.1.1 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 are 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, then 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.

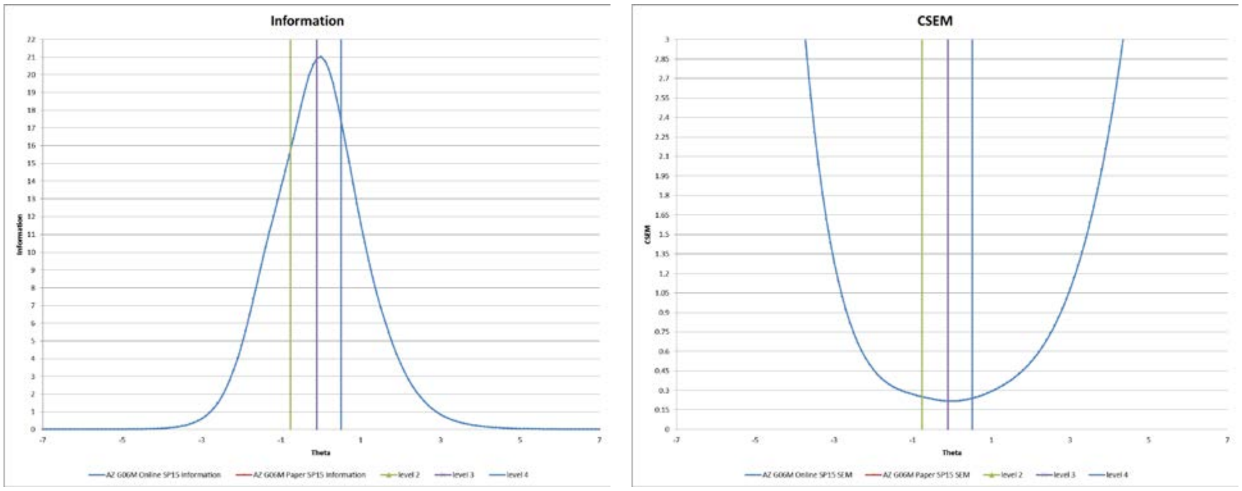
Exhibit 4.6.1.1 Test Characteristics Curve Differences



For the base year, test construction targets were based on the likely locations of performance standards as identified via linkage to Utah’s statewide assessments. For fixed form assessments, it is desirable to target test information near critical cut scores, especially the proficient cut score, and to ensure reasonable precision of measurement across the range of performance standards. Utah adopted proficient level performance standards that were consistent with the NAEP proficient level performance standard and the ACT college and career ready benchmark for high school EOC assessments, and thus were considered rigorous and expected to be consistent with performance standards recommended by Arizona educators. Subsequent to the base year, construction of AzMERIT test forms target test information to maximize precision of measurement near the Arizona performance standards. As illustrated in Exhibit 4.6.1.2, by evaluating test characteristics in reference to the likely location of important cut scores, test developers aimed to develop test forms that measured with precision in the locations where students were likely to be classified into performance levels.

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

Exhibit 4.6.1.2 Test Information and Standard Errors Relative to Performance Standards



4.6.2 ASSEMBLING TEST FORMS

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

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

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

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

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

- The items are sequentially ordered.
- Each item of the same type is presented in a consistent manner.
- The listing of the options for the multiple-choice items is consistent.
- The answer options are lettered with A, B, C, and D.
- 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 were 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 ADE for review and approval.

4.6.3 EMBEDDED FIELD TEST SLOTS

Each operational test form contains designated slots for administration of items that did not contribute to students' test scores. For the base year test forms, these included AIMS items to establish linkages to the previous assessment scale, operational test items from adjacent grades to support development of a vertical scale, field-testing of AIRCore items for construction of spring 2016 operational test forms, and PISA items to support international benchmarking for standard setting activities. Exhibit 4.6.3.1 provides a count of items administered in the EFT slots in base year forms. Because ADE wished to establish linkages to AIMS only for purposes of providing benchmarks to help guide standard setting panelists in their deliberations about the location of performance standards for the new AzMERIT assessments, the number of AIMS items included in the linking design was reduced by half from the original proposal which included a complete operational test form. To establish linkages to previous AIMS assessment, a set of AIMS linking items were also included for administration in the embedded field test slots of the spring 2015 operational test forms.

Exhibit 4.6.3.1 Composition of Online Embedded Items in Operational Test Forms

Row Labels	AIMS Linking	Field Test	Vertical Linking	PISA
ELA 3	13	78	38	--
ELA 4	9	75	41	--
ELA 5	13	77	40	--
ELA 6	14	75	38	--
ELA 7	11	80	40	--
ELA 8	10	76	38	--
ELA 9	--	104	40	--
ELA 10	9	99	42	29
ELA 11	--	113	--	--
Math 3	14	75	45	--
Math 4	15	75	45	--
Math 5	15	75	47	--
Math 6	15	75	47	--
Math 7	16	75	47	--
Math 8	16	75	62	--
Algebra I	14	100	31	19
Geometry	7	100	--	7
Algebra II	3	100	--	4

For online test administrations, AIR employed our field test engine to administer test items in the embedded slots. As described previously, the field-test algorithm randomly assigned both the field-test items and the field-test item position, ensuring that

- a random sample of students were administered each item; and
- for any given item, the students were sampled with equal probability.

AIR's field-test algorithm yields a representative, randomized sample of student responses for each item. The field-test algorithm also leads to randomization of item position and the context in which items appear. Field-testing each item in many positions and contexts rendered the resulting statistics more robust to these factors.

For paper assessments, AIR staff constructed fixed EFT blocks. Selection of items for EFT slots was designed to ensure proportional representation of AIMS and AIRCore items. Items selected for paper EFT slots were also submitted to ADE for review and approval. Inclusion of AIRCore fieldtest items on paper test forms was intended to support the establishment of an independent paper-based item bank, and subsequent construction of test forms from that pool, in the event that investigation of mode differences necessitated it. Following the spring 2015 test administration, this proved unnecessary.

5. TEST ADMINISTRATION

5.1 ELIGIBILITY

Arizona public school students in Grade 3 and above were required to participate in AzMERIT testing. Additionally, any student enrolled in a private school or Bureau of Indian Education school and any students that are home schooled 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 who are enrolled in high school level English language arts 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. For AzMERIT's spring 2015 administration, students in Grades 3 - 8 who were also enrolled in high school level mathematics or ELA courses were required to participate in both test levels.

Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them eligible for the alternate assessment for ELA and Mathematics were excluded from AzMERIT.

5.2 ADMINISTRATION PROCEDURES

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

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

Prior to the beginning of the 2014–2015 assessment, statewide District Test Coordinator training sessions were conducted to provide information regarding both the paper and computer-based test administrations. The training also provided an overview of the Test Delivery System (TDS), Online Reporting System (ORS), and Test Information Distribution Engine (TIDE). Recorded training sessions and narrated training videos were posted online. The Test Coordinator Manual and Test Administration Directions were shipped to every testing district. Additionally, test administrators were required to complete the online TA Certification Course before administering a computer-based test. District Test Coordinators and School Test Coordinators were responsible ensuring that all test administration personnel (paper and computer-based) were properly trained using the various resources prior to the start of testing.

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

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

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

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

The *AzMERIT Test Administration Manuals* 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, starting and pausing test sessions. Similar guidance is provided for the paper testing environment, including instructions for the paper testing session, monitoring sessions, verifying student information, and assigning test accommodations. Additional instructions for administering tests to students using Braille accommodated test booklets are provided in the *Supplemental Instructions for Braille* documents.

Certified 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 ensured that the School Test Coordinators in each school were appropriately trained and aware of policies and procedures, and that they were trained to use the reporting system.

School Test Coordinators 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. School Test Coordinators were primarily responsible for identifying and training Test Administrators. They also created or approved testing schedules and procedures for the school. If the school administered AzMERIT online, the School Test Coordinators worked with Technology Coordinators to ensure that the necessary secure browsers were installed and any other technical issues were resolved. During the testing window, School Test Coordinators needed to monitor testing progress, ensure that all students participate as appropriate, and handle testing incidents as necessary.

Test Administrators 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 out during testing. They were required to administer AzMERIT tests following the directions found in the *AzMERIT Test Administration Manuals*. 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 ADE.

Test Administrators who administered computer-based AzMERIT tests conducted a training test session using the AzMERIT Sample Tests.

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

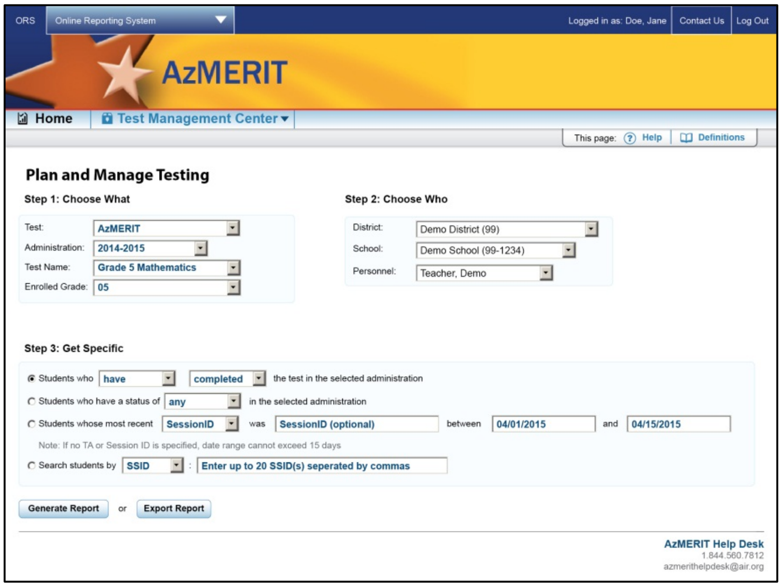
For the computer-based ELA Reading tests, headphones or earbuds were required. There were no technical specifications for headphones or earbuds. The equipment was to be checked to ensure they worked with the computer or device the students would use for the assessment prior to the first day of testing.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

For the paper-based AzMERIT tests, Test Administrators needed to ensure that students used No. 2 pencils to record their responses. School Test Coordinators provided TAs with the materials needed to administer each test session. Secure materials were delivered or picked up immediately before the beginning of each test session. During 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 to the School Test Coordinator.

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

MANAGING TESTING

To help schools manage their test schedule, allocate testing resources, and prioritize testing, the AzMERIT online reporting system, which is described in detail in Chapter 6, offered participation reports for online testers. Within the online reporting system, educators can generate up-to-the-minute reports showing students' test status. In addition, users can set testing schedules, monitor testing progress across schools, and track students' participation based on their performance on previous tests.



Plan and Manage Testing

Step 1: Choose What

Test: AzMERIT

Administration: 2014-2015

Test Name: Grade 5 Mathematics

Enrolled Grade: 05

Step 2: Choose Who

District: Demo District (99)

School: Demo School (99-1234)

Personnel: Teacher, Demo

Step 3: Get Specific

● Students who have completed the test in the selected administration

○ Students who have a status of any in the selected administration

○ Students whose most recent SessionID was SessionID (optional) between 04/01/2015 and 04/15/2015

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

○ Search students by SSID : Enter up to 20 SSID(s) seperated by commas

AzMERIT Help Desk
1.844.560.7812
azmerithelpdesk@air.org

5.3 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.

5.3.1 UNIVERSAL TEST ADMINISTRATION CONDITIONS

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

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- 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 individualized education plan 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.

5.3.2 UNIVERSAL TESTING TOOLS FOR COMPUTER BASED TESTERS

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

Exhibit 5.3.2.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 5.3.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to 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	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	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	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	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

5.3.3 SUBJECT AREA TOOLS FOR CBT AND PBT

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 5.3.3.1.

Exhibit 5.3.3.1 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): scientific calculators are acceptable EOC (entire test): 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.

5.3.4 ACCOMMODATIONS

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

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

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section. Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student's individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact ADE for guidance.

Allowable accommodations are described below.

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. There are no specific CBT tools to support these accommodations.

Exhibit 5.3.4.1 Accomodations for Students with an Injury

Accommodation	Description
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	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. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Rest/Breaks	Student may take breaks during testing sessions to rest.

ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNER (ELL) 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. This includes English Language Learner (ELL) students and students withdrawn from English language services at parent request. Reclassified Fluent English Proficient (FEP) students are monitored for two school years. These FEP Year 1 and FEP Year 2 students also may use, as appropriate, any of the universal test administration conditions and any of the following accommodations.

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

Exhibit 5.3.4.2 summarizes accommodations that may be provided for ELL and FEP students.

Exhibit 5.3.4.2 Allowable Accommodations for ELL and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.

Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Exact oral translation, in the student's native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word published, paper translation dictionary. Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 5.3.4.3, as designated in their IEP or 504 plan.

Exhibit 5.3.4.3 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. PBT – Provide a Large Print test booklet. A PBT Large Print test booklet requires Adult Transcription on a regular size paper test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.

Paper Test Booklet	CBT – Provide a regular size paper test booklet for a student at a school administering the CBT. If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Sign Test Content	Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the Math test. Signing the content of the Reading portion of the ELA test is prohibited.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

5.4 SYSTEM SECURITY

5.4.1 SECURE SYSTEM DESIGN

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

AIR's hosting provider, Rackspace, has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely. Rackspace partners with 9 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 data are recoverable at all times. Each disk array is internally redundant, with multiple disks containing each data element. Immediate recovery from failure of any individual disk is performed by accessing the redundant data on another disk. AIR maintains support and maintenance agreements through our hosting provider for all of the hardware used by our systems.

5.4.2 SYSTEM SECURITY COMPONENTS

AIR has built-in security controls in all of its data stores and transmissions. 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.

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

NETWORK SECURITY

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

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

SOFTWARE SECURITY

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

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

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

AIR psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior, blueprint match rates, and item exposure rates, and also provide cheating analysis reports. These reports are described more completely in Section 5.5 on Test Security.

5.5 TEST SECURITY

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

The test security procedures involve the following:

- Procedures to ensure security of test materials
- Procedures to investigate test irregularities

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

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 were 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) were required to be returned immediately to the School Test Coordinator and placed in locked storage. Secure materials were never to be left unsecured and were not to remain in classrooms or be taken off the school's campus overnight. Secure materials were 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 needed to be turned off.

It is unethical and shall be 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 since 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 paper-based testing include:

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

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

Throughout the testing window, test administrators were to report any test incidents (e.g., disruptive students, loss of Internet connectivity) to the School Test Coordinator immediately. A test incident could include testing that was interrupted for an extended period of time due to a local technical malfunction or severe weather. School Test

Coordinators notified District Test Coordinators of any test irregularities that were reported. District Test Coordinators were responsible for submitting requests for test invalidations to the Department of Education via AIR's Test Information Distribution Engine, or TIDE. The Department of Education 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.

5.6 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 include clear test administration policies, effective test administrator training, and tools to identify possible irregularities in test administrations.

For online administrations, quality assurance (QA) reports are generated during and after the test windows. These are geared toward detection of 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-and-pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations) as requested by the state. AIR's test delivery system (TDS) captures all of this information.

Unlike with paper assessments where data analysis must await the close of test window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout each test administration window, following the first operational administration. Following the base year, the analyses used to detect the testing anomalies can be run anytime within the testing window. Evidence evaluated included changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school. Spring 2015 reports were reviewed following scoring of operational tests.

5.6.1 CHANGES IN STUDENT PERFORMANCE

Beginning in the 2015-2016 school year, for both online and paper test takers, it will be possible to examine score changes between years using a regression model. For between-year comparisons, the scores between past and current years are compared, with the current-year score regressed on the test score from the previous year. Between-year comparisons are performed starting with the second year of the test administration.

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 studentized *t* residuals are greater than $|3|$.

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

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

where *s* = standard deviation of residuals in an aggregate unit; *n* = number of students in an aggregate unit (e.g., testing session or test administrator); 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.

If the aggregate unit size is 1–5 students, the aggregate unit was flagged if the percentage of flagged students was greater than 50%. 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.

5.6.2 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.

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

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

5.6.3 INCONSISTENT ITEM RESPONSE PATTERN (PERSON FIT)

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

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

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

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

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

where *s* = standard deviation of l_z values in an aggregate unit and *n* = number of students in an aggregate unit. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit (e.g., test session, test administrator, school).

5.6.4 RESPONSE CHANGE AND RESPONSE SIMILARITY

Response Change in Paper-based Tests

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

Response Pattern Similarity in Computer-based Tests

In fixed-form assessment environments, students may more readily copy from one another than would be possible in a CAT 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.

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 5.6.4.1 provides sample output for the response similarity analysis. Each record indicates a pair of students flagged for overly similar patterns of responding. Access to a seating chart increases the power of this approach significantly, since students with overly similar response patterns who are known to have been seated in close proximity, obviously have greater opportunity to copy their responses. This method is also useful for detecting cheating rings, where the same students are identified across multiple flagged pairs. This is evident in Exhibit 5.6.4.1, where a common group of students are each flagged in multiple comparisons.

Exhibit 5.6.4.1 Sample Roster Flagging Student Pairs with Excessively Similar Responses

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 LastName	Student1 FirstName	Student2 Barcode	Student2 LastName	Student2 FirstName
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

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 LastName	Student1 FirstName	Student2 Barcode	Student2 LastName	Student2 FirstName
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

6. REPORTING AND INTERPRETING AZMERIT SCORES

A set of score reports is provided for each administration that summarizes student performance in each grade and content area. Score reports provide data on the performance of individual students and on the aggregated performance of students at various levels—such as state, districts, schools, and teachers. The test data are based on all students who participated in the AzMERIT assessment for the 2014-2015 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 mailed directly to families, detailing student performance on overall tests and subscores. In addition, Arizona offers detailed individual and aggregate level data to educators via AIR's Online Reporting System (ORS), which provides score data for each AzMERIT test, both computer-based and paper-based. The ORS allows users to compare score data between individual students and the school, district, or overall state, and also provides information about performance on subscore categories.

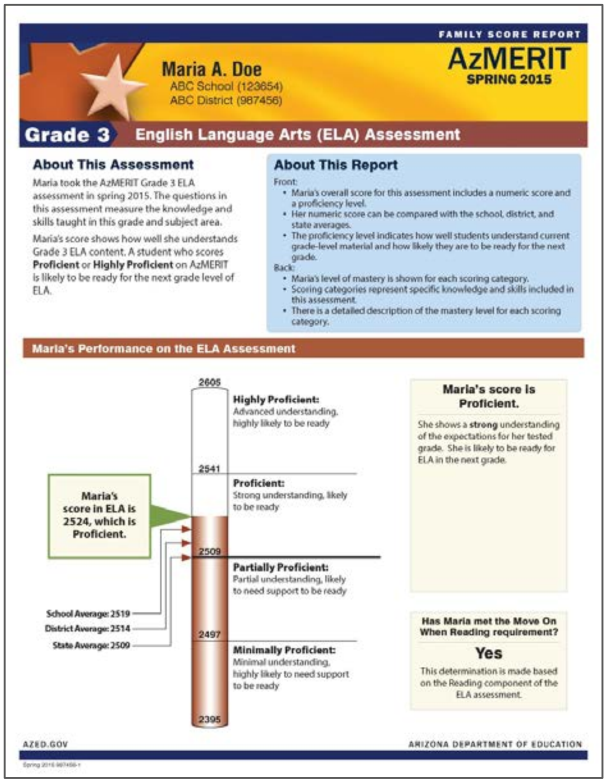
6.1 APPROPRIATE USES FOR SCORES AND REPORTS

The state provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning and classroom instruction. All reporting systems for the AzMERIT, both paper and online, are designed with stakeholders, such as teachers, parents and students, who are not technical measurement experts, in mind and ensure that test results are used in ways that lead to valid inferences about student achievement and contribute to student learning. 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 azmeritportal.org. The sections below provide additional guidance for interpreting results.

6.2 REPORTS PROVIDED

6.2.1 FAMILY REPORTS



FAMILY SCORE REPORT

AzMERIT
SPRING 2015

Maria A. Doe
ABC School (123654)
ABC District (987456)

Grade 3 English Language Arts (ELA) Assessment

About This Assessment

Maria took the AzMERIT Grade 3 ELA assessment in spring 2015. The questions in this assessment measure the knowledge and skills taught in this grade and subject area.

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

About This Report

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

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

Maria's Performance on the ELA Assessment

2605
Highly Proficient: Advanced understanding, highly likely to be ready
2541
Proficient: Strong understanding, likely to be ready
2509
Partially Proficient: Partial understanding, likely to need support to be ready
2497
Minimally Proficient: Minimal understanding, highly likely to need support to be ready
2395
Maria's score in ELA is 2524, which is Proficient.
School Average: 2519
District Average: 2514
State Average: 2509

Maria's score is Proficient.

She shows a **strong** understanding of the expectations for her tested grade. She is likely to be ready for ELA in the next grade.

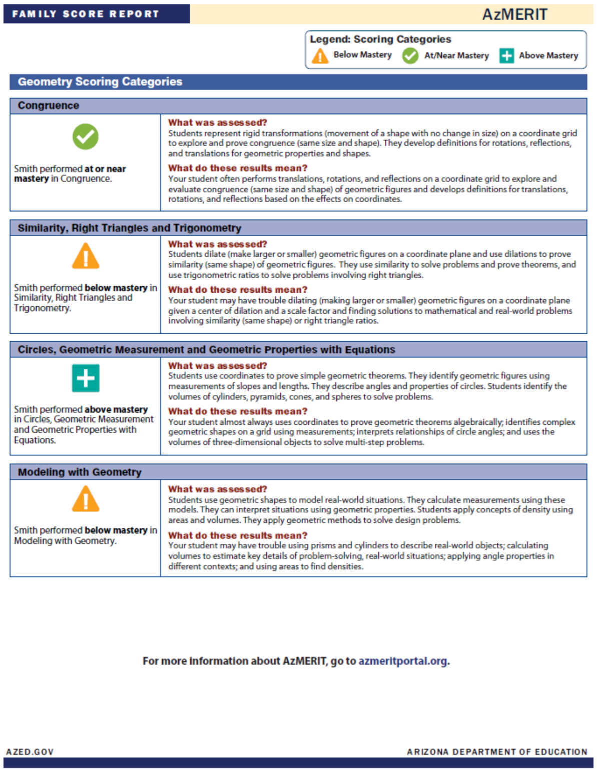
Has Maria met the Move On When Reading requirement?

Yes

This determination is made based on the Reading component of the ELA assessment.

AZED.GOV ARIZONA DEPARTMENT OF EDUCATION

Spring 2015 987456-1



FAMILY SCORE REPORT AzMERIT

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

Geometry Scoring Categories

Congruence

✓ Smith performed **at or near mastery** in Congruence.

What was assessed?
Students represent rigid transformations (movement of a shape with no change in size) on a coordinate grid to explore and prove congruence (same size and shape). They develop definitions for rotations, reflections, and translations for geometric properties and shapes.

What do these results mean?
Your student often performs translations, rotations, and reflections on a coordinate grid to explore and evaluate congruence (same size and shape) of geometric figures and develops definitions for translations, rotations, and reflections based on the effects on coordinates.

Similarity, Right Triangles and Trigonometry

⚠ Smith performed **below mastery** in Similarity, Right Triangles and Trigonometry.

What was assessed?
Students dilate (make larger or smaller) geometric figures on a coordinate plane and use dilations to prove similarity (same shape) of geometric figures. They use similarity to solve problems and prove theorems, and use trigonometric ratios to solve problems involving right triangles.

What do these results mean?
Your student may have trouble dilating (making larger or smaller) geometric figures on a coordinate plane given a center of dilation and a scale factor and finding solutions to mathematical and real-world problems involving similarity (same shape) or right triangle ratios.

Circles, Geometric Measurement and Geometric Properties with Equations

➕ Smith performed **above mastery** in Circles, Geometric Measurement and Geometric Properties with Equations.

What was assessed?
Students use coordinates to prove simple geometric theorems. They identify geometric figures using measurements of slopes and lengths. They describe angles and properties of circles. Students identify the volumes of cylinders, pyramids, cones, and spheres to solve problems.

What do these results mean?
Your student almost always uses coordinates to prove geometric theorems algebraically; identifies complex geometric shapes on a grid using measurements; interprets relationships of circle angles; and uses the volumes of three-dimensional objects to solve multi-step problems.

Modeling with Geometry

⚠ Smith performed **below mastery** in Modeling with Geometry.

What was assessed?
Students use geometric shapes to model real-world situations. They calculate measurements using these models. They can interpret situations using geometric properties. Students apply concepts of density using areas and volumes. They apply geometric methods to solve design problems.

What do these results mean?
Your student may have trouble using prisms and cylinders to describe real-world objects; calculating volumes to estimate key details of problem-solving, real-world situations; applying angle properties in different contexts; and using areas to find densities.

For more information about AzMERIT, go to azmeritportal.org.

AZED.GOV ARIZONA DEPARTMENT OF EDUCATION

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, and
  - The exclamation symbol indicates a student is performing below mastery in a scoring category
- rubric scores for the writing portion of the ELA test, including descriptions of what those rubric scores mean; and
- school, district, and state average scores for comparative purposes.

In addition, following the initial year of administration, reports will include longitudinal data that is designed to allow parents to track student achievement over time.

6.2.2 ONLINE REPORTING SYSTEM FOR EDUCATORS

AzMERIT results are reported using AIR's Online Reporting System, 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 answer 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 online score reports were produced after the completion of the standard setting process and were available to students, parents, teachers, and districts beginning in October.

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 percent 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 for the selected entity displays. For more detailed information for a subject and a grade, the user must select that subject and grade.

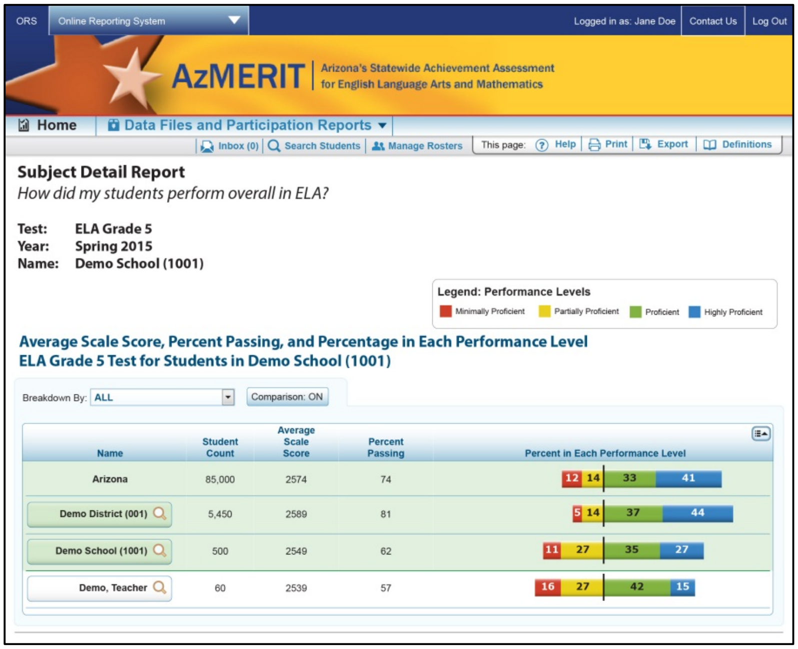
On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district the school belongs to are provided above the school summary results so that the school performance can be compared with 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 6.2.2.1 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 6.2.2.1 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 Percent Passing) across grades and subjects or course
Subject Detail	District	Average scale score, percent passing, and percent at each performance level for a district and each school within that district; ability to disaggregate data by subgroup
	School	Average scale score, percent passing, and percent at each performance level for a school and each teacher within that school; ability to disaggregate data by subgroup
	Teacher	Average scale score, percent passing, and percent 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; a 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



Subject Detail Report

How did my students perform overall in ELA?

Test: ELA Grade 5
Year: Spring 2015
Name: Demo School (1001)

Legend: Performance Levels
🟥 Minimally Proficient 🟨 Partially Proficient 🟩 Proficient 🟦 Highly Proficient

Average Scale Score, Percent Passing, and Percentage in Each Performance Level
ELA Grade 5 Test for Students in Demo School (1001)

Breakdown By: ALL
Comparison: ON

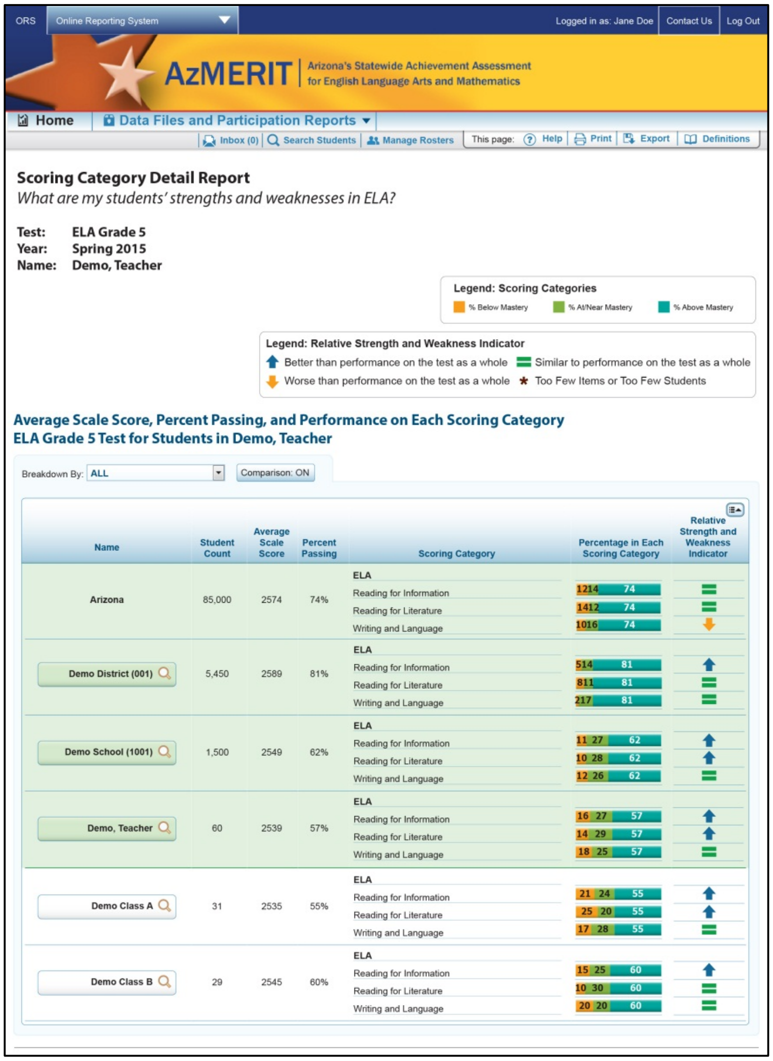
Name	Student Count	Average Scale Score	Percent Passing	Percent in Each Performance Level
Arizona	85,000	2574	74	12 14 33 41
Demo District (001)	5,450	2589	81	5 14 37 44
Demo School (1001)	500	2549	62	11 27 35 27
Demo, Teacher	60	2539	57	16 27 42 15

Aggregated subject reports show average performance for the state, districts, schools, teachers, and classes. Bar charts displays 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 who completed the selected test
- **Average Scale Score:** Average scale score of students who completed the selected test
- **Percent Passing:** The percent of tested students reaching the proficient threshold on the selected test
- **Percent at Each Performance Level:** The distribution of students across each of the four performance levels

SCORING CATEGORY DETAIL REPORTS



Scoring Category Detail Report

What are my students' strengths and weaknesses in ELA?

Test: ELA Grade 5
Year: Spring 2015
Name: Demo, Teacher

Legend: Scoring Categories
🟧 % Below Mastery 🟩 % At/Near Mastery 🟦 % Above Mastery

Legend: Relative Strength and Weakness Indicator
↑ Better than performance on the test as a whole = Similar to performance on the test as a whole
↓ Worse than performance on the test as a whole ★ Too Few Items or Too Few Students

Average Scale Score, Percent Passing, and Performance on Each Scoring Category
ELA Grade 5 Test for Students in Demo, Teacher

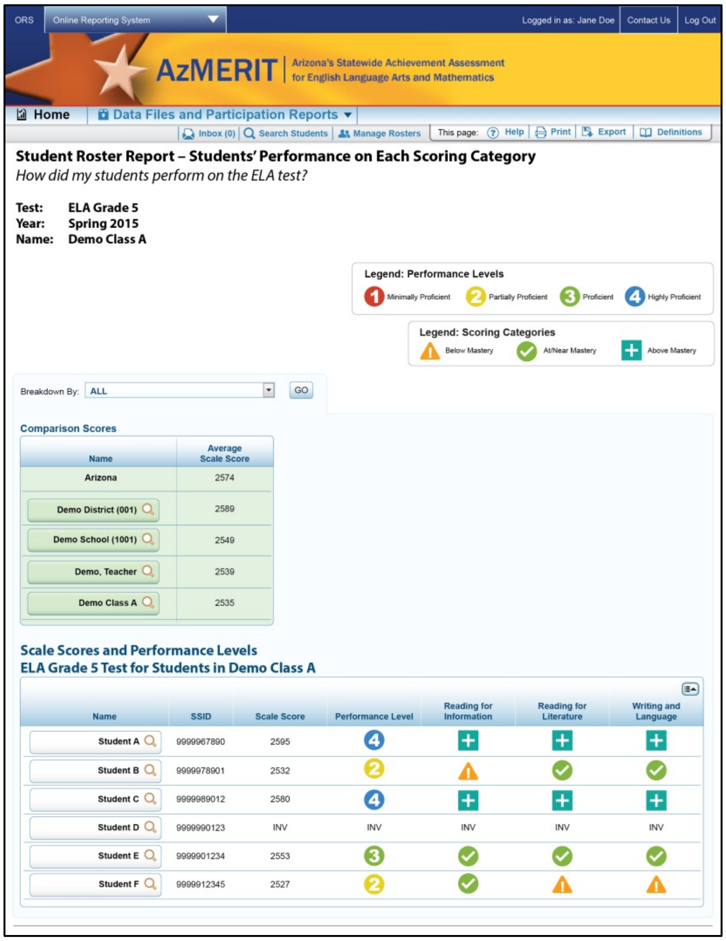
Breakdown By: ALL Comparison: ON

Name	Student Count	Average Scale Score	Percent Passing	Scoring Category	Percentage in Each Scoring Category	Relative Strength and Weakness Indicator
Arizona	85,000	2574	74%	ELA		
				Reading for Information	1214 74	=
				Reading for Literature	1412 74	=
				Writing and Language	1016 74	↓
Demo District (001)	5,450	2589	81%	ELA		
				Reading for Information	514 81	↑
				Reading for Literature	811 81	=
				Writing and Language	217 81	=
Demo School (1001)	1,500	2549	62%	ELA		
				Reading for Information	11 27 62	↑
				Reading for Literature	10 28 62	↑
				Writing and Language	12 26 62	=
Demo, Teacher	60	2539	57%	ELA		
				Reading for Information	16 27 57	↑
				Reading for Literature	14 29 57	↑
				Writing and Language	18 25 57	=
Demo Class A	31	2535	55%	ELA		
				Reading for Information	21 24 55	↑
				Reading for Literature	25 20 55	↑
				Writing and Language	17 28 55	=
Demo Class B	29	2545	60%	ELA		
				Reading for Information	15 25 60	↑
				Reading for Literature	10 30 60	=
				Writing and Language	20 20 60	=

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 a group of students' performance 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 minus sign for a scoring category does not imply a lack of proficiency. Instead, it simply communicates that these students' performance on that scoring category was statistically lower than their performance across all other scoring categories put together. Although the students are doing well, an educator may want to focus instruction on these areas.

STUDENT ROSTER REPORTS



Student Roster Report – Students' Performance on Each Scoring Category

How did my students perform on the ELA test?

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

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

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

Breakdown By: ALL

Comparison Scores

Name	Average Scale Score
Arizona	2574
Demo District (001)	2589
Demo School (1001)	2549
Demo, Teacher	2539
Demo Class A	2535

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

Name	SSID	Scale Score	Performance Level	Reading for Information	Reading for Literature	Writing and Language
Student A	9999967890	2595	4	+	+	+
Student B	9999978901	2532	2	⚠	✓	✓
Student C	9999989012	2580	4	+	+	+
Student D	9999990123	INV	INV	INV	INV	INV
Student E	9999901234	2553	3	✓	✓	✓
Student F	9999912345	2527	2	✓	⚠	⚠

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

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

- **Scale score:** The score of each student who completed the test.
- **Performance level:** Represents levels of overall subject mastery with respect to the Arizona College and Career Ready Standards

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

INDIVIDUAL STUDENT REPORTS

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

6.3 INTERPRETATION OF SCORES

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

Performance levels represent levels of mastery with respect to the Arizona College and Career Ready Standards 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 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 ADE.

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

The student’s performance in each content area assessment is summarized in an overall test score referred to as a scale score. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale score is 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 2014–2015 school year indicates the same level of achievement as an average scale score of 2450 for grade 3 students in the 2015–2016 school year even though the test may include a slightly different set of items.

As described in Section 8 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 6.3.1.

Exhibit 6.3.1 AzMERIT Scale Score Ranges

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

In addition to overall ELA test scores, a subscale score for the reading subject area of ELA is computed for grade 3 students, based on both reading for information and reading for literature items. The Grade 3 Reading score is reported with respect to a “Move on When Reading” performance level classification. For the 2015 AzMERIT administration, Arizona students in Grade 3 who did not meet the MOWR requirement are required to receive extra help during school year 2015-2016. The scale score cut for the “Move on When Reading” is 2446; if the student’s reading subscale score meets or exceeds this cut, the student is classified as meeting the performance standard. If the student’s reading subscale score is below that, they do not meet the standard.

7. PERFORMANCE STANDARDS

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

7.1 STANDARD SETTING PROCEDURES

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the Arizona College and Career Ready Standards. 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 College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Arizona College and Career Ready 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 standard 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 College and Career Ready Standards. The workshop consisted of a series of standardized and rigorous procedures that Arizona educators, serving as standard setting panelists, followed to recommend performance standards. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists use their expert knowledge of the Arizona College and Career Ready Standards and student achievement to map the performance level descriptors adopted by the Arizona State Board of Education onto an ordered item book (OIB) based on the first operational test form administered to students in spring 2015.

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

Panelists were also provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade level assessment related to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces 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.

7.1.1 PERFORMANCE LEVEL DESCRIPTORS

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

- Minimally Proficient
- Partially Proficient
- Proficient
- Highly Proficient

Prior to convening the standard setting workshops, AIR, in consultation with 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 College and Career Ready Standards. Following a cycle of revisions to the draft PLDs, 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 azed.gov.

7.2 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 7.2.1 presents the performance standard associated with panelist-recommended OIB page numbers in logit value, as well as the percentage of students classified as meeting or exceeding each standard. Following the standard setting workshop, panelist recommendations were submitted to Arizona's State Board of Education; the Board formally adopted the standards in August 2015.

Exhibit 7.2.1 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 7.2.2 shows the percentage of student 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 7.2.2 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%

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
8	43%	24%	20%	13%
Algebra I	45%	23%	23%	9%
Geometry	47%	24%	24%	6%
Algebra II	47%	24%	23%	6%

Exhibit 7.2.3 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 2014 field test administration. As Exhibit 7.2.3 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 7.2.3 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

	Percent of Students Meeting Standard			
Grade/ Course	**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%

8. SCALING AND EQUATING

Item parameters for the assessments were calibrated following the initial spring administration in 2015 and a vertical scale was established for reporting both ELA and math. In addition, a series of linking studies were performed to allow comparison of performance on the AzMERIT with 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 ranged from 3395 to 3839 across grade level and high school End of Course tests.

8.1 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 scale for scoring and reporting. This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by AIR, ADE, and HumRRO, which acts as a third party quality assurance contractor.

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

8.1.1 CALIBRATION OF AZMERIT ITEM BANKS

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

In spring 2015, the base year of AzMERIT, 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 subsequent administrations of AzMERIT, pre-equated item parameters will be used to score student test records.

8.1.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

AzMERIT is scored using maximum likelihood estimation. As described previously, parameter estimates are calibrated using the Rasch model for dichotomously scored items and Masters' partial credit model for polytomous items.

LIKELIHOOD FUNCTION

The likelihood function for generating the MLEs is based on a mixture of items types and can therefore be expressed as:

$$\mathrm{L}(\theta) = \mathrm{L}(\theta)^{\mathrm{MC}}\mathrm{L}(\theta)^{\mathrm{CR}}$$

where:

$$\mathrm{L}(\theta)^{\mathrm{MC}} = \prod_{\mathrm{i}=1}^{\mathrm{N}} \left[\frac{1}{1+\exp[-\mathrm{D}(\theta-\mathrm{b_i})]}\right]^{\mathrm{x_i}} \left[1+\frac{1}{1+\exp[-\mathrm{D}(\theta-\mathrm{b_i})]}\right]^{1-\mathrm{x_i}}$$

$$\mathrm{L}(\theta)^{\mathrm{CR}} = \prod_{\mathrm{i}=1}^{\mathrm{N}} \frac{\exp\sum_{\mathrm{k}=1}^{\mathrm{x_i}} \mathrm{D}(\theta-\delta_{\mathrm{ki}})}{1+\sum_{\mathrm{j}=1}^{\mathrm{m_i}} \exp\sum_{\mathrm{k}=1}^{\mathrm{j}} \mathrm{D}(\theta-\delta_{\mathrm{ki}})}$$

and where $\mathrm{b_i}$is the location parameter, $\mathrm{x_i}$ is the observed response to the item, i indexes item and δ_{ki} is the kth step for item i with m total categories.

We subsequently find $\arg\max\theta\ \mathrm{L}(\theta)$ as the student's theta (i.e. MLE) given the set of items administered to the student.

DERIVATIVES

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

$$\frac{\partial \ln L(\theta)^{MC}}{\partial\theta} = \sum_{i=1}^{N} x_i - \left[\frac{1}{1+\exp[-(\theta-b_i)]}\right]$$

$$\frac{\partial \ln L(\theta)^{CR}}{\partial\theta} = \sum_{i=1}^{N} x_i - \left[\frac{\sum_{j=1}^{m_i} j \exp\sum_{k=1}^{x_i}(\theta-\delta_{ki})}{1+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}(\theta-\delta_{ki})}\right]$$

$$\frac{\partial^2 \ln L(\theta)^{MC}}{\partial^2\theta} = -\sum_{i=1}^{N}\left(1-\left[\frac{1}{1+\exp[-(\theta-b_i)]}\right]\right)\left[\frac{1}{1+\exp[-(\theta-b_i)]}\right]$$

$$\frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2\theta} = \sum_{i=1}^{N}\left[\frac{\sum_{j=1}^{m_i} j \exp\sum_{k=1}^{x_i}(\theta-\delta_{ki})}{1+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}(\theta-\delta_{ki})}\right]^2 - \left[\frac{\sum_{j=1}^{m_i} j^2 \exp\sum_{k=1}^{x_i}(\theta-\delta_{ki})}{1+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}(\theta-\delta_{ki})}\right]$$

Hence, the estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} \bigg/ \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t}$$

where

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)^{MC}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} = \frac{\partial^2 \ln L(\theta)^{MC}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta}$$

and where θ_t denotes the estimated θ at iteration t.

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 (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 test score.

8.2 ESTABLISHING A VERTICAL SCALE IN ELA AND MATH

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade level assessment (g) were embedded in field test slots of the assessment in the grade below (g-1). 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.

8.2.1 LINKING ITEMS

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

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

However, AzMERIT assessments of high school 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, since Geometry is not a focus of instruction in Algebra I courses. An alternative approach, and the one adopted by ADE, was to link the grade 8 mathematics scale to both the Algebra I and Geometry EOC scales, since the grade 8 assessment includes items measuring both algebra and geometry. Because Algebra II builds on the knowledge and skills assessed in Algebra I, all Algebra II items were used to link the Algebra I assessments to the Algebra II scale.

8.2.2 LINKING ANALYSIS

When feasible, it is desirable to establish linkages using both concurrent calibrations and chain-linking approaches to ensure that results are consistent across methods. An important advantage of chain linking approaches is that, because item response theory (IRT) calibrations proceed by establishing the within-grade scale, the achievement construct intended by the blueprint and enacted in the operational test form is preserved. Unfortunately, however, at each step in the linking chain, the linking error accumulates, so that linking constants for grades more distant from the reference grade are less precise than are linking constants for grades in closer proximity to the reference grade. Concurrent calibrations do not accrue linking error across grade levels, so that linking constants are similarly precise between all grade levels. However, the calibrations resulting from this approach measure the construct that is common across the linked assessments, which may be different from the intended achievement construct at each grade level, especially for subjects such as 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

To facilitate the development of a vertical scale that will be sensitive to student growth over time, we first evaluated the performance of vertical linking items between the grade levels in which they were administered to identify any items that were more difficult for students in the intended grade than they were for students in the lower grade. For math, items that showed proportion correct scores lower in the intended grade than in the lower grade were dropped from the final vertical linking set. This resulted in dropping on average just over two items per linking set, with a maximum of six items dropped for the linkage between grade 6 and grade 7 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 8.2.2.1 Number of Items Dropped and Remaining in the Final Vertical Linking Set

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

CHAIN-LINKING

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

CONCURRENT CALIBRATION

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

Exhibit 8.2.2.2 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. To more directly examine the magnitude of gains across grade level assessments, Exhibit 8.2.2.3 shows the difference between linking constants between each of the grade levels assessed. Relative gains are also represented graphically in Exhibit 8.2.2.4 and Exhibit 8.2.2.5 for mathematics and ELA, 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 end of course assessments. For the ELA assessments, the cross grade gains are more modest, and tend to diminish in the higher grade levels.

Exhibit 8.2.2.2 Vertical Linking Constants Resulting from Chain-Linking Within Grade Scales and Concurrent Calibration of Items Across Grades

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

Exhibit 8.2.2.3 Linking Constant Differences Between Each of the Grade Level Scales

	Vertical Linking Constant Differences			
	Mathematics		Reading	
Linkage	Chain-Linked	Concurrent	Chain-Linked	Concurrent
G3→G4	1.32	1.3	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

Linking constants resulting from the chain-linking and concurrent calibration approach are quite consistent, indicating that both approaches converge on a common growth scale. Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain-linking method preserves the within grade measurement construct, and was therefore selected as a preliminary vertical scale for the purpose of recommending performance standards. We note that ordered item books for the standard setting workshop were based on the within grade scales, so any modifications to the vertical scale will 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, with the exception of grade 6 to grade 7, which showed just over a half standard deviation gain. Similar 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.

Exhibit 8.2.2.4 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: Mathematics

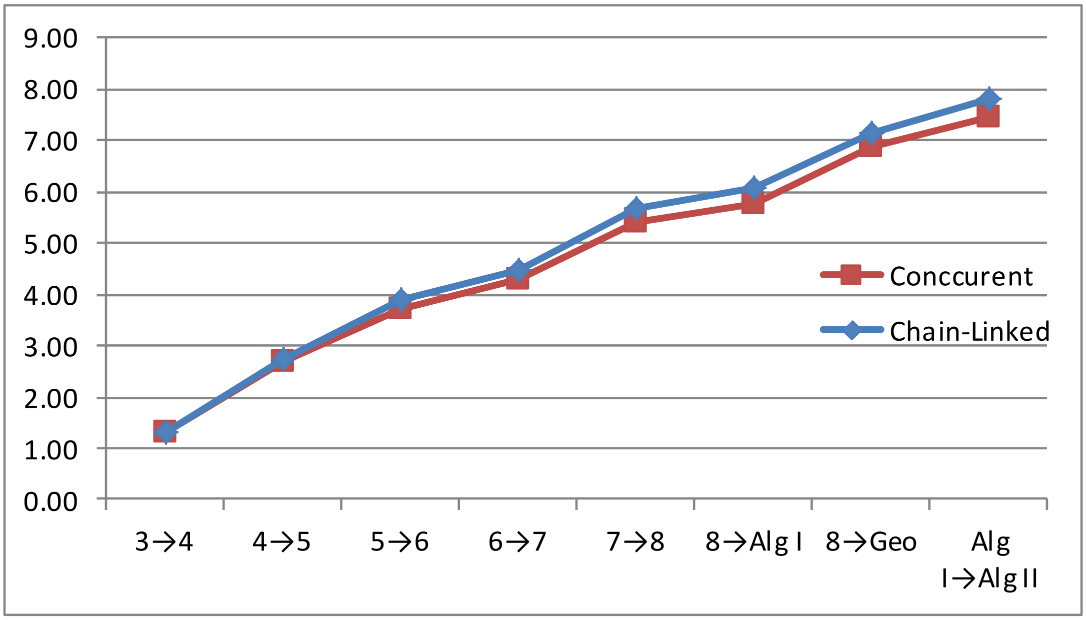
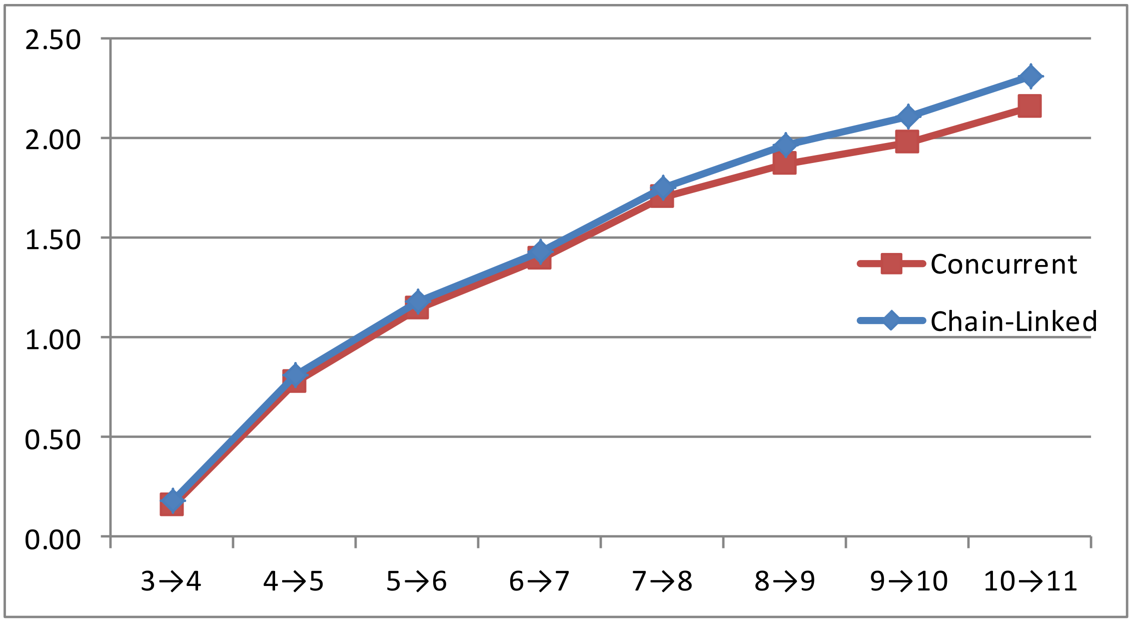


Exhibit 8.2.2.5 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: ELA



8.3 AZMERIT REPORTING SCALE (SCALE SCORES)

The AzMERIT assessments are reported on common scales within each subject (ELA and math). The IRT vertical scale is 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.

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

where a=30, b=2500 for ELA tests, and a=30, b=3500 for Mathematics tests.

$\theta_{V=}\theta + VL\ Constant$, θ is the on-grade ability estimate and VL constant is listed below for each of the tests, as described above. For reporting, the on-grade ability estimate is truncated at +/- 3.5.

After transforming theta ability estimates to the vertical AzMERIT reporting scale, the observable scale scores nearest each of the performance standard cut scores are evaluated. If the observable scale score nearest the performance standard is below the cut score, the scale score is rounded up to be equal to the cut score. If the observable scale score nearest the performance standard is above the cut score, no special rounding rule is applied.

Overall scale scores for the AzMERIT are mapped into 4 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 8.3.1 shows the scale score ranges for the performance levels for each test.

Exhibit 8.3.1 Scale Score Ranges for Performance Levels

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

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
Grade 8 Math	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

8.4 EQUATING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to conducting the standard setting workshops and reporting test scores for the spring 2015 assessments, AIR and ADE separately performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration. The AIR study also identified the linking constants necessary to place item parameter estimates across modes on a common scale for test scoring and reporting.

AIR used a matched samples design (Way, Davis, and Fitzpatrick, 2006) 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. The regression analysis identified for each student a predicted score on the paper AzMERIT assessment from previous year achievement, covarying demographic variables that included gender, ethnicity, income level status, English language learner (ELL) 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 calibrated independently for the matched online and paper test administration mode samples. The linking constant necessary to bring the matched sample paper item parameters onto the matched sample online scale was then computed. The linkages were computed in two ways. Mean linking was taken as the difference between the average item difficulty estimates from the matched sample paper calibration and the average item difficulty estimates from the matched sample online item parameter estimates. Mean-sigma linking equating was also used to place the paper item parameters on the online scale. The resulting linking constants indicated that parameter estimates resulting from the independent calibrations of the paper and online assessments are quite comparable. The largest identified mode effect was for items on the grade 3 ELA assessment which were, on average, slightly more difficult for students who were administered the assessment online. Examination of the linked item parameter estimates indicated that items with the greatest discrepancy between online and paper were not isolated within a particular content standard or item type.

ADE also independently investigated mode comparability using a strategy based on the operational test administration statewide (Scott, 2015). In her study, Scott (2015) first identified which Arizona schools elected to administer AzMERIT online and which on paper, and then examined the two samples of schools for any differences in performance on the spring 2014 paper administration of AIMS. Having found no difference in mean 2014 performance between the two groups, there would be no expectation for performance differences on AzMERIT except as a function of test administration mode. Following the spring 2015 administration of AzMERIT, ADE examined the performance of schools participating online and on paper, and again found performance on the AzMERIT to be comparable between the two sets of schools.

The mode comparability studies examined the comparability of item parameters and resulting test scores from the online and paper administrations of the spring 2015 AzMERIT assessments in ELA (including reading and writing) and math. The matched samples analyses revealed generally that item difficulty estimates and resulting student ability estimates were comparable across test administration modes. Small mode effects were identified for some grades in the ELA assessments, with items in the grade 3 assessment proving slightly more difficult when administered online. Even for the largest effect in grade 3 ELA, the magnitude of the mode difference was quite small, amounting to just under one raw score point (approximately four point on the AzMERIT scale), impacting the proficient rate by about 1.6%. Given the generally strong comparability of item difficulty across mode, ADE, with support of their Technical Advisory Committee and approval of the State Board of Education, adopted a single set of bank parameters for scoring student responses on the AzMERIT across the online and paper test administration modes.

A summary of AIR's mode comparability study, including description of matched samples, linking constants, and estimated effect of applying mode corrections in student performance, and Scott's complete 2015 mode comparability study are provided in the 2015 performance standard setting technical report.

8.5 LINKING THE AZMERIT TO OTHER SCALES FOR PERFORMANCE COMPARISON

8.5.1 ESTABLISHING LINKAGES TO AIMS, SAGE, SMARTER BALANCED, PISA

To facilitate comparisons of Arizona achievement to other national and international benchmarks, a number of external linking sets were embedded in the 2015 AzMERIT field test pools. 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. 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 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, and
- Utah's SAGE via common items in the operational test form.

Subsequent to calibration of the AzMERIT operational items and establishment of the reference scale, parameter estimates for those items were anchored to their reference values and all items administered in the embedded field test (EFT) blocks were calibrated under that constraint, placing parameter estimates for all field test and external linking item sets on the same AzMERIT scale defined by the operational item parameters. All external linking items had two sets of item parameters: a) external scale, and b) AzMERIT scale. To identify the location of external scale performance standards on the AzMERIT scale, AIR identified the linking constants necessary to transform item parameters from the external reference scale to the AzMERIT scale. Where the external scale was calibrated using the Rasch model, such as with AIMS, mean-sigma equating was used to identify the location of external performance standards on the AzMERIT scale. For external scales calibrated using more general IRT models, Stocking-Lord equating was used to identify the location of external scale performance standards on the AzMERIT scale.

In the context of standard setting, this procedure enabled ADE to identify a location in the AzMERIT 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.

8.5.2 IDENTIFYING THE LOCATION OF THE ACT COLLEGE-READY CUT ON AZMERIT

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the ACT college ready cuts were identified on the AzMERIT scale and provide to panelists during performance standards workshops in 2015. In order to identify the location of the ACT college ready cuts for the AzMERIT End-of-Course assessments, a two-step approach was used to first identify the location of the ACT college ready benchmark on the AIMS scale, and then use the linkages between AIMS and AzMERIT to map the ACT college ready benchmark onto the AzMERIT scale(s). For this purpose, ADE provided ACT and AIMS scores for a recent cohort of students.

A selected sample of fewer than half of Arizona students took the ACT. Therefore, a two-step approach was used to handle missing data in the analysis of the relationship between the AIMS and ACT test scores. The approach was similar to the strategy employed by the National Assessment of Educational Progress (NAEP) in which student demographic information was used to impute achievement of students absent during the NAEP administration (McLaughlin, Scarloss, Stancavage, & Blankenship, 2005).

In the first step, all records with missing ACT test scores were excluded from the analysis. Then the target ACT score for the remaining students were regressed onto the corresponding AIMS score and demographic variables. Stepwise selection was used to identify the prediction model.

To help refine the prediction model, in a second step, the missing ACT test scores were imputed to better account for the percentage of students who did not take the ACT prior to graduation. To impute missing ACT scores, it must be assumed that the relationship between AIMS and ACT scores identified in the first step can be generalized to the students who did not participate in an administration of the ACT test and that missing ACT scores can be imputed using the available demographic information and the AIMS scores. Using the entire cohort, including both observed and imputed ACT test scores, ACT scores was again regressed on AIMS scale scores and demographic variables.

9. RELIABILITY

9.1 ESTIMATING RELIABILITY

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

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

9.2 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 9.2.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and mathematics assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 9.2.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade/Course	ELA	Math
3	0.90	0.91
4	0.89	0.91
5	0.89	0.91
6	0.90	0.90
7	0.89	0.91
8	0.90	0.91
9/Algebra I	0.88	0.90
10/Geometry	0.89	0.88
11/Algebra II	0.88	0.86

9.3 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.

For the AzMERIT assessments scored using MLE, the mathematical statement of the CSEM for student *i* is:

$$CSEM(\hat{\theta}_i) = \frac{1}{\sqrt{I(\hat{\theta}_i)}}$$

where $I(\hat{\theta}_i)$ is the Fisher information at the MLE and is calculated:

$$I(\hat{\theta}) = -\frac{\partial^2 l(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}.$$

In general, the second derivative for the *i*th 1PL item is

$$\frac{\partial^2 log([p(\theta)]^{z_i}[q(\theta)]^{1-z_i})}{\partial\theta^2} = \begin{cases} -D^2\frac{q_i(\theta)\left(p_i^3(\theta)\right)}{p_i^2(\theta)} \text{ if } z_i = 1 \\ -D^2 q_i(\theta)(p_i(\ \theta\ ))\ \text{ if } z_i = 0 \end{cases}$$

The second derivative for the *i*th Master’s Partial Credit Model item is

$$\frac{\partial^2 log\left(P(z_i|\theta)\right)}{\partial\theta^2} = D^2\frac{\left[\sum_{j=1}^{m_i} j\text{Exp}(\sum_{k=1}^{j} D(\theta - b_{ki}))\right]^2}{\left[1+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}D(\theta-b_{ki})\right]^2} - D^2\frac{\sum_{j=1}^{m_i} j^2\text{Exp}(\sum_{k=1}^{j}D(\theta-b_{ki}))}{1+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}D(\theta-b_{ki})}$$

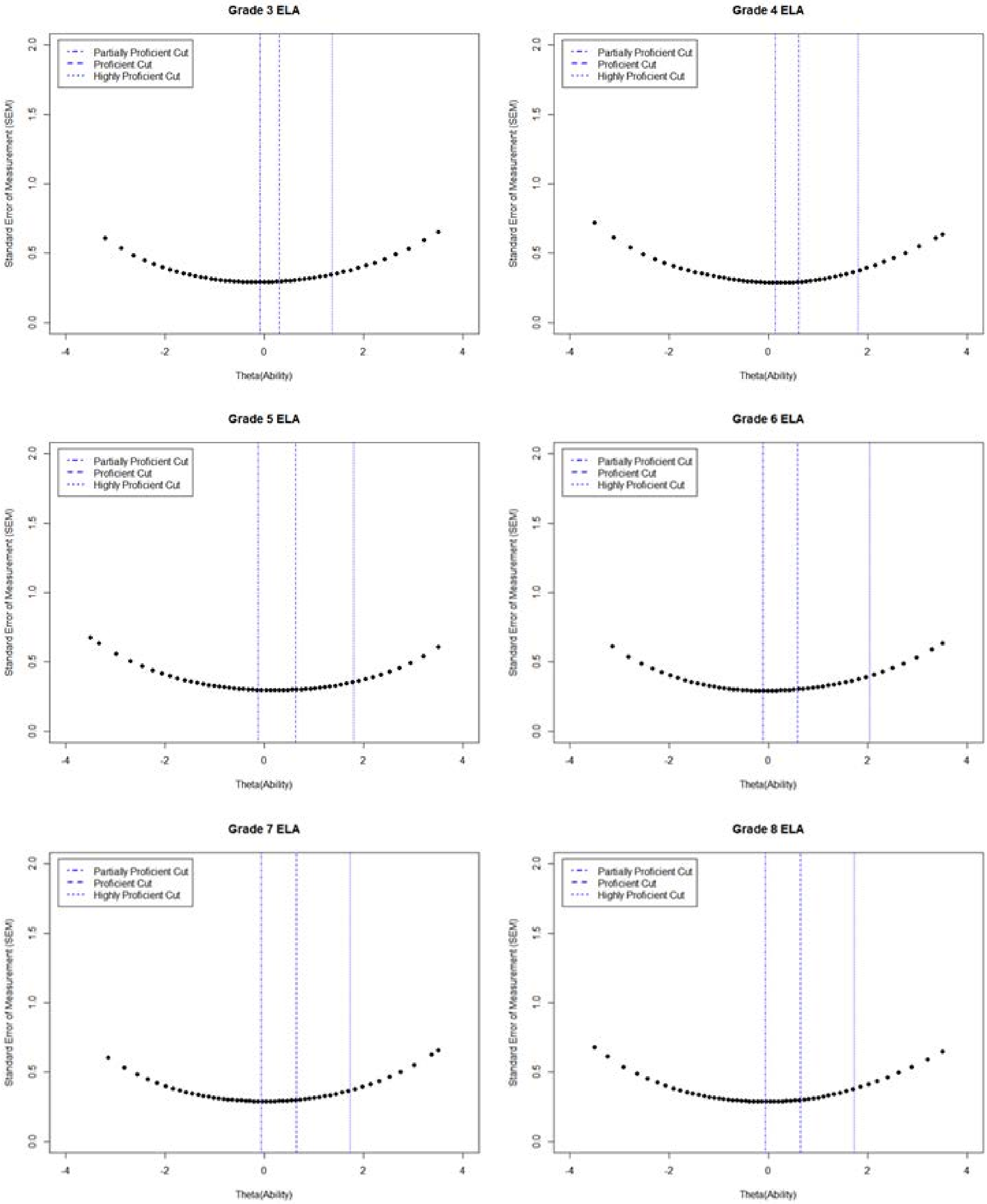
Standard errors of the MLEs are transformed to be placed onto the reporting scale. This transformation is:

$$SE_{vs} = a * SE_{\theta_i}$$

where SE_θ is the standard error of the ability estimate on the θ scale; and a is the slope of the scaling constants that take θ to the reporting scale. For both ELA and Mathematics tests, a=30.

The figures in Exhibit 9.3.1 and Exhibit 9.3.2 present graphically the standard errors of measurement for the AzMERIT ELA and mathematics assessments. Each figure also includes the location of the three AzMERIT performance standards. As the figures indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Approaching and Proficient performance standards. Test scores near the tails of the ability distribution are somewhat less precise as expected. An SEM of .3 on the theta metric is consistent with an internal consistency of 0.9. The tables in Appendix B show the mean SEMs for students scoring in each of the performance levels on the AzMERIT reporting scale. While these tables also 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 9.3.1 Overall Standard Error of Measurement for ELA



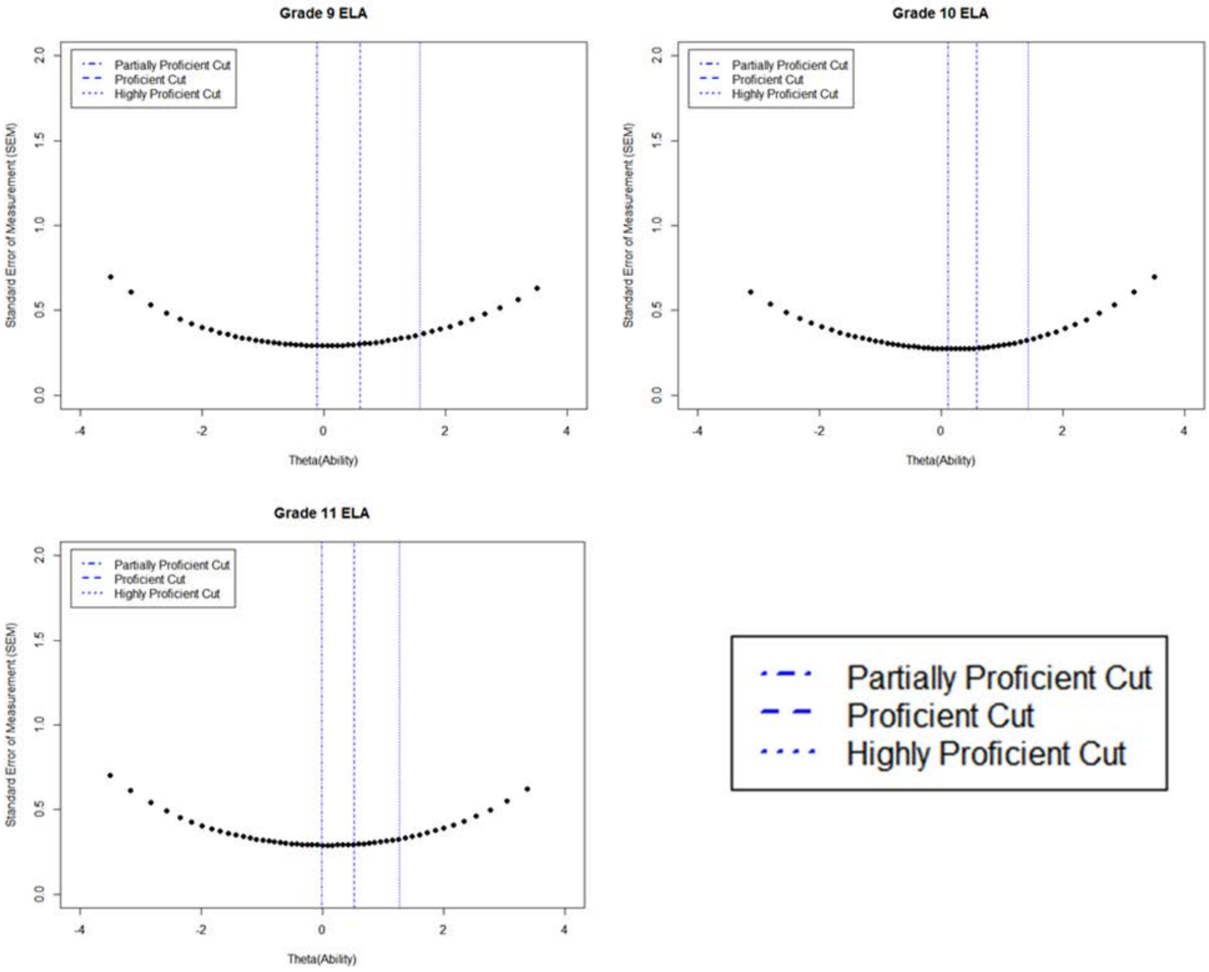
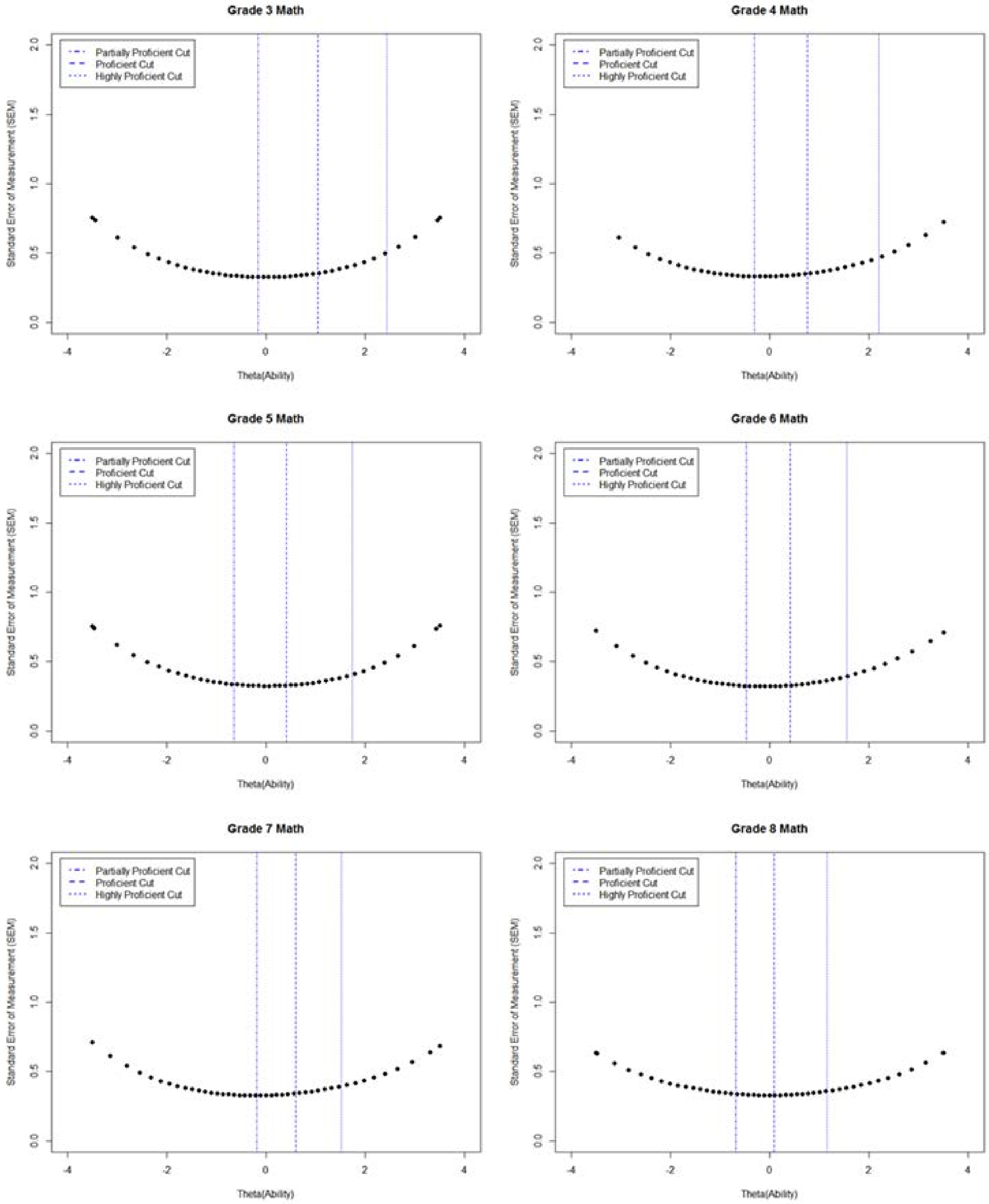
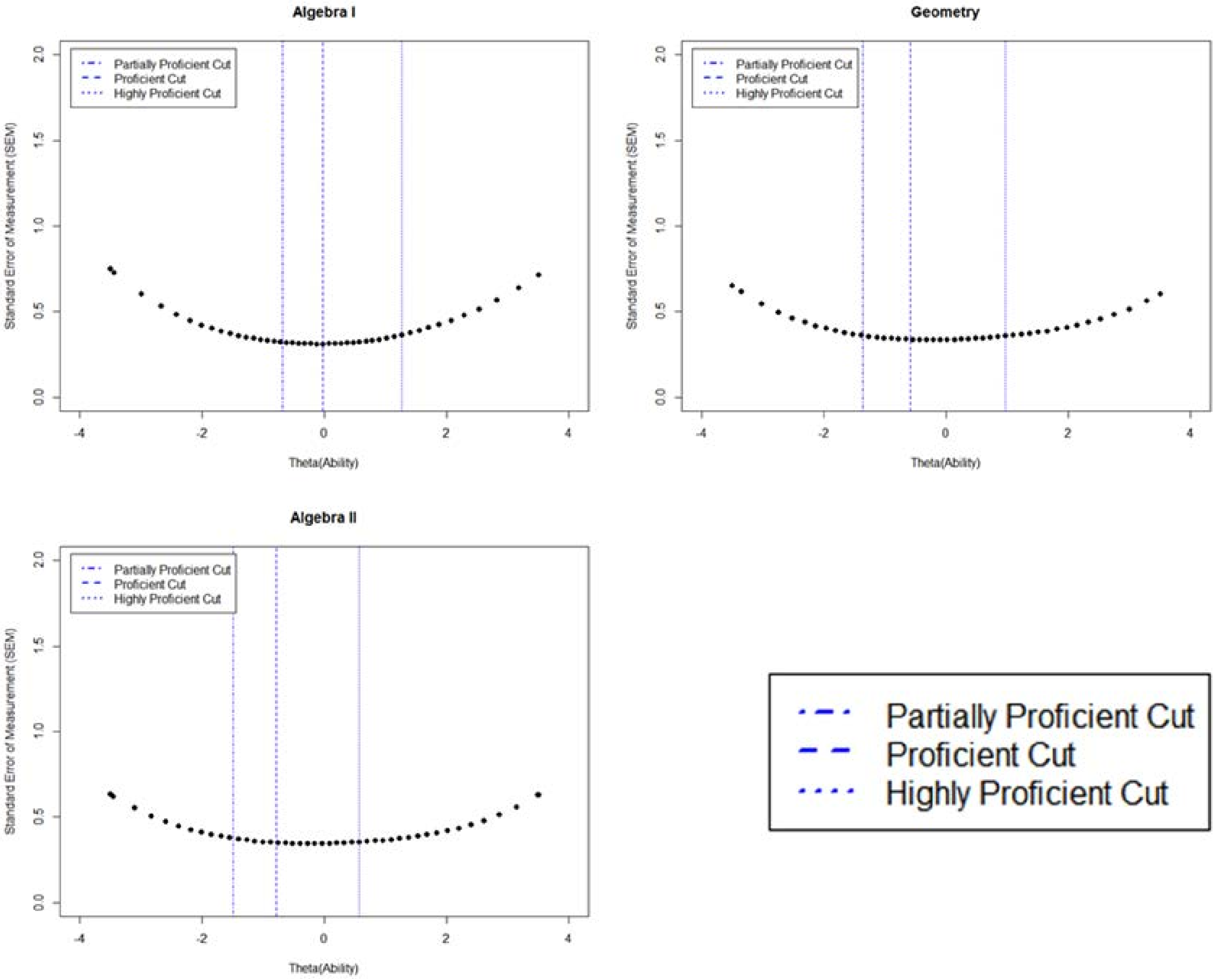

Grade 9 ELA
Grade 10 ELA
Grade 11 ELA
Partially Proficient Cut
Proficient Cut
Highly Proficient Cut
Standard Error of Measurement (SEM)
0.0
0.5
1.0
1.5
2.0
-4
-2
0
2
4
Theta(Ability)
Partially Proficient Cut
Proficient Cut
Highly Proficient Cut

Exhibit 9.3.2 Overall Standard Error of Measurement for Mathematics





9.4 STUDENT CLASSIFICATION CONSISTENCY

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 2.15 in the *Standards for Educational and Psychological Testing* (AERA, 1999). This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

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

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

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

For the *i*th student, the student's estimated ability is $\hat{\theta}_i$ with a standard error, $se\left(\hat{\theta}_i\right)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N\left(\theta_i, se\left(\hat{\theta}_i\right)\right)$, assuming a normal distribution, where $\theta_i$is the unknown true ability of the *i*th student. The probability of the true score *at or above* the cut score is estimated as

$$p(\theta_i \geq \theta_c) = p\left(\frac{\theta_i - \hat{\theta}_i}{e\left(\hat{\theta}_i\right)} \geq \frac{\theta_c - \hat{\theta}_i}{e\left(\hat{\theta}_i\right)}\right) = p\left(\frac{\hat{\theta}_i - \theta_i}{e\left(\hat{\theta}_i\right)} < \frac{\hat{\theta}_i - \theta_c}{e\left(\hat{\theta}_i\right)}\right) = \Phi\left(\frac{\hat{\theta}_i - \theta_c}{e\left(\hat{\theta}_i\right)}\right).$$

Similarly, the probability of the true score being *below* the cut score is estimated as

$$p(\theta_i < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta}_i - \theta_c}{e\left(\hat{\theta}_i\right)}\right).$$

9.4.1 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se\left(\hat{\theta}_i\right)\right)$, we can estimate the above probabilities directly using the likelihood function. The likelihood function of theta given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated 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 being classified *at or above* the cut score, θ_c, given the student's item scores $\mathbf{z} = (z_1, \cdots, z_I)$ and item parameters $\mathbf{b} = (\mathbf{b}_1, \cdots, \mathbf{b}_I)$, based on the *I* items administered, can be estimated as $P(\theta \geq \theta_c | \mathbf{z}, \mathbf{b}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}$, where, the likelihood function is

$$L(\theta \,|\, \mathbf{z}, \mathbf{b}) = \prod_{i \in MC}\left(\frac{Exp(z_i\theta - b_i z_i)}{1 + Exp(\theta - b_i)}\right)\prod_{i \in CR}\left(\frac{Exp(z_i\theta - \sum_{k=1}^{z_i} b_{ik})}{1 + \sum_{i=1}^{K_i}\left(Exp(\sum_{k=1}^{i}(\theta - b_{ik}))\right)}\right)$$

and $\mathbf{b}_i = (b_i)$ if the *i*th item is a dichotomous item, or $\mathbf{b}_i = \left(b_{i1}, \cdots, b_{iK_i}\right)$ if the *i*th item is a polytomous item with the maximum possible score K_i.

Similarly, we can estimate the probability of *below* the cut score as:

$$P(\theta < \theta_c|\mathbf{z},\mathbf{b}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}$$

In Exhibit 9.4.1.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassification (*false positive*) occurs when, for example, a student who actually achieved Meets proficiency on the true score, is classified incorrectly as achieving Exceeds proficiency. 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 9.4.1.1 Classification Accuracy

		Classification on a Form Actually Taken	
		Above the Cut Score	**Below the Cut Score**
Classification on True Score	**At or Above the Cut Score**	N11 (Truly above the cut)	N10 (False negative)
	Below the Cut Score	N01 (False positive)	N00 (Truly below the cut)

Where

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

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

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

$$N_{10} = \sum\nolimits_{i \in N_0} P(\theta_i \geq \theta_c \,|\, \mathbf{z},\mathbf{b})$$

Where N_1 contains 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 $\frac{N_{11}+N_{00}}{N}$, with $N = N_1 + N_0$.

9.4.2 CLASSIFICATION CONSISTENCY

As shown in Exhibit 9.4.2.1, 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 9.4.2.1 Classification Consistency

		Classification on the 2nd Form Taken	
		Above the Cut Score	**Below the Cut Score**
Classification on the 1st 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)

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

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

Similarly, the probability of consistency and inconsistency can be estimated based on a student's item scores and the item parameters.

The probability of consistency for *at or above* the cut score is estimated as

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

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

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

The probability of inconsistency is estimated as

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

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

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

$$N_{11} = \sum_{i=1}^{N} P\left(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c \mid \mathbf{z}, \mathbf{b}\right)$$

$$N_{10} = \sum_{i=1}^{N} P\left(\theta_{i,1} \geq \theta_c, \theta_{i,2} < \theta_c \mid \mathbf{z}, \mathbf{b}\right)$$

$$N_{00} = \sum_{i=1}^{N} P\left(\theta_{i,1} < \theta_c, \theta_{i,2} < \theta_c \mid \mathbf{z}, \mathbf{b}\right)$$

$$N_{01} = \sum_{i=1}^{N} P\left(\theta_{i,1} < \theta_c, \theta_{i,2} \geq \theta_c \mid \mathbf{z}, \mathbf{b}\right)$$

and

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

Exhibit 9.4.3 presents the decision accuracy and consistency indexes for spring 2015 administration of AzMERIT. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, while the accuracy a single test score and the true score, which does not include measurement error.

Exhibit 9.4.3 Decision Accuracy and Consistency Indexes for Performance Standards

	Accuracy (%)			Consistency (%)		
Grade	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
			ELA			
3	0.91	0.92	0.96	0.88	0.88	0.94
4	0.91	0.91	0.97	0.87	0.87	0.96
5	0.91	0.92	0.98	0.87	0.88	0.97
6	0.91	0.92	0.98	0.88	0.89	0.97
7	0.91	0.92	0.98	0.87	0.89	0.97
8	0.91	0.92	0.97	0.88	0.89	0.96
9	0.90	0.92	0.97	0.87	0.89	0.96
10	0.91	0.92	0.96	0.87	0.89	0.95
11	0.91	0.92	0.96	0.87	0.90	0.95
			Mathematics			
3	0.93	0.93	0.95	0.91	0.90	0.93
4	0.92	0.92	0.96	0.89	0.89	0.94
5	0.92	0.93	0.96	0.89	0.90	0.94
6	0.91	0.93	0.97	0.88	0.90	0.95
7	0.92	0.94	0.96	0.89	0.91	0.95
8	0.91	0.93	0.97	0.87	0.90	0.95
Algebra I	0.90	0.94	0.97	0.86	0.91	0.95
Geometry	0.89	0.94	0.98	0.85	0.91	0.97
Algebra II	0.90	0.94	0.98	0.85	0.91	0.97

9.5 RELIABILITY FOR SUB-GROUPS IN THE POPULATION

Exhibit 9.5.1 and 9.5.2 shows the mean reliability for each of the subgroups: African Americans, Asian, Native Hawaiians/Pacifica Islanders, Hispanic/Latinos, American Indian or Alaskans, Whites, Multiple Ethnicities and females and males (regardless of racial or ethnic group). Each racial and/or ethnic group was composed of approximately equal numbers of males and females. As the Exhibit indicates, internal consistency reliabilities are consistent across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups.

Exhibit 9.5.1 Internal Consistency Reliability by Subgroup– ELA

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All Students	0.90	0.89	0.89	0.90	0.89	0.90	0.88	0.89	0.88
Female	0.90	0.89	0.88	0.89	0.89	0.89	0.87	0.88	0.88
Male	0.90	0.89	0.89	0.89	0.89	0.90	0.88	0.89	0.88
African American	0.88	0.88	0.87	0.88	0.88	0.89	0.88	0.86	0.84
Asian	0.90	0.88	0.89	0.90	0.90	0.90	0.87	0.88	0.89
Native Hawaiian/Pacific Islander	0.89	0.86	0.90	0.89	0.87	0.87	0.82	0.86	0.84
Hispanic/Latino	0.88	0.88	0.87	0.87	0.87	0.88	0.86	0.86	0.84
American Indian or Alaskan	0.85	0.85	0.83	0.84	0.84	0.86	0.81	0.82	0.77
White	0.89	0.88	0.88	0.89	0.88	0.89	0.87	0.88	0.88
Multiple Ethnicities	0.90	0.89	0.87	0.89	0.88	0.89	0.88	0.88	0.88

Exhibit 9.5.2 Internal Consistency Reliability by Subgroup – Mathematics

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
All Students	0.91	0.90	0.91	0.90	0.91	0.91	0.90	0.88	0.86
Female	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.87	0.85
Male	0.91	0.90	0.91	0.90	0.91	0.91	0.90	0.89	0.87
African American	0.90	0.89	0.89	0.88	0.88	0.88	0.85	0.81	0.77
Asian	0.87	0.89	0.91	0.91	0.91	0.92	0.91	0.91	0.91
Native Hawaiian/Pacific Islander	0.90	0.89	0.90	0.90	0.90	0.91	0.89	0.87	0.85
Hispanic/Latino	0.90	0.89	0.89	0.88	0.89	0.89	0.87	0.83	0.79
American Indian or Alaskan	0.89	0.88	0.87	0.86	0.86	0.86	0.80	0.75	0.70
White	0.89	0.89	0.91	0.90	0.91	0.91	0.91	0.90	0.88
Multiple Ethnicities	0.90	0.90	0.91	0.90	0.91	0.91	0.90	0.89	0.87

9.6 RELIABILITY FOR SUBSCALES

Coefficient alpha internal consistency reliability estimates associated with the subscales for the 2015 operational forms are presented in Exhibit 9.6.1-9.2.6. As indicated in the Exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT.

Exhibit 9.6.1 Subscale Reliabilities – ELA Grades 3-11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 3	0.72	0.77	0.77
Grade 4	0.75	0.78	0.69
Grade 5	0.73	0.72	0.76
Grade 6	0.78	0.71	0.76
Grade 7	0.75	0.70	0.78
Grade 8	0.76	0.74	0.76
Grade 9	0.77	0.68	0.73
Grade 10	0.72	0.73	0.76
Grade 11	0.74	0.65	0.77

Exhibit 9.6.2 Subscale Reliabilities – Mathematics Grades 3-5

	Numbers & Operations-Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
Grade 3	0.65	0.71	0.83
Grade 4	0.73	0.58	0.83
Grade 5	0.78	0.68	0.82

Exhibit 9.6.3 Subscale Reliabilities – Mathematics Grades 6 & 7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
Grade 6	0.75	0.72	0.69	0.59
Grade 7	0.75	0.67	0.66	0.73

Exhibit 9.6.4 Subscale Reliabilities – Mathematics Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability and the Number System
Grade 8	0.79	0.69	0.71	0.53

Exhibit 9.6.5 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.83	0.77	0.54
Algebra II	0.66	0.59	0.67

Exhibit 9.6.6 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.60	0.63	0.35	0.72

10. VALIDITY

10.1 VALIDITY OF AZMERIT TEST SCORE INTERPRETATIONS

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support 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 may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

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

In 2010, Arizona adopted new academic content standards in English language arts (ELA) and math. The Arizona College and Career Ready Standards are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate. In spring 2015, the Arizona Department of Education (ADE) administered Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) to assess proficiency on the new Arizona College and Career Ready Standards for the first time. The AzMERIT measures English language arts (ELA) and mathematics in grades 3-8 and following completion of high school coursework in ELA Grade 9, ELA Grade 10, ELA Grade 11, Algebra I, Geometry, and Algebra II.

Because directly measuring student achievement against each benchmark in the Arizona College and Career Ready 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 ACCRS. To ensure that each student is assessed on the intended breadth and depth of the ACCRS, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark. Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzMERIT ELA and math test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels.

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. 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
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

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

In addition, the AzMERIT test delivery system provides a range of accessibility tools and accommodations for reducing construct irrelevant barriers to accessing test content for virtually all students. The range of accommodations provided in the online testing environment far exceed the typical accommodations made available in paper based test administrations. Exhibits 10.1.1-10.1.5 list the accommodations and accessibility supports currently available for students taking the AzMERIT assessments online. Paper test forms are available as an accommodation for students testing in online schools should the accommodations provided online not be sufficient to remove barriers to accessing test content. These include both large print and Braille forms, which are also available, for students who need them, in schools administering AzMERIT as a paper-based assessment.

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

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,

- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- 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 individualized education plan 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 10.1.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 10.1.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to 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	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	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	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	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 10.1.2.

Exhibit 10.1.2 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): scientific calculators are acceptable EOC (entire test): 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.

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

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

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section. Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student's individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact 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 10.1.3 Accomodations for Students with an Injury

Accommodation	Description
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	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. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Rest/Breaks	Student may take breaks during testing sessions to rest.

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

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

Exhibit 10.1.4 summarizes accommodations that may be provided for ELL and FEP students.

Exhibit 10.1.4 Allowable Accommodations for ELL and FEP Students

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

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 10.1.5, as designated in their IEP or 504 plan.

Exhibit 10.1.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.

Braille Test Booklet	Provide a paper Braille test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. PBT – Provide a Large Print test booklet. A PBT Large Print test booklet requires Adult Transcription on a regular size paper test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Paper Test Booklet	CBT – Provide a regular size paper test booklet for a student at a school administering the CBT. If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.

10.2 EVIDENCE BASED ON TEST CONTENT

Because the AzMERIT are designed to measure student progress toward achievement of the ACCRS the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the academic standards.

Alignment of content standards for spring 2015 AzMERIT was achieved through a rigorous test development process that proceeded from the content standards and referred back to those standards in a highly iterative test development process that included the state department of education, test developers, and educator and stakeholder committees. Because most of the items used to construct the spring 2015 AzMERIT test forms were drawn from Utah's Student Assessment of Growth and Excellence (SAGE) item banks, item development generally proceeded from the Utah Core Standards (UCS), and the review process described below was with respect to those standards. However, prior to form development activities for AzMERIT, these items were subjected to an additional round of reviews by content experts and educators in Arizona to ensure the alignment of item content to the ACCRS and the appropriateness of test content for Arizona students.

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. AzMERIT blueprints specify the range and depth with which each of the content strands and standards will be covered in each test administration. Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical.

With the desired alignment of AzMERIT's blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet the blueprints. Developing test forms is difficult because test blueprints can be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. And 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.

As detailed in chapter 4, all items included in the AzMERIT underwent a rigorous item development and review process, beginning with multiple rounds of review by test vendor content experts, then proceeding to client review, and finally to review by panels of educators that are assessing both the content of the item, and whether the item complies with bias or fairness and sensitivity guidelines.

Items successfully passing through this committee review process were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is 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 have to pass a three-stage review to be included in the final item pool from which operational forms are created. In the first stage of this review, a team of psychometricians reviews all flagged items to ensure that the data are accurate, properly analyzed, response keys are correct, and that there are no other obvious problems with the items.

Committees are then reconvened to re-evaluate flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance in association with the content the item assesses, the committees recommend that flagged items be either rejected or deemed eligible for inclusion in operational test administrations.

10.2.1 ARIZONA REVIEW

Although the standards frameworks embodied by the ACCRS and UCS share much in common, they are not the same. It is important to emphasize Arizona adopted a standards framework independently from that adopted by Utah, and the AzMERIT test blueprint, which specifies how achievement of the ACCRS is to be measured, was also developed independently of the SAGE blueprints and in fact there are marked differences. Thus, the SAGE item bank was not developed to measure achievement of the ACCRS as implemented in the AzMERIT test design. However, because SAGE was designed as a system of adaptive assessments, the item banks associated with each assessment are relatively large, and most AzMERIT blueprint elements could be met by items in the SAGE banks. There were, however, some AzMERIT specifications that could not be met by items in the SAGE banks, and in those cases items were pulled either from the AIMS item pool or from AIR's proprietary item pool. A total of twelve AIMS items were needed to construct the AzMERIT test forms: five items for each of Algebra I and Geometry forms, and two items on the Algebra II form. Nineteen AIR Core items were utilized across tests, including one item each on the grade three and eight mathematics tests, seven items on the geometry form, and 10 items on the Algebra II form.

Moreover, the SAGE items were not developed for administration in Arizona and had therefore not been reviewed by Arizona educators or stakeholders. ADE therefore instituted a review process to ensure that each item eligible for inclusion in an AzMERIT test form had been reviewed both by the Department and by Arizona educators. During these review meetings, Arizona educators evaluated each item that would be considered eligible for constructing the spring 2015 AzMERIT test forms. Only those items approved by the Arizona review committees were eligible for inclusion in AzMERIT test forms.

Arizona educators reviewed a total of 870 operational items at these meetings, 434 in ELA and 436 in math. In ELA, the committees rejected one passage and 8 additional items across grades, all of which were replaced with items and passages that the committees did approve. In math, the committees rejected 24 total items across all grades. These items were also replaced during the meetings, and the replacement items were reviewed and approved by the committees.

10.3 EVIDENCE FOR INTERPRETATION OF PERFORMANCE

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

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

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. For each assessment, panelists were provided the approximate location of performance standards for other important assessment systems. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

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

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

Based on the adopted performance standards, Exhibit 10.3.1 shows the estimated percentage of students meeting the AzMERIT proficient standard for each assessment in spring 2015. Exhibit 10.3.1 also shows the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. Exhibit 10.3.1 also presents the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2014 field test administration. As Exhibit 10.3.1 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.

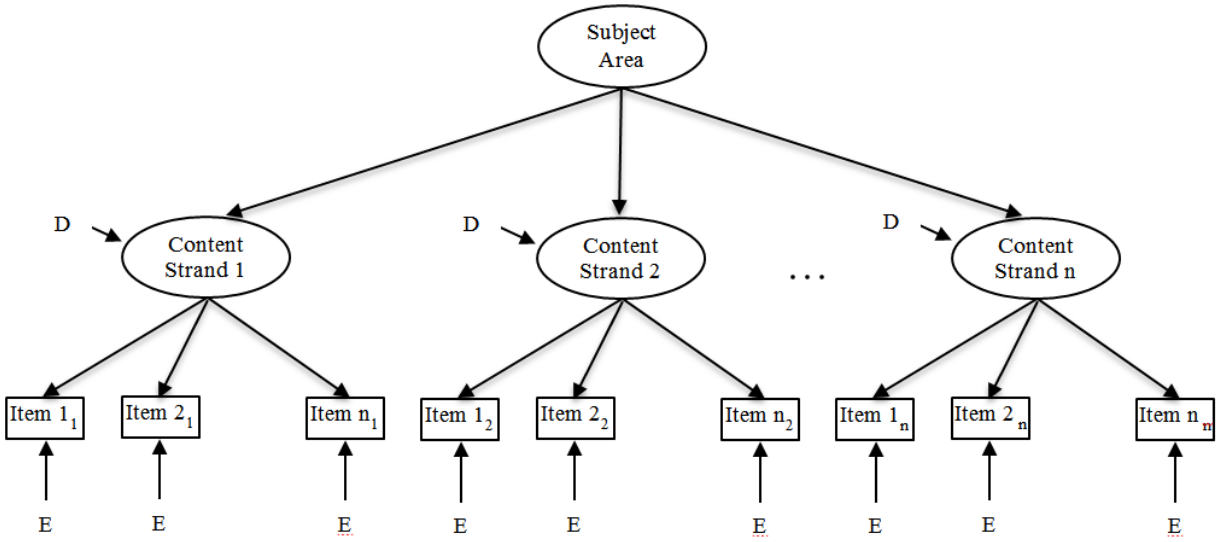
Exhibit 10.3.1 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

	Percent of Students Meeting Standard			
Grade/ Course	**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%

10.4 EVIDENCE BASED ON INTERNAL STRUCTURE

Arizona's 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 10.4.1. 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 10.4.1 Second-Order Structural Model for AzMERIT Assessments



Confirmatory factor analysis was used to evaluate the fit of this structural model to student response data from the AzMERIT test administrations. For each of test forms administered in spring 2015, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ2 statistic, with good model fit indicated by a non-significant χ2 statistic. The χ2 statistic is sensitive to sample size, however, so even well-fitting models will demonstrate highly significant χ2 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), the Root Mean Square of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR) were also used to evaluate model fit.

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

Exhibit 10.4.2 Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Index	Indication of Good Fit
CFI	≥ .95
TLI	≥ .95
RMSEA	≤ .05

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

10.4.1 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 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 10.4.1.1. All of the statistics indicate the general achievement factor model fit the data well. This pattern was true across all grades. The CFI and TLI values were generally equal to or greater than .95, and the RMSEA values were all below .05, indicating good fit for the base model.

Exhibit 10.4.1.1 Goodness-of-Fit for the AzMERIT ELA First-Order Model

First-Order Models			
Grade	CFI	TLI	RMSEA
3	0.934	0.931	0.047
4	0.949	0.946	0.033
5	0.966	0.964	0.039
6	0.955	0.953	0.043
7	0.974	0.972	0.037
8	0.964	0.963	0.048
9	0.924	0.921	0.039
10	0.948	0.945	0.042
11	0.928	0.925	0.034

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 10.4.1.2. All of the statistics indicate 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 equal to or greater than .95, with RMSEA values well below the .05 threshold used to indicate good fit.

Exhibit 10.4.1.2 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Second-Order Models			
Grade	CFI	TLI	RMSEA
3	0.958	0.956	0.038
4	0.970	0.969	0.025
5	0.980	0.979	0.030
6	0.973	0.972	0.033
7	0.983	0.982	0.029
8	0.980	0.979	0.036
9	0.962	0.960	0.028
10	0.972	0.970	0.031
11	0.949	0.947	0.029

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 10.4.1.3. We note that model fit for first-order model of general achievement are also very high and provide 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 general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 10.4.1.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	13560.7	3	$p < .001$
4	8460.9	3	$p < .001$
5	10944.7	3	$p < .001$
6	12019.8	3	$p < .001$
7	8848.6	3	$p < .001$
8	15590.1	3	$p < .001$
9	8896.6	3	$p < .001$
10	9084.7	3	$p < .001$
11	4412.8	3	$p < .001$

10.4.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 10.4.2.1. Across all grades, results indicate the second-order models posited by the AzMERIT assessments fit the data well. The CFI and TLI values were all .97 to .99, RMSEA values are all .03 or lower. SRMR values between .02 and .04, well below the values used to indicate good fit.

Exhibit 10.4.2.1 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Second-Order Models			
Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.99	0.99	0.02
6	0.98	0.98	0.03

7	0.99	0.99	0.02
8	0.99	0.99	0.02
9	0.98	0.98	0.02
10	0.98	0.97	0.02
11	0.98	0.98	0.02

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 10.4.2.2. The chi-square difference test shows that across grade levels, the DOK-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 10.4.2.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	21402.6	4	*p* < .001
4	12053.6	4	*p* < .001
5	17102.9	4	*p* < .001
6	18192.1	4	*p* < .001
7	16351.4	4	*p* < .001
8	25454.7	4	*p* < .001
9	14989.3	4	*p* < .001
10	14920.9	4	*p* < .001
11	8075.1	4	*p* < .001

10.4.3 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 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 10.4.3.1. All of the statistics indicate the general achievement factor model fit 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 unidimensional model.

Exhibit 10.4.3.1 Goodness-of-Fit for the AzMERIT Mathematics First-Order Model

	First-Order Models		
grade	CFI	TLI	RMSEA
3	0.975	0.973	0.027
4	0.976	0.975	0.024
5	0.976	0.975	0.026
6	0.975	0.973	0.023
7	0.982	0.981	0.021
8	0.969	0.967	0.026
Algebra I	0.976	0.975	0.023
Algebra II	0.973	0.971	0.021
Geometry	0.986	0.985	0.018

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 10.4.3.2. The models show very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their .05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 10.4.3.2 Goodness-of-Fit for the AzMERIT Mathematics Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.979	0.978	0.024
4	0.978	0.977	0.024
5	0.978	0.977	0.025
6	0.976	0.975	0.023
7	0.983	0.982	0.021
8	0.970	0.969	0.026
Algebra I	0.978	0.977	0.022
Algebra II	0.974	0.972	0.020
Geometry	0.987	0.986	0.017

The results of the comparison between the second-order, strand-based model and the first-order, general achievement model are presented in Exhibit 10.4.3.3. Again, model fit for the general achievement first-order model is very high, providing evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether knowledge of the depth of knowledge level of items provides information beyond that provided by the more general model. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 10.4.3.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	3225.0	3	$p < .001$
4	1326.3	3	$p < .001$
5	1427.0	3	$p < .001$
6	1036.2	4	$p < .001$
7	559.8	4	$p < .001$
8	1039.3	4	$p < .001$
Algebra I	750.9	3	$p < .001$
Algebra II	246.5	3	$p < .001$
Geometry	269.7	4	$p < .001$

10.4.4 MATHEMATICS DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 10.4.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their.05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 10.4.4.1 Goodness-of-Fit for the AzMERIT Mathematics Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.99	0.99	0.02
Geometry	0.97	0.97	0.02

The results of the comparison between the second-order, DOK-based model and the first-order, general achievement model are presented in Exhibit 10.4.4.2. 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 10.4.4.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	331.4	3	*p* < .001
4	309.5	3	*p* < .001
5	14.9	3	*p* < .001
6	14.5	3	*p* < .001
7	236.6	3	*p* < .001
8	79.2	3	*p* < .001
Algebra I	20.1	3	*p* < .001
Algebra II	26.4	3	*p* < .001
Geometry	20.9	3	*p* < .001

10.5 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 10.5.1-10.5.6. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability. The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = r_{xy} / SQRT(r_{xx} * r_{yy}),$$

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 underling construct.

Exhibit 10.5.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 3 to 11

		Observed Correlation		Disattenuated Correlation	
Grade	**Subscale**	**Informational Text**	**Literature**	**Informational Text**	**Literature**
3	Literature	0.72		0.97	
	Writing & Language	0.64	0.68	0.86	0.88
4	Literature	0.72		0.94	
	Writing & Language	0.61	0.63	0.85	0.86
5	Literature	0.69		0.95	
	Writing & Language	0.65	0.62	0.87	0.84
6	Literature	0.70		0.94	
	Writing & Language	0.68	0.64	0.88	0.87
7	Literature	0.69		0.95	
	Writing & Language	0.68	0.66	0.89	0.89
8	Literature	0.70		0.93	
	Writing & Language	0.66	0.66	0.87	0.88
9	Literature	0.66		0.91	
	Writing & Language	0.62	0.56	0.83	0.79
10	Literature	0.68		0.94	
	Writing & Language	0.63	0.62	0.85	0.83
11	Literature	0.66		0.95	
	Writing & Language	0.65	0.60	0.86	0.85

Exhibit 10.5.2 Subscale Intercorrelations– Mathematics Grade 3 to 5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.67		0.99	
	OAT_NBT	0.69	0.79	0.94	1.00
4	MDG	0.67		1.00	
	OAT_NBT	0.77	0.71	0.99	1.00
5	MDG	0.70		0.96	
	OAT_NBT	0.79	0.72	0.99	0.96

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

Exhibit 10.5.3 Subscale Intercorrelations– Mathematics Grade 6 & 7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.73			0.99		
	RP	0.71	0.71		0.99	1.00	
	GSP	0.65	0.65	0.63	0.98	1.00	1.00
7	NS	0.73			1.00		
	RP	0.74	0.69		1.00	1.00	
	GSP	0.73	0.67	0.71	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 10.5.4 Subscale Intercorrelations– Mathematics Grade 8

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	G	SPNS	EE	G	SPNS
8	Functions (F)	0.74			1.00		
	Geometry(G)	0.73	0.69		1.00	1.00	
	SPNS	0.63	0.61	0.60	1.00	1.00	1.00

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

Exhibit 10.5.5 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.77		1.00	
	Statistics	0.64	0.63	1.00	1.00
Algebra II	Functions	0.66		1.00	
	Statistics	0.71	0.65	1.00	1.00

Exhibit 10.5.6 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.63			1.00		
	Modeling with Geometry (MG)	0.65	0.65		1.00	1.00	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.65	0.67	0.70	1.00	1.00	1.00

Note: CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations;

10.6 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.

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

Exhibit 10.6.1 shows the correlations between student test scores on the statewide AzMERIT assessment with corresponding test scores on a district-wide administration of the 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 10.6.1 Correlations between AzMERIT and Locally Administered NWEA Test Scores

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

Exhibit 10.6.2 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 10.6.2 Correlations between AzMERIT and Locally Administered NWEA Test Scores

	AzMERIT ELA/NWEA Math		AzMERIT Math/NWEA Reading	
Grade	**Sample Size**	**Correlation**	**Sample Size**	**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, 2016). These researchers reported the mean correlations between 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 grade levels 3 through 8. These results likewise show good convergence between AzMERIT and other locally administered assessments purporting to measure the same constructs.

10.7 SUMMARY

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

11. CONSTRUCTED-RESPONSE SCORING

The AzMERIT assessments in ELA and mathematics utilize a variety of item types to assess students' mastery of the Arizona College and Career Ready Standards (ACCRS). ADE leverages AIR's item scoring technology to machine-score student responses to most items, including traditional selected-response (multiple choice) item types, and machine-scored constructed response (MSCR) items types. These item types are designed to capture and score a variety of response types, such as graphing, drawing or arranging graphic regions, selecting or rearranging sentences or phrases within passages, or entering equations or words, allowing AzMERIT items to assess a wide range of student knowledge and skills. In most cases, constructed response machine-scored items that are developed for online administration are adapted for paper and responses are captured in a format that allows machine-scoring.

In addition, some constructed response items are scored by human raters, also referred to as hand-scored. The writing components of ELA assessments consist of one essay prompt; in 2015, all writing essay responses were hand-scored. In addition, mathematics assessments that are administered on paper included a small number of items that are scored by human raters, generally items that required students to produce an equation. The reading components of the ELA assessments, both online and paper, and the mathematics assessments administered online are machine-scored in their entirety.

AIR partners with Measurement, Incorporated (MI) to fulfill all hand-scoring requirements. AIR provides the automated electronic scoring and MI provides all hand-scoring 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 reliability and validity of item scores.

11.1 MACHINE SCORING

As part of the item development process for machine-scored item types other than multiple-choice, a rubric validation process is 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. However, because the items had not previously been administered to Arizona students, the rubric validation process was conducted for operational test items following the spring 2015 administration of AzMERIT. 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, and randomly selected responses from the population.

In its base year of 2015, in addition to field-test items, all operational machine-scored AzMERIT items underwent an additional rubric validation process prior to the final scoring of items and student tests. Generally, rubric validation will be conducted following the initial field test administration of each item.

11.2 HAND-SCORING

Measurement Incorporated (MI) partners with AIR to fulfill all hand-scoring needs for AzMERIT. For items that are scored by human raters, each student response was scored by at least one reader (Reader 1). Ten percent of all responses receive a second reading (Reader 2) for the purpose of monitoring and maintaining sufficient inter-rater reliability, as discussed below. 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 was based on the resolution score, when present, or else on the initial read.

For 2015, MI was provided with all materials required for scoring, including rubrics, anchor papers, and training materials, as developed during administration in Utah. For hand-scored mathematics items, MI was provided with a list of acceptable responses for each item. No additional rangefinding activities were necessary.

11.2.1 HANDSCORING PROCESS

The Arizona hand-scoring efforts used MI's image-based system, VSC™. VSC is composed of two primary subsystems: VSC Capture™ and VSC Score™. MI used the VSC Score subsystem. 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 percent of items scored)
- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
    - Accommodating paper-based scores when images are of insufficient quality
    - Outputting data easily into MI's score reporting applications

Paper-pencil tests were scanned into VSC. The images were displayed to trained and qualified scorers who score the images online. Scorers had access only to those items for which they had been qualified to score. Online assessment responses were also converted into images and displayed in an identical manner to paper-pencil student responses using the same VSC scoring application.

When logging onto VSC Score, scorers were presented with a scoring set, which is the images-scoring equivalent of a physical packet of student responses. The scoring set was generated by randomly selecting student responses from the pool of non-scored student responses. The resultant set of responses was checked out to the scorer. The images they receive 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. Condition codes were assigned to student responses by scoring leadership per Arizona specifications, such as a code noting that the response was left blank; condition codes are summarized in Exhibit 11.2.1.1.

Exhibit 11.2.1.1 Handscoring Non-Scorable Condition Codes

Code	Description
A	Blank
B	Undecipherable or illegible
C	Non-English (not applicable for math responses)
D	Off-topic (response not related to task, response that student refuses to answer item, response that student does not understand given item)
E	Original content is too limited to evaluate (response is only a copy or part of item stem or is an unrelated drawing or marking)

In situations where the scorer was unable to score an item because there may be no response, or the response is not meaningful or readable for a number or reasons, the scorer flagged the item and it was routed to a scoring director for a condition code assignment. These codes were determined in the scoring rules and were assigned to the response by the scoring director.

After scoring all of the responses in a set, the scorer reviewed any of the responses and modified the scores before committing them to the system. Once the scores had been committed, the set was checked in and responses are 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.

11.2.2 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 responses.

11.2.3 HAND-SCORING 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 hand-scores, MI puts scoring directors, team leaders, and scorers through a rigorous screening and training process.

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 through the use of validation sets that are incorporated into the ongoing live scoring.

All of the Arizona training materials provided for the initial operational ELA scoring were scoring guides composed of anchor responses as well as training, qualifying, and recalibration sets approved for use by the state as a result of approval of existing documentation from AIR's Item Tracking System (ITS), which is the repository for all item attributes, including scoring rubrics. In subsequent years, 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.

MI's scoring directors ensured ELA scoring guides had detailed annotations to explain how the scoring criteria are applied to each response's specific features and why it should be assigned a particular score. The approach was to focus on the precise scoring rationale and 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 are allowed to begin live scoring. MI worked with Arizona to finalize the number of training and qualifying sets for each item and determine the appropriate qualifying percentage. All scoring decisions and supplemental responses were submitted more than one month before the start of scoring for review and approval by the state.

MI employs an online training interface for the scorer training in all scoring sites. Using the online training interface allows scorers to be engaged and communicate with scoring leadership throughout the training process. The interface allows for scorer viewing of each item, passages and other associated item stimuli, and scoring guides.

MI's online training interface also allowed observers from Arizona to witness training in real-time. Through the use of TurboMeeting software, observers were able to visually see the responses being trained and discussed as each training set progressed. Observers were also allowed to hear the training through the software's audio function. In addition to observing the training of leadership virtually, representatives from Arizona also traveled to individual scoring sites to observe training in-person. This allowed Arizona to observe MI's training techniques and interact with project leadership. The State was able to provide additional guidance on scoring rationale during the training process. These observations allowed MI to further ensure reliability in the hand-scoring 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 following our vigorous training process. Each staff member was also expected to maintain a consistent level of scoring quality throughout the scoring effort or he or she was released from the project. MI continually monitored performance in order to guarantee scoring accuracy.

For math, MI trained scorers to hand-score a limited number of mathematics items from the paper assessment that could not be machine-scored. Scoring leadership reviewed all hand-scored mathematics items prior to training. Using the scoring rubrics provided from ITS, leadership provided feedback and questions to both AIR and Arizona to ensure consistency in training methodology. Mathematics items were trained and scored individually with the use of the provided scoring rubrics. Qualified mathematics scorers received training that included all possible answers to each individual item.

Mathematics hand-scoring was monitored in the same was 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.

11.2.4 RATER EFFECTS

To ensure reliability, MI tracked scorer statistics through the VSC system, and reported the number of first and second reads for each item. The reports also break down perfect, adjacent, and nonadjacent agreement rates per reader and team for each item throughout the scoring project.

The rater agreement reports show percentages of perfect agreement (Equal), adjacent agreement (Adj Low or Adj High) and nonadjacent agreement (Low or High). IRR reports also detail mismatched scores when there is a difference involving nonscorable condition codes, (Mismatch CC) or a nonscorable/scorable mix (MM CC/Score).

Exhibit 11.2.4.1 shows the rater agreement for each of the writing prompts administered on the AzMERIT.

Exhibit 11.2.4.1 ELA Writing Prompt Rater Agreement Report

				Rater Agreement						
		Total	**Second**		**Adj**		**Adj**		**Mismatch**	**MM**
Grade	**Dimension**	**Reads**	**Reads**	**Low**	**Low**	**Equal**	**High**	**High**	**CC**	**CC/Score**
3	Purpose/Organization	96,441	18,121	0.4	7	**84.9**	7	0.4	0	0.2
	Evidence/Elaboration	96,441	18,121	0.7	7	**84.3**	7	0.7	0	0.2
	Conventions	96,441	18,121	0.3	6.4	**86.2**	6.5	0.3	0	0.2
4	Purpose/Organization	94,767	17,271	0.5	7	**85.2**	6.9	0.5	0	0
	Evidence/Elaboration	94,767	17,271	0.4	5.9	**87.3**	5.9	0.4	0	0
	Conventions	94,767	17,271	0.1	6.3	**87.3**	6.3	0.1	0	0
5	Purpose/Organization	94,325	17,239	0.4	8.2	**82.8**	8.1	0.4	0	0
	Evidence/Elaboration	94,325	17,239	0.4	8.1	**83**	8.1	0.4	0	0
	Conventions	94,325	17,239	0.2	8.7	**82.1**	8.7	0.2	0	0
6	Purpose/Organization	93,915	17,132	0.5	14	**70.9**	14	0.5	0	0.1
	Evidence/Elaboration	93,915	17,132	0.6	14.1	**70.5**	14.1	0.6	0	0.1
	Conventions	93,915	17,132	0.3	12.4	**74.5**	12.4	0.3	0	0.1
7	Purpose/Organization	92,301	16,868	0.7	13.4	**71.8**	13.4	0.7	0	0
	Evidence/Elaboration	92,301	16,868	0.7	13.6	**71.3**	13.6	0.7	0	0
	Conventions	92,301	16,868	0.4	9.5	**80.2**	9.5	0.4	0	0
8	Purpose/Organization	92,564	16,653	0.3	12	**75.1**	12.2	0.3	0	0.1
	Evidence/Elaboration	92,564	16,653	0.4	12.8	**73.4**	13	0.4	0	0.1
	Conventions	92,564	16,653	0.2	10.4	**78.8**	10.5	0.2	0	0.1
9	Purpose/Organization	89,568	16,389	0.2	3.6	**92.5**	3.6	0.2	0	0
	Evidence/Elaboration	89,568	16,389	0.3	3.5	**92.5**	3.5	0.3	0	0
	Conventions	89,568	16,389	0.2	3.5	**92.6**	3.5	0.2	0	0
10	Purpose/Organization	81,394	14,910	0.4	13.1	**73.1**	13.1	0.4	0	0
	Evidence/Elaboration	81,394	14,910	0.6	12.2	**74.5**	12.1	0.6	0	0
	Conventions	81,394	14,910	0.1	10.5	**78.7**	10.5	0.1	0	0
11	Purpose/Organization	68,858	12,937	0.3	7.1	**85.2**	7.1	0.3	0	0
	Evidence/Elaboration	68,858	12,937	0.2	6.8	**85.9**	6.8	0.2	0	0
	Conventions	68,858	12,937	0.1	7.4	**85.1**	7.4	0.1	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

12. QUALITY CONTROL PROCEDURES

Quality assurance procedures are enforced through all stages of AzMERIT test development, administration, and scoring and reporting of results. This section describes quality assurance procedures associated with

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

Because quality assurance procedures pervade all aspects of test development, we note that discussion of quality assurance procedures is not limited to this section, but is also included in sections describing all phases of test development and implementation.

12.1 QUALITY ASSURANCE IN TEST CONSTRUCTION

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

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

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

The FormBuilder also plots the distribution of item difficulties, both classical and IRT indices, to both flag extremely easy or difficult items and to ensure that the distribution of item difficulties is consistent across test forms and with the bank. As test developers build forms, FormBuilder generates TCCs, TICs, and CSEMs for the reference (previously used) form and the target (new) form(s) on the screen. The TCCs and SEMCs are plotted using a different color trace line for each prototype form. Using FormBuilder, our content specialists select test items that match the blueprint and are of appropriate difficulty. Beginning with content considerations and supplementing those considerations with statistical considerations, AIR creates alternate, parallel test forms by comparing TCCs for the form that is being created with TCCs from previous forms. To the degree that the TCC for the total test is the same as for previous tests, the raw score required for meeting any performance standard will remain as close to the same as it was on previous forms.

When submitting test forms for review by ADE, AIR produces a form evaluation workbook that includes an evaluation summary checklist, as well as summary statistics and test characteristic graphs.

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 is reviewing a test form for content, in addition to making sure all the benchmark/indicator item requirements are met, test developers must also make 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.

Once the items and passages for the form have been 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 lettered with A, B, C, and D.
- All graphics are consistently presented.
- All tables and charts have titles and are consistently formatted.
- The number of the answer choice letters should be approximately equal across the form.
- The answer key should be 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 have been 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 form is approved by the psychometric team, the test developer uploads the item list into FormBuilder. After operational forms were defined in FormBuilder, all bookmaps (test maps), key files, and conversion tables were produced directly from FormBuilder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents were 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.

Bookmaps include any item attribute stored in ITS, so that in addition to form-level attributes such as test administration and item position, item attributes such as content standard, benchmark, indicator, complexity, item release status, point value, weight, keyed response, and more are included in the bookmap. The bookmap feature in FormBuilder was customized to AzMERIT.

As a further layer of quality assurance for printed test booklets, both during the blueline production phase prior to printing and again following the final printing of all test forms, two AIR technical team staff members independently took all test forms. Responses to the test forms were compared to the answer keys for each form to confirm the accuracy of scoring keys. In addition, the printed forms were compared against ITS and FormBuilder for content and item ordering to ensure that no changes to the form were introduced prior to printing.

12.2 QUALITY ASSURANCE IN TEST PRODUCTION

The production of computer-delivered assessments involves two distinct types of products, each of which follows an appropriate quality assurance process:

1. Content for online delivery shares some processes with paper versions, but also requires additional, unique steps.

2. Online test delivery software must deliver the content reliably (and, with the right tools, the accommodations, layouts, etc.).

The AzMERIT test delivery system also has a real-time quality-monitoring component built in. As students are administered assessments, data flow through the test delivery system's 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.

12.2.1 PRODUCTION OF CONTENT

While the online workflow requires some additional steps, it actually removes a substantial amount of work from the time critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. A paper form must await final forms construction before blackline proofs can show how the item will look in the booklet. Online, the appearance of the item screen can be known with certainty before the final test form is ever constructed. This characteristic of online forms enables us to lock down the final presentation of each item well before forms are constructed. In turn, this moves the final blueline review of items much earlier in the process, removing it from the critical path.

The production of computer-based tests includes five key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.

2. Forms are finalized using the process described in Section 4.6, and final forms are approved in our FormBuilder software.

3. Complete test packages are created with our test packager, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.

4. Forms are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the field.

5. The final system is deployed to a staging environment accessible to ADE for user acceptance testing and final review.

12.2.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The Item Tracking System (ITS) integrates directly with the test delivery system (TDS) display module, and displays each item exactly as it will appear to the student. This process is called web preview, and web preview is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent web preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will see it. Layout changes can be made after this process in two ways:

1. Content can be revised and re-approved for web display.

2. Online style sheets can change to revise the layout of all items on the test.

Both of these processes are subject to strict change control protocols to ensure that accidental changes are not introduced. Below, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the benefit of allowing final layout review much earlier in the process, reducing the work that must be done during the very busy period just before tests go live.

12.2.3 APPROVAL OF FINAL FORMS

Section 4.6.1 describes our process for constructing operational test forms, including the approval of test forms by ADE. The forms are built in FormBuilder (a component of our ITS), and upon approval, they are ready for preliminary publication.

12.2.4 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.

12.2.5 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.

12.2.6 USER ACCEPTANCE TESTING AND FINAL REVIEW

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the test delivery system serves both a software evaluation and content approval role. The UAT period provides the Department with an opportunity to interact with the exact test with which the students will interact.

12.2.7 FUNCTIONALITY AND CONFIGURATION

The items, both in themselves and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of our system:

1. Testing on the system itself to ensure function, performance, and capacity

2. Capacity planning

3. Continuous monitoring

AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

12.3 QUALITY ASSURANCE IN DOCUMENT PROCESSING

12.3.1 SCANNING ACCURACY

When test documents are scanned, a quality control sample of documents consisted of ten test cases per document type (normally between five and six hundred 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 testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner, editing process (validation and data correction), transfer to the project database, and scoring were accurate according to the reporting rules provided by ADE.

12.3.2 QUALITY ASSURANCE IN EDITING AND DATA INPUT

At a minimum, MI implemented, maintained, and constantly updated the following quality assurance controls:

- Score key verification Post analysis of item keys
- Response analyses to determine score frequency distribution by item verification of bank values of item statistics
- Live data checks to verify that data/results conform to approved specifications comprehensive software test plan
- Double data entry correction process to verify student response and demographic information report data verification
- Reviewed and proofread all electronic and printed report deliverables

MI utilized a double data correction process to achieve the highest level of quality and accuracy in Arizona CBT and PBT assessment student data. Data correction operators used their sophisticated Data Inspection, Correction and Entry (DICE) application, which retrieved flagged data records and highlighted the problem field on a computer screen for resolution. The operator compared the highlighted data on the answer document template, retrieved the original document for resolution, and made any necessary correction.

After an operator corrected a flagged record, the same flagged record was routed to a second data correction operator who repeated the data correction process. After a flagged record was edited by two independent operators, the data correction application checked to verify that both operators made identical corrections. If the two corrections differed, the record was routed to a supervisor for a third and final resolution. Agreement rate statistics were generated for the individual data correction operators, allowing the supervisor to monitor their job performance. This process continued until all flagged records are examined and resolved.

Thorough training significantly improves the accuracy of data correction. To ensure that goal, MI trained their data correction staff on the use of the data correction application and on the specific validation errors and procedures associated with the specific project. Practice sets generated by the programming staff allowed data correction staff to learn on samples of answer documents that simulated the kinds of errors they were expected to correct for the actual assessment prior to actually processing live data. Additionally, each user had an electronic copy of the data correction user's guide for reference.

MI developed verification routines as part of their standard data validation to detect duplicate student tests in the assessment, whether in a single LEA (local educational agency) or across LEAs, and student moves between schools. MI staff then worked closely with ADE to resolve these discrepancies through processes called Barcode Processing and Tested Roster. These processes and the business rules governing them were described in a set of requirements developed in conjunction with ADE. They involved direct data transfer in several steps between the MI and ADE databases, with the goal of ensuring that each student final report was sent to the school where the test was taken, that it had accurate demographic data, and that the test reported was the correct test per the business rules.

12.4 QUALITY ASSURANCE IN DATA PREPARATION

AIR's test delivery system has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. For example, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the sorts of checks that are done for data review, but (a) they are done on operational data and (b) they are conducted in real time so that our psychometricians can catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM to the database of record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to ADE and their quality assurance contractor. AIR psychometricians ensure that data in the extract files matches the DoR prior to delivery to ADE.

12.5 QUALITY ASSURANCE IN TEST FORM EQUATING

Item information necessary for statistical and psychometric analyses is provided to ADE and HumRRO, ADE's independent quality assurance contractor, prior to test administration. Item information is published as part of the configuration of the online assessment system that AIR employs for administering, scoring, and reporting test scores. Information contained in these workbooks includes, but is not limited to, unique item ID used for item tracking, test form ID, location on the test form, correct answer, item difficulty, and information about the strand, standard, and benchmark each item measures. These item files are used in quality control checks of the assessment data scoring and analysis.

To ensure security, all data is shared using ADE's SFTP site.

Prior to operational work, AIR produces simulated datasets for the purpose of testing software and analysis procedures, and shares with 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 resolution of questions that arise, analysis specifications are finalized.

12.6 QUALITY ASSURANCE IN SCORING AND REPORTING

12.6.1 QUALITY ASSURANCE IN HAND SCORING

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.

VSC provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer's performance every day to ensure that he or she is on target and conduct one-on-one retraining sessions when necessary. MI's quality assurance procedures allow scoring staff to identify struggling scorers very early and begin retraining immediately.

We monitor their scoring intensively to ensure 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 as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be culled 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 are available that show both daily and cumulative (project-to-date) data. 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 ARIZONA DEPARTMENT OF EDUCATION

ADE also directly observes MI activities, both onside and virtually. MI provides virtual access to the training activities through the online training interface, as well as on-site training and on-site scoring. Arizona monitors the scoring process through the Client Command Center (CCC) with access to view and run specific reports during the scoring process. This ability to attend the training, qualification, and initial scoring virtually provides Arizona the most efficient use of oversight by reducing the travel requirements for on-site attendance for ADE staff.

IDENTIFYING, EVALUATING, AND INFORMING THE STATE ON ALERT PAPERS

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the examinee. 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.

12.6.2 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 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 also 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, we merge item response data with the demographic information taken either from previous year assessment data, or if current year enrollment data is available by the time simulated data files are created, we can verify online reporting using current year testing information. By populating the simulated data files with real school information, it is possible to verify that special 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 the Department each year. Although 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 test administration window, a series of Quality Assurance Reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. In the context of adaptive test administrations, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to specifications.

An additional set of cheating analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school level. The quality assurance reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the test window to ensure that test administrations conform to blueprint and 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 12.6.2.1 presents an overview of the quality assurance (QA) reports.

Exhibit 12.6.2.1 Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	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)
Blueprint Match Rates	To monitor unexpected low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

ITEM ANALYSIS REPORT

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For multiple-choice items, the proportion of students selecting each of 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 selected-response items and polyserial correlations for polytomous constructed response, performance, and technology 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.

We begin by defining $p_{ij} = pr(z_{ij} = 1)$, representing the probability that student i responds correctly to item j (z_{ij} represents the student's score on the item). For selected-response items we use the 3PL IRT model to calculate the expected score on item j for student i with estimated ability $\hat{\theta}$ as

$$E(z_{ij}) = c_j + (1 - c_j)\frac{\exp\left(Da_j(\hat{\theta}_i - b_j)\right)}{1 + \exp\left(Da_j(\hat{\theta}_i - b_j)\right)}$$

For constructed response, performance, or technology items, using the Generalized Partial Credit model, the expected score for student i with estimated ability $\hat{\theta}$ on an item j with a maximum possible score of K_j is calculated as

$$E(z_{ij}) = \sum_{l=1}^{K_j} \frac{l\exp\left(Da_j \sum_{k=1}^{l}(\hat{\theta}_i - b_{j,k})\right)}{1 + \sum_{m=1}^{K_j} \exp\left(Da_j \sum_{k=1}^{m}(\hat{\theta}_i - b_{j,k})\right)}$$

For each item *j*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij})$$

The statistic δ is aggregated across students of different abilities for each item,

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

The report can be configured to report all items or flag and report only those items where the fit index is above a given threshold (e.g., items could be flagged when

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > 1.96$$

where $\bar{\delta}_j = \frac{SD(\delta_{ij})}{\sqrt{n}}$.

CHEATING ANALYSIS

Another component in the suite of QA reports is geared toward detection of possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies. The cheating detection component of the QA reports are described in detail in Section X.X. 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 can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

12.6.3 REPORTING

Scores for online assessments are assigned by automated systems in real time. For machine scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field-testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The hand-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Hand-scored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the quality monitor (QM) system where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies client-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 there is only one place where the “official” record is stored. Only after scores have passed the QM checks and are uploaded to the DoR are they passed to the Online Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Online Reporting System until it passes all of the QM system’s validation checks.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

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

Council of Chief State School Officers. 2014. Criteria for Procuring and Evaluating High-Quality Assessments. Retrieved from www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf.

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*(1), 67–86.

Estrada, S., Burnham, C., Feld, J. K., Bergan, J. R., & Bergan, J. R. (2016). *Research Commentary: Can Local Assessment Data be Successfully Used as Part of an Arizona A-F Accountability System?* Assessment Technology Incorporated.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, *21*(3), 187-206.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, *4*, 231-246.

Karkee, T., Lewis, D. M., Hoskens, 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). A user's guide to WINSTEPS: Rasch-Model Computer Program. Chicago, IL: MESA Press.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measuremen*t, *16*(4), 247-260.

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.

Masters, G. N. (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*(2), 443-451.

McLaughlin, D., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005). Using State Assessments to Impute Achievement of Students Absent from NAEP: An Empirical Study in Four States. Washington, DC: American Institutes for Research. Retrieved from www.air.org/files/McLaughlin_AbsentStudents.pdf.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*, 115-135.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Phillipine Statistician*, *52*(1–4), 81–92.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265-276.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10.

Scott, L. (2015). *Analysis of Mode Comparability of AzMERIT's Online and Paper Administrations for Spring 2015.* Unpublished manuscript, Arizona Department of Education.

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.

Webb, N. L. (2002). *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States.* Washington, DC: Council of Chief State School Officers.

Webb, N. (2005, April). Issues related to judging the alignment of curriculum standards and assessments. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909-921.

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

English Language Arts Assessment Blueprint

Grade 3

Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 4

Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 5

Strands	Min	Max
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 6

Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 7

Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 8

Strands	Min	Max
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 9

Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 10

Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 11

Strands	Min	Max
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Listening Standards will only be assessed on the computer-based assessment.

In Grades 3-5 some items in the Reading and Language Strands will also be aligned to the standards for Reading: Foundational Skills.

Percentage of Points by Depth of Knowledge Level

Grade	DOK Level 1	DOK Level 2	DOK Level 3	DOK Level 4
3-11	10%-20%	50%-60%	15%-25%	16%-19% (Writing)

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

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	25%	29%
Expressions and Equations	29%	33%
Geometry, Statistics and Probability	17%	21%

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	32%	36%
Functions	21%	25%
Geometry	23%	27%
Statistics and Probability and The Number System	15%	19%

Algebra I

Conceptual Categories	Min.	Max.
Algebra	40%	44%
Functions	36%	40%
Statistics	17%	21%

Geometry

Domain	Min.	Max.
Congruence	23%	27%
Similarity, Right Triangles and Trigonometry	27%	31%
Circles , Geometric Measurement and Geometric Properties with Equations	23%	27%
Modeling with Geometry	17%	21%

Algebra II

Conceptual Categories	Min.	Max.
Algebra	34%	38%
Functions	32%	36%
Statistics	27%	31%

Percentage of Points by Depth of Knowledge Level

Grade	DOK Level 1	DOK Level 2	DOK Level 3
3-11	10%-20%	60%-70%	12%-30%

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

Revised by ADE on 8/19/15

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

Appendix B.1-- Standard Errors of Measurement at Performance Level Cuts – ELA

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 ELA	9.61	8.87	9.40	11.76	9.68
Grade 4 ELA	9.50	8.62	9.46	12.74	9.55
Grade 5 ELA	9.82	8.87	9.37	11.43	9.48
Grade 6 ELA	9.48	8.83	9.83	13.34	9.62
Grade 7 ELA	9.32	8.67	9.62	13.22	9.46
Grade 8 ELA	9.16	8.69	9.66	12.88	9.51
Grade 9 ELA	9.45	8.83	9.59	11.74	9.43
Grade 10 ELA	9.00	8.21	8.76	11.27	9.03
Grade 11 ELA	9.41	8.79	9.27	11.17	9.47

Appendix B.2-- Standard Errors of Measurement at Performance Level Cuts – Mathematics

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 Math	10.62	9.99	11.91	18.52	12.09
Grade 4 Math	10.74	10.04	11.57	17.28	11.60
Grade 5 Math	11.49	9.87	10.55	15.92	11.40
Grade 6 Math	10.84	9.70	10.48	14.82	10.95
Grade 7 Math	10.86	9.91	10.70	14.11	11.11
Grade 8 Math	11.57	9.95	10.08	12.94	11.10
Algebra I	11.00	9.50	9.73	13.66	10.66
Geometry	12.84	10.52	10.19	11.51	11.55
Algebra II	13.77	10.96	10.43	11.34	12.25

Appendix C.1 -- Number of Students in Each Assessment by Gender and Ethnicity – ELA Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
Female	21240	21344	20614	19721	19252	19418	13067	12179	10398
Male	22016	21969	21749	20740	20076	20168	13520	12808	11113
Unknown	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
African American	2221	2194	2334	2229	2137	2198	1373	1251	1037
Asian	1072	1085	1064	951	909	953	652	622	571
Native Hawaiian/Pacific Islander	147	107	99	98	84	95	51	70	58
Hispanic/Latino	19586	19364	18764	17984	17580	17556	9700	9153	7363
American Indian or Alaskan	2732	2762	2725	2439	2193	2145	1786	1749	1456
White	16025	16504	16250	15733	15466	15625	11885	11099	9505
Multiple Ethnicities	1195	1068	920	809	748	778	612	595	497

Appendix C.2 -- Number of Students in Each Assessment by Gender and Ethnicity – ELA Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
Female	21365	20710	20724	21887	21530	21466	25067	22328	17945
Male	21811	21272	21850	22199	21879	21877	25566	22507	18118
Unknown	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
African American	2129	2050	2122	2192	2328	2260	2588	2255	1841
Asian	1262	1323	1446	1425	1392	1412	1616	1454	1259
Native Hawaiian/Pacific Islander	135	113	114	117	126	139	118	139	100
Hispanic/Latino	18992	18196	18361	18622	18366	18425	22879	20072	15598
American Indian or Alaskan	2140	2212	2115	2127	2347	2210	2174	1918	1575
White	16922	16737	17144	18383	17696	17801	19669	17650	14271
Multiple Ethnicities	1238	1030	988	948	920	872	877	738	512

Appendix C.3 -- Number of Students in Each Assessment by Gender and Ethnicity – Mathematics Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
Female	21349	21468	20695	19807	19553	19608	15157	11742	10466
Male	22191	22116	21866	20842	20467	20393	16108	12167	10652
Unknown	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
African American	2241	2218	2350	2239	2181	2209	1582	1142	1026
Asian	1077	1090	1068	958	924	958	714	655	564
Native Hawaiian/Pacific Islander	147	108	100	98	84	96	55	63	56
Hispanic/Latino	19698	19480	18852	18067	17945	17728	12291	8787	7316
American Indian or Alaskan	2755	2787	2742	2453	2273	2191	2097	1804	1516
White	16128	16579	16303	15788	15615	15781	13331	10546	9608
Multiple Ethnicities	1204	1068	926	814	761	781	652	491	498

Appendix C.4 -- Number of Students in Each Assessment by Gender and Ethnicity – Mathematics Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
Female	21417	20777	20736	21898	21542	21399	24838	22043	18578
Male	21923	21398	21909	22267	21947	21905	25889	22183	17562
Unknown		1				1	1	1	1
African American	2132	2063	2121	2196	2328	2269	2688	2215	1792
Asian	1267	1323	1449	1425	1396	1406	1569	1508	1399
Native Hawaiian/Pacific Islander	137	112	114	118	125	139	146	114	104
Hispanic/Latino	19056	18257	18385	18627	18429	18421	23513	19474	15227
American Indian or Alaskan	2159	2222	2132	2140	2333	2248	2285	1921	1423
White	16973	16833	17163	18429	17706	17726	18845	17649	15019
Multiple Ethnicities	1244	1039	993	953	921	870	781	785	556

Appendix D.1 -- Operational Item Parameter Estimates – Grade 3 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	0.08349				0.08349
2	MC4	0.39681				0.39681
3	HT	-0.90182				-0.90182
4	MC4	0.21483				0.21483
5	MC4	-0.29083				-0.29083
6	MC4	-0.72282				-0.72282
7	MC4	1.21233				1.21233
8	MC4	-0.37951				-0.37951
9	MC4	-1.11123				-1.11123
10	HT	-0.11016				-0.11016
11	MS6	2.45575				2.45575
12	MC4	0.98710				0.98710
13	MC4	0.80043				0.80043
14	MC4	1.04463				1.04463
15	ETC	-1.26547	0.16809			-0.54869
16	ETC	-0.20048				-0.20048
17	ETC	-0.89613	0.70043			-0.09785
18	MC4	-0.10541				-0.10541
19	MC4	-1.43336				-1.43336
20	EBSR4	-0.76650	-0.75832			-0.76241
21	MC4	-1.18875				-1.18875
22	MC4	0.64185				0.64185
23	MC4	-0.87930				-0.87930
24	NL	-0.35113	0.49111			0.06999
25	MC4	0.05610				0.05610
26	MC4	-0.90057				-0.90057
27	HT	-0.47704	4.83258			2.17777
28	MC4	0.28528				0.28528
29	MC4	0.22526				0.22526
30	HT	-2.40053	0.53641			-0.93206
31	MC4	-0.55613				-0.55613
32	MC4	-1.89438				-1.89438
33	HT	1.06391				1.06391
34	MC4	-0.34720				-0.34720
35	MC4	0.60496				0.60496
36	MC4	-1.35966				-1.35966
37	MC4	-0.14111				-0.14111
38	ETC	-1.40248	0.43274			-0.48487

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
39	ETC	0.42891				0.42891
40	ETC	1.19670				1.19670
41	ETC	-0.59287				-0.59287
42 C	ER	-1.10577	0.03845			-0.53366
42 E	ER	-0.06377	1.20975	2.80538		1.31712
42 O	ER	-0.36967	1.18207	2.82345		1.21195

*****Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.2 -- Operational Item Parameter Estimates – Grade 4 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-0.90911				-0.90911
2	MC4	-1.32735				-1.32735
3	HT	-0.27702				-0.27702
4	MC4	-1.25226				-1.25226
5	MC4	-1.32249				-1.32249
6	MC4	-0.24199				-0.24199
7	MC4	-1.81594				-1.81594
8	MC4	-0.26043				-0.26043
9	MC4	-1.26807				-1.26807
10	HT	1.94801	1.18503			1.56652
11	MC4	-0.37670				-0.37670
12	MC4	0.84714				0.84714
13	HT	0.08530	0.22484			0.15507
14	MC4	1.09074				1.09074
15	MS4	-0.20667				-0.20667
16	ETC	-1.56097				-1.56097
17	ETC	0.25414				0.25414
18	ETC	-0.17084				-0.17084
19	MC4	-0.73947				-0.73947
20	MC4	0.16954				0.16954
21	MC4	-0.79490				-0.79490
22	MC4	-1.92826				-1.92826
23	HT	-0.24860	2.31384			1.03262
24	EBSR4	0.68017	0.18633			0.43325
25	MC4	-0.10446				-0.10446
26	MC4	-0.63858				-0.63858
27	MC4	0.38273				0.38273
28	MC4	-0.76702				-0.76702
29	MC4	0.37730				0.37730
30	MC4	0.19280				0.19280
31	HT	3.11384				3.11384
32	EBSR4	0.91216	0.19824			0.55520
33	MC4	0.61265				0.61265
34	MC4	1.08464				1.08464
35	MC4	0.08857				0.08857
36	HT	0.56641	-0.44751			0.05945

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.59331				0.59331
38	EBSR4	1.23482	-0.11392			0.56045
39	ETC	-0.61901				-0.61901
40	ETC	-0.15883				-0.15883
41	ETC	0.18441				0.18441
42 C	ER	-2.04018	0.45214			-0.79402
42 E	ER	0.16147	2.15609	4.16976		2.16244
42 O	ER	-0.20285	1.83718	4.41845		2.01759

***Note**: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.3 -- Operational Item Parameter Estimates – Grade 5 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	0.19926				0.19926
2	MC4	-0.85880				-0.85880
3	MC4	-0.83991				-0.83991
4	MC4	0.28762				0.28762
5	MC4	0.44313				0.44313
6	MC4	0.23160				0.23160
7	MC4	0.43824				0.43824
8	HT	1.82691	1.71529			1.77110
9	HT	0.43134	1.05498			0.74316
10	MC4	-0.20234				-0.20234
11	MC4	1.14918				1.14918
12	MC4	-0.35350				-0.35350
13	HT	1.09502	0.96996			1.03249
14	HT	-0.26056	2.12278			0.93111
15	MS4	1.33077				1.33077
16	MC4	-0.01324				-0.01324
17	MC4	-0.80353				-0.80353
18	MC4	1.14445				1.14445
19	MC4	0.23690				0.23690
20	MC4	-1.08181				-1.08181
21	MC4	-0.79078				-0.79078
22	HT	-1.61111				-1.61111
23	MC4	-0.38291				-0.38291
24	MC4	-0.19644				-0.19644
25	MC4	-0.26032				-0.26032
26	ETC	-0.76704	1.08608			0.15952
27	ETC	-1.90697				-1.90697
28	ETC	-0.67833	0.90083			0.11125
29	HT	-3.67375	-1.45309			-2.56342
30	MC4	-1.19943				-1.19943
31	MC4	-0.36305				-0.36305
32	MC4	0.93489				0.93489
33	MC4	-0.30561				-0.30561
34	MC4	0.13798				0.13798
35	MS5	-0.31067				-0.31067
36	MS4	-0.02295				-0.02295

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	-0.56409				-0.56409
38	ETC	1.26613				1.26613
39	ETC	0.11731				0.11731
40	ETC	-0.24450				-0.24450
41	ETC	-1.47971	0.01885			-0.73043
42 C	ER	-1.93742	0.58950			-0.67396
42 E	ER	0.16319	2.02538	3.24737		1.81198
42 O	ER	0.18317	1.90208	3.31976		1.80167

*****Note: The last three items show the parameters for the one writing item that is scored on three dimensions: C for Conventions, E for Elaboration and O for Organization.

Appendix D.4 -- Operational Item Parameter Estimates – Grade 6 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.59932				-1.59932
2	MC4	-1.56323				-1.56323
3	MC4	-0.34797				-0.34797
4	MC4	-0.64981				-0.64981
5	MC4	-1.26228				-1.26228
6	EBSR4	-0.58026	4.69176			2.05575
7	MC4	-0.91956				-0.91956
8	MC4	-0.15280				-0.15280
9	MC4	0.88120				0.88120
10	MC4	-0.88145				-0.88145
11	MC4	0.06794				0.06794
12	MC4	0.48034				0.48034
13	MC4	0.22169				0.22169
14	ETC	0.10150				0.10150
15	ETC	-0.42930	1.32996			0.45033
16	ETC	-1.22470	0.98388			-0.12041
17	MC4	-0.08313				-0.08313
18	EBSR4	0.57372	-1.62084			-0.52356
19	MC4	0.61112				0.61112
20	MC4	0.25555				0.25555
21	MC4	-0.02386				-0.02386
22	MC4	-0.43346				-0.43346
23	MC4	0.34652				0.34652
24	MC4	1.50941				1.50941
25	MC4	-0.04865				-0.04865
26	MS5	-0.02264				-0.02264
27	MC4	-0.13178				-0.13178
28	MC4	0.08787				0.08787
29	HT	-0.20619	1.93121			0.86251
30	MC4	1.49410				1.49410
31	EBSR4	0.94361	1.61463			1.27912
32	MC4	0.28446				0.28446
33	MC4	-0.82151				-0.82151
34	MC4	-0.77650				-0.77650
35	MC4	0.55465				0.55465
36	MC4	-0.22039				-0.22039

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	-0.54162				-0.54162
38	ETC	-2.51833				-2.51833
39	ETC	-1.57982	0.21870			-0.68056
40	ETC	0.30964				0.30964
41	ETC	0.30667				0.30667
42 C	ER	-1.34556	0.22150			-0.56203
42 E	ER	-0.09270	1.51128	3.06534		1.49464
42 O	ER	-0.58754	1.19987	3.07716		1.22983

*****Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.5 -- Operational Item Parameter Estimates – Grade 7 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.82204				-1.82204
2	HT	-0.26090				-0.26090
3	HT	-0.92512	1.44132			0.25810
4	MC4	1.22419				1.22419
5	MC4	-0.58597				-0.58597
6	HT	1.00581				1.00581
7	MS5	0.49772				0.49772
8	MC4	-0.17470				-0.17470
9	MC4	0.36345				0.36345
10	MC4	-0.52540				-0.52540
11	HT	-0.10313				-0.10313
12	MS4	1.78000				1.78000
13	MC4	-0.43478				-0.43478
14	ETC	-1.09927	0.89291			-0.10318
15	ETC	-0.77970				-0.77970
16	ETC	0.08194				0.08194
17	ETC	-0.98128				-0.98128
18	MC4	0.33204				0.33204
19	MC4	-0.59480				-0.59480
20	MS5	0.73816				0.73816
21	MC4	-0.72887				-0.72887
22	MC4	-0.12352				-0.12352
23	MC4	-0.39080				-0.39080
24	MC4	0.06401				0.06401
25	MC4	-0.65753				-0.65753
26	MC4	-1.09155				-1.09155
27	MC4	-0.22566				-0.22566
28	MC4	0.49703				0.49703
29	MC4	1.10958				1.10958
30	MC4	0.51878				0.51878
31	MC4	-0.70900				-0.70900
32	EBSR4	0.40699	0.91501			0.66100
33	MC4	0.03598				0.03598
34	HT	-2.13791	0.20105			-0.96843
35	HT	0.88061				0.88061
36	EBSR4	0.96247	-0.27019			0.34614

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MS6	1.38411				1.38411
38	MC4	-0.75189				-0.75189
39	ETC	-1.10696	0.01964			-0.54366
40	ETC	-1.08392	0.78122			-0.15135
41	ETC	-0.46422				-0.46422
42 C	ER	-1.73815	-0.67491			-1.20653
42 E	ER	-0.50585	1.37176	3.07024		1.31205
42 O	ER	-0.61360	1.38951	3.08854		1.28815

*****Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.6 -- Operational Item Parameter Estimates – Grade 8 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-2.36894				-2.36894
2	MC4	-0.51753				-0.51753
3	MC4	0.31509				0.31509
4	EBSR4	1.07591	-1.73311			-0.32860
5	MC4	0.02807				0.02807
6	MS5	-0.14024				-0.14024
7	MS5	1.15696				1.15696
8	MC4	0.75205				0.75205
9	MC4	0.05707				0.05707
10	MC4	-0.06763				-0.06763
11	MC4	-0.38230				-0.38230
12	MC4	0.37134				0.37134
13	MS6	1.40328				1.40328
14	MC4	0.06190				0.06190
15	ETC	0.45404				0.45404
16	ETC	-2.29297				-2.29297
17	ETC	-1.53228				-1.53228
18	ETC	-0.65169	2.53769			0.94300
19	EBSR4	-0.70818	-1.24338			-0.97578
20	MC4	0.23357				0.23357
21	HT	-1.10565	3.30365			1.09900
22	MC4	0.02405				0.02405
23	MC4	0.18439				0.18439
24	MC4	-0.36972				-0.36972
25	HT	2.90776	-1.28216			0.81280
26	MC4	-0.06570				-0.06570
27	MC4	-0.74500				-0.74500
28	MC4	0.35395				0.35395
29	MC4	0.10919				0.10919
30	MC4	0.78355				0.78355
31	MC4	-0.86763				-0.86763
32	MC4	-0.10644				-0.10644
33	MC4	0.11732				0.11732
34	MC4	0.88343				0.88343
35	MC4	1.00994				1.00994
36	EBSR4	-0.13976	0.89940			0.37982

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.66011				0.66011
38	MS4	-0.03646				-0.03646
39	ETC	-1.36971	-0.99521			-1.18246
40	ETC	-0.11034	0.82358			0.35662
41	ETC	-0.69022				-0.69022
42 C	ER	-2.06546	-0.52800			-1.29673
42 E	ER	-1.44149	0.67610	3.67416		0.96959
42 O	ER	-1.97084	-0.01623	3.32653		0.44649

*****Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.7 -- Operational Item Parameter Estimates – Grade 9 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.04899				-1.04899
2	MC4	-0.96946				-0.96946
3	MS5	0.13386				0.13386
4	MC4	-0.11641				-0.11641
5	NL	-0.20347				-0.20347
6	MC4	-0.32889				-0.32889
7	MC4	-0.87974				-0.87974
8	MC4	-0.54847				-0.54847
9	MC4	-0.39379				-0.39379
10	MC4	0.38045				0.38045
11	MC4	0.15985				0.15985
12	MC4	-0.18218				-0.18218
13	MC4	0.01430				0.01430
14	MC4	-0.22230				-0.22230
15	MC4	0.44926				0.44926
16	ETC	-0.74854	0.76546			0.00846
17	ETC	-2.15025	0.80673			-0.67176
18	ETC	-0.98422				-0.98422
19	MC4	0.91452				0.91452
20	MC4	-0.07659				-0.07659
21	MC4	-0.28086				-0.28086
22	HT	0.12199	0.51459			0.31829
23	MC4	-0.07907				-0.07907
24	MC4	0.65177				0.65177
25	MS6	1.56141				1.56141
26	MC4	-0.74584				-0.74584
27	MC4	-1.21848				-1.21848
28	MC4	0.63067				0.63067
29	MS6	0.78122				0.78122
30	MC4	0.36911				0.36911
31	EBSR4	-0.77709	2.32017			0.77154
32	MC4	0.80007				0.80007
33	MC4	0.05750				0.05750
34	MC4	0.35139				0.35139
35	MC4	-0.82593				-0.82593
36	HT	-0.56327				-0.56327

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	HT	0.68140				0.68140
38	MC4	-0.15288				-0.15288
39	MS4	0.28481				0.28481
40	ETC	-0.46997				-0.46997
41	ETC	-0.93231				-0.93231
42	ETC	0.78252				0.78252
43	ETC	-1.69358	0.61556			-0.53901
44 C	ER	-1.99081	-0.24495			-1.11788
44 E	ER	0.44000	2.20456	3.46515		2.03657
44 O	ER	-1.08168	1.67788	3.64219		1.41280

*****Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.8 -- Operational Item Parameter Estimates – Grade 10 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-0.17713				-0.17713
2	MC4	-0.51272				-0.51272
3	MC4	-1.84934				-1.84934
4	HT	-1.84911	0.71335			-0.56788
5	MC4	0.71020				0.71020
6	EBSR4	0.83107	0.00851			0.41979
7	EBSR4	1.86753	-1.28293			0.29230
8	MC4	-0.20740				-0.20740
9	EBSR4	0.01742	0.38768			0.20255
10	MS5	1.15828				1.15828
11	MC4	0.00806				0.00806
12	MC4	-0.05557				-0.05557
13	MC4	-0.42075				-0.42075
14	ETC	-0.60054				-0.60054
15	ETC	0.67202				0.67202
16	ETC	0.18833				0.18833
17	ETC	0.57411	1.84561			1.20986
18	MC4	0.11487				0.11487
19	MC4	-0.26494				-0.26494
20	MC4	0.17104				0.17104
21	MC4	0.06903				0.06903
22	MC4	0.08195				0.08195
23	MC4	0.05540				0.05540
24	MC4	0.02694				0.02694
25	MC4	0.29358				0.29358
26	MC4	-0.53789				-0.53789
27	MC4	-0.94852				-0.94852
28	MC4	0.31653				0.31653
29	MC4	-0.12248				-0.12248
30	MC4	0.02380				0.02380
31	MC4	0.16204				0.16204
32	MC4	-0.52380				-0.52380
33	MC4	-0.39119				-0.39119
34	MC4	-0.31335				-0.31335
35	MC4	0.88721				0.88721
36	HT	1.84659	0.10145			0.97402

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.52920				0.52920
38	MC4	0.60183				0.60183
39	MC4	-0.10195				-0.10195
40	ETC	-0.33161				-0.33161
41	ETC	-0.97145				-0.97145
42	ETC	-0.25493				-0.25493
43	ETC	-0.89397	1.74965			0.42784
44 C	ER	-2.08750	-0.19028			-1.13889
44 E	ER	-1.46115	0.40789	2.40362		0.45012
44 O	ER	-2.05950	0.36424	2.43185		0.24553

*Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.9 -- Operational Item Parameter Estimates – Grade 11 ELA

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-0.61028				-0.61028
2	HT	-0.18435	0.29935			0.05750
3	MC4	-1.66781				-1.66781
4	MC4	-1.17087				-1.17087
5	NL	1.20794	0.81834			1.01314
6	MC4	0.70197				0.70197
7	MC4	0.12808				0.12808
8	MC4	-0.71121				-0.71121
9	MC4	-0.39031				-0.39031
10	MC4	-0.00192				-0.00192
11	MC4	-0.30674				-0.30674
12	MC4	-0.08661				-0.08661
13	MC4	0.55756				0.55756
14	MC4	-0.18694				-0.18694
15	HT	2.01632				2.01632
16	ETC	-0.27674				-0.27674
17	ETC	-0.81620	0.39514			-0.21053
18	ETC	0.11783				0.11783
19	ETC	-0.10187				-0.10187
20	MC4	-1.42046				-1.42046
21	MC4	-0.35030				-0.35030
22	MC4	0.89684				0.89684
23	MC4	-0.95855				-0.95855
24	MC4	0.56994				0.56994
25	MC4	-0.07161				-0.07161
26	HT	2.02362				2.02362
27	MC4	0.10930				0.10930
28	MC4	-0.29090				-0.29090
29	MC4	-0.82221				-0.82221
30	MC4	-0.02144				-0.02144
31	MC4	-0.72981				-0.72981
32	EBSR4	0.94971	1.31219			1.13095
33	MC4	-0.40717				-0.40717
34	MC4	0.81854				0.81854
35	MC4	0.47951				0.47951
36	MC4	-0.20582				-0.20582

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	-0.35209				-0.35209
38	MC4	-0.03791				-0.03791
39	MS4	1.42382				1.42382
40	ETC	-0.01790				-0.01790
41	ETC	0.00441				0.00441
42	ETC	-0.68645	0.37809			-0.15418
43	ETC	-1.32097				-1.32097
44 C	ER	-2.66117	-0.11579			-1.38848
44 E	ER	-0.30179	1.30583	3.10427		1.36944
44 O	ER	-1.78033	1.38473	2.95415		0.85285

*Note: The last three items are the parameters for the one writing item that is scored on three dimensions: C is Conventions, E is Elaboration and O is Organization

Appendix D.10 -- Operational Item Parameter Estimates – Grade 3 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-2.01816				-2.01816
2	MC4	-1.45851				-1.45851
3	MC4	-1.01221				-1.01221
4	EQ	-0.55434				-0.55434
5	MC4	0.23613				0.23613
6	GI	0.33487				0.33487
7	MC4	0.12079				0.12079
8	MC4	0.75911				0.75911
9	MC4	-0.36538				-0.36538
10	MC4	0.59716				0.59716
11	MC4	1.38605				1.38605
12	MC4	-0.57358				-0.57358
13	MC4	1.22547				1.22547
14	EQ	1.85614				1.85614
15	MC4	1.12171				1.12171
16	MC4	0.19825				0.19825
17	MC4	0.24003				0.24003
18	MC4	0.16931				0.16931
19	MC4	0.47456				0.47456
20	MC4	-0.20907				-0.20907
21	MC4	-0.68463				-0.68463
22	MC4	-0.94326				-0.94326
23	MC4	-1.33286				-1.33286
24	MC4	-1.65549				-1.65549
25	MC4	-1.01175				-1.01175
26	MC4	-0.90569				-0.90569
27	MC4	-0.72661				-0.72661
28	MC4	-0.64939				-0.64939
29	MC4	0.43228				0.43228
30	MC4	-0.26263				-0.26263
31	EQ	1.43606				1.43606
32	MC4	0.07620				0.07620
33	MC4	0.91894				0.91894
34	EQ	1.83709				1.83709
35	MC4	1.66505				1.66505
36	MC4	0.38018				0.38018

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.63258				0.63258
38	MC4	0.53976				0.53976
39	MC4	0.40147				0.40147
40	MC4	0.18888				0.18888
41	GI	0.08079				0.08079
42	MC4	-0.31377				-0.31377
43	MC4	-0.82445				-0.82445
44	MC4	-0.72357				-0.72357
45	MC4	-1.08350				-1.08350

Appendix D.11 -- Operational Item Parameter Estimates – Grade 4 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.28691				-1.28691
2	MC4	-1.15157				-1.15157
3	MC4	-1.00494				-1.00494
4	MC4	-1.19898				-1.19898
5	MC4	-0.29313				-0.29313
6	MC4	0.74507				0.74507
7	MC4	-0.32106				-0.32106
8	MC4	0.18490				0.18490
9	MC4	0.30756				0.30756
10	MC4	0.52853				0.52853
11	MC4	1.62313				1.62313
12	MC4	1.49491				1.49491
13	EQ	1.94682				1.94682
14	MC4	1.67854				1.67854
15	EQ	1.52638				1.52638
16	MC4	0.26603				0.26603
17	MC4	0.18352				0.18352
18	MC4	-0.21052				-0.21052
19	MC4	-0.52591				-0.52591
20	MC4	-0.69238				-0.69238
21	MC4	-1.02215				-1.02215
22	MC4	-0.73582				-0.73582
23	MC4	-1.04527				-1.04527
24	MC4	-2.21249				-2.21249
25	MC4	-1.17364				-1.17364
26	MC4	-0.95266				-0.95266
27	MC4	-0.54301				-0.54301
28	MC4	-0.40512				-0.40512
29	MC4	-0.56351				-0.56351
30	MC4	-0.16944				-0.16944
31	MC4	0.63803				0.63803
32	MC4	0.24510				0.24510
33	MC4	0.95644				0.95644
34	EQ	2.97069				2.97069
35	EQ	1.70501				1.70501
36	MC4	0.74327				0.74327

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.75943				0.75943
38	MC4	-0.16424				-0.16424
39	MC4	-0.31287				-0.31287
40	EQ	0.72975				0.72975
41	MC4	0.30409				0.30409
42	MC4	-0.43922				-0.43922
43	MC4	-0.84677				-0.84677
44	MC4	-0.99191				-0.99191
45	MC4	-1.27370				-1.27370

Appendix D.12 -- Operational Item Parameter Estimates – Grade 5 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-2.53507				-2.53507
2	MC4	-0.89232				-0.89232
3	MC4	-0.72229				-0.72229
4	EQ	0.22161				0.22161
5	MC4	-0.71692				-0.71692
6	EQ	0.27143				0.27143
7	MC4	0.03113				0.03113
8	MC4	1.10899				1.10899
9	MC4	-0.33160				-0.33160
10	MC4	1.02935				1.02935
11	GI	1.70411				1.70411
12	MC4	1.29829				1.29829
13	MC4	0.91512				0.91512
14	MC4	-0.02629				-0.02629
15	EQ	0.28574				0.28574
16	EQ	0.75464				0.75464
17	MC4	0.01605				0.01605
18	MC4	0.37091				0.37091
19	MC4	-0.90844				-0.90844
20	MC4	-0.36294				-0.36294
21	MC4	-0.81576				-0.81576
22	MC4	-0.99664				-0.99664
23	MC4	-1.84209				-1.84209
24	MC4	-1.23338				-1.23338
25	MC4	-0.39584				-0.39584
26	MC4	-0.29925				-0.29925
27	MC4	-0.13330				-0.13330
28	MC4	-0.23270				-0.23270
29	MC4	-0.77154				-0.77154
30	MC4	0.55243				0.55243
31	MC4	0.55605				0.55605
32	EQ	1.45341				1.45341
33	MC4	0.80129				0.80129
34	MC4	1.45381				1.45381
35	EQ	1.57902				1.57902
36	MC4	0.30925				0.30925

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	EQ	1.23694				1.23694
38	EQ	0.62712				0.62712
39	MC4	0.15752				0.15752
40	MC4	0.18933				0.18933
41	MC4	0.04623				0.04623
42	MC4	-0.72874				-0.72874
43	MC4	-0.85675				-0.85675
44	MC4	-1.10731				-1.10731
45	MC4	-1.06058				-1.06058

Appendix D.13 -- Operational Item Parameter Estimates – Grade 6 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-2.12659				-2.12659
2	MC4	-0.98870				-0.98870
3	MC4	-0.82658				-0.82658
4	MC4	-0.42727				-0.42727
5	MC4	-0.18287				-0.18287
6	MC4	-0.41033				-0.41033
7	MC4	0.25739				0.25739
8	EQ	0.98562				0.98562
9	MC4	-0.68754				-0.68754
10	MC4	0.19647				0.19647
11	MC4	-0.23423				-0.23423
12	MC4	-1.34984				-1.34984
13	MC4	0.94937				0.94937
14	MC4	0.31352				0.31352
15	MC4	0.44047				0.44047
16	MC4	1.27656				1.27656
17	MC4	0.02708				0.02708
18	MC4	0.64021				0.64021
19	MC4	-0.35862				-0.35862
20	MC4	0.28544				0.28544
21	MC4	-0.29381				-0.29381
22	MC4	-0.19138				-0.19138
23	MC4	-1.18940				-1.18940
24	MC4	-1.04596				-1.04596
25	EQ	2.93461				2.93461
26	MC4	-0.93883				-0.93883
27	MC4	-1.43481				-1.43481
28	MC4	-0.67826				-0.67826
29	MC4	-0.70572				-0.70572
30	MC4	0.10615				0.10615
31	MC4	0.72417				0.72417
32	MC4	-0.03451				-0.03451
33	MC4	0.07986				0.07986
34	MC4	0.64733				0.64733
35	MS6	1.68123				1.68123
36	EQ	3.33588				3.33588

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	EQ	1.43004				1.43004
38	GI	1.38921				1.38921
39	EQ	1.14526				1.14526
40	MC4	0.07459				0.07459
41	MC4	-0.10651				-0.10651
42	MC4	0.29520				0.29520
43	MC4	-0.73435				-0.73435
44	MC4	-0.37687				-0.37687
45	MC4	-0.96980				-0.96980
46	MC4	-1.56632				-1.56632
47	MC4	-1.35655				-1.35655

Appendix D.14 -- Operational Item Parameter Estimates – Grade 7 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	0.02010				0.02010
2	EQ	2.69271				2.69271
3	MC4	0.18543				0.18543
4	EQ	0.99921				0.99921
5	GI	1.28140				1.28140
6	MS6	1.64080				1.64080
7	MC4	-1.31221				-1.31221
8	MC4	0.00036				0.00036
9	MC4	-0.52339				-0.52339
10	MC4	-0.58716				-0.58716
11	MC4	-0.59949				-0.59949
12	MC4	-2.08036				-2.08036
13	MC4	-1.18217				-1.18217
14	MC4	-1.65160				-1.65160
15	MC4	-0.38740				-0.38740
16	MC4	-1.27632				-1.27632
17	MC4	-1.11464				-1.11464
18	MC4	-0.73811				-0.73811
19	EQ	0.83340				0.83340
20	EQ	0.12846				0.12846
21	MC4	-0.41277				-0.41277
22	MC4	0.58520				0.58520
23	MC4	-0.56655				-0.56655
24	EQ	0.93085				0.93085
25	EQ	2.72334				2.72334
26	EQ	2.53654				2.53654
27	MC4	0.11751				0.11751
28	MC4	0.53696				0.53696
29	EQ	0.82104				0.82104
30	GI	2.15779				2.15779
31	EQ	0.10463				0.10463
32	MC4	-0.45743				-0.45743
33	EQ	0.16667				0.16667
34	MC4	-0.53342				-0.53342
35	MC4	-0.66411				-0.66411
36	MC4	-0.27228				-0.27228

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	-1.68977				-1.68977
38	MC4	-1.44006				-1.44006
39	MC4	-1.23368				-1.23368
40	MC4	-0.89416				-0.89416
41	MC4	-0.76502				-0.76502
42	MC4	-0.25524				-0.25524
43	EQ	-0.40696				-0.40696
44	MC4	0.11267				0.11267
45	MC4	-0.29337				-0.29337
46	MC4	1.24555				1.24555
47	MI	1.51706				1.51706

Appendix D.15 -- Operational Item Parameter Estimates – Grade 8 Mathematics

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.59759				-1.59759
2	MC4	-2.24926				-2.24926
3	MC4	-1.30732				-1.30732
4	MC4	-1.23521				-1.23521
5	MC4	-0.86394				-0.86394
6	EQ	-1.07246				-1.07246
7	MC4	-0.67147				-0.67147
8	EQ	1.81609				1.81609
9	EQ	0.70333				0.70333
10	EQ	0.69767				0.69767
11	GI	0.96460				0.96460
12	EQ	0.87585				0.87585
13	EQ	3.20576				3.20576
14	EQ	0.78129	-0.77477			0.00326
15	EQ	2.16186				2.16186
16	EQ	1.42064				1.42064
17	MC4	0.22818				0.22818
18	MI	-0.30785	1.11589			0.40402
19	MC4	-0.62068				-0.62068
20	MC4	-0.12124				-0.12124
21	MC4	-1.44484				-1.44484
22	MC4	-1.73823				-1.73823
23	MC4	-2.26975				-2.26975
24	MC4	-1.11140				-1.11140
25	MC4	-2.74800				-2.74800
26	MC4	0.96252				0.96252
27	MC4	-2.64276				-2.64276
28	MC4	-0.68220				-0.68220
29	GI	-0.97059	-0.07175			-0.52117
30	EQ	0.58142				0.58142
31	MC4	-0.72208				-0.72208
32	EQ	1.14410				1.14410
33	GI	0.55368				0.55368
34	MS5	1.89528				1.89528
35	MC4	0.66009				0.66009
36	GI	2.65920				2.65920

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	MC4	0.66284				0.66284
38	GI	1.67902				1.67902
39	MC4	1.99404				1.99404
40	EQ	1.60945				1.60945
41	MC4	0.82245				0.82245
42	MC4	-0.00591				-0.00591
43	GI	0.71648				0.71648
44	MC4	-0.35161				-0.35161
45	MC4	-0.89887				-0.89887
46	MC4	-1.59365				-1.59365
47	MC4	-1.95217				-1.95217

Appendix D.16 -- Operational Item Parameter Estimates – Algebra I

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.13846				-1.13846
2	MC4	-1.18249				-1.18249
3	MC4	-0.68595				-0.68595
4	MC4	-1.17858				-1.17858
5	MC4	-0.28972				-0.28972
6	EQ	0.17576				0.17576
7	MC4	0.27113				0.27113
8	GI	0.98952				0.98952
9	MC4	-0.18662				-0.18662
10	MC4	-0.68869				-0.68869
11	MI	0.41420				0.41420
12	EQ	0.53197				0.53197
13	MC4	-0.86850				-0.86850
14	MC4	-0.78065				-0.78065
15	MC4	-0.53134				-0.53134
16	EQ	0.40788				0.40788
17	MC4	0.05606				0.05606
18	EQ	1.02504				1.02504
19	GI	2.81360				2.81360
20	MC4	-0.03451				-0.03451
21	MC4	-0.33266				-0.33266
22	MC4	-1.00720				-1.00720
23	MC4	-0.99020				-0.99020
24	MC4	-1.57015				-1.57015
25	MC4	-0.98113				-0.98113
26	MC4	-0.80088				-0.80088
27	MC4	-0.53190				-0.53190
28	MC4	0.10244				0.10244
29	MC4	-0.63800				-0.63800
30	EQ	0.94908				0.94908
31	MC4	0.27894				0.27894
32	GI	1.16287				1.16287
33	MC4	0.47232				0.47232
34	EQ	1.31614	0.04418			0.68016
35	MC4	-0.57492				-0.57492
36	MC4	-0.31995				-0.31995

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	EQ	2.15413				2.15413
38	MC4	0.00362				0.00362
39	MC4	-0.28406				-0.28406
40	EQ	2.20092				2.20092
41	MC4	-0.57470				-0.57470
42	MC4	0.09400				0.09400
43	MC4	-0.02796				-0.02796
44	MC4	-0.50059				-0.50059
45	GI	2.77030				2.77030
46	MC4	-0.56858				-0.56858
47	MC4	-0.28556				-0.28556

Appendix D.17 -- Operational Item Parameter Estimates – Geometry

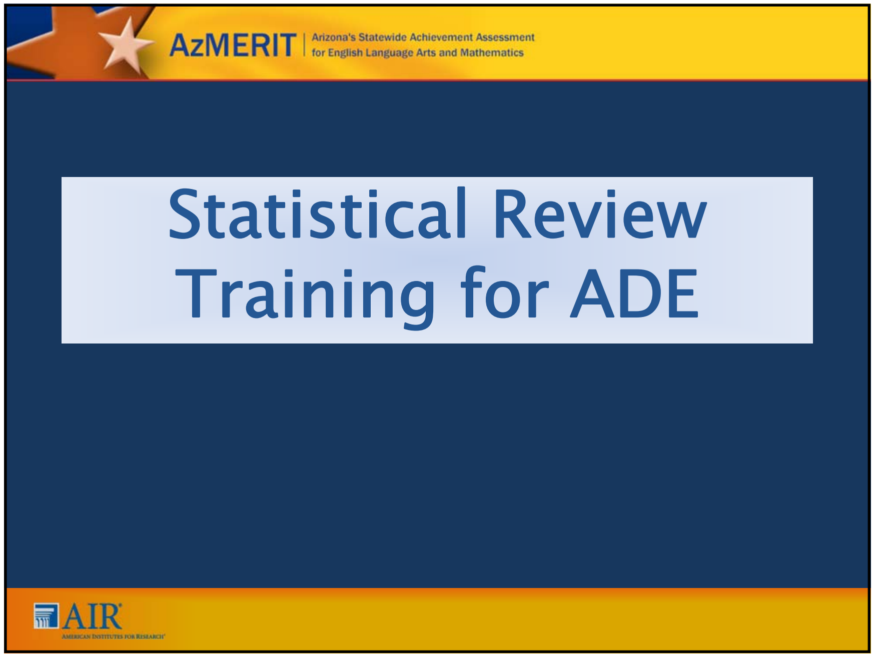
Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	0.27467				0.27467
2	MC4	-1.36745				-1.36745
3	MC4	-0.86434				-0.86434
4	MC4	-1.02017				-1.02017
5	EQ	0.16054				0.16054
6	EQ	1.80627				1.80627
7	GI	0.27462				0.27462
8	MS5	2.37713				2.37713
9	EQ	-0.27107				-0.27107
10	MC4	-1.64475				-1.64475
11	GI	-1.64777	1.76869			0.06046
12	MC4	-2.16615				-2.16615
13	MC4	-1.34398				-1.34398
14	EQ	1.43738				1.43738
15	MC4	-0.94892				-0.94892
16	EQ	0.40245	2.87469			1.63857
17	EQ	-0.40479				-0.40479
18	MC4	-0.12041				-0.12041
19	EQ	2.27972				2.27972
20	MC4	-1.26194				-1.26194
21	MC4	-1.54054				-1.54054
22	GI	-0.83485				-0.83485
23	MC4	-0.32968				-0.32968
24	MC4	-2.34500				-2.34500
25	MC4	-1.63536				-1.63536
26	EQ	-1.36610				-1.36610
27	MC4	-1.29970				-1.29970
28	EQ	-0.76390				-0.76390
29	EQ	0.23509				0.23509
30	MC4	-0.49515				-0.49515
31	GI	2.00423				2.00423
32	MC4	-1.15631				-1.15631
33	EQ	1.64866				1.64866
34	EQ	0.95865				0.95865
35	GI	0.86569				0.86569
36	MS5	1.12317				1.12317

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	GI	1.50301				1.50301
38	MC4	0.77861				0.77861
39	EQ	1.14689				1.14689
40	GI	0.70095	2.94754			1.82425
41	EQ	2.65676				2.65676
42	MC4	-0.57664				-0.57664
43	EQ	1.13759				1.13759
44	MC4	-1.02449				-1.02449
45	MC4	-1.15482				-1.15482
46	EQ	-0.10367				-0.10367
47	MC4	-0.15177				-0.15177

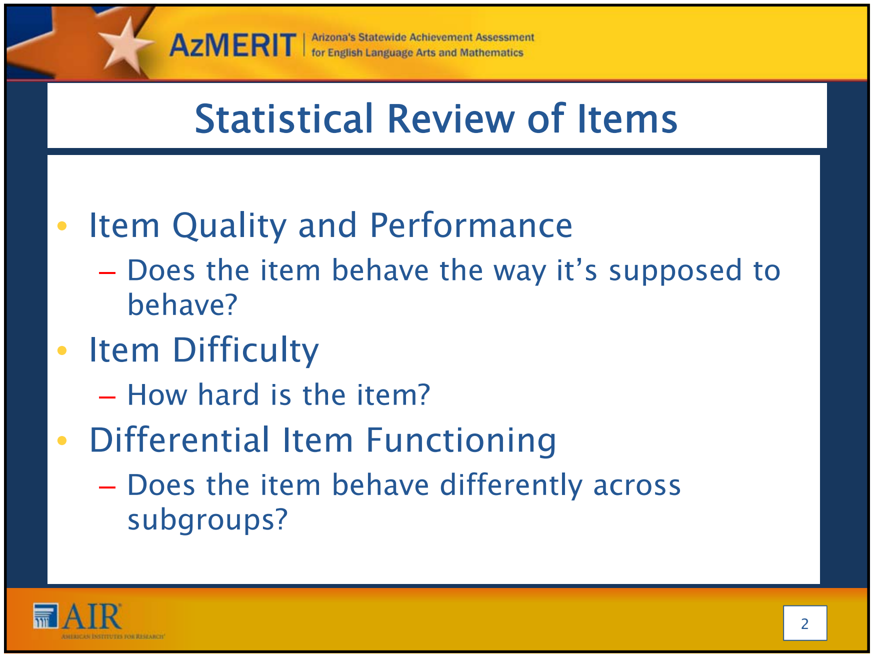
Appendix D.18 -- Operational Item Parameter Estimates – Algebra II

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
1	MC4	-1.14790				-1.14790
2	MC4	-2.46835				-2.46835
3	MC4	-2.32344				-2.32344
4	MC4	-2.28541				-2.28541
5	MC4	-0.77417				-0.77417
6	EQ	1.17831				1.17831
7	MC4	0.21503				0.21503
8	EQ	1.40995				1.40995
9	EQ	0.33501				0.33501
10	MC4	-1.70655				-1.70655
11	MS5	0.80334				0.80334
12	EQ	0.84135				0.84135
13	EQ	-1.49590				-1.49590
14	MC4	-0.54801				-0.54801
15	MC4	-0.56631				-0.56631
16	EQ	1.96209				1.96209
17	MC4	-1.17324				-1.17324
18	EQ	0.97523				0.97523
19	EQ	3.00930				3.00930
20	MC4	-2.02063				-2.02063
21	MC4	-0.59620				-0.59620
22	MC4	-1.30859				-1.30859
23	EQ	-1.01899				-1.01899
24	MS6	2.22258				2.22258
25	EQ	-0.29880				-0.29880
26	MC4	-1.98646				-1.98646
27	MC4	-0.85034				-0.85034
28	EQ	-0.94491				-0.94491
29	MS6	0.97139				0.97139
30	GI	2.02801				2.02801
31	EQ	1.17110				1.17110
32	EQ	-0.68265				-0.68265
33	MC4	1.16516				1.16516
34	EQ	1.83440				1.83440
35	EQ	2.07159				2.07159
36	MC4	-0.90329				-0.90329

Item	Item Type	Item Parameter Estimates				Average Rasch Value
		Step 1	Step 2	Step 3	Step 4	
37	EQ	1.12750				1.12750
38	EQ	0.56780				0.56780
39	MS5	-0.17509				-0.17509
40	EQ	1.72323				1.72323
41	EQ	2.06971				2.06971
42	MC4	-1.43147				-1.43147
43	EQ	1.67485				1.67485
44	MC4	0.06342				0.06342
45	EQ	-0.19149				-0.19149
46	EQ	-0.84907	-0.24521			-0.54714
47	MC4	-1.97502				-1.97502



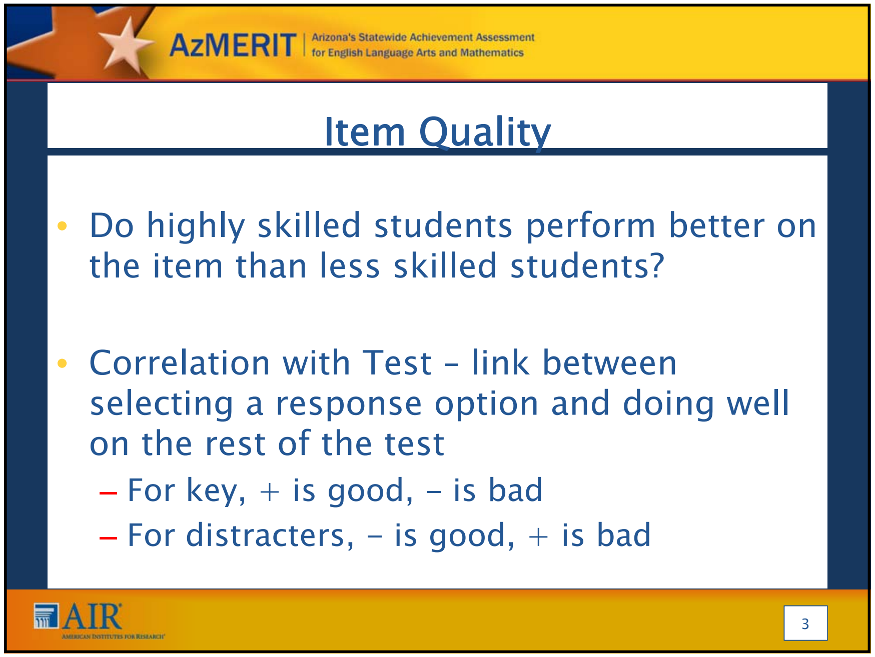
Statistical Review Training for ADE



Statistical Review of Items

- Item Quality and Performance
  - Does the item behave the way it's supposed to behave?
- Item Difficulty
  - How hard is the item?
- Differential Item Functioning
  - Does the item behave differently across subgroups?

2



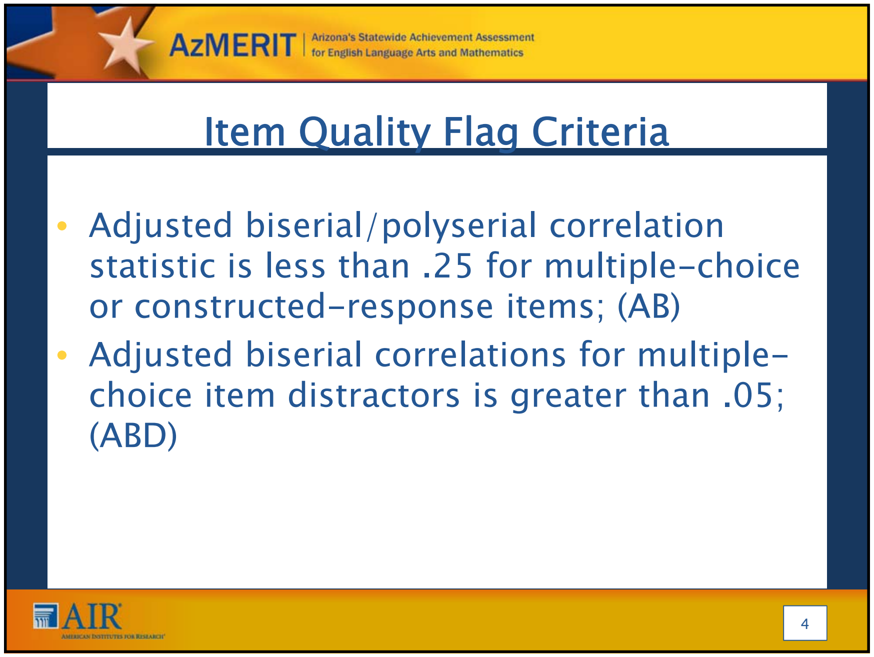
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Item Quality

- Do highly skilled students perform better on the item than less skilled students?
- Correlation with Test – link between selecting a response option and doing well on the rest of the test
  - For key, + is good, – is bad
  - For distracters, – is good, + is bad

AIR AMERICAN INSTITUTES FOR RESEARCH

3



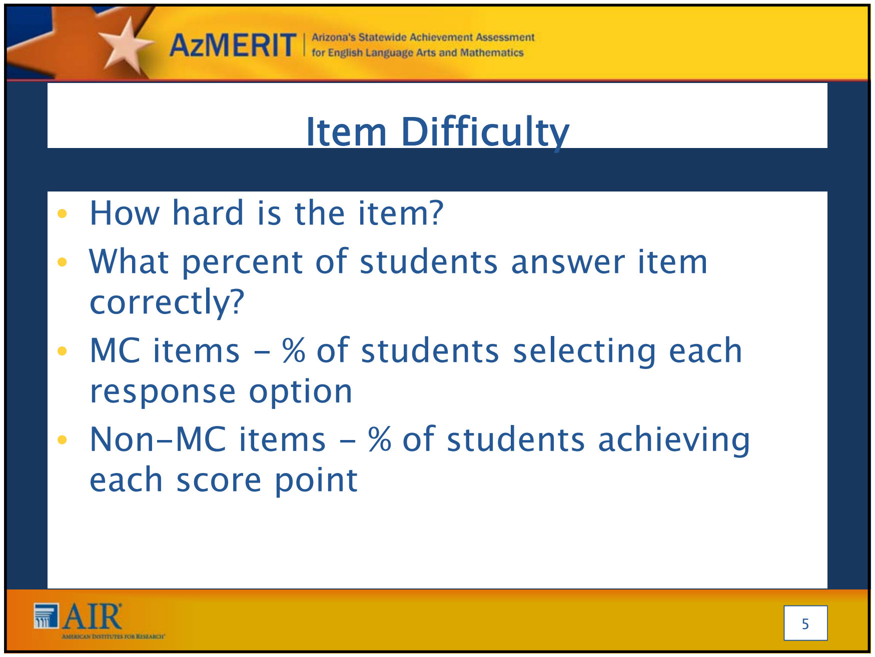
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Item Quality Flag Criteria

- Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items; (AB)
- Adjusted biserial correlations for multiple-choice item distractors is greater than .05; (ABD)

AIR AMERICAN INSTITUTES FOR RESEARCH

4



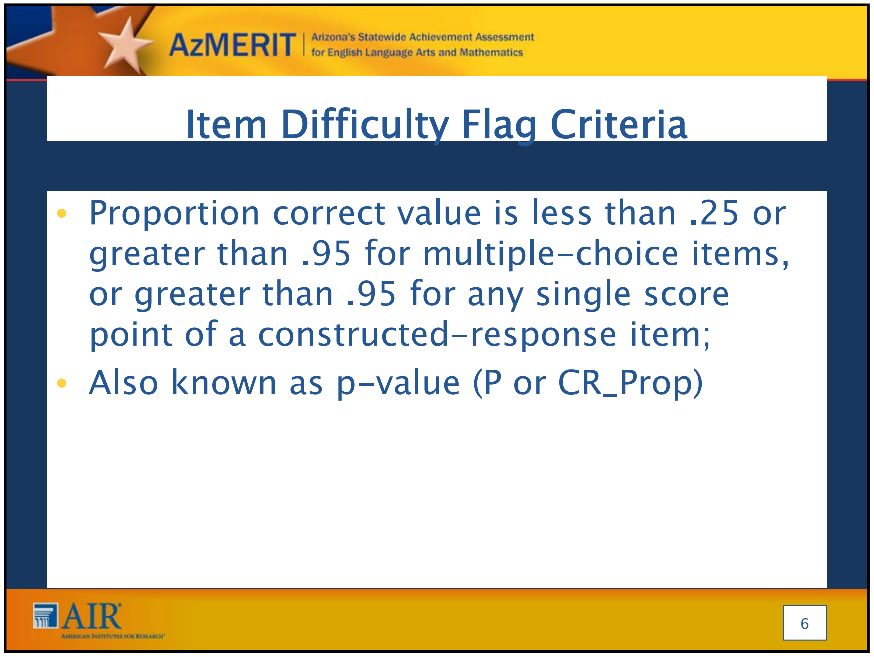
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Item Difficulty

- How hard is the item?
- What percent of students answer item correctly?
- MC items – % of students selecting each response option
- Non-MC items – % of students achieving each score point

AIR American Institutes for Research

5



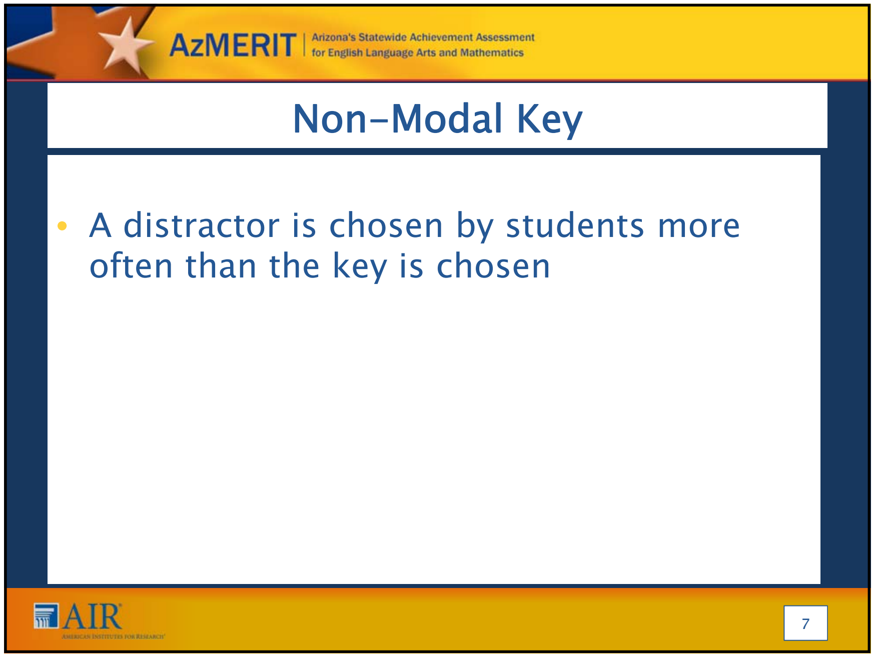
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Item Difficulty Flag Criteria

- Proportion correct value is less than .25 or greater than .95 for multiple-choice items, or greater than .95 for any single score point of a constructed-response item;
- Also known as p-value (P or CR_Prop)

AIR American Institutes for Research

6



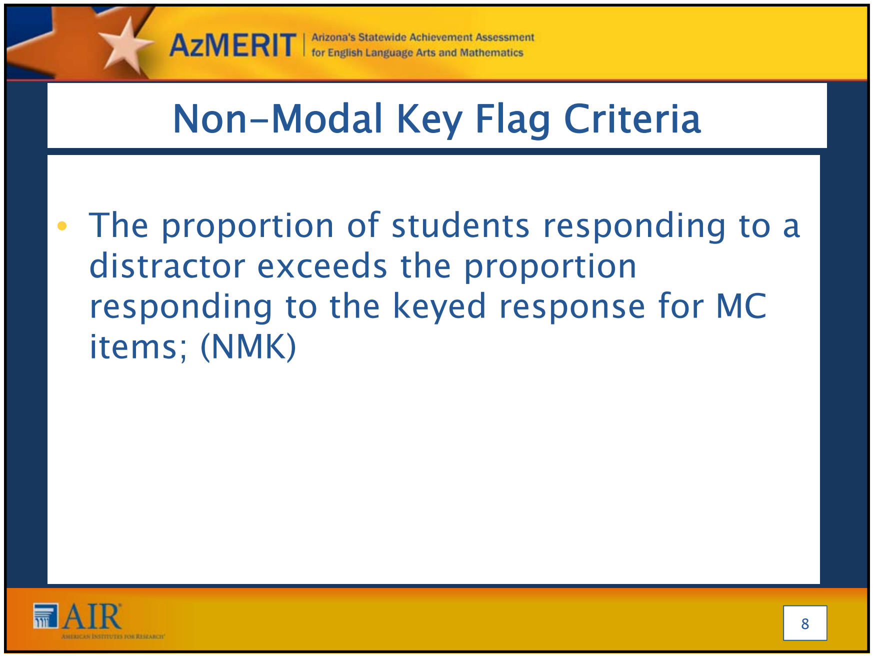
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Non-Modal Key

- A distractor is chosen by students more often than the key is chosen

AIR American Institutes for Research

7



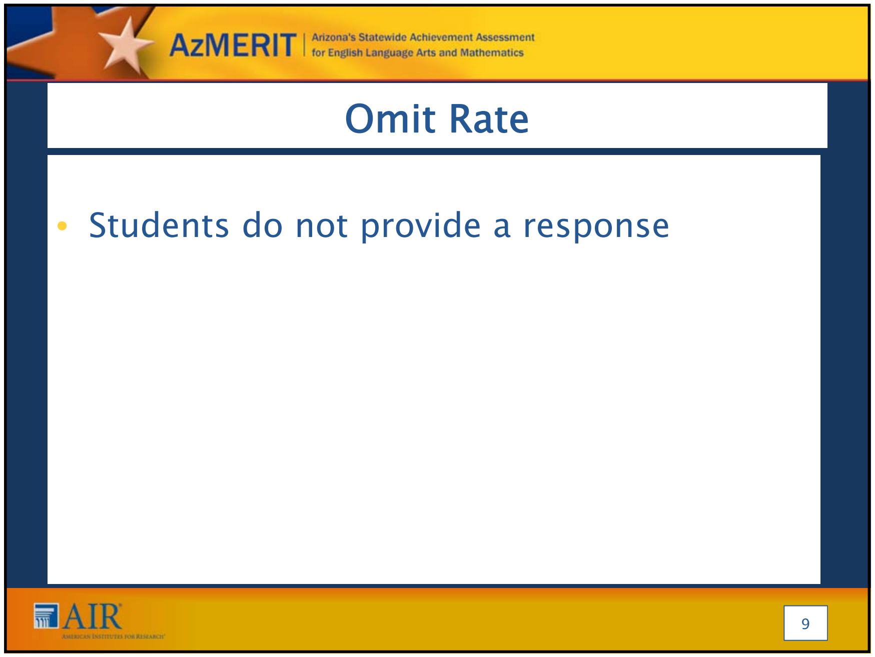
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Non-Modal Key Flag Criteria

- The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items; (NMK)

AIR American Institutes for Research

8



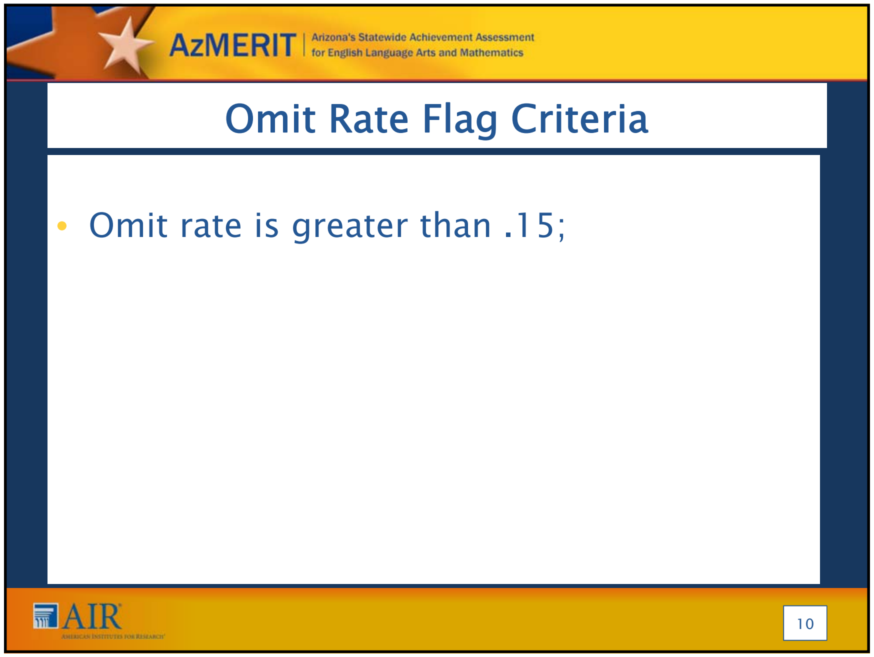
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Omit Rate

- Students do not provide a response

AIR American Institutes for Research

9



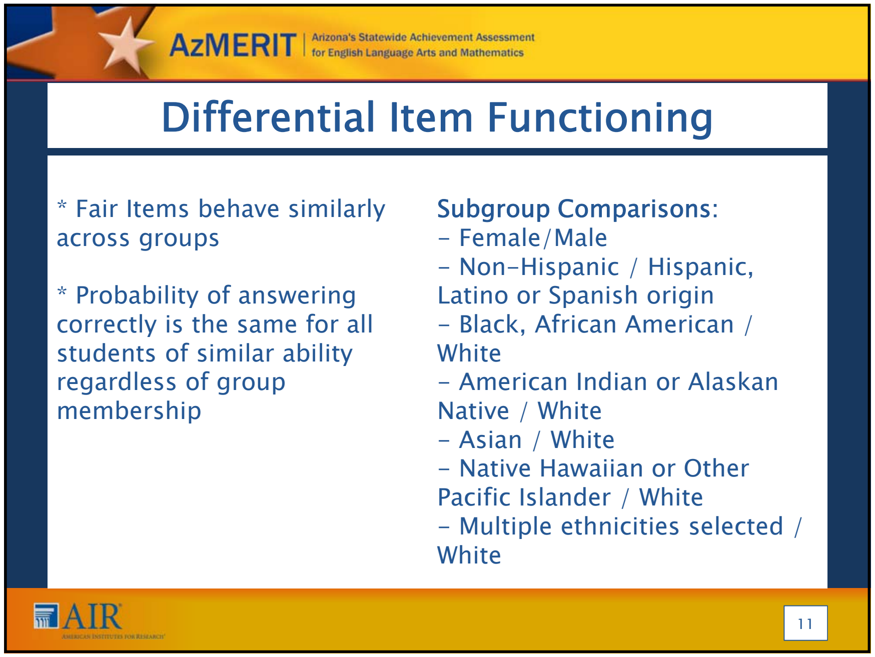
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Omit Rate Flag Criteria

- Omit rate is greater than .15;

AIR American Institutes for Research

10



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

Differential Item Functioning

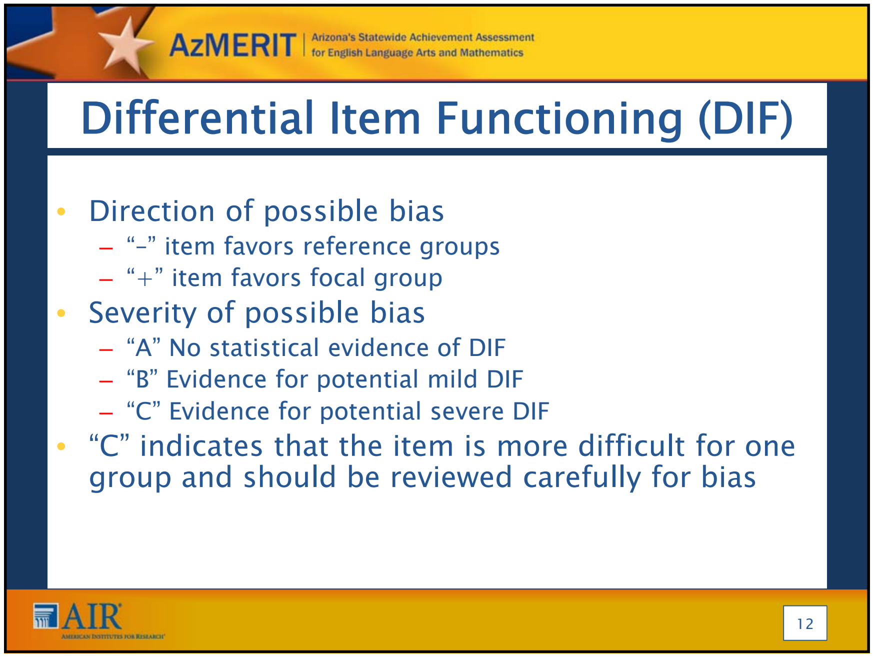
* Fair Items behave similarly across groups

* Probability of answering correctly is the same for all students of similar ability regardless of group membership

Subgroup Comparisons:
- Female/Male
- Non-Hispanic / Hispanic, Latino or Spanish origin
- Black, African American / White
- American Indian or Alaskan Native / White
- Asian / White
- Native Hawaiian or Other Pacific Islander / White
- Multiple ethnicities selected / White

AIR

11



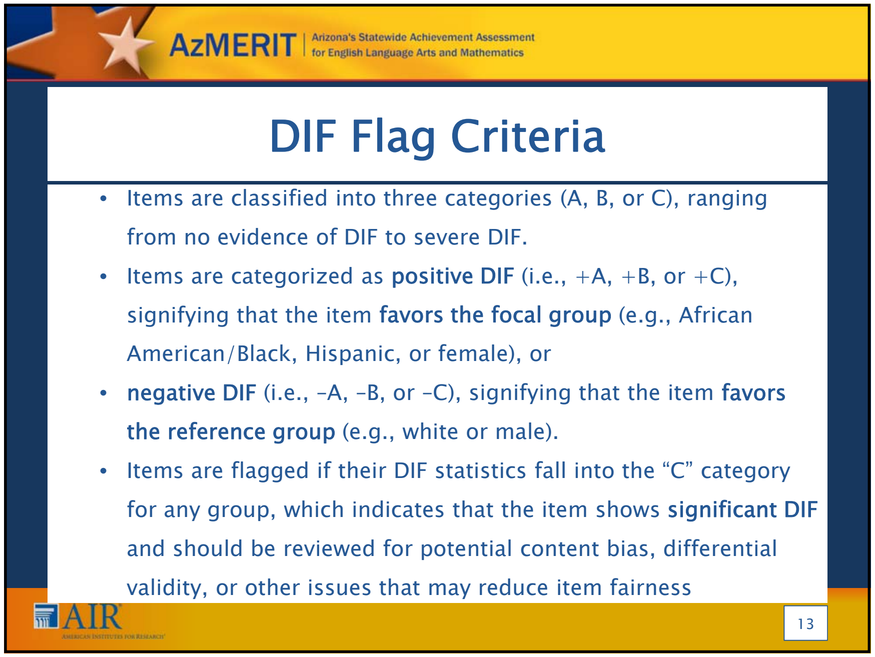
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

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



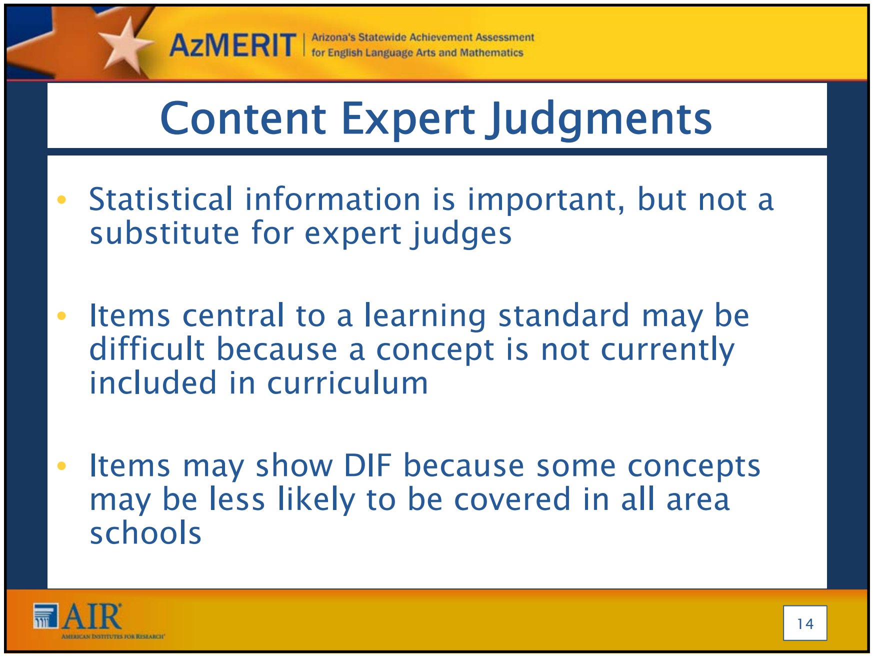
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

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 American Institutes for Research

13



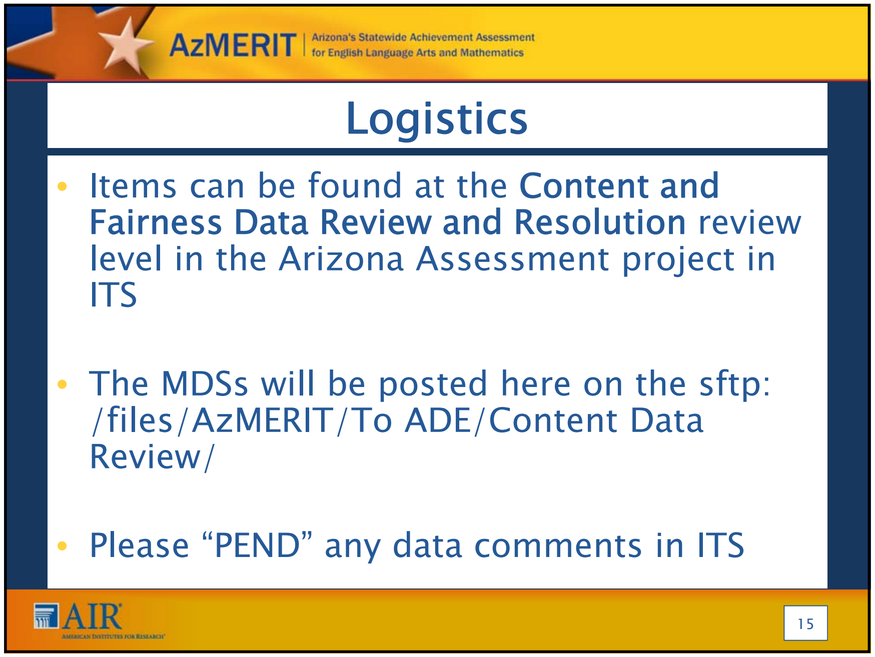
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Content Expert Judgments

- Statistical information is important, but not a substitute for expert judges
- Items central to a learning standard may be difficult because a concept is not currently included in curriculum
- Items may show DIF because some concepts may be less likely to be covered in all area schools

AIR American Institutes for Research

14



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

Logistics

- Items can be found at the **Content and Fairness Data Review and Resolution** review level in the Arizona Assessment project in ITS
- The MDSs will be posted here on the sftp: /files/AzMERIT/To ADE/Content Data Review/
- Please “PEND” any data comments in ITS

AIR American Institutes for Research

15