Cambium Assessment

Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Mathematics

2020–2021 School Year

September 2021

ARIZONA STATEWIDE ASSESSMENT

ARIZONA'S STATEWIDE ACHIEVEMENT ASSESSMENT (AZM2)

ENGLISH LANGUAGE ARTS GRADES 3–11

MATHEMATICS GRADES 3–8, GRADE 10

2020–2021 ANNUAL TECHNICAL REPORT

AUGUST 2021

Prepared by Cambium Assessment, Inc. (CAI) in collaboration with
the Arizona Department of Education (ADE)

TABLE OF CONTENTS

APPENDICES

1 INTRODUCTION: THE VALIDITY OF AZM2 TEST SCORE INTERPRETATIONS

1.1 OVERVIEW

The purpose of this technical report is to document the evidence supporting the claims made for how Arizona's Statewide Achievement Assessment (AzM2)[1] test scores may be interpreted. Evidence for the validity of test score interpretations is central to claims that AzM2 test scores can be used to evaluate the effectiveness with which Arizona districts and schools teach students the Arizona State Standards and if individual students have achieved those standards by the end of each school year. Thus, this report begins with a review of the validity evidence evaluated to date. Evidence for the validity of test score interpretations is expected to accrue over time; therefore, this section of the technical report will expand as more evidence is gained.

Chapter 2 describes the design and development of the AzM2 assessment system, including the Arizona State Standards, which define the content domain to be assessed by AzM2; the development of test specifications, including blueprints, that ensure that the assessments adequately sampled the breadth and depth of the content domain; and test development procedures that ensure alignment of test forms with the blueprint specifications.

Chapter 3 shows the results of the spring 2021 administration of the full AzM2 assessment system, including end-of-course (EOC) assessments in English language arts (ELA) and mathematics for grades 3–8 and high school. These chapters provide summaries of the test-taking student population and their performance on the assessments. Additionally, these chapters describe administration-specific evidence for the reliability of the AzM2 assessments, including internal consistency reliability, standard errors of measurement, and the reliability of performance-level classifications.

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

Chapter 4 describes the item development process, specifically the sequence of reviews that each item must pass through before being eligible for AzM2 test administration. This chapter also describes the procedures for constructing test forms from items successfully passing through the review process. Chapter 5 documents the test administration procedures, including eligibility for participation in the AzM2 assessments; testing conditions, including accessibility tools and accommodations; systems security for assessments administered online; and test security procedures for all test administrations. Chapter 6 provides a description of the score reporting system and the interpretation of test scores.

Chapter 7 describes the procedures that the Arizona Department of Education (ADE) uses to identify and adopt performance standards for AzM2 assessments. Chapter 8 describes the procedures used to scale and equate the AzM2 assessments for scoring and reporting. Chapter 9 describes the procedures for scoring constructed-response items, both machine-scored and handscored, and it provides summary rater agreement results. Chapter 10 provides an overview of the quality assurance (QA) processes used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

[1] Beginning with the 2019–2020 school year, Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) was renamed Arizona's Statewide Achievement Assessment (AzM2).

1.2 VALIDITY EVIDENCE

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

The evidence required to support the validity of test score interpretations depends 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 instead it’s of test score interpretations. Some test score interpretations are supported by validity evidence, while others are not. Thus, the test itself is not considered valid or invalid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Determining whether the test measures the intended construct is central to evaluating the validity of test score interpretations. This type of evaluation requires a clear definition of the measurement construct. For the AzM2, the Arizona State Standards define the measurement construct.

In 2010, Arizona adopted new academic content standards in ELA and mathematics. The Arizona State Standards are designed to ensure that students across grades receive the instruction they need to be on track for college and careers by the time they graduate.[2] In spring 2015, the ADE administered AzM2 to assess proficiency on the new Arizona State Standards for the first time. The AzM2 measured ELA and mathematics in grades 3–8 and, for high school students, follows the completion of coursework in ELA grades 9–11, as well as Algebra I, geometry, and Algebra II.

AzM2 measures students’ knowledge in the content areas of English language arts (ELA) and mathematics in grades 3–8 and 10 (Cohort 2023). Cohort is calculated for every student based on their initial enrollment in grade 9. Students who were enrolled in grade 9 in the academic year 2019–2020 belong to Cohort 2023. Each AzM2 test is aligned to Arizona’s College and Career Ready Standards (AZCCRS). AzM2 is available as a computer-based test (CBT) or as a paper-based test (PBT).

Because measuring student achievement directly against each benchmark in the Arizona State Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the Arizona State Standards.[3] To ensure that each student is assessed on the intended breadth and depth of the Arizona State Standards, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark.[4] 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

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

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

[4] Standard 4.1: Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended test-taker population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

determines how student achievement of the Arizona State Standards is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzM2 ELA and mathematics test blueprints that specify the distribution of items across reporting strands and Depth of Knowledge (DOK) levels. The ELA and mathematics blueprints are also provided in Appendix B.

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 provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement.[5] Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

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

Test development specialists receive extensive training on the principles of universal design and apply them 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 AzM2 Test Delivery System (TDS) provides a range of accessibility tools and accommodations to virtually all students for reducing construct-irrelevant barriers to accessing test content.[6] The range of accommodations provided in the online testing environment far exceeds the typical accommodations available in paper-based testing (PBT) administrations. Exhibits 1.2.1–1.2.5 list the accommodations and accessibility supports currently available for students taking the AzM2 assessments online. Paper-pencil test forms are available as an accommodation for students testing in online schools

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

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

should the accommodations provided online be insufficient to remove barriers to accessing test content. These include both large print and braille forms. Section 5.3 describes the available testing tools and accommodations for students testing online and, on a paper-pencil form.

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

- Testing in a small group, testing one-on-one, or testing in a separate location or study carrel
- Being seated in a specific location within the testing room or being seated using special furniture
- Having the test administered by a familiar TA
- Using a special pencil or pencil grip
- Using a placeholder
- Using devices that allow the student to see the test, such as eyeglasses, contact lenses, magnification, and special lighting
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT)
- Using devices that allow the student to hear the test directions, such as hearing aids and amplification tools
- Wearing noise buffers after the scripted directions have been read
- Signing the scripted directions using American Sign Language (ASL)
- Repeating the scripted directions at student request
- Answering questions about the scripted directions or the directions that students read on their own
- Reading the test quietly to himself/herself, as long as other students are not disrupted
- Providing extended time (the testing session must be competed in the same school day it was started; no student is expected to need more than twice the estimated testing time)

While some of the items listed as universal test administration conditions might be included in a student’s Individualized Education Plan (IEP) as an accommodation, for AzM2 testing purposes, these are not considered testing accommodations and are available to any student who needs them, not just to students with IEPs.

Exhibit 1.2.1 summarizes the universal testing tools available to all students in all AzM2 tests; these features cannot be disabled by TAs.

Exhibit 1.2.1 Universal Testing Tools for CBT Available to All Students

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

AzM2 testing requires specific subject-area tools or resources for certain portions of AzM2. The required tools are described in Exhibit 1.2.2.

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

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

Note: The details of the AzM2 calculator guidance are presented in Appendix A.

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

Research indicates that more accommodations is not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education (SPED) need, or language need and the accommodation(s) that are provided to the student during educational activities, including assessment. TAs are instructed to make accommodation decisions based on individual needs and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation that is not already used regularly in the classroom may be put in place for an AzM2 test.

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide clues or suggestions, verbal or otherwise, that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students during AzM2 testing are generally limited to those listed in the *AzM2 Testing Conditions, Tools, and Accommodations Guidance* manual and summarized in this section. The ADE takes care to ensure that allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzM2. If a student's IEP calls for a testing accommodation that is not listed, TAs are instructed to contact the ADE for guidance.

Students with an injury, such as a broken hand or arm, which would make it difficult to participate in AzM2, may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. There are no specific CBT tools to support these accommodations.

Exhibit 1.2.3 Accommodations for Injured Students

Accommodation	Description of Use
Adult Transcription	If a student with an injury is testing at a CBT school and cannot enter his or her responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided, verbally or by gestures, directly into the Data Entry Interface (DEI) or onto the paper-pencil booklet and then into the DEI. If a student with an injury at a PBT school cannot write his or her responses in a booklet, an adult must transfer the student's responses exactly as provided verbally or by gestures.
Assistive Technology	Assistive technology may be used for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copies must be shredded. Any electronic copies must be deleted. This accommodation also requires adult transcription (refer to the appropriate entry in this table for rules on adult transcription).
Rest/Breaks	Students may take breaks during testing sessions.

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

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

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

Exhibit 1.2.4 Allowable Accommodations for EL and FEP Students

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

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

Exhibit 1.2.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus for any AzM2 mathematics test without restrictions.
Adult Transcription	If a student testing at a CBT school has an IEP indicating that he or she cannot enter his or her responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided verbally or by gestures, directly into the DEI or onto the paper-pencil booklet and then into the DEI. If a student testing at a PBT school has an IEP indicating Adult Transcription, an adult must transfer the student's responses exactly as provided verbally or by gestures onto the paper-pencil booklet.
Assistive Technology	This is the use of assistive technology for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copies must be shredded. Any electronic copies must be deleted. This accommodation requires Adult Transcription (refer to the appropriate entry in this table for rules on Adult Transcription).
Braille Test Booklet	Provide a paper braille test booklet. This accommodation requires Adult Transcription (refer to the appropriate entry in this table for rules on Adult Transcription).
Large Print Test Booklet	CBT: Either increase default zoom settings when a student participates in CBT or provide a PBT Large Print test booklet. PBT: Provide a Large Print test booklet. PBT: Large Print test booklet requires Adult Transcription into the DEI (refer to the appropriate entry in this table for rules on Adult Transcription).
Paper-Pencil Test Booklet	CBT: Student's IEP must indicate that the student cannot enter his or her responses on the computer and requires a paper-pencil test or Adult Transcription. The school will provide a Special Paper Version booklet for the student. The student's responses must be entered directly into the DEI or transcribed onto the paper-pencil booklet and then entered into the DEI (refer to the appropriate entry in this table for rules on Adult Transcription).

1.3 EVIDENCE BASED ON TEST CONTENT

Because the AzM2 assessments are designed to measure student progress toward achieving the Arizona State Standards, the validity of AzM2 test score interpretations critically depend on the degree to which test content is aligned with the expectations for student learning specified in the academic standards.[7]

Alignment of content standards is achieved through a rigorous test-development process that proceeds from the content standards and refers to those standards in a highly iterative process that includes the ADE, test developers, and educator committees. Since spring 2016, the items used to develop operational test forms were drawn from custom Arizona item development and CAI's Independent College and Career Readiness (ICCR) item bank. Both custom Arizona items and ICCR

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

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

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards will be covered in each test administration.[8] Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the Arizona State Standards is evaluated, the alignment of test blueprints with the content standards is critical.

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

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

- **Preliminary Review.** The item is reviewed by a group of Cambium Assessment, Inc. (CAI) content-area experts.
- **Content Review 1.** The item is reviewed by a CAI content specialist.
- **Editorial Review.** A copy editor checks the item for correct grammar/usage.
- **Senior Content Review.** The item is reviewed by a lead content expert.

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

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

[8] Standard 4.1: Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended test-taker population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

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

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

Items passing through the internal review process are sent to the ADE for review. At this stage, items may be further revised based on any edits or changes requested by the ADE, or they may be rejected outright. Items passing through the ADE's review must then pass through a stakeholder review in which a committee of educators reviews each item's accuracy, alignment to the intended standard and DOK level, and item fairness and language sensitivity. Thus, all items considered for inclusion in the AzM2 item pools were initially reviewed by an educator committee, which checked to ensure that each item and associated stimulus materials was

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

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

Additionally, rubric-scored items, both machine-scored and human-scored, are validated following field test administration. Machine-scored items go through a rubric validation process wherein samples of student responses are reviewed, along with resulting scores, to ensure that rubrics are enacted as intended. This process is described in Section 9.1.1. Human-scored items go through a rangefinding process before scoring in which samples of item responses are used to create scorer training materials and ensure that the scoring rubric is appropriate, as described in Section 9.1.2.

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

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

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

1.4 EVIDENCE FOR INTERPRETATION OF PERFORMANCE STANDARDS

The alignment of test content to the Arizona State Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the detailed learning expectations. However, the interpretation of AzM2 test scores rests fundamentally upon how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the Arizona standards. AzM2 test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the Arizona State Standards. The cut score establishing the Proficient level of performance is the most critical because it indicates that students are meeting grade-level expectations for achievement of the Arizona standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Therefore, procedures used to adopt performance standards for the AzM2 assessments are central to the validity of test score interpretations.[9]

Following the first operational administration of the AzM2 in spring 2015, a standard-setting workshop was conducted to recommend a set of performance standards for reporting student achievement of the Arizona State Standards to the Arizona State Board of Education. Arizona educators, serving as standard-setting panelists, followed a standardized and rigorous procedure to recommend performance-level cut scores. The workshops employed the Bookmark procedure, a widely used method in which standard-setting panelists used their expert knowledge of the Arizona State Standards and student achievement to map the Performance-Level Descriptors (PLDs) adopted by Arizona to an ordered-item booklet (OIB) comprising the spring 2015 operational test form and augmented with items administered in the embedded field-test slots to minimize information gaps in the operational test form.[10]

Panelists were provided with contextual information to inform their primarily content-driven cut-score recommendations. In addition, for each assessment, panelists were provided with the approximate location of performance standards for other important assessment systems. The panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant American College Testing (ACT) college-ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and geometry assessments. Panelists recommending performance standards for the grades 3–8 summative assessments were provided with the approximate location of relevant performance standards for the National Assessment of Educational Progress (NAEP) at grades 4 and 8, and interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced

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

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

performance standards for the grades 3–8 and 11 assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided with the corresponding locations for the previous performance standards for Arizona's Instrument to Measure Standards (AIMS). They were asked to consider the location of these benchmarks when making their content-based cut-score recommendations. When panelists can use benchmark information to locate performance standards that converge across assessment systems, the validity of test score interpretation is bolstered.

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

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

Based on the adopted performance standards, Exhibit 1.4.1 shows the estimated percentage of students meeting the AzM2 proficient standard for each assessment in spring 2015. Exhibit 1.4.1 also shows the approximate percentage of Arizona students expected to meet the ACT college-ready standards and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. It also shows the expected proficient rate for the Smarter Balanced assessments, system-wide, based on the spring 2014 field-test administration. As indicated, the performance standards recommended for AzM2 assessments are quite consistent with relevant ACT college-ready standards, and NAEP and Smarter Balanced proficient benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 1.4.1 Percentage of Students Meeting AzM2 and Benchmark Proficient Standards

		Percentage of Students Meeting Standards		
Test	**AzM2 Proficient**	**Arizona ACT College-Ready**	**Arizona NAEP Proficient**	**Projected SBAC**
		ELA		
Grade 3	41%			38%
Grade 4	38%		28%	41%
Grade 5	30%			44%
Grade 6	34%			41%
Grade 7	33%			38%
Grade 8	32%		28%	41%
Grade 9	27%			
Grade 10	30%			
Grade 11	25%	34%		41%
		Mathematics		
Grade 3	42%			39%
Grade 4	42%		42%	38%
Grade 5	40%			33%
Grade 6	32%			33%
Grade 7	31%			33%
Grade 8	33%		32%	32%

Percentage of Students Meeting Standards				
Test	AzM2 Proficient	Arizona ACT College-Ready	Arizona NAEP Proficient	Projected SBAC
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

Although CAI previously identified ACT college-ready cut scores on the AzM2 ELA and mathematics scales for the standard-setting committee's use in 2015, that study involved an indirect linkage. In that study, student performance on the grade 10 AIMS was used to predict subsequent student performance on the ACT tests. Then, a linking study between the AIMS and AzM2 allowed for the identification of the ACT cut scores on the AIMS scale to be represented on the AzM2 scale.

To directly examine the relationships between the AzM2 and ACT assessments, the ADE obtained the ACT test scores for Arizona students graduating high school in spring 2016. More details of the direct linking study using AzM2 and ACT data are shown in Section 8.5.2 of this technical report.

Exhibit 1.4.2 shows the location of the ACT college-ready cut scores for mathematics and reading on the AzM2 scale. The first column shows the location as identified via indirect linkage through AIMS, and this was provided as benchmark information to AzM2 standard-setting panelists. The second column shows the location of the ACT college-ready cut scores as identified via direct linkage between ACT and AzM2 described here. The third column shows the location of the AzM2 Meets Performance Standards on the Algebra II and grade 11 ELA assessments. As indicated in the table, the location of the ACT college-ready cut scores on the AzM2 scale was reasonably consistent across methods, especially for ELA. Notably, the results affirm that the location of adopted AzM2 performance standards is consistent with the ACT college-ready criteria.

Exhibit 1.4.2 Locations of the ACT College-Ready Cut Scores on the AzM2 Scales

	Location of ACT College-Ready Cut Scores on AzM2 Scale		AzM2 Meets Performance Standard
	Via Indirect Linkage Through AIMS	Via Direct Linkage with AzM2	
Algebra II	3,704	3,727	3,711
Grade 11 ELA	2,579	2,585	2,585

The equipercentile equating method was used to verify the linkage between ACT and AzM2 test scores. The AzM2 scale score associated with the ACT college-ready cut scores in reading was 2,585 on the AzM2 ELA scale. The location of the ACT college-ready cut score in mathematics was 3,727 for the AzM2 mathematics scale. Results from the equipercentile approach were thus consistent with the cut scores identified using regression models.

1.4.1 AZM2 GRADE 10 MATHEMATICS PERFORMANCE STANDARDS

Although originally scheduled for spring 2020, in spring 2021, the ADE began to transition away from the end-of-course (EOC) assessments in high school. Instead, they administered a grade 10 summative assessment in spring 2021, and in spring 2022, they have adopted the ACT college entrance exam as the high school accountability assessment. Although not an EOC test, the new grade 10 summative mathematics test did not comprehensively sample the full breadth of the high school mathematics standards. Rather, the revised test design measured student achievement of Algebra I and geometry academic content standards, reflecting the standard configuration of high school mathematics coursework, with instruction in Algebra I typically completed by grade 9 and instruction in geometry typically completed by grade 10.

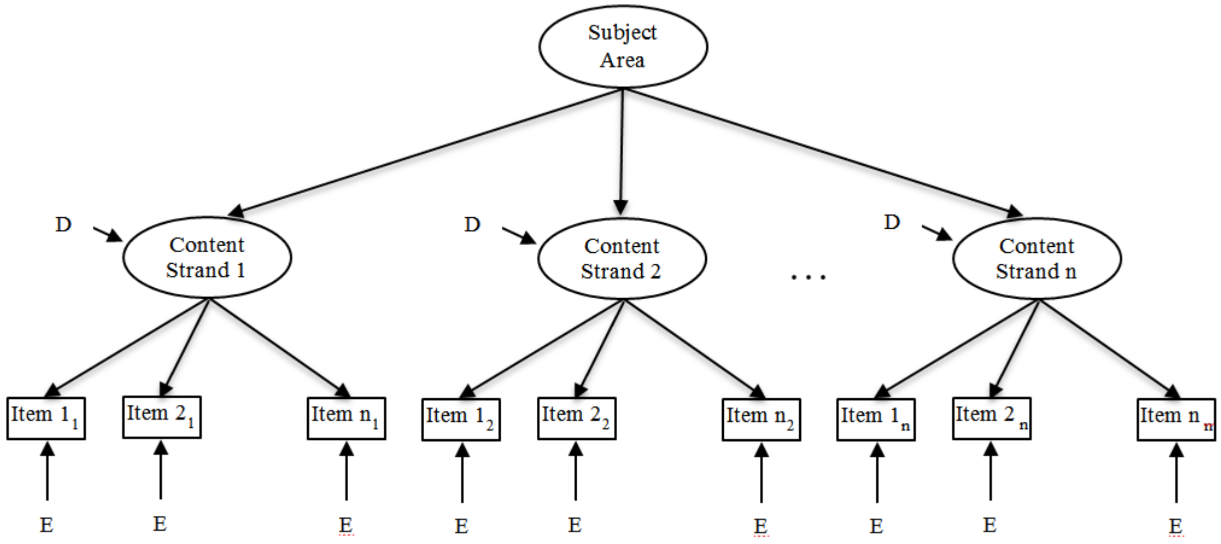
Although the mathematics standards measured in the grade 10 summative mathematics test are the same standards covered separately by the previous Algebra I and geometry assessments, the revised blueprint and the displacement of the algebra content from course instruction substantially altered the measurement constructs assessed in the new grade 10 summative assessment. To account for changes to the measurement construct, ADE had originally planned to conduct a standard-setting workshop to recommend new performance standards for the grade 10 summative mathematics test following the spring 2020 test administration.

However, with the cancellation of the spring 2020 assessments and only a single administration of the new summative assessment scheduled for spring 2021, ADE saw merit in maintaining the current geometry performance standards for the one-year administration of the grade 10 summative mathematics assessment. With this approach, following the spring 2021 administration of the grade 10 summative mathematics assessment, geometry items, comprising approximately 40% of grade 10 summative test items, were anchored to their bank values with the algebra items calibrated under this constraint, placing all items on the geometry scale. Linking the summative test items to the geometry scale allowed ADE to report summative assessment performance standards that reflect the same level of general ability or rigor as those reported for the geometry end-of-course assessment. However, test score interpretations would be quite different given the change in test design.

1.5 EVIDENCE BASED ON INTERNAL STRUCTURE

The AzM2 assessment represents a structural model of student achievement in grade-level and course-specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., Reading Information, Reading Literature, language, writing). Content strands within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Exhibit 1.5.1. As the exhibit illustrates, each item is an indicator of an academic content strand. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content strand serves as an indicator of achievement in a subject area. As at the item level, the content strands include an error term indicating that the content strands are not pure indicators of overall achievement in the subject area. The paths from the content strands to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content strand construct. Similarly, the paths from subject-area achievement to the content strands represent the second-order factor loading, indicating the degree to which academic content strand constructs are correlated with the underlying construct of subject-area achievement.

Exhibit 1.5.1 Second-Order Structural Model for AzM2 Assessments



Confirmatory factor analysis (CFA) was used to evaluate the fit of this structural model to student response data.[11] We examined the goodness of fit between the structural model and the operational test data. The goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size; however, even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices, such as the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis Index (Tucker & Lewis, 1973), and the root mean square error of approximation (RMSEA) were also used to evaluate model fit. The guidelines for evaluating the goodness of fit are presented in Exhibit 1.5.2.

The AzM2 assessments also claim to measure subject-area achievement using test items that probe student knowledge and skills across multiple DOKs. As with the content standards, the classification of items by DOK also represents a structural model that can be evaluated using CFA.[12] In this case, each item is an indicator of a DOK level first-order factor, and each DOK level is an indicator of subject-area achievement. Thus, CFA was used to evaluate the fit of this DOK structural model to student response data from the spring 2019 AzM2 test administration.

Exhibit 1.5.2 Guidelines for Evaluating Goodness of Fit

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

In addition to testing the fit of the hypothesized AzM2 second-order CFA model, we examined the degree to which the second-order model improved fit over the more general one-factor model of academic achievement in each subject area. Because the one-factor, general-achievement model was nested within the second-order model, a simple likelihood ratio test was used to determine whether the added information provided by the structure of the Arizona State Standards frameworks improved model fit over a general-achievement model. Results indicating improved model fit for the second-order factor model support the interpretation of content standard performance above that provided by the overall subject-area score.[13]

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

1.5.1 ELA CONTENT MODEL

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

Exhibit 1.5.1.1 Goodness of Fit for the AzM2 ELA First-Order Model

Grade	CFI	TLI	RMSEA
3	0.97	0.96	0.04
4	0.97	0.97	0.03
5	0.97	0.97	0.03
6	0.97	0.97	0.03
7	0.97	0.97	0.03
8	0.97	0.97	0.03
10	0.96	0.96	0.03

The goodness-of-fit statistics for the hypothesized AzM2 second-order models in ELA are shown in Exhibit 1.5.1.2. All the statistics indicate that the second-order models posited by the AzM2 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 above.95, with RMSEA values well below the .05 threshold used to indicate good fit.

The results of the comparison between the hypothesized AzM2 model and the general-achievement model are presented in Exhibit 1.5.1.3. We note that model fit for the first-order, general-achievement model was also very high and provides evidence for the unidimensionality of the subject-area assessments. The purpose of these analyses is to determine whether the posited second-order reporting model adds information beyond that provided by the first-order model. The chi-square difference test shows that, across grade levels, the strand-based, second-order model showed significantly better fit than the first-order, general-achievement model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.5.1.2 Goodness of Fit for the AzM2 ELA Second-Order Model

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

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

Grade	χ^2	*df*	*p* value
3	11,104.664	3	p < .001
4	8,710.343	3	p < .001
5	11,209.327	3	p < .001
6	7,277.245	3	p < .001

Grade	χ^2	*df*	*p* value
7	6,496.039	3	p < .001
8	9,434.533	3	p < .001
10	2,080.738	3	p < .001

1.5.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzM2 second-order models in ELA are shown in Exhibit 1.5.2.1. Across all grades, results indicate that the second-order models posited by the AzM2 assessments fit the data well. The CFI and TLI values were .98–.99 and the RMSEA values were all .02.

Exhibit 1.5.2.1 Goodness of Fit for the AzM2 ELA Second-Order Model

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

The results of the comparison between the hypothesized AzM2 model and the general-achievement model are shown in Exhibit 1.5.2.2. The chi-square difference test shows that, across grade levels, the DOK-based second-order model showed significantly better fit than the first-order, general-achievement model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

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

Grade	χ^2	*df*	*p* value
3	10,941.713	4	*p* < .001
4	9,541.961	4	*p* < .001
5	9,820.848	4	*p* < .001
6	8,350.609	4	*p* < .001
7	6,979.488	4	*p* < .001
8	10,244.295	4	*p* < .001
10	5,643.834	4	*p* < .001

1.5.3 MATHEMATICS CONTENT MODEL

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

Exhibit 1.5.3.1 Goodness of Fit for the AzM2 Mathematics First-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.95	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.03
7	0.99	0.98	0.02
8	0.97	0.97	0.03
10	0.99	0.98	0.02

The goodness-of-fit statistics for the strand-based, second-order models are shown in Exhibit 1.5.3.2. The models show very good fit, with the CFI and TLI fit indices above .95, and RMSEA estimates well below their .05 cut-off values. These statistics indicate that the second-order models are a good fit for the data.

Exhibit 1.5.3.2 Goodness of Fit for the AzM2 Mathematics Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.96	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.02
7	0.99	0.99	0.02
8	0.97	0.97	0.03
10	0.99	0.98	0.02

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

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

Grade	χ^2	*df*	*p* value
3	4,858.475	2	$p < .001$
4	7,470.266	2	$p < .001$
5	6,475.997	3	$p < .001$
6	2,124.797	4	$p < .001$
7	1,269.169	4	$p < .001$
8	6,948.457	3	$p < .001$
10	242.674	4	$p < .001$

1.5.4 MATHEMATICS DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 1.5.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .95 and RMSEA estimates well below their .05 cut-off values. These statistics indicate that the second-order models are a good fit for the data.

Exhibit 1.5.4.1 Goodness of Fit for the AzM2 Mathematics Second-Order Model

Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.95	0.95	0.04
5	0.97	0.97	0.03
6	0.98	0.98	0.02
7	0.99	0.98	0.02
8	0.97	0.97	0.03
10	0.99	0.98	0.02

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

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

Grade	χ^2	*df*	*p* value
3	276.254	3	$p < .001$
4	1,296.511	3	$p < .001$
5	1,064.235	3	$p < .001$
6	2,275.704	3	$p < .001$
7	127.198	3	$p < .001$
8	2,819.923	3	$p < .001$
10	260.426	3	$p < .001$

1.6 EVIDENCE FOR RELATIONSHIPS WITH CONCEPTUALLY RELATED CONSTRUCTS

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and the variables of interest 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.[14]

Observed correlations between alternate indicators of student achievement of course objectives, such as locally administered assessments of student achievement and AzM2, should be limited only by the unreliability of the measures. When both assessments measure student achievement in common subject areas, such as with locally administered and statewide assessments of mathematics achievement, we expect test scores among the common subject-area assessments to be substantially correlated. Additionally, we expect that the magnitude of observed correlations among test scores in different subject areas will be lower than those in a common subject area. Because the content domains assessed in ELA and mathematics tests are quite different, AzM2 ELA test scores should correlate less well with locally administered

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

assessments of mathematics than ELA. However, it is important to note that test scores across subject areas and test systems nevertheless are expected to be highly correlated. This expectation is because, even though subject-area test scores measure different academic content domains, student achievement across subject areas is influenced by internal (e.g., general intelligence) and external (e.g., socioeconomic status) factors. These factors contribute to student achievement across all academic subject areas so that student test scores across subject areas tend to be highly intercorrelated. Therefore, while we certainly expect correlations among test scores across subject areas to be lower than correlations among test scores within a subject area, we nevertheless expect correlations among test scores across subject areas to be quite high.

Exhibit 1.6.1 shows the correlations among student test scores on the spring 2015 statewide AzM2 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 0.82–0.84 between AzM2 ELA (assessing reading, writing, and listening) and NWEA reading. Correlations between AzM2 and NWEA mathematics scores are even higher, ranging from 0.85–0.89.

Exhibit 1.6.1 Correlations Between AzM2 and Locally Administered NWEA Test Scores

Grade	ELA Sample Size	ELA Correlation	Mathematics Sample Size	Mathematics Correlation
3	1,426	0.82	1,429	0.86
4	1,214	0.84	1,214	0.88
5	1,303	0.84	1,303	0.88
6	1,119	0.82	1,115	0.85
7	1,081	0.82	1,082	0.89
8	1,090	0.82	1,091	0.89

Exhibit 1.6.2 shows the discriminant correlations between AzM2 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 measure domain-specific knowledge and skills in addition to common factors underlying student achievement.

Exhibit 1.6.2 Discriminant Correlations Between AzM2 and Locally Administered NWEA Test Scores

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

Convergent correlations between AzM2 and locally administered assessments were also reported by Estrada and colleagues (Estrada, Burnham, Feld, Bergan, & Bergan, 2015). These researchers reported the mean correlations among various local assessments and AzM2 test scores for ELA and mathematics assessments in grades 3–8. The mean correlations between AzM2 and various local assessments of ELA ranged from .77–.79 across the grade levels investigated and the mean correlations between AzM2 and local mathematics assessments ranged from .71–.75 across grades 3–8. These results show good convergence among AzM2 and other locally administered assessments purporting to measure the same constructs.

1.7 MEASUREMENT INVARIANCE ACROSS SUBGROUPS

Measurement invariance occurs when the likelihood of responding correctly conforms to the measurement model and is independent of group membership and when the parameters of a measurement model are statistically equivalent across groups.[15] The parameters of interest in measurement invariance testing are the factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that it is not necessary to demonstrate invariance across groups on these parameters. In general, measurement invariance testing can be conducted using a series of multiple-group CFA models, which impose identical parameters across groups. The measurement model parameters—including factor patterns (configural invariance), factor loadings (metric or weak invariance), latent intercepts/thresholds (scalar or strong invariance), and unique or residual factor variances (strict invariance)—are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups.

Appendix C shows the results of measurement invariance testing by subgroups for ELA and mathematics. The full set of tables associated with these analyses is provided for each grade-level and subject-area assessment. The series “a” tables (e.g., Tables B.1a, B.2a) show the global model fit indices for the measurement invariance tests for each assessment. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using the χ2 difference test (at $\alpha \leq 0.05$) and the examination of significant differences of the root mean square error of approximation (RMSEA, change in RMSEA ≤ 0.015; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across the following subgroups: gender (Model A); ethnicity including African American vs. White (Model B-1), Hispanic vs. White (Model B-2), Asian vs. White (Model B-3), American Indian vs. White (Model B-4), and Multi-Ethnic vs. White (Model B-5); special education program status (SPED; Model C); economic disadvantage status (Low Income; Model D); limited English proficiency status (LEP; Model E); and accommodated test forms (Accommodation, Model F). Invariance tests of subgroups were investigated separately for each grade-level and subject-area test. Because in each ELA assessment, students were randomly assigned to one of six writing prompts for administration, the missing responses on the writing items resulted in unsuccessful model convergence. Thus, to achieve model convergence, we included the students who took a common writing prompt for online and paper-pencil tests in each ELA assessment.

The null hypothesis of the χ2 difference test is that the more restricted invariance model (e.g., metric) fits the data equally and the less restricted invariance model (e.g., configural). Given the sensitivity of the χ2 difference tests to sample size, we examined additional significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retaining the more restricted invariance model (Chen, 2007).

The “b” series tables in Appendix C (e.g., Tables C.1b, C.2b) show the model fit indices of scalar invariance models assuming the same factor pattern + identical factor loadings + identical latent intercept/threshold across subgroups. Global model fit indices included the CFI and RMSEA. CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.947 to 0.989, and RMSEA ranged from 0.013 to 0.035. For mathematics, CFI values ranged from 0.943 to 0.991, and RMSEA ranged from 0.011 to 0.043.

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

Although the χ2 difference test ideally should be nonsignificant, all χ2 difference tests were significant at α = .05 due to large sample sizes. Despite significant χ2 difference tests for most models, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0.000 to 0.002 for both ELA and mathematics). Based on the similar magnitudes of the RMSEA (i.e., no material change across all tested models; Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, ELA and mathematics test scores have the same measurement structure across gender, ethnicity (African American vs. White, Hispanic vs. White, Asian vs. White, American Indian vs. White, and Multi-Ethnic vs. White), SPED status, economically disadvantaged status, limited English proficiency status, and accommodation test forms.

1.8 DIFFERENTIAL MODE EFFECTS ACROSS SUBGROUPS

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

To identify possible differential effects of mode across subgroups, we used the observed test score as the dependent variable and then covaried the predicted test score to isolate the impact of mode. The demographic variables of interest include gender, EL status, SPED, free or reduced-price lunch (FRL) status, migrant status, and six ethnicity subgroups as predictors. We created dummy-coded variables to represent those non-white ethnicities with 0 as no and 1 as yes. Additionally, gender was coded as 0 for male and 1 for female. EL was coded as 1 for students as EL and 0 for non-EL. SPED was coded as 1 for students in a SPED program and 0 for students not attending any SPED program. FRL (or Social Economic Status [SES]) was coded as 1 for students having FRL and 0 as non-FRL. Migrant was coded as 1 for students from a migrant family and 0 for non-migrant students. Significant interactions between the mode of test administration and the demographic subgroup comparisons indicate differential mode effects among the specified demographic subgroups.

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

Results indicated that mode effects were more pronounced for SPED students relative to the general education population. Especially for the high school EOC tests, AzM2 tests were more difficult for SPED students when administered a paper-pencil test than an online test.

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

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

Exhibit 1.8.1 Parameter Estimates for Differential Mode Effects by Subgroups Interactions

Test	Gender	White	Black	Asian	Native Hawaiian/Pacific Islander	Hispanic/Latino	American Indian	Special Education	Limited English Proficiency	Free/Reduced-Lunch	Migrant
ELA											
Grade 3E	0.49									0.27	
Grade 4E											
Grade 5E											
Grade 6E								-0.61			
Grade 7E								0.50			
Grade 8E					1.66	-0.34					
Grade 9E	0.45							-0.74			
Grade 10E								-1.23		-0.41	
Grade 11E	-0.33					0.36		-0.58			
Mathematics											
Grade 3M								0.57			
Grade 4M									0.52	-	-4.46
Grade 5M							-0.89			0.34	
Grade 6M		1.15	0.96				0.69		0.60	-0.31	
Grade 7M	-0.26									0.25	-2.87
Grade 8M		0.89					0.86		-0.58		
Algebra I						0.73		-0.80	-0.95	0.50	
Geometry						-0.44		-1.32		1.11	
Algebra II							-1.07	-0.75		0.63	

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

1.9 EVIDENCE FOR STUDENT GROWTH—OVERALL AND BY SUBGROUPS

The AzM2 assessments report student test scores on a vertical scale, allowing families and teachers to make inferences about student growth across school years. The validity of test score interpretations about student growth over time depends strongly on the vertical linking design used to develop the vertical scale. But even when test score interpretations are appropriate to the scaling design, it is important to examine whether student gains may be interpreted consistently across subgroups or whether differential gain rates across subgroups limit the inferences that can be made about these gains over time.[16] To address this issue, we examined student growth rates across student gender, race/ethnicity, SPED, limited English proficiency (LEP), and low-income status (Low Income).

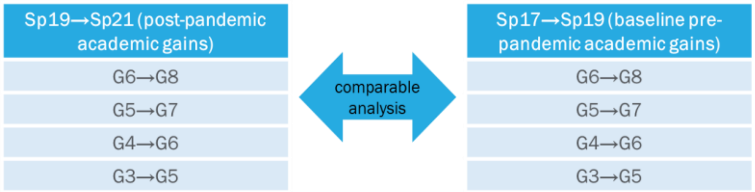
[16] Standard 3.15: Test developers and publishers who claim that a test can be used with test takers from specific subgroups are responsible for providing the information necessary to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Exhibit 1.9.1 shows the mean test scores on the spring 2019 and the spring 2021 administrations of AzM2 for students participating in both test administrations and the correlation between test scores across the two assessment occasions. Correlations between test scores are quite high and indicate substantial consistency in rank ordering of student achievement between the two test administrations.

Exhibit 1.9.1 Test Score Stability and Performance Gains Overall

Assessment 2019→2021	N	Spring 2019 Scale Score		Spring 2021 Scale Score		Change from 2019 to 2021		Percentage Scoring Lower		Correlation
		Mean	Std. Dev	Mean	Std. Dev	Mean	IRT based Standard Error	Expected	Observed	
ELA										
G3E→G5E	65,888	2506	30.92	2536	38.02	30	14.74	0.13	0.08	0.81
G4E→G6E	68,533	2525	32.18	2542	32.53	17	14.34	0.24	0.18	0.81
G5E→G7E	69,133	2544	36.87	2549	33.73	6	15.35	0.41	0.39	0.81
G6E→G8E	69,592	2547	32.44	2557	36.28	10	14.40	0.35	0.31	0.83
Mathematics										
G3M→G5M	67,116	3529	43.82	3573	42.86	44	17.06	0.10	0.07	0.77
G4M→G6M	69,673	3559	44.83	3604	44.20	45	16.67	0.09	0.06	0.79
G5M→G7M	70,534	3590	42.28	3628	42.41	38	16.25	0.11	0.07	0.81
G6M→G8M	70,994	3619	43.94	3650	40.49	31	16.03	0.15	0.11	0.82

Longitudinal analyses were conducted to examine the differential gains in student academic achievement across years. The spring 2021 summative subject-area test scores were regressed onto prior student area achievement and demographic variables. Since the 2020 test data are not available, the spring 2021 scores were regressed to spring 2019 scores, representing a two-year growth. This two-year growth baseline was created to detect if there was any difference between the pre-pandemic and the post-pandemic growth. The spring 2017 to spring 2019 score gain, i.e., the pre-pandemic growth, was used as the baseline for the two-year growth comparison. The following graph shows the design of the cohort comparison.


Sp19→Sp21 (post-pandemic academic gains)
G6→G8
G5→G7
G4→G6
G3→G5
comparable analysis
Sp17→Sp19 (baseline pre-pandemic academic gains)
G6→G8
G5→G7
G4→G6
G3→G5

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

To examine if there was a differential cohort effect in the student academic growth, we combined the testing data from the two cohorts (2017–2019 and 2019–2021) into one dataset. For example, in the grade 3 to grade 5 growth model, the spring 2021 grade 5 and spring 2019 grade 5 scores were combined as the dependent variable, and the spring 2019 grade 3 and spring 2017 grade 3 scores were combined as an independent variable. A dummy variable was created to represent cohort: 1 for the records in the 2019–2021 cohort; 0 for the records in the 2017–2019 cohort. The grade g score is the dependent variable in the regression model. The grade g-2 score is included as an independent variable. To compare ethnic subgroup performance, we created six dummy variables contrasting white students with each of the other ethnic groups (e.g., Hispanic vs. White, African American vs. White, Hawaiian/Pacific Islander vs. White, American Indian vs. White, Multiple Race vs. White, Asian vs. White). Gender was coded 1 for female. SPED, LEP, and Low-Income students were coded as 1 to contrast with students who were not identified with those needs and were coded as 0.

In addition, the dummy coded cohort variable and the interaction between the cohort variable and each of the predictors were also included in the regression model as predictors. This cohort regression model allowed us to examine whether there were any differential gains between the two cohorts and determine which demographic groups might have been differentially impacted.

Exhibit 1.9.2 and Exhibit 1.9.3 show the standardized regression coefficient estimates and partial R-squared of the differential effect on students' growth across subgroups. Although many individual effects attained conventional levels of statistical significance due to large sample sizes, we focused only on highly significant effects ($p < 0.0001$) and non-zero partial squared associated with more practically significant effect sizes that may point to trends across grade-level and/or subject-area assessments. Appendix D shows the regression model parameter estimates of differential growth for the ELA and mathematics assessments, including standardized and unstandardized coefficients, the standard error of the unstandardized coefficient, *p*-value, and R-squared regardless of the significance level.

The pre-pandemic differential growth across demographic subgroups is shown under the "Intercept" section. The results indicate that females generally performed better than males for ELA across grades in the spring 2019 test scores. With respect to ethnicity, African American students, Hispanic students, and American Indian students generally performed less well than white students in both ELA and mathematics. Asian students generally performed better than white students in both ELA and mathematics. For all other ethnic group comparisons, the focal groups generally performed less well than whites. Special education (SPED) students, limited English proficient (LEP) students, and low-income students all performed less well than the general education population in both ELA and mathematics.

Differential growth between the pre-pandemic cohort and the post-pandemic cohort across demographic subgroups is presented under the "Cohort by Intercept" section. The results indicated a smaller gain in post-pandemic growth compared to pre-pandemic growth in both ELA and mathematics except in grade 6 and grade 8 ELA assessment. Looking at the standardized coefficient estimates and partial R-squared across growth models, the largest decline in students' growth between the two cohorts is found in the grade 3 to grade 5 growth model for ELA and mathematics, and the achievement loss between the two cohorts got smaller as the grade level increased. No significant differential growth between the pre-pandemic and post-pandemic cohorts is observed for any of the demographic subgroups.

Exhibit 1.9.2 Standardized Regression Coefficient and Partial R-squared of Differential Growth Across Subgroups: ELA

Effect	2019 G3E -> 2021 G5E		2019 G4E -> 2021 G6E		2019 G5E -> 2021 G7E		2019 G6E -> 2021 G8E	
	β	Partial R^2	β	Partial R^2	β	Partial R^2	β	Partial R^2
Intercept	0.00	.	0.00	.	0.00	.	0.00	.
Female vs. Male	0.04	0.01	0.04	0.01	0.06	0.02	0.08	0.02
LEP vs. non-LEP	-0.10	0.09	-0.07	0.07	-0.06	0.07	-0.06	0.07
SPED vs. non-SPED	-0.09	0.11	-0.09	0.10	-0.07	0.11	-0.08	0.11
Low Income vs. non-Low Income	-0.02	0.04	-0.03	0.04	-0.02	0.04	-0.03	0.03
Hispanic vs. White	-0.05	0.02	-0.03	0.02	-0.03	0.02	-0.04	0.02
African American vs. White	-0.03	0.01	-0.03	0.01	-0.02	0.01	-0.02	0.01
Hawaiian/Pacific Islander vs. White								
American Indian vs. White	-0.05	0.02	-0.04	0.02	-0.03	0.02	-0.04	0.02
Multiple Race vs. White								
Asian vs. White					0.04	0.01	0.03	0.01
Cohort by Intercept	-0.06	0.01	-0.05	0.01	-0.03	0.00	-0.04	0.00
Female vs. Male								
LEP vs. non-LEP								
SPED vs. non-SPED								
Low Income vs. non-Low Income								
Hispanic vs. White								
African American vs. White								
Hawaiian/Pacific Islander vs. White								
American Indian vs. White								
Multiple Race vs. White								
Asian vs. White								

Note: β = Standardized regression coefficient. R^2=R squared. For the effect of special groups, the coefficient represents the difference compared to their contrast group; SPED = Special Education Status *vs*. Non-SPED. LEP = Limited English Proficiency *vs*. Non-LEP, Low Income = Low Income *vs.* Non-Low Income. For the effect of ethnic groups, the coefficient represents differential growth rate compared to White students.

Exhibit 1.9.3 Standardized Regression Coefficient and Partial R-squared of Differential Growth Across Subgroups: Mathematics

Effect	**2019 G3M -> 2021 G5M**		**2019 G4M -> 2021 G6M**		**2019 G5M -> 2021 G7M**		**2019 G6M -> 2021 G8M**	
	β	**Partial R^2**	***β***	**Partial R^2**	***β***	**Partial R^2**	***β***	**Partial R^2**
Intercept	0.00	.	0.00	.	0.00	.	0.00	.
Female vs. Male	0.02	0.00						
LEP vs. non-LEP	-0.06	0.06	-0.06	0.06	-0.08	0.07	-0.05	0.05
SPED vs. non-SPED	-0.08	0.09	-0.08	0.09	-0.08	0.09	-0.07	0.08
Low Income vs. non-Low Income	-0.04	0.04	-0.03	0.04	-0.03	0.04	-0.02	0.03
Hispanic vs. White	-0.03	0.02	-0.05	0.03	-0.08	0.03	-0.03	0.03
African American vs. White	-0.04	0.01	-0.04	0.02	-0.04	0.02	-0.01	0.01
Hawaiian/Pacific Islander vs. White								
American Indian vs. White	-0.04	0.02	-0.04	0.02	-0.05	0.02	-0.03	0.02
Multiple Race vs. White								
Asian vs. White	0.03	0.01	0.03	0.01	0.02	0.01	0.02	0.01
Cohort by Intercept	-0.10	0.04	-0.09	0.03	-0.07	0.01	-0.04	0.01
Female vs. Male								
LEP vs. non-LEP								
SPED vs. non-SPED								
Low Income vs. non-Low Income								
Hispanic vs. White								
African American vs. White								
Hawaiian/Pacific Islander vs. White								
American Indian vs. White								
Multiple Race vs. White								
Asian vs. White								

Note: ***β*** = Standardized regression coefficient. R^2=R squared. For the effect of special groups, the coefficient represents the difference compared to their contrast group; SPED = Special Education Status *vs*. Non-SPED. LEP = Limited English Proficiency *vs*. Non-LEP, Low Income = Low Income *vs.* Non-Low Income. For the effect of ethnic groups, the coefficient represents differential growth rate compared to White students.

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

Administration of the new AzM2 online tests is untimed, so schools may flexibly schedule students to take the tests in computer labs throughout the testing window. Thus, students taking the same grade-level or EOC test are not required to test on the same day. Because the days and times on which tests can be administered are variable, the possibility arises that performance factors associated with the time of day or day of the week may influence student test scores.

A series of regression models were developed to predict student performance using the day of the week and the time of the day variables and the duration of the test administration from test start to test end. The dependent variable for these analyses was the spring 2016 AzM2 scale score. To control for student achievement, we first covaried the previous achievement using spring 2015 AzM2 test scores. Because of the need to covary the previous achievement, the analyses were limited to students participating in the grades 4–8 and high school EOC assessments in mathematics and ELA tests for whom 2015 test scores were available. The day of the week was coded as 1 to 5 (1 for Monday, 2 for Tuesday, and so on). For the regression analyses, the time of day and the duration were continuous variables using the actual time. Time-of-day effects were further evaluated using paired comparisons among early morning, late morning, early afternoon, and late afternoon.

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

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

Time-of-day effects were less consistent. For high school students taking ELA assessments, morning start times were associated with better performance than afternoon start times. For middle school students, later morning start times were associated with poorer performance than early morning or late afternoon start times. In grade 6, ELA tests with morning start times were associated with lower scores than tests with afternoon start times.

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

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

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

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

Tests ending early in the afternoon were generally associated with higher scores than tests ending earlier in the day. However, grade 6 ELA proved an exception, with tests ending in the early morning associated with the highest scores. In addition, longer test administrations were associated with higher performance.

1.11 ARIZONA GLOSSARY STUDY

Construct-irrelevant barriers to accessing test content limit the validity of test score interpretations. When the use of vocabulary that is not relevant to the measured construct interferes with a student's ability to understand a test item, the item is not accurately assessing the intended construct. To evaluate the validity of testing accommodations such as glossaries, we expect that reducing access barriers will improve student performance for the disadvantaged group without having an effect on the general education population. However, if there is a main effect of the accommodation on all groups, the accommodation is likely modifying the measurement construct.

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

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

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

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

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

The second part of the study focused on the effectiveness and validity of glossaries. Glossary entries, if effective and valid, should increase the performance on items with glossaries for EL students but should have no effect on the performance of native speakers. In a randomized control trial, the pop-up glossaries were administered to students taking the Arizona spring 2017 ELA and mathematics state assessments. Approximately 60,000 students in each grade participated in the study. EL students ranged from about 1,000 to 8,000 per grade, with more in the lower grades. The participants were randomly assigned into three conditions: English glossary only; English glossary and Spanish translation; and no glossary. Exhibit 1.11.1 summarizes the number of students selected for the study by grade, subject, EL status, and experimental condition.

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

Grade	Glossary	ELA			Mathematics		
		non-EL	EL	Total	non-EL	EL	Total
3	ENG Only	19,385	2,535	21,920	19,442	2,569	22,011
	ENG+SP	19,780	2,449	22,229	19,874	2,481	22,355
	No Gloss	19,616	2,532	22,148	19,678	2,563	22,241
	Total	58,781	7,516	66,297	58,994	7,613	66,607
4	ENG Only	19,800	2,425	22,225	19,897	2,450	22,347
	ENG+SP	20,014	2,520	22,534	20,121	2,545	22,666
	No Gloss	20,140	2,350	22,490	20,249	2,375	22,624
	Total	59,954	7,295	67,249	60,267	7,370	67,637
5	ENG Only	19,802	1,924	21,726	19,898	1,935	21,833
	ENG+SP	20,182	1,928	22,110	20,235	1,941	22,176
	No Gloss	20,046	1,906	21,952	20,133	1,920	22,053
	Total	60,030	5,758	65,788	60,266	5,796	66,062
6	ENG Only	19,682	1,380	21,062	19,716	1,397	21,113
	ENG+SP	20,016	1,343	21,359	20,083	1,361	21,444
	No Gloss	19,906	1,393	21,299	19,939	1,410	21,349
	Total	59,604	4,116	63,720	59,738	4,168	63,906
7	ENG Only	19,841	1,241	21,082	19,472	1,251	20,723
	ENG+SP	20,092	1,307	21,399	19,712	1,306	21,018
	No Gloss	19,954	1,316	21,270	19,635	1,323	20,958
	Total	59,887	3,864	63,751	58,819	3,880	62,699
8	ENG Only	20,098	1,044	21,142	17,018	1,048	18,066
	ENG+SP	20,419	1,118	21,537	17,365	1,108	18,473
	No Gloss	20,370	1,029	21,399	17,315	1,025	18,340
	Total	60,887	3,191	64,078	51,698	3,181	54,879
9 / Algebra I	ENG Only	16,243	548	16,791	18,482	561	19,043
	ENG+SP	16,477	589	17,066	18,676	595	19,271
	No Gloss	16,430	530	16,960	18,604	513	19,117
	Total	49,150	1667	50,817	55,762	1,669	57,431
10 / Geometry	ENG Only	15,224	326	15,550	15,460	334	15,794
	ENG+SP	15,482	372	15,854	15,727	410	16,137
	No Gloss	15,279	323	15,602	15,688	357	16,045
	Total	45,985	1,021	47,006	46,875	1,101	47,976
11 / Algebra II	ENG Only	13,897	183	14,080	14,124	182	14,306
	ENG+SP	14,029	218	14,247	14,163	175	14,338
	No Gloss	13,990	209	14,199	14,082	208	14,290
	Total	41,916	610	42,526	42,369	565	42,934

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

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

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

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

β_j effect of item j,

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

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

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

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

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

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

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

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

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

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

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

1.12 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretations is ongoing. Nevertheless, sufficient evidence currently exists to support the principal claims for the test scores, including that AzM2 test scores indicate the degree to which students have achieved the Arizona State Standards at each grade level, and that students scoring at the proficient level or higher demonstrate levels of achievement consistent with national benchmarks indicating that they are on track to college readiness. These claims are supported by evidence of a test-development process that ensures alignment of test content to the Arizona State Standards, a standard-setting process that yielded performance standards consistent with those of rigorous, national benchmarks. Confirmatory factor analyses indicate that the subject-area assessments are unidimensional and therefore consistent with the measurement model. The CFAs also show that the hypothesized reporting strand structure of the AzM2 provides significant additional information about student achievement. In addition, test scores on the AzM2 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 Statewide Achievement Assessment (AzM2) to measure student mastery of the Arizona academic standards and progress toward college and career readiness. The AzM2 measures student progress in English language arts (ELA) and mathematics in grades 3–8 and 10. The Arizona Department of Education (ADE) worked with Cambium Assessment, Inc. (CAI) to develop and administer the AzM2 beginning in the spring of 2015. In accordance with state requirements, the AzM2 was designed to[17]:

- Align to the academic standards adopted by the Arizona State Board of Education in 2016 (Arizona State Standards);
- Supply criterion-referenced summative assessments for grades 3–8, and criterion-referenced end-of-course (EOC) assessments in identified high school mathematics and ELA courses for implementation beginning in the 2014–2015 school year;
- Assess, without bias, a range of basic knowledge and lower-level cognitive skills and higher-order, analytical thinking skills in writing, analysis, and problem-solving across subjects, using multiple assessment methods;
- Provide valid, reliable, and timely data to educators and policymakers to advance the academic success of Arizona students and inform the state's accountability measures;
- Communicate results to students, parents, and educators in a clear and timely manner to guide instruction;
- Provide an accurate perspective of the quality of learning occurring in classrooms and schools;
- Offer educators, students, and families critical tools to improve student achievement, including, but not limited to, formative and interim assessments, sample items, and practice tests;
- Allow meaningful national or multistate comparisons of school and student achievement;
- Use 21st-century technology to deliver the assessment, as available infrastructure allows;
- Ensure clarity, transparency, accuracy, and security in all aspects of assessment development, deployment, scoring, and reporting;
- Provide for content and psychometric evaluation and validation;
- Establish the involvement of Arizona stakeholders—educators, students, parents, and institutions of higher education, and business—in the development of the test, test-related materials, and achievement levels indicative of college and career readiness;
- Demonstrate accessibility for all students, with optimal access for English learners (ELs) and students with special needs;
- Respect Arizona's local control of the selection of classroom instructional materials; and
- Satisfy assessment goals in a cost-efficient manner.

The AzM2 was first administered in spring 2015, assessing proficiency in ELA in grades 3–11, in mathematics in grades 3–8, and following completion of Algebra I, geometry, and Algebra II (or similar) coursework. Following the initial administration, the AzM2 for grades 3–8 has been administered in the spring of each academic year; tests assessing high

[17] Standard 7.1: The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.
Standard 7.2: The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

school end-of-course (EOC) tests were administered in the fall, spring, and summer of each academic year. There was no testing in the fall or summer for SY2020–2021.

The Rasch model and Masters' (1982) partial credit model, an extension of the one-parameter Rasch model that allows for graded responses, were used to estimate item parameters for the AzM2. Item pools for grade-level summative and EOC assessments were calibrated following the first operational administration in spring 2015 and then adjusted for parameter drift following the spring 2016 administration. A vertical linking design was also implemented to produce a common vertical scale across grade levels to monitor student growth across grades 3–8 and high school EOC assessments. In subsequent years, pre-equated bank item parameter estimates have been applied directly for final scoring and reporting, a strategy that allows for more rapid reporting of tests administered online.

2.1 DEVELOPMENT OF ARIZONA STATE STANDARDS

In 2016, the Arizona State Board of Education adopted new academic content standards in ELA and mathematics that reflect high expectations of all Arizona students and strive to ensure that high school graduates are college- and career-ready. The Arizona State Standards in mathematics describe expectations for learning in grades K–8 and the first three high school courses (Algebra I, geometry, Algebra II; Mathematics 1, 2, 3) plus specific standards that could be included in a fourth high school credit mathematics course. The Arizona State Standards in ELA describe the reading, writing, language, speaking, and listening skills that students should acquire from grades K–12. The standards can be found on ADE's website.

2.2 AZM2 TEST DESIGN

The AzM2 is a series of fixed-form assessments intended to be administered online, but it is offered as a dual mode, online computer-based test (CBT) and paper-based test (PBT) to accommodate schools that are not yet ready to transition to the online testing environment. A common, operational base form is administered to all students within a given test grade and subject. Each assessment is composed of two to three discrete test sessions. The AzM2 operational item pools include various selected-response items, machine-scored constructed-response (MSCR) items, and some handscored, constructed-response items in the paper-pencil mathematics forms where MSCR items could not readily be rendered for paper-based testing (PBT) administration. AzM2 also includes essay responses. In spring 2016, a sample of online writing responses was handscored (100% double scoring with resolution of all discrepancies) to develop statistical models to machine-score the remaining online responses.

Five types of MSCR items were included in the AzM2 forms: graphic-response, natural-language, equation-response, hot-text, and table-input items. The graphic-response item types require students to place or move around objects in the answer space. A student can also plot points, draw lines, and draw shapes. The natural-language item types require students to type an English-language answer. The equation-response items require students to enter a value or equation. Hot-text items ask students to select or rearrange sentences or phrases in a passage. The table-input item types require students to input numerical values into a table.

The validity of computer-assigned scores for constructed-response items was evaluated following the spring 2015 online administration of the embedded field-test items. Rubric validation for all operational test items was completed before test construction and was based on the previous field-test administration of those items.

Each ELA assessment included one writing essay prompt that required an extended essay response. For the online test administrations, students were randomly administered one of two writing tasks. A random sample of student responses to each writing task was selected for handscoring. Two human raters scored these responses on three distinct scoring dimensions or rubrics: Statement of Purpose/Focus and Organization, Evidence/Elaboration, and Conventions/Editing, with any discrepancies adjudicated in a resolution score. This sample of essay responses and writing scores was used to develop the statistical models for machine-scoring the remaining online essay responses. All essays administered on paper-pencil tests were handscored. In addition, handscoring was required for a subset of mathematics items administered on paper, generally equation items, for which it was not possible to represent the item on paper in a way that allowed machine scoring.

3 SUMMARY OF SPRING 2021 OPERATIONAL TEST ADMINISTRATION

The following Arizona's Statewide Achievement Assessment (AzM2) assessments were administered in spring 2021:

- ELA (reading and writing) in grades 3–8 and 10
- Mathematics in grades 3–8 and 10

Online administration of the AzM2 occurred from April 5–16, 2021, for grade 3 writing; April 5–23, 2021, for grades 4–8 and 10 writing; April 5–30, 2021, for grade 3 reading and mathematics; and April 5–May 14, 2021, for grades 4–8 and 10 reading and mathematics. The paper-pencil version of the AzM2 was administered from April 5–14, 2021, for grade 3; and from April 5–21, 2021, for grades 4–8 and 10.

In the spring 2015 administration, item parameters for the mathematics assessments were calibrated following the online administration to establish the AzM2 bank scale. In the spring 2016 administration, all field-test items were placed on the AzM2 bank scale by concurrent calibrations of operational and field-test items. In spring 2021, the mathematics tests were scored using pre-equated item parameter estimates following the spring 2016 test administration of AzM2. Thus, no post-equating activities were conducted before the scoring and reporting of the mathematics tests in spring 2021 except the grade 10 mathematics. The new grade 10 AzM2 summative mathematics test employed a revised test design that measures student achievement of Algebra I and geometry academic content standards. To place all items in the grade 10 mathematics test on the geometry scale, geometry items were anchored to their bank value to calibrate the parameters for the algebra items. The post-equated algebra parameters were used along with the geometry bank parameters to score the test.

In the spring 2015 administration, item parameters for the English language arts (ELA) assessments were calibrated following the online administration to establish the AzM2 bank scale. In spring 2016, students were randomly assigned one of six writing prompts for administration in each ELA online assessment. Following the spring 2016 test administration, all operational items, including reading and writing, were concurrently calibrated and linked back to the AzM2 bank scale using the mean-mean equating method. In addition, all field-test items were concurrently calibrated with the mean-mean equated operational items. In spring 2021, students were assigned one of two items associated with the two writing rubrics (Informative-Explanatory or Opinion for grades 3–5 or Informative-Explanatory or Argumentative for grades 6–11). The pre-equated parameters calibrated following the spring 2016 test administration of AzM2 were used for the spring 2021 final scoring and reporting except for the new grade 10 mathematics. This section of the technical report summarizes the operational test results for the spring 2021 administration of the AzM2. 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 attempt requirements for scoring and reporting. The demographic composition of students taking the AzM2 in ELA and mathematics is presented in Exhibit 3.1.1 and Exhibit 3.1.2 by assessment and subgroup.[18] The tables in Appendix F show the demographic composition of test takers by mode of test administration.

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

Exhibit 3.1.1 Number of Students Participating in ELA Assessments by Subgroups: Spring 2021

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 10
All Students	71,748	72,741	73,102	75,531	76,172	76,233	63,048
Female	35,495	36,132	35,882	36,887	37,382	37,297	31,188
Male	36,253	36,609	37,220	38,644	38,790	38,936	31,860
African American	3,650	3,779	3,836	3,947	3,848	3,843	3,020
Asian	2,322	2,258	2,162	2,304	2,330	2,245	1,823
Native Hawaiian/Pacific Islander	215	296	246	230	222	231	172
Hispanic/Latino	32,810	33,493	33,867	34,997	35,782	35,390	27,016
American Indian or Alaskan	1,965	1,990	1,963	2,071	2,069	2,144	1,721
White	27,136	27,317	27,416	28,544	28,460	28,976	26,169
Multiple Ethnicities	3,650	3,608	3,612	3,439	3,461	3,404	3,127
Limited English Proficiency	6,875	6,327	6,233	6,157	6,533	6,140	3,382
Special Education	8,671	9,178	9,036	8,761	8,296	8,017	5,791
Free or Reduced-Price Lunch	28,037	28,554	28,565	29,495	29,302	28,659	22,427
Accommodation	2,303	2,510	2,481	2,429	2,149	2,078	728

Exhibit 3.1.2 Number of Students Participating in Mathematics Assessments by Subgroups: Spring 2021

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Math 10
All Students	73,048	73,750	74,270	76,780	77,724	77,872	64,417
Female	36,000	36,544	36,406	37,436	38,089	38,026	31,696
Male	37,048	37,206	37,864	39,344	39,635	39,846	32,721
African American	3,765	3,846	3,958	4,043	3,970	3,970	3,122
Asian	2,344	2,273	2,177	2,326	2,342	2,262	1,828
Native Hawaiian/Pacific Islander	223	304	250	233	230	236	179
Hispanic/Latino	33,546	34,031	34,469	35,735	36,684	36,286	27,886
American Indian or Alaskan	2,074	2,070	2,091	2,157	2,165	2,227	1,805
White	27,388	27,572	27,682	28,798	28,807	29,405	26,401
Multiple Ethnicities	3,708	3,654	3,643	3,489	3,526	3,486	3,196
Limited English Proficiency	7,098	6,458	6,380	6,350	6,749	6,365	3,550
Special Education	8,942	9,411	9,267	8,980	8,548	8,235	6,034
Free or Reduced-Price Lunch	28,695	29,026	29,048	29,974	29,940	29,338	23,112
Accommodation	2,334	2,498	2,463	2,452	2,144	2,051	696

3.2 CLASSICAL ITEM ANALYSIS

Because AzM2 is an online assessment system, the classical item analysis statistics for selected-response and constructed-response items reported here are calculated based on all online student responses. Classical item analysis statistics are used to monitor item behavior and investigate item-scoring irregularities throughout the testing window for online assessments, and follow the processing of answer documents for paper-based testing (PBT) administrations. Classical item analyses examine the degree to which the items function as intended with respect to the underlying scales. For online and paper-based test administrations, quality assurance (QA) reports provide the required item and test statistics for each selected-response and constructed-response item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item during test administration. Key statistics computed and examined include biserial/polyserial correlations for item discrimination, biserial correlations for distractors for selected-response items, and proportion correct for item difficulty.

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

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

Error! Reference source not found. present the average proportion of students responding correctly and average point biserial/polyserial correlations from spring 2021 online administrations of AzM2, respectively. As indicated, the items on the mathematics assessments were somewhat more difficult than those on the ELA assessment. While the mean difficulty of ELA items is relatively consistent across grade-level assessments, the average difficulty of mathematics items increases across grade levels and course assessments. Mean biserial correlations are reasonably high and consistent across assessments. Exhibit 3.2.2 shows the number of items flagged for proportion correct value, biserial/polyserial correlation, and distractor biserial/polyserial for the operational items in the spring 2021 online forms. The flagging criteria are presented in Section 4.5.1.

Exhibit 3.2.1 Average Proportion Correct and Point Biserial Correlations for Operational Test Items Administered Online: Spring 2021

Grade	Average *p*-Value	*p*-Value SD	Average Point-Biserial	Point-Biserial SD
		ELA		
3	0.44	0.14	0.48	0.14
4	0.51	0.17	0.51	0.12
5	0.52	0.17	0.56	0.12
6	0.50	0.18	0.49	0.12
7	0.50	0.17	0.49	0.12
8	0.50	0.17	0.54	0.13
10	0.51	0.14	0.47	0.13
		Mathematics		
3	0.52	0.18	0.65	0.12
4	0.49	0.18	0.67	0.12
5	0.42	0.16	0.62	0.14
6	0.40	0.18	0.62	0.12

Grade	Average *p*-Value	*p*-Value SD	Average Point-Biserial	Point-Biserial SD
7	0.44	0.19	0.61	0.12
8	0.37	0.16	0.58	0.14
10	0.35	0.19	0.55	0.14

Exhibit 3.2.2 Number of Items Flagged For *p*-Value, Biserial/Polyserial, or DIF for Operational Test Items Administered Online: Spring 2021

Grade	Proportion Correct	Biserial/Polyserial Correlation	Biserial Correlation for Distractor
		ELA	
3	0	3	2
4	0	1	0
5	0	1	0
6	0	1	3
7	0	1	1
8	0	1	2
10	0	1	2
		Mathematics	
3	0	0	0
4	0	0	1
5	0	0	1
6	0	2	2
7	0	1	3
8	0	0	1
10	0	1	4

3.3 ITEM RESPONSE THEORY ANALYSIS

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

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

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

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (also known as the Rasch model) is used to calibrate dichotomously scored AzM2 items and takes the form

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

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

For items with multiple, ordered-response categories (i.e., partial credit items), AzM2 items are calibrated using the Rasch-family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category *i* for an item with m_j categories can be written as

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

The tables in Appendix E provide Rasch and Masters' partial credit model item parameter estimates for the spring 2021 operational test items. Because AzM2 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. The selected-response items include traditional four-option multiple-choice items; technology-enhanced selected-response items, which may require students to select one or more options; and MSCR items, for which students' constructed-response items are scored electronically using explicit rubrics. In addition, the average Rasch difficulty is presented for each scoring dimension of the writing prompt administered at each grade. As illustrated in Exhibit 3.3.1, selected-response items are, on average, less difficult than the constructed-response item types. Within the constructed-response items, Evidence and Elaboration within the writing prompts were, on average, consistently found to be the most difficult.

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

Grade/ Course	SR			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
				ELA					
3	39	0.06	0.81	-	-	-	1.59	1.58	-1.16
4	41	0.13	0.61	-	-	-	3.62	4.00	-0.09
5	41	0.10	0.84	-	-	-	2.39	3.07	-0.85
6	41	0.05	0.75	-	-	-	2.28	2.95	-1.21
7	41	0.06	0.86	-	-	-	2.36	2.76	-1.56
8	41	0.06	0.93	-	-	-	0.97	1.16	-1.62
10	43	0.07	0.83	-	-	-	0.84	1.22	-2.03
				Mathematics					
3	22	-0.11	1.14	23	0.31	1.18	-	-	-
4	12	-0.31	1.31	33	0.16	1.11	-	-	-
5	15	-0.41	0.95	30	0.30	0.84	-	-	-
6	21	-0.34	1.26	26	0.35	0.98	-	-	-
7	21	-0.58	0.86	26	0.61	0.95	-	-	-
8	25	-0.56	1.09	22	0.33	0.75	-	-	-

Grade/ Course	SR			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
10	32	-0.87	1.11	18	-0.24	0.9	-	-	-

Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). The rule of thumb is that items with good model-data-fit have Infit and Outfit within the range of 0.7–1.3. Exhibit 3.3.2 summarizes the number of online administered operational test items with Infit and Outfit statistics below, within, and above the range of .7–1.3.

Exhibit 3.3.2 Summary of Item Fit Statistics for Items Administered Online

Grade/Course	Infit			Outfit		
	Below 0.7	Between .7–1.3	Above 1.3	Below 0.7	Between .7–1.3	Above 1.3
ELA						
3	0	44	1	1	39	5
4	0	46	1	2	43	2
5	0	44	3	0	43	4
6	0	47	0	3	39	5
7	0	46	1	0	44	3
8	0	46	1	3	36	8
10	0	48	1	0	47	2
Mathematics						
3	1	41	3	2	33	10
4	0	43	2	0	41	4
5	0	43	2	2	35	8
6	0	44	3	2	39	6
7	0	45	2	5	37	5
8	0	47	0	2	36	9
10	1	48	1	4	41	5

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 to Exhibit 3.4.3. The AzM2 bank scale was established based on the spring 2015 assessments in which the item calibrations were centered on items rather than persons, resulting in operational test forms with a mean difficulty of 0 and standard deviation of 1. 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: Combined Online and Paper-Based Testing

Test/Grade	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
ELA					
3	71,748	2,495	32.97	2,605	2,395
4	72,741	2,519	33.26	2,610	2,400
5	73,102	2,535	38.23	2,629	2,419
6	75,531	2,541	32.70	2,641	2,431
7	76,172	2,549	33.94	2,648	2,438
8	76,233	2,556	36.47	2,658	2,448
10	63,048	2,563	30.62	2,668	2,458
Mathematics					
3	73,048	3,509	47.34	3,605	3,395
4	73,750	3,541	47.82	3,645	3,435
5	74,270	3,572	42.96	3,688	3,478
6	76,780	3,603	44.23	3,722	3,512
7	77,724	3,627	42.46	3,739	3,529
8	77,872	3,649	40.41	3,776	3,566
10	64,417	3,674	37.36	3,819	3,609

Exhibit 3.4.2 Test Score Summary Statistics: Online Testing

Test/Grade	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
ELA					
3	62,976	2,493	32.63	2,605	2,395
4	64,006	2,517	32.93	2,610	2,400
5	64,351	2,534	38.34	2,629	2,419
6	65,049	2,539	32.15	2,641	2,431
7	65,916	2,546	33.31	2,648	2,438
8	66,021	2,553	35.58	2,658	2,448
10	56,609	2,562	30.32	2,668	2,458
Mathematics					
3	64,167	3,507	46.99	3,605	3,395
4	64,978	3,539	47.59	3,645	3,435
5	65,431	3,570	42.43	3,688	3,478
6	66,277	3,601	43.57	3,722	3,512
7	67,335	3,624	41.63	3,739	3,529
8	67,468	3,646	39.25	3,776	3,566
10	57,871	3,672	35.85	3,819	3,609

Exhibit 3.4.3 Test Score Summary Statistics: Paper-Based Testing

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Observed Max.	Observed Min.
		ELA			
3	8,772	2,509	32.21	2,605	2,395
4	8,736	2,531	33.05	2,610	2,422
5	8,751	2,548	35.05	2,629	2,420
6	10,484	2,555	32.70	2,641	2,431
7	10,256	2,564	33.69	2,648	2,450
8	10,213	2,574	37.07	2,658	2,448
10	6,442	2,576	30.38	2,668	2,458
		Mathematics			
3	8,884	3,523	47.60	3,605	3,395
4	8,773	3,555	47.43	3,645	3,435
5	8,839	3,588	43.79	3,688	3,478
6	10,505	3,622	43.83	3,722	3,512
7	10,389	3,646	42.95	3,739	3,529
8	10,405	3,667	43.26	3,776	3,566
10	6,550	3,697	42.53	3,819	3,609

The percentage of students in each performance level by grade and content area and the percentage of students at or above Proficient are presented in Exhibit 3.4.4 to Exhibit 3.4.6.

Exhibit 3.4.4 Percentage of Students in Performance Levels: Combined Online and Paper-Based Testing

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
			ELA			
3	71,748	52	13	25	10	35
4	72,741	40	15	33	12	45
5	73,102	34	21	29	17	45
6	75,531	39	24	31	6	37
7	76,172	43	20	29	8	37
8	76,233	45	20	23	11	35
10	63,048	51	17	24	9	32
			Mathematics			
3	73,048	38	26	24	12	36
4	73,750	41	24	25	10	35
5	74,270	43	26	22	10	32
6	76,780	51	19	19	11	30
7	77,724	53	18	17	13	30
8	77,872	56	17	15	11	26
10	64,417	54	20	21	5	26

Exhibit 3.4.5 Percentage of Students in Performance Levels: Online Testing

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	62,976	54	13	24	9	32
4	64,006	42	15	32	11	43
5	64,351	36	21	28	16	44
6	65,049	42	24	29	5	34
7	65,916	46	20	27	7	34
8	66,021	48	21	22	10	32
10	56,609	53	17	22	8	30
Mathematics						
3	64,167	40	26	23	11	34
4	64,978	43	24	24	9	33
5	65,431	45	26	21	9	30
6	66,277	54	19	17	10	27
7	67,335	55	18	16	11	27
8	67,468	59	17	14	10	24
10	57,871	56	20	20	4	24

Exhibit 3.4.6 Percentage of Students in Performance Levels: Paper-Based Testing

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
3	34	13	35	18	53	34
4	24	14	43	19	63	24
5	21	20	36	23	59	21
6	23	22	44	12	55	23
7	24	20	40	16	56	24
8	26	20	32	22	54	26
10	33	15	36	16	52	33
Mathematics						
3	28	24	29	19	48	28
4	30	23	32	15	47	30
5	28	26	30	16	46	28
6	33	21	27	20	46	33
7	34	19	23	24	47	34
8	38	19	22	21	43	38
10	31	21	33	15	48	31

3.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 3.5.1 through Exhibit 3.5.4 presents the number and percentage, respectively, of students in each grade and subject at each performance level, by gender (female, male) and ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian, White, Multiple Ethnicities), and by other demographic information, such as special education status (SPED), limited English proficiency (LEP), eligibility for free or reduced-price lunch (FRL), and accommodation.

Exhibit 3.5.1 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based ELA

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	37,231	17,800	19,430	2,408	534	120	21,265	1,520	9,733	1,650	6,935	6,164	18,609	2,029
	Partially Proficient	9,451	4,772	4,678	438	290	27	4,102	213	3,846	534	594	344	3,529	122
	Proficient	18,034	9,212	8,822	632	852	55	6,014	201	9,250	1,030	877	324	4,824	137
	Highly Proficient	7,034	3,711	3,323	172	646	13	1,429	31	4,307	436	265	43	1,075	15
4	Minimally Proficient	28,740	13,424	15,315	2,044	327	112	17,431	1,324	6,384	1,117	6,809	5,395	15,361	2,059
	Partially Proficient	10,924	5,399	5,525	603	248	37	5,415	278	3,830	513	875	463	4,575	229
	Proficient	24,356	12,560	11,793	928	1,003	105	8,742	340	11,813	1,422	1,207	422	7,282	195
	Highly Proficient	8,729	4,750	3,976	204	680	42	1,905	49	5,290	556	287	47	1,337	27
5	Minimally Proficient	24,978	11,025	13,953	1,792	281	82	15,180	1,162	5,489	992	6,726	4,941	13,334	1,950
	Partially Proficient	14,991	7,426	7,563	801	336	61	7,557	414	5,105	715	1,145	794	6,386	310
	Proficient	20,997	10,765	10,232	919	754	70	8,113	299	9,703	1,139	827	410	6,531	168
	Highly Proficient	12,140	6,666	5,472	324	791	33	3,017	88	7,119	766	338	88	2,314	53
6	Minimally Proficient	29,604	12,822	16,781	2,075	334	99	17,662	1,346	6,965	1,122	6,996	5,159	15,384	2,009
	Partially Proficient	17,809	9,004	8,804	949	431	52	8,441	430	6,688	817	1,054	655	7,131	267
	Proficient	23,327	12,350	10,976	821	1,107	68	7,912	273	11,898	1,247	632	310	6,291	143
	Highly Proficient	4,796	2,712	2,084	102	432	11	982	22	2,994	253	79	33	690	10
7	Minimally Proficient	32,959	14,298	18,660	2,161	357	88	19,314	1,425	8,313	1,300	6,904	5,549	16,517	1,800
	Partially Proficient	15,046	7,850	7,192	741	334	51	7,265	334	5,645	672	745	632	5,799	218
	Proficient	21,907	11,749	10,156	807	973	68	7,877	275	10,754	1,151	549	330	6,035	118
	Highly Proficient	6,267	3,485	2,782	139	666	15	1,326	35	3,748	338	98	22	951	13
8	Minimally Proficient	34,291	14,650	19,640	2,241	371	113	19,830	1,513	8,932	1,290	6,874	5,422	16,531	1,833
	Partially Proficient	15,606	8,038	7,567	731	359	48	7,154	357	6,265	691	662	457	5,786	154
	Proficient	17,609	9,450	8,157	617	736	48	6,289	226	8,759	932	384	221	4,882	75
	Highly Proficient	8,734	5,159	3,573	254	779	22	2,118	48	5,020	491	97	40	1,460	16
10	Minimally Proficient	32,096	14,351	17,745	1,953	403	97	16,783	1,304	10,217	1,339	5,195	3,071	14,356	660
	Partially Proficient	10,719	5,588	5,130	452	239	29	4,434	217	4,823	524	320	170	3,584	41
	Proficient	14,856	8,081	6,771	498	688	41	4,705	170	7,838	912	225	115	3,657	23
	Highly Proficient	5,385	3,168	2,217	119	493	5	1,094	30	3,292	352	52	26	832	4

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

Exhibit 3.5.2 Number of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based Mathematics

Grade	Performance Level	Overall	Female	Male	African American	Asian	Hawaiian/ Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	28,115	14,412	13,702	2,177	318	94	17,246	1,365	5,713	1,201	5,769	5,417	14,984	1701
	Partially Proficient	18,896	9,525	9,371	909	377	68	9,150	474	6,930	988	1,695	1,179	7,732	437
	Proficient	17,367	8,315	9,051	518	757	45	5,565	196	9,288	997	1,071	424	4,704	158
	Highly Proficient	8,675	3,750	4,925	161	892	16	1,586	39	5,458	523	408	78	1,276	38
4	Minimally Proficient	30,374	15,448	14,923	2,371	276	109	18,460	1,425	6,473	1,257	6,575	5,192	16,224	1,935
	Partially Proficient	17,701	9,008	8,692	793	364	76	8,380	404	6,792	891	1,569	833	6,963	369
	Proficient	18,355	8,860	9,494	557	926	72	5,810	198	9,730	1,061	968	364	4,821	168
	Highly Proficient	7,327	3,229	4,097	125	707	47	1,381	44	4,577	445	299	69	1,019	26
5	Minimally Proficient	31,638	15,655	15,981	2,440	292	104	19,143	1,463	6,858	1,336	6,931	5,144	16,563	1,939
	Partially Proficient	18,981	9,607	9,371	940	428	76	8,735	416	7,444	939	1,392	832	7,248	355
	Proficient	16,425	7,967	8,458	475	749	53	5,223	175	8,806	944	720	340	4,211	146
	Highly Proficient	7,232	3,177	4,054	103	708	17	1,368	37	4,574	424	224	64	1,026	23
6	Minimally Proficient	39,445	19,492	19,952	2,854	406	133	23,258	1,677	9,538	1,578	7,480	5,496	19,753	2,108
	Partially Proficient	14,481	7,342	7,138	630	396	41	6,319	277	6,086	731	816	517	5,179	204
	Proficient	14,255	6,893	7,361	421	674	38	4,459	154	7,803	705	492	257	3,632	97
	Highly Proficient	8,605	3,710	4,894	138	850	21	1,699	49	5,372	475	192	80	1,411	43
7	Minimally Proficient	40,811	20,504	20,306	2,831	413	114	24,111	1,685	9,981	1,675	7,370	6,001	20,031	1,909
	Partially Proficient	13,867	6,866	7,000	554	318	54	6,054	270	5,957	659	595	455	4,865	141
	Proficient	12,981	6,276	6,705	400	533	41	4,385	152	6,818	652	397	221	3,425	68
	Highly Proficient	10,068	4,443	5,624	185	1,078	21	2,134	58	6,051	540	186	72	1,619	26
8	Minimally Proficient	43,844	21,553	22,289	2,952	429	136	25,111	1,781	11,702	1,731	7,366	5,780	20,554	1,876
	Partially Proficient	13,482	6828	6,654	520	319	42	5,562	257	6,138	644	488	374	4,459	111
	Proficient	11,722	5,738	5,983	328	535	34	3,701	136	6,381	606	252	159	2,934	35
	Highly Proficient	8,828	3,907	4,921	170	979	24	1,913	53	5,184	505	129	52	1,391	29
10	Minimally Proficient	34,715	16,739	17,976	2,274	333	108	18,627	1,389	10,503	1,481	5,478	3,165	15,661	648
	Partially Proficient	12,805	6,815	5,989	479	292	39	5,110	248	5,974	662	352	245	4,189	34
	Proficient	13,501	6,724	6,777	334	681	25	3,641	154	7,850	816	175	121	2,880	13
	Highly Proficient	3,402	1,418	1,983	37	522	7	509	14	2,075	237	31	19	384	1

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

Exhibit 3.5.3 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based ELA

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	52	50	54	66	23	56	65	77	36	45	80	90	66	88
	Partially Proficient	13	13	13	12	12	13	13	11	14	15	7	5	13	5
	Proficient	25	26	24	17	37	26	18	10	34	28	10	5	17	6
	Highly Proficient	10	10	9	5	28	6	4	2	16	12	3	1	4	1
	At or Above Proficient	35	36	34	22	65	32	23	12	50	40	13	5	21	7
4	Minimally Proficient	40	37	42	54	14	38	52	67	23	31	74	85	54	82
	Partially Proficient	15	15	15	16	11	13	16	14	14	14	10	7	16	9
	Proficient	33	35	32	25	44	35	26	17	43	39	13	7	26	8
	Highly Proficient	12	13	11	5	30	14	6	2	19	15	3	1	5	1
	At or Above Proficient	45	48	43	30	75	50	32	20	63	55	16	7	30	9
5	Minimally Proficient	34	31	37	47	13	33	45	59	20	27	74	79	47	79
	Partially Proficient	21	21	20	21	16	25	22	21	19	20	13	13	22	12
	Proficient	29	30	27	24	35	28	24	15	35	32	9	7	23	7
	Highly Proficient	17	19	15	8	37	13	9	4	26	21	4	1	8	2
	At or Above Proficient	45	49	42	32	71	42	33	20	61	53	13	8	31	9
6	Minimally Proficient	39	35	43	53	14	43	50	65	24	33	80	84	52	83
	Partially Proficient	24	24	23	24	19	23	24	21	23	24	12	11	24	11
	Proficient	31	33	28	21	48	30	23	13	42	36	7	5	21	6
	Highly Proficient	6	7	5	3	19	5	3	1	10	7	1	1	2	0
	At or Above Proficient	37	41	34	23	67	34	25	14	52	44	8	6	24	6
7	Minimally Proficient	43	38	48	56	15	40	54	69	29	38	83	85	56	84
	Partially Proficient	20	21	19	19	14	23	20	16	20	19	9	10	20	10
	Proficient	29	31	26	21	42	31	22	13	38	33	7	5	21	5
	Highly Proficient	8	9	7	4	29	7	4	2	13	10	1	0	3	1
	At or Above Proficient	37	41	33	25	70	37	26	15	51	43	8	5	24	6
8	Minimally Proficient	45	39	50	58	17	49	56	71	31	38	86	88	58	88
	Partially Proficient	20	22	19	19	16	21	20	17	22	20	8	7	20	7
	Proficient	23	25	21	16	33	21	18	11	30	27	5	4	17	4
	Highly Proficient	11	14	9	7	35	10	6	2	17	14	1	1	5	1
	At or Above Proficient	35	39	30	23	67	30	24	13	48	42	6	4	22	4
10	Minimally Proficient	51	46	56	65	22	56	62	76	39	43	90	91	64	91
	Partially Proficient	17	18	16	15	13	17	16	13	18	17	6	5	16	6

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	Proficient	24	26	21	16	38	24	17	10	30	29	4	3	16	3
	Highly Proficient	9	10	7	4	27	3	4	2	13	11	1	1	4	1
	At or Above Proficient	32	36	28	20	65	27	21	12	43	40	5	4	20	4

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

Exhibit 3.5.4 Percentage of Students at Each Performance Level by Gender, Ethnicity, and Other Demographic Information: Combined Online and Paper-Based Mathematics

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
3	Minimally Proficient	38	40	37	58	14	42	51	66	21	32	65	76	52	73
	Partially Proficient	26	26	25	24	16	30	27	23	25	27	19	17	27	19
	Proficient	24	23	24	14	32	20	17	9	34	27	12	6	16	7
	Highly Proficient	12	10	13	4	38	7	5	2	20	14	5	1	4	2
	At or Above Proficient	36	34	38	18	70	27	21	11	54	41	17	7	21	8
4	Minimally Proficient	41	42	40	62	12	36	54	69	23	34	70	80	56	77
	Partially Proficient	24	25	23	21	16	25	25	20	25	24	17	13	24	15
	Proficient	25	24	26	14	41	24	17	10	35	29	10	6	17	7
	Highly Proficient	10	9	11	3	31	15	4	2	17	12	3	1	4	1
	At or Above Proficient	35	33	37	18	72	39	21	12	52	41	13	7	20	8
5	Minimally Proficient	43	43	42	62	13	42	56	70	25	37	75	81	57	79
	Partially Proficient	26	26	25	24	20	30	25	20	27	26	15	13	25	14
	Proficient	22	22	22	12	34	21	15	8	32	26	8	5	14	6
	Highly Proficient	10	9	11	3	33	7	4	2	17	12	2	1	4	1
	At or Above Proficient	32	31	33	15	67	28	19	10	48	38	10	6	18	7
6	Minimally Proficient	51	52	51	71	17	57	65	78	33	45	83	87	66	86
	Partially Proficient	19	20	18	16	17	18	18	13	21	21	9	8	17	8
	Proficient	19	18	19	10	29	16	12	7	27	20	5	4	12	4
	Highly Proficient	11	10	12	3	37	9	5	2	19	14	2	1	5	2
	At or Above Proficient	30	28	31	14	66	25	17	9	46	34	8	5	17	6

Grade	Performance Level	Percentage of Students in Each Grade and Subject at Each Performance Level													
		Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
7	**Minimally Proficient**	53	54	51	71	18	50	66	78	35	48	86	89	67	89
	Partially Proficient	18	18	18	14	14	23	17	12	21	19	7	7	16	7
	Proficient	17	16	17	10	23	18	12	7	24	18	5	3	11	3
	Highly Proficient	13	12	14	5	46	9	6	3	21	15	2	1	5	1
	At or Above Proficient	30	28	31	15	69	27	18	10	45	34	7	4	17	4
8	**Minimally Proficient**	56	57	56	74	19	58	69	80	40	50	89	91	70	91
	Partially Proficient	17	18	17	13	14	18	15	12	21	18	6	6	15	5
	Proficient	15	15	15	8	24	14	10	6	22	17	3	2	10	2
	Highly Proficient	11	10	12	4	43	10	5	2	18	14	2	1	5	1
	At or Above Proficient	26	25	27	13	67	25	15	8	39	32	5	3	15	3
10	**Minimally Proficient**	54	53	55	73	18	60	67	77	40	46	91	89	68	93
	Partially Proficient	20	22	18	15	16	22	18	14	23	21	6	7	18	5
	Proficient	21	21	21	11	37	14	13	9	30	26	3	3	12	2
	Highly Proficient	5	4	6	1	29	4	2	1	8	7	1	1	2	0
	At or Above Proficient	26	26	27	12	66	18	15	9	38	33	3	4	14	2

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

3.6 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance-level classifications. It essentially addresses how likely a student is to achieve the same score or be classified in the same performance level across multiple administrations of equivalently constructed and administered test forms. The reliability of test scores and performance classifications is evaluated from various perspectives as part of each test administration. Test score reliability is traditionally estimated using both classical and IRT approaches. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the IRT framework, measurement error varies across the range of abilities. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the test information function (TIF) represents the standard error of measurement (SEM). The SEM is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale at the middle of the test distribution and greater on scaled values farther away from the middle.

The reliability evidence of the AZM2 test scores is provided with reliability, SEM, and classification accuracy and consistency in each achievement level.

3.6.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale for all students. The marginal reliability coefficients are nearly identical or close to coefficient alpha. For our analysis, the marginal reliability coefficients were computed using operational items.

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

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

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

Exhibit 3.6.1.1 presents the marginal reliability coefficients for all students. The reliability coefficients for all subjects and grades range from 0.90–0.94.

Exhibit 3.6.1.1 Overall Reliabilities by Subject/Test for AzM2 Scores

Grade	ELA		Mathematics	
	Reliability	Variance	Reliability	Variance
3	0.90	1,065	0.93	2,208
4	0.91	1,084	0.94	2,265
5	0.92	1,470	0.92	1,800
6	0.90	1,034	0.93	1,899
7	0.90	1,109	0.92	1,733
8	0.92	1,266	0.92	1,540
10	0.90	919	0.91	1,285

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

3.6.2 STANDARD ERROR OF MEASUREMENT

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

Exhibit 3.6.2.1 and Exhibit 3.6.2.2 present the CSEM for the AzM2 ELA and mathematics assessments with respect to the four AzM2 performance-level cuts. These tables also include associated CSEM around cut score. As the tables indicate, the AzM2 test scores are most precise near the middle of the ability distribution, especially near the Partially Proficient and Proficient performance standard cuts.[19] Test scores near the tails of the ability distribution are somewhat less precise, as expected. While these numbers indicate that the AzM2 test scores are slightly 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

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

classifications. Exhibit 3.6.2.3 through Exhibit 3.6.2.16 present the CSEMs and corresponding performance levels for each scale score for the AzM2 ELA and mathematics assessments.

Exhibit 3.6.2.1 Performance Level and Associated CSEMs Spring 2021: ELA

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

Exhibit 3.6.2.2 Performance Level and Associated CSEMs Spring 2021: Mathematics

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

Exhibit 3.6.2.3 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 3 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,395	22	2,497	9	2,510	9	2,543	11
2,408	18	2,499	9	2,513	9	2,547	11
2,417	16	2,502	9	2,516	9	2,551	12
2,425	15	2,505	9	2,519	10	2,556	12
2,432	14	2,508	9	2,522	10	2,561	13
2,437	13			2,525	10	2,567	13
2,443	12			2,529	10	2,573	14
2,447	12			2,532	10	2,580	15
2,451	11			2,536	10	2,588	16
2,455	11			2,539	11	2,598	18
2,459	11					2,605	20
2,463	10						
2,466	10						
2,470	10						
2,473	10						
2,476	10						
2,479	9						
2,482	9						
2,485	9						
2,488	9						
2,491	9						
2,493	9						

Note: For Grade 3 ELA = writing prompt 13022 administered

Exhibit 3.6.2.4 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 4 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,408	22	2,511	9	2,524	9	2,559	11
2,421	18	2,514	9	2,527	9	2,564	12
2,431	16	2,516	9	2,530	9	2,568	12
2,438	15	2,519	9	2,533	9	2,573	13
2,445	13	2,522	9	2,536	9	2,579	14
2,451	13			2,538	10	2,586	15
2,456	12			2,542	10	2,594	16
2,460	12			2,545	10	2,603	18
2,465	11			2,548	10	2,610	19
2,469	11			2,552	10		
2,472	10			2,555	11		
2,476	10						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,479	10						
2,482	10						
2,486	10						
2,489	9						
2,492	9						
2,494	9						
2,497	9						
2,500	9						
2,503	9						
2,506	9						
2,508	9						

Note: For Grade 4 ELA = writing prompt 13,119 administered

Exhibit 3.6.2.5 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 5 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,419	23	2,520	9	2,544	10	2,578	12
2,421	22	2,522	9	2,547	10	2,582	13
2,435	18	2,525	9	2,550	10	2,588	13
2,445	16	2,528	9	2,553	10	2,594	14
2,452	15	2,531	9	2,557	10	2,601	15
2,459	14	2,534	9	2,561	11	2,609	16
2,465	13	2,537	10	2,564	11	2,618	17
2,470	12	2,540	10	2,569	11	2,629	19
2,475	12			2,573	12		
2,479	11						
2,483	11						
2,487	11						
2,491	10						
2,494	10						
2,498	10						
2,501	10						
2,504	10						
2,508	10						
2,511	10						
2,514	9						
2,517	9						

Note: For Grade 5 ELA = writing prompt 13246 administered

Exhibit 3.6.2.6 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 6 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,444	18	2,532	9	2,554	10	2,597	13
2,454	16	2,534	9	2,557	10	2,601	13
2,462	15	2,537	9	2,560	10	2,607	14
2,469	14	2,539	9	2,564	10	2,614	15
2,474	13	2,542	9	2,567	10	2,623	16
2,479	12	2,545	9	2,570	10	2,632	18
2,484	12	2,548	9	2,574	11	2,641	20
2,489	11	2,551	9	2,578	11		
2,493	11			2,582	11		
2,496	11			2,586	12		
2,500	10			2,591	12		
2,503	10						
2,507	10						
2,510	10						
2,513	10						
2,516	10						
2,519	9						
2,522	9						
2,525	9						
2,528	9						

Note: For Grade 6 ELA = writing prompt 13306 administered

Exhibit 3.6.2.7 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 7 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,438	21	2,543	10	2,562	10	2,600	12
2,447	19	2,546	10	2,566	10	2,603	13
2,458	16	2,549	10	2,569	10	2,609	13
2,466	15	2,553	10	2,573	10	2,615	14
2,473	14	2,556	10	2,576	11	2,623	15
2,479	13	2,559	10	2,580	11	2,631	16
2,484	13			2,584	11	2,641	18
2,489	12			2,589	11	2,648	19
2,494	12			2,593	12		
2,498	11						
2,502	11						
2,506	11						
2,510	10						
2,514	10						

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,517	10						
2,521	10						
2,524	10						
2,527	10						
2,530	10						
2,534	10						
2,537	10						
2,540	10						

Note: For Grade 7 ELA = writing prompt 13401 administered

Exhibit 3.6.2.8 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 8 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,448	20	2,551	9	2,573	10	2,605	11
2,453	18	2,553	9	2,576	10	2,610	12
2,463	16	2,556	9	2,580	10	2,615	12
2,471	15	2,559	9	2,583	10	2,620	13
2,477	14	2,562	9	2,586	10	2,626	13
2,483	13	2,564	9	2,590	10	2,632	14
2,488	12	2,567	9	2,593	11	2,639	15
2,493	12	2,570	9	2,597	11	2,648	17
2,497	11			2,601	11	2,658	19
2,501	11						
2,505	11						
2,509	10						
2,512	10						
2,516	10						
2,519	10						
2,522	10						
2,525	9						
2,528	9						
2,531	9						
2,534	9						
2,537	9						
2,539	9						
2,542	9						
2,545	9						
2,548	9						

Note: For Grade 8 ELA = writing prompt 13439 administered

Exhibit 3.6.2.9 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 10 ELA

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
2,458	21	2,567	9	2,581	9	2,608	10
2,467	19	2,569	9	2,583	9	2,611	11
2,477	16	2,572	9	2,586	9	2,615	11
2,485	15	2,575	9	2,589	9	2,619	11
2,492	14	2,577	9	2,592	9	2,624	12
2,498	13			2,595	10	2,628	12
2,503	12			2,598	10	2,633	13
2,508	12			2,601	10	2,639	13
2,513	11			2,604	10	2,645	14
2,517	11					2,652	15
2,521	11					2,660	17
2,524	10					2,668	18
2,528	10						
2,531	10						
2,535	10						
2,538	10						
2,541	10						
2,544	9						
2,547	9						
2,550	9						
2,553	9						
2,555	9						
2,558	9						
2,561	9						
2,564	9						

Note: For Grade 10 ELA = writing prompt 13637 administered

Exhibit 3.6.2.10 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 3 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,395	22	3,495	10	3,531	11	3,573	14
3,408	19	3,498	10	3,534	11	3,580	15
3,418	17	3,501	10	3,537	11	3,588	17
3,427	15	3,505	10	3,542	11	3,598	19
3,434	14	3,508	10	3,546	12	3,605	20
3,440	13	3,512	10	3,550	12		
3,446	13	3,515	10	3,555	12		
3,451	12	3,519	10	3,561	13		
3,456	12	3,522	10	3,566	13		

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,460	11	3,526	11				
3,465	11						
3,469	11						
3,473	11						
3,476	11						
3,480	11						
3,484	10						
3,487	10						
3,491	10						

Exhibit 3.6.2.11 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 4 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,435	21	3,532	10	3,563	11	3,608	13
3,444	18	3,535	10	3,567	11	3,614	14
3,454	16	3,538	10	3,571	11	3,621	15
3,462	15	3,542	10	3,575	11	3,629	16
3,469	14	3,545	10	3,579	11	3,639	19
3,475	13	3,549	10	3,583	11	3,645	20
3,480	13	3,552	10	3,587	12		
3,485	12	3,556	10	3,592	12		
3,490	12	3,559	10	3,597	12		
3,494	11			3,602	13		
3,499	11						
3,503	11						
3,507	11						
3,510	11						
3,514	10						
3,518	10						
3,521	10						
3,525	10						
3,528	10						

Exhibit 3.6.2.12 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 5 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,478	23	3,563	10	3,595	10	3,635	12
3,481	22	3,565	10	3,597	10	3,638	12
3,494	18	3,568	10	3,600	10	3,644	13

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,504	16	3,572	10	3,603	10	3,650	14
3,512	15	3,575	10	3,607	10	3,656	15
3,519	14	3,578	10	3,610	10	3,664	16
3,525	13	3,581	10	3,614	10	3,674	18
3,530	12	3,584	10	3,617	11	3,687	22
3,535	12	3,587	10	3,621	11	3,688	22
3,539	11	3,590	10	3,625	11		
3,543	11			3,629	11		
3,547	11						
3,551	11						
3,555	10						
3,558	10						

Exhibit 3.6.2.13 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 6 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,512	21	3,602	10	3,629	10	3,663	11
3,521	19	3,606	10	3,631	10	3,668	12
3,531	17	3,609	10	3,635	10	3,672	12
3,539	15	3,612	10	3,638	10	3,677	12
3,546	14	3,615	10	3,641	10	3,683	13
3,552	13	3,619	10	3,645	10	3,689	14
3,558	13	3,622	10	3,648	10	3,696	15
3,563	12	3,625	10	3,652	10	3,704	16
3,568	12			3,655	11	3,714	19
3,573	11			3,659	11	3,722	21
3,577	11						
3,581	11						
3,585	11						
3,588	11						
3,592	10						
3,596	10						
3,599	10						

Exhibit 3.6.2.14 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 7 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,529	22	3,629	10	3,652	10	3,680	11
3,543	18	3,632	10	3,654	10	3,685	12

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,553	16	3,635	10	3,658	10	3,689	12
3,561	15	3,638	10	3,661	10	3,694	13
3,567	14	3,641	10	3,665	10	3,700	13
3,574	13	3,644	10	3,668	11	3,706	14
3,579	12	3,648	10	3,672	11	3,713	15
3,584	12			3,676	11	3,721	16
3,589	12					3,731	19
3,593	11					3,739	21
3,597	11						
3,601	11						
3,605	11						
3,608	10						
3,612	10						
3,615	10						
3,619	10						
3,622	10						
3,625	10						

Exhibit 3.6.2.15 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 8 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,566	20	3,650	10	3,674	9	3,705	11
3,572	19	3,654	10	3,677	10	3,707	11
3,582	16	3,657	10	3,681	10	3,711	11
3,590	15	3,660	9	3,684	10	3,716	11
3,597	14	3,663	9	3,687	10	3,720	12
3,603	13	3,666	9	3,690	10	3,725	12
3,608	12	3,669	9	3,693	10	3,731	13
3,613	12	3,672	9	3,697	10	3,737	14
3,618	12			3,700	10	3,744	15
3,622	11					3,752	16
3,626	11					3,762	19
3,630	11					3,776	22
3,634	10						
3,637	10						
3,641	10						
3,644	10						
3,647	10						

Exhibit 3.6.2.16 Conditional Standard Error of Measurement (CSEM) for Scale Score: Spring 2021 – Grade 10 Mathematics

Minimally Proficient		Partially Proficient		Proficient		Highly Proficient	
Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
3,609	16	3,674	10	3,697	9	3,744	11
3,615	15	3,677	9	3,700	9	3,748	11
3,622	14	3,678	10	3,703	9	3,752	12
3,628	13	3,680	9	3,706	9	3,757	12
3,633	12	3,683	9	3,709	9	3,762	13
3,638	12	3,686	9	3,711	9	3,768	14
3,643	11	3,689	9	3,714	9	3,774	15
3,647	11	3,692	9	3,717	9	3,782	16
3,651	11	3,694	9	3,720	10	3,792	18
3,654	10			3,723	10	3,805	22
3,658	10			3,726	10	3,819	27
3,661	10			3,730	10		
3,665	10			3,733	10		
3,668	10			3,736	10		
3,671	10			3,740	11		

3.6.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed to estimate the likelihood of consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).[20] This index considers the consistency of classifications for the percentage of test takers that would, hypothetically, be classified in the same category on an alternate, equivalent form.

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

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

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

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

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

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

where $\Phi(\cdot)$ is the cumulative function of standard normal distribution. Similarly, the probability of the true score being below the cut score is estimated as

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

3.6.4 CLASSIFICATION ACCURACY

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

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

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

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

where the likelihood function is

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

and $P(x_i|\theta)$ is calculated from the Rasch model or partial credit model based on the estimated item parameters.

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

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

Mathematically, we have

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

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

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

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

where N_1 consists of the students with estimated $\hat{\theta}_i$ being at and above the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being below the cut score. The accuracy index is then computed as

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

In Exhibit 3.6.4.1, accurate classifications occur when the decision made based on the true score agrees with the decision made based on the form taken. Misclassifications, false positives, and false negatives occur when students' true score classifications are different from students' observed scores (e.g., a student whose true score results in a classification as Proficient, but whose observed score results in an incorrect classification as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 3.6.4.1 Classification Accuracy

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

3.6.5 CLASSIFICATION CONSISTENCY

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

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

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

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

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

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

The probability of inconsistency is estimated as

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

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

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

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

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

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

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

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

As shown in Exhibit 3.6.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 3.6.5.1 Classification Consistency

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

3.6.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

Exhibit 3.6.6.1 shows the classification accuracy and consistency indexes for the spring 2021 administration of the AzM2. Exhibit 3.6.6.2 and Exhibit 3.6.6.3 present the classification accuracy and consistency indexes for each of the identified subgroups: gender (female and male), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with SPED, FRL, and accommodations). Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because the consistency index assumes two test scores, both of which include measurement error. In contrast, the accuracy index assumes only a single test score plus the true score, which does not include measurement error.

Exhibit 3.6.6.1 Classification Accuracy and Consistency Estimates for Performance Standards Overall

Grade	Accuracy			Consistency		
	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
ELA						
3	0.92	0.93	0.96	0.89	0.90	0.95
4	0.92	0.92	0.96	0.89	0.89	0.94
5	0.94	0.93	0.94	0.91	0.90	0.92
6	0.91	0.92	0.97	0.88	0.89	0.96
7	0.91	0.92	0.97	0.87	0.88	0.95
8	0.93	0.93	0.96	0.90	0.91	0.94
10	0.92	0.92	0.96	0.88	0.89	0.94
Mathematics						
3	0.94	0.94	0.96	0.91	0.92	0.94
4	0.94	0.94	0.97	0.91	0.91	0.95
5	0.93	0.94	0.97	0.90	0.92	0.96
6	0.93	0.95	0.97	0.91	0.93	0.96
7	0.93	0.95	0.97	0.91	0.92	0.95
8	0.93	0.95	0.97	0.91	0.93	0.96
10	0.92	0.95	0.98	0.89	0.93	0.98

Exhibit 3.6.6.2 Classification Accuracy and Consistency Estimates for Performance Standards Across Subgroups: ELA

Grade	Subgroup	Accuracy			Consistency		
		Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
3	Overall	0.92	0.93	0.96	0.89	0.90	0.95
	Female	0.92	0.93	0.96	0.89	0.89	0.94
	Male	0.92	0.93	0.96	0.89	0.90	0.95
	African American	0.92	0.94	0.98	0.89	0.92	0.97
	Hispanic/Latino	0.92	0.94	0.98	0.89	0.91	0.97
	Asian	0.92	0.91	0.92	0.89	0.88	0.89
	White	0.92	0.91	0.94	0.88	0.88	0.92
	Hawaiian/Pacific	0.90	0.92	0.97	0.87	0.89	0.96

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

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

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

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

Exhibit 3.6.6.3 Classification Accuracy and Consistency Estimates for Performance Standards Across Subgroups: Mathematics

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

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

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

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

3.6.7 RELIABILITY FOR SUBGROUPS IN THE POPULATION

Exhibit 3.6.7.1 and Exhibit 3.6.7.2 show the reliability for each of the identified subgroups: gender (female and male), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with Individualized Education Plans [IEPs], SPED[21], FRL, and accommodations). As the exhibits indicate, reliabilities are generally stable across subgroups, meaning that the AzM2 assessments measure a common underlying achievement dimension across all subgroups, and that test scores are similarly precise across demographic subgroups. For subgroups where the reliability coefficients are attenuated, there is a corresponding decrease in the subgroup variance relative to the overall student population, indicating that attenuation of reliability in subgroups is due to a restriction of range.

Exhibit 3.6.7.1 Internal Consistency Reliability by Subgroup: ELA

Grade	Statistic	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
3	Reliability	0.90	0.90	0.90	0.88	0.90	0.88	0.88	0.84	0.90	0.90	0.86	0.77	0.88	0.79
	Variance	1065	1057	1067	889	1106	805	864	685	1042	1071	817	522	849	561
4	Reliability	0.91	0.90	0.91	0.89	0.90	0.90	0.89	0.87	0.89	0.90	0.87	0.81	0.89	0.82
	Variance	1084	1077	1082	894	1155	1063	895	723	1023	1079	833	533	862	563
5	Reliability	0.92	0.92	0.92	0.91	0.90	0.91	0.91	0.89	0.91	0.92	0.88	0.85	0.91	0.85
	Variance	1470	1436	1481	1292	1390	1270	1262	1065	1361	1461	1045	780	1261	780
6	Reliability	0.90	0.90	0.90	0.88	0.90	0.90	0.88	0.85	0.90	0.89	0.82	0.79	0.88	0.80
	Variance	1034	1008	1037	833	1104	1094	847	655	1042	986	609	506	826	546
7	Reliability	0.90	0.89	0.90	0.89	0.89	0.87	0.88	0.86	0.89	0.89	0.84	0.81	0.88	0.81
	Variance	1109	1032	1155	994	1198	882	942	800	1088	1097	787	640	937	649
8	Reliability	0.92	0.92	0.92	0.91	0.92	0.91	0.91	0.88	0.91	0.92	0.85	0.83	0.91	0.83
	Variance	1266	1220	1260	1149	1352	1103	1096	859	1225	1277	736	628	1066	641
10	Reliability	0.90	0.89	0.90	0.89	0.89	0.88	0.89	0.86	0.89	0.90	0.83	0.81	0.89	0.82
	Variance	919	860	948	868	891	733	818	690	878	942	616	561	819	599

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

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

Exhibit 3.6.7.2 Internal Consistency Reliability by Subgroup: Mathematics

Grade	Statistic	Overall	Female	Male	African American	Asian	Hawaiian/Pacific	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodations
3	Reliability	0.93	0.93	0.93	0.92	0.92	0.93	0.92	0.90	0.92	0.93	0.92	0.88	0.92	0.89
	Variance	2208	2082	2323	1791	2269	1915	1752	1396	1994	2185	2090	1263	1794	1399
4	Reliability	0.94	0.93	0.94	0.92	0.92	0.94	0.93	0.91	0.93	0.93	0.92	0.89	0.92	0.89
	Variance	2265	2158	2365	1907	2007	2349	1861	1613	2011	2180	2046	1425	1857	1490
5	Reliability	0.92	0.92	0.92	0.88	0.92	0.91	0.90	0.86	0.92	0.92	0.87	0.82	0.90	0.83
	Variance	1800	1701	1894	1299	1994	1592	1391	1147	1742	1790	1346	999	1397	1063
6	Reliability	0.93	0.92	0.93	0.89	0.93	0.92	0.90	0.86	0.93	0.93	0.85	0.82	0.90	0.84
	Variance	1899	1747	2041	1377	2103	1923	1438	1156	1917	1891	1224	979	1464	1083
7	Reliability	0.92	0.92	0.93	0.89	0.92	0.91	0.90	0.87	0.92	0.92	0.85	0.81	0.90	0.80
	Variance	1733	1635	1824	1340	2031	1486	1347	1100	1716	1749	1052	829	1349	823
8	Reliability	0.92	0.91	0.92	0.87	0.93	0.91	0.89	0.84	0.92	0.92	0.79	0.74	0.88	0.76
	Variance	1540	1413	1661	1053	2108	1469	1156	856	1643	1622	748	579	1144	657
10	Reliability	0.91	0.90	0.92	0.85	0.93	0.89	0.87	0.83	0.92	0.92	0.73	0.75	0.87	0.68
	Variance	1285	1162	1404	818	1750	1069	949	733	1397	1441	533	565	935	460

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

3.6.8 SUBSCALE RELIABILITY

Reliability estimates associated with the subscales for the 2019 operational forms are presented in Exhibit 3.6.8.1 through Exhibit 3.6.8.5. As indicated in the exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzM2.

Exhibit 3.6.8.1 Subscale Reliabilities: ELA Grades 3–8 and 10

Grade	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
3	0.73	0.73	0.81
4	0.75	0.75	0.80
5	0.79	0.80	0.77
6	0.76	0.73	0.77
7	0.78	0.72	0.73
8	0.79	0.77	0.81
10	0.80	0.67	0.79

Exhibit 3.6.8.2 Subscale Reliabilities: Mathematics Grades 3–5

	Numbers & Operations-Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
3	0.69	0.77	0.88
4	0.81	0.66	0.89
5	0.74	0.75	0.84

Exhibit 3.6.8.3 Subscale Reliabilities: Mathematics Grades 6 & 7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
6	0.81	0.74	0.70	0.59
7	0.73	0.65	0.73	0.75

Exhibit 3.6.8.4 Subscale Reliabilities: Mathematics Grade 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability & the Number System
8	0.79	0.68	0.48	0.71

Exhibit 3.6.8.5 Subscale Reliabilities: Mathematics Grade 10

	Algebra	Functions	Statistics & Quantitative Reasoning	Congruence & Geometric Properties with Equations	Similarity, Right Triangles, and Trigonometry & Circles and Geometric Measurement
10	0.72	0.58	0.65	0.63	0.52

3.7 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibit 3.7.1 through Exhibit 3.7.5. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.[22] The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

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

where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y. When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct. The disattenuated correlation equals 1 when the disattenuated correlation is greater than 1.

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

Exhibit 3.7.1 Subscale Intercorrelations: ELA Grades 3–8 and 10

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
3	Literature	0.69		0.95	
	Writing & Language	0.67	0.68	0.87	0.88
4	Literature	0.74		0.99	
	Writing & Language	0.69	0.70	0.90	0.90
5	Literature	0.78		0.98	
	Writing & Language	0.69	0.70	0.89	0.89
6	Literature	0.73		0.98	
	Writing & Language	0.67	0.65	0.87	0.86
7	Literature	0.72		0.96	
	Writing & Language	0.66	0.64	0.87	0.88
8	Literature	0.75		0.96	
	Writing & Language	0.71	0.70	0.89	0.89
10	Literature	0.66		0.90	
	Writing & Language	0.71	0.59	0.89	0.82

Exhibit 3.7.2 Subscale Intercorrelations: Mathematics Grades 3–5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.74		1.01	
	OAT_NBT	0.76	0.83	0.93	1.01
4	MDG	0.72		0.98	
	OAT_NBT	0.78	0.77	1.02	1.01
5	MDG	0.74		0.99	
	OAT_NBT	0.80	0.76	1.01	0.96

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

Exhibit 3.7.3 Subscale Intercorrelations: Mathematics Grade 6 & 7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	NS	0.80			1.04		
	RP	0.77	0.79	0.77	1.03	1.09	
	GSP	0.70	0.71	0.70	1.02	1.08	1.03
7	NS	0.77			1.11		
	RP	0.79	0.75		1.08	1.09	
	GSP	0.75	0.73	0.76	1.02	1.05	1.02

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

Exhibit 3.7.4 Subscale Intercorrelations: Mathematics Grade 8

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
8	F	0.76			1.04		
	G	0.68	0.61		1.10	1.07	1.10

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
	SPNS	0.78	0.71	0.65	1.04	1.02	1.11

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

Exhibit 3.7.5 Subscale Intercorrelations and Reliability Estimates: Mathematics Grade 10

	Subscale	Observed Correlations				Disattenuated Correlations			
		A	C_GPE	F	SRTT_C GM	A	C_GPE	F	SRTT_C GM
Grade	C_GPE	0.72				1.11			
	F	0.78	0.70			1.14	1.14		
	SRTT_CGM	0.71	0.69	0.70		1.06	1.14	1.09	
	S_NQ	0.65	0.59	0.63	0.59	1.14	1.02	1.11	1.03

Note: A = Algebra; F = Functions; S_QR = Statistics & Quantitative Reasoning; C_GPE = Congruence & Geometric Properties with Equations; SRTT_CGM = Similarity, Right Triangles, and Trigonometry & Circles and Geometric Measurement

3.8 HANDSCORING AGREEMENT RATE

For grades in which statistical models were constructed for machine scoring of essay responses, Measurement, Inc. (MI) handscored responses per prompt, with each response double scored and any discrepant scores routed for a final resolution score. At each grade, students responded to one of two randomly selected writing tasks. Exhibit 3.8.1 shows the summary of the rater agreement for the writing prompts administered on the AzM2 spring 2021 online tests. The rater agreement reports show percentages of exact agreement (Equal), adjacent scores (Adj. Low or Adj. High), and nonadjacent scores (Non-Adj Low or Non-Adj High). The tables also identify mismatched scores when there is a difference involving nonscorable condition codes (Mismatch NS) or a nonscorable/scorable mix (MM NS/Score). Exhibit 3.8.1 summarizes those results, showing the mean exact agreement rate for dimension scores across grades. Generally, exact agreement rates ranged from 51%–77%, with little variability across the essay prompts.

Exhibit 3.8.1 ELA Writing Prompt Rater Agreement Report: Spring 2021 Administration

Grade	Dimension	Total Read	Second Read	Non Adj Low	Adj Low	Equal	Adj High	Non Adj High	Mismatch NS	MM NS/Score
3	Purpose/Organization	9,895	1,680	1.9	19.0	54.5	19.0	1.9	0.0	3.7
	Evidence/Elaboration	9,887	1,680	1.4	19.2	55.0	19.2	1.4	0.0	3.7
	Conventions	10,164	1,680	0.3	17.6	60.5	17.6	0.3	0.0	3.7
4	Purpose/Organization	9,749	1,722	1.0	20.4	56.8	20.1	1.0	0.0	0.3
	Evidence/Elaboration	9,756	1,722	1.4	18.6	59.7	18.6	1.4	0.0	0.3
	Conventions	10,051	1,722	1.0	17.5	62.6	17.5	1.0	0.0	0.3
5	Purpose/Organization	9,854	1,750	1.1	17.9	61.7	17.9	1.1	0.0	0.1
	Evidence/Elaboration	9,847	1,750	0.7	18.2	62.1	18.2	0.7	0.0	0.1
	Conventions	10,098	1,750	0.6	14.5	69.7	14.5	0.6	0.0	0.1
6	Purpose/Organization	11,803	2,100	0.6	20.4	57.7	20.4	0.6	0.0	0.3
	Evidence/Elaboration	11,804	2,100	0.6	19.3	59.8	19.3	0.6	0.0	0.3
	Conventions	12,130	2,100	0.8	15.4	67.4	15.4	0.8	0.0	0.3
7	Purpose/Organization	11,729	2,074	1.2	19.8	57.8	19.8	1.2	0.0	0.3
	Evidence/Elaboration	11,733	2,074	1.4	19.6	57.7	19.6	1.4	0.0	0.3
	Conventions	12,000	2,074	0.6	13.7	71.2	13.7	0.6	0.0	0.3

Grade	Dimension	Total Read	Second Read	Non Adj Low	Adj Low	Equal	Adj High	Non Adj High	Mismatch NS	MM NS/Score
8	**Purpose/Organization**	11,660	2,064	1.9	21.9	52.2	21.9	1.9	0.0	0.3
	Evidence/Elaboration	11,670	2,064	2.4	22.1	50.8	22.1	2.4	0.0	0.3
	Conventions	11,856	2,064	0.8	10.6	76.9	10.6	0.8	0.0	0.3
10	**Purpose/Organization**	7,557	1,310	1.7	22.2	51.9	22.2	1.7	0.0	0.3
	Evidence/Elaboration	7,566	1,310	2.4	21.8	51.5	21.8	2.4	0.0	0.3
	Conventions	7,747	1,310	1.2	14.8	67.6	14.8	1.2	0.0	0.3

4 ITEM DEVELOPMENT AND TEST CONSTRUCTION

Arizona's Statewide Achievement Assessments (AzM2) are rigorously examined in accordance with the guidelines provided in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence that is necessary to validate assessment scores for their intended purposes based on these standards.

The AzM2 assessments were designed to measure student progress toward achievement of the Arizona State Standards. Although the validity of AzM2 test score interpretations is evaluated along several dimensions, as a criterion-referenced system of tests, the meaning of test scores is critically evaluated by the degree to which test content was aligned with the Arizona State Standards.[23]

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards. This process refers to those standards in a highly iterative test development process that includes the Arizona Department of Education (ADE), test developers, and educator and stakeholder committees. Items used to develop the spring 2015 operational test forms were drawn mainly from the Independent College and Career Readiness (ICCR) item pool, which was developed to align with the Common Core State Standards (CCSS). The development process for the summer 2016 and fall 2016 operational tests was the same for the spring 2016 operational test and is described in the 2016 AzM2 technical report. The items were all reviewed by Arizona content experts and educators before field testing in spring 2016 and subsequent operational test administration in spring 2017. Only items that aligned well with the Arizona State Standards were used. A few previously developed Arizona items that also aligned to the Arizona State Standards were used to supplement the AzM2 item pool.

The items used for spring 2019 operational test forms were reused for the spring 2021 operational forms. Items used to develop the spring 2019 and 2021 operational test forms were drawn from custom Arizona item development and CAI's ICCR item bank. Both custom Arizona items and ICCR items were developed to align with the   CCSS. These items were all reviewed by the ADE, Arizona content experts and educators, and Arizona community members before field testing in spring 2016, spring 2017, spring 2018, and spring 2019, and subsequent operational test administration in spring 2017, spring 2018, spring 2019, and spring 2021. Only items that aligned well with the Arizona State Standards and were free of bias or sensitivity concerns were used.

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

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

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

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

4.1 ITEM DEVELOPMENT PROCESS

The content development process for AzM2 is managed within CAI's Item Tracking System (ITS), which acts as a content development and management tool, item bank, and publication system supporting both paper-pencil and online publication. This item-development workflow leads items from inception through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes and rationales at each review and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence and that the review is conducted (or recorded in the case of committee review) by an authorized person. As items travel through Arizona's extensive review process, every version of every item is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

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

The AzM2 item development process is predicated on a high level of interaction between test developers at CAI and the ADE, as well as with Arizona educators and stakeholders. CAI's ITS manages item content throughout the entire life cycle of an item from inception through a series of agreed-on 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 CAI management on-demand reports of the content and status of the inventory of items. Each item is directed through a sequence of reviews and sign-offs by CAI and ADE staff before it is locked in for field-test or operational administration.

The ITS is integrated with the item display engine used by the AzM2 online Test Delivery System (TDS). This feature, combined with a "web approval" process, allows the display of online items to be "locked" well before test forms are constructed and ensures that only approved items are administered to Arizona students.

4.1.1 ITEM WRITING

Test development experts use item specifications to guide the item development process.[24] These item specifications, developed by content experts at CAI and the ADE, strategically guide the item development process. They are detailed documents that specify content limits, model tasks, and response types for a specific standard. Item writers use these specifications while developing items to make the best use of the available item types.

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

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

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

Detailed item specifications include the following:

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

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

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

- **DOK 1—Recall.** Students recall basic mathematical ideas, perform basic arithmetic operations using established algorithms, and identify examples of general mathematics principles.
- **DOK 2—Skill/Concept.** Students apply their basic knowledge (DOK 1) and extend their thinking to problem solve, identify relationships, and draw conclusions.
- **DOK 3—Strategic Thinking.** Students go beyond basic problem solving (e.g., word problems) to extend their thinking to nonroutine problem solving, hypothesize, and critique arguments or problem-solving strategies.
- **DOK 4—Extended Thinking.** At this highest level, students engage in extended problem-solving activities, which require the 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 test construction blueprint indicates the percentage of test items aligned to each DOK level for each assessment. 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.

4.1.1.1 ELA

ELA item development often begins with the development of reading passages. AzM2 passages represent a variety of genres and topics. CAI’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 AzM2 utilizes both single passages and 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 analyze each passage in depth. The passage specifications call for a close examination of both quantitative measures, such as word counts and Lexile readabilities, and qualitative measures, such as passage structure and levels of meaning, all of which are defined as important measures of text complexity.

AzM2'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 various stimuli (at least two to three per task) that serve as a springboard for an informed piece of writing. Students are given research articles, charts and graphs, and narratives to serve as the basis for their written responses. 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 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 and from primary and secondary sources. Students develop a main idea and a primary focus as they relate facts, details, and examples.

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

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

4.1.1.2 MATHEMATICS

Calculators are not allowed for assessments at grades 3–6, while students participating in the grade 10 assessment are allowed continual access to specific calculator functions. For the grades 7 and 8 assessments, where calculator usage is allowable for some item types, the test items are grouped into two segments, administered separately to students: calculator and no calculator. The construct of the items dictates in which section they are to be assessed.

4.2 MACHINE-SCORED CONSTRUCTED-RESPONSE ITEM DEVELOPMENT TOOLS

AzM2 includes several machine-scored constructed-response (MSCR) items that leverage a sophisticated system that allows for a large variety of item types expecting varied student responses to be developed and scored efficiently and economically.

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

In addition, throughout the item development and review process, test developers can mimic the many different possible student responses and review how the rubric is applied to those responses. Test developers can evaluate 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 multi-step items. Test developers can customize the equation palette to show the appropriate functions. Just as the keypad is customizable, the answer spaces are, as well. Additional answer spaces can be added as needed by the item writer. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer.

Such tools are integrated into the ITS, providing test developers with the power and flexibility to use technology to create sophisticated AzM2 items.

4.2.1 ITEM TYPES

AzM2 includes a wide variety of item types designed around a broad and growing catalog of response mechanisms. In addition to selected-response items, which have traditional multiple-choice and more advanced multi-select and two-part items, AzM2 tests utilize various item types, including those with the following response mechanisms:

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

AzM2 items use technology to measure deeper knowledge and the student's application of knowledge in a more open-ended way and to machine-score many such items. All MSCR items administered in AzM2 are accessible. There may be occasions where it is necessary to sacrifice accessibility for some populations to measure a critical standard, but test development staff would need to consider the measurement benefit carefully before developing that item.

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

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

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

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

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

4.3 ITEM REVIEW

This section describes the multi-step item review process that items travel through–from inception to several rounds of review by test developers, the ADE, and educators; and to field testing and final review–prior to inclusion on operational test forms.[25] The items used for the spring 2019 operational test forms were reused for the spring 2021 operational forms. Items used to develop the spring 2019 and spring 2021 operational test forms were drawn from custom Arizona item development and CAI's ICCR item bank. Both custom Arizona items and ICCR items were developed to align with the CCSS. These items were all reviewed by the ADE, Arizona content experts and educators, and Arizona community members, before field testing in spring 2016, spring 2017, spring 2018, and spring 2019, and subsequent operational test administration in spring 2017, spring 2018, spring 2019, and spring 2021. Only items that aligned well with the Arizona State Standards and were free of bias or sensitivity concerns were used.

The item review procedures used to develop and review AzM2 test items are designed to ensure item accuracy and alignment with the intended Arizona State Standards. Following a standard item review process, item reviews proceed initially through a series of internal reviews before items are eligible for review by the ADE's content experts. Most of CAI's content staff who are responsible for conducting internal reviews are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through four internal review steps before it is eligible for review by the ADE. Those steps include:

- **Preliminary review.** This review is conducted by a group of CAI content-area experts.
- **Content Review 1.** This review is performed by a CAI content specialist.
- **Editorial Review.** In this step, a copyeditor checks the item for correct grammar/usage.
- **Senior Content Review.** This review is performed by the lead content expert.

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

At every stage of the item review process, beginning with preliminary review, CAI's test developers analyze each item to ensure that it meets the following criteria:

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

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

Items passing through the internal review process are sent to the ADE for review. At this stage, items may be further revised based on any edits or changes requested by the ADE or rejected outright. Items passing through the ADE's review then pass through a stakeholder review, in which educators review each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzM2 item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials were:

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

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

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

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

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

4.4 FIELD TESTING

To establish a pool of items for constructing future AzM2 test forms, newly developed test items were embedded in the spring 2016, spring 2017, spring 2018, and spring 2019 AzM2 test forms for field testing. Embedding field-test items in operational assessments yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations. Several factors that may influence item difficulty in the context of operational test administrations may be less relevant in stand-alone field-test contexts. For example, in a high-stakes test, such as high school end-of-course (EOC) exams where test performance may impact student grades, students may be motivated to expend greater effort to achieve maximum performance. Conversely, the high-stakes assessments may also be more likely to elicit anxiety in some students, thus impairing their performance on the tests. Even when assessments are low stakes for students, schools often work to convey to students the importance of statewide assessments in ways that are likely not done for independent field tests. While the impact of contextual factors may not be great, embedded field testing ensures that all aspects of the operational testing context influencing item difficulty are incorporated into the resulting item parameter estimates.

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

A potential drawback of the EFT approach is the increased assessment burden placed on students and schools. For this reason, AzM2 utilizes EFT designs for purposes of item bank maintenance. Arizona uses CAI’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:[26]

- 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

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

The embedded field-testing algorithm consists of two different algorithms—one for identifying which field-test items will be administered to which student (the distribution algorithm), and one for selecting the position on the test for each item administered to the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field-test items designated for administration to each student. This randomization ensures that (a) each item is seen by a representative sample of Arizona students, and (b) every item is as likely as every other item to appear in a class or school, minimizing clustering effects.

In addition, a fixed block of field-test items was also embedded in paper-pencil AzM2 test forms so that the number of items responded to by students did not vary between assessment modes.

In the spring 2015 administrations, item parameters for the ELA and mathematics assessments were calibrated following the online administration to establish the AzM2 bank scale. Following the spring 2016 and spring 2017 test administrations, the free calibration was performed on the operational items on each of the ELA and mathematics tests. Then, the free calibrated item parameters were linked back to the 2015 spring scale using the mean-mean equating method. The field-test item calibration was conducted by anchoring on the post-equated operational item parameters for all the ELA and mathematics tests. However, only the ELA spring 2016 operational tests were scored using the post-equated item parameters.

4.5 ITEM STATISTICS

Following the close of spring testing windows, CAI 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.[27] Analysis of field-test items includes classical item statistics and the item response theory (IRT) item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subsgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to 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 CAI and ADE psychometric and content staff for possible miskey or scoring errors. Items that pass through CAI and ADE statistical review are accepted for future operational use. Appendix G provides the slide presentation used to train reviewers for item data review. The training is designed to ensure that all reviewers understand how items are evaluated and that they are interpreting item statistics correctly.

4.5.1 CLASSICAL STATISTICS

Classical item analyses ensured that the field-test items function as intended with respect to the AzM2's underlying scales. CAI's analysis program computed the required item and test statistics for each selected-response (SR) and constructed-

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

response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are either extremely difficult or extremely easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer (*p*-value) and those selecting the incorrect responses is computed. For CR items, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the number of points possible). Items are flagged for review if the *p*-value was less than .05.

The item discrimination index indicates the extent to which each item differentiates between those test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for dichotomous items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response was between .23 and .27. Items with biserials less than .23 were automatically rejected.

Distractor analysis for the dichotomous items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response is greater than 0. In addition, items are flagged if the proportion of students responding to a distractor exceeds the proportion selecting the keyed response. Although non-modal response keys are typically observed with difficult items, in combination with poor item discrimination, it may indicate a miskeyed item.

4.5.2 ITEM RESPONSE THEORY STATISTICS

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

4.5.3 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 CAI and the ADE's staff to make a final decision about whether the item should be excluded from the pool of potential items given its performance in field testing potential items.

CAI 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/focus group):

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

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

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

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed in Exhibit 4.5.3.1. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focus 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, 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 4.5.3.1. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focus or reference groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992; Camilli & Shepard, 1994; Muniz, Hambleton, & Xing, 2001; Sireci & Rios, 2013).

Exhibit 4.5.3.1 DIF Classification Rules

Item Type	Category	Rule
Dichotomous Items	C	MH χ^2 is significant and $\lvert\Delta_{\text{hat } MH}\rvert \geq 1.5$
	B	MH χ^2 is significant and $\lvert\Delta_{\text{hat } MH}\rvert < 1.5$

	A	MH χ^2 is not significant
Polytomous Items	C	MH χ^2 is significant and $\lvert SMD \rvert / \lvert SD \rvert \geq .25$
	B	MH χ^2 is significant and $\lvert SMD \rvert / \lvert SD \rvert < .25$
	A	MH χ^2 is not significant

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, CAI content specialists begin the process of constructing test forms. Operational passages and items qualified for operational forms are those that meet all the criteria established by the ADE in terms of content, fairness review, and data characteristics.

4.6.1 OPERATIONAL FORM CONSTRUCTION

Each AzM2 form is built to exactly match the detailed test blueprint and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the DOK with which it is covered, the type of items that measure the constructs, and every other content-relevant aspect of the test. The statistical targets, which are held constant across years and across modes, ensure that students receive scores of similar precision, regardless of which form of the test they receive.[28]

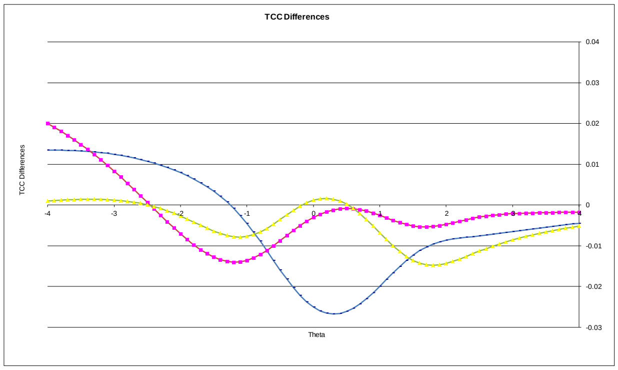
CAI's test developers used Form Builder software to help construct operational forms. Form Builder interfaces with CAI's Item Tracking System (ITS) to extract test information and interactively create test characteristics curves (TCCs), test information curves, and standard error of measurement curves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

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

As test developers built forms, the Form Builder-generated TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 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 across a range of abilities. These differences reflect different emphases in test information across forms at these ability levels. If the difficulty and error structure for the target forms is virtually identical to the reference form, as in the sample TCC and SEM curves, the item selection process concludes with multiple, parallel test forms. Once the goal of parallel forms is achieved, the information is entered into ITS, which tracks item usage and generates bookmaps (test maps) for use in scoring, forms development, and other processes.

[28] Standard 4.12: Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

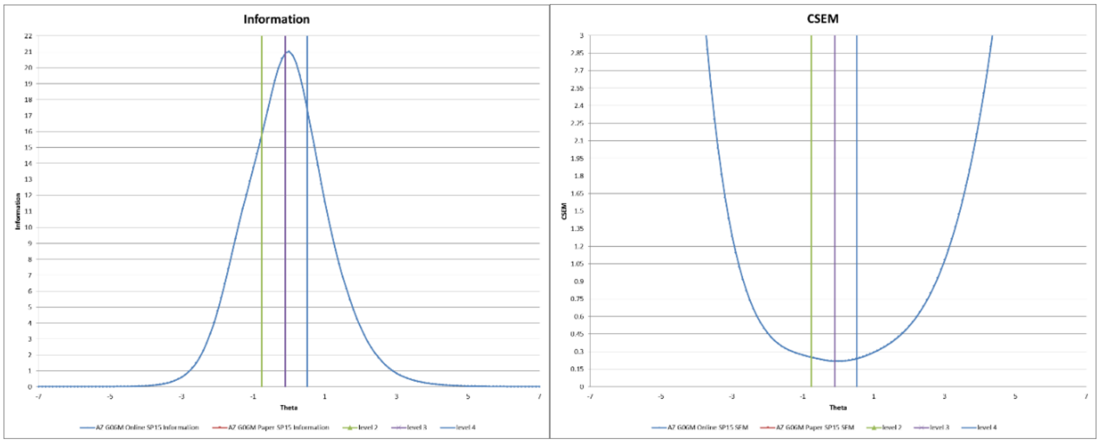
Exhibit 4.6.1.1 Test Characteristics Curve Differences



The reference form for each assessment is the operational test form administered in spring 2015. As illustrated in Exhibit 4.6.1.2, by evaluating test characteristics in reference to the base year forms, students are administered tests each year that are equivalent in difficulty across the range of ability. The TCC and SEM graphs used to evaluate the spring 2021 operational test forms are presented in Appendix H.

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

Exhibit 4.6.1.2 Test Information and Standard Errors Relative to Performance Standards



4.6.2 TEST INFORMATION FUNCTION

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

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

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

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

To better depict measurement error at various points along the scale, which is congruent with the *Standards for Educational and Psychological Testing,* the graphs and the values of test information function (TIF) for the spring 2018 online forms and the spring 2021 online forms are presented in Appendix I. Additionally, the graph and the values of the ratio for information function between the spring 2018 online forms and the spring 2021 online forms are presented in Appendix I.

4.6.3 ASSEMBLING TEST FORMS

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

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

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

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

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

- The items are sequentially ordered.
- Each item of the same type is presented in a consistent manner.
- The listing of the options for the multiple-choice items is consistent.
- The answer options are labeled correctly.
- All graphics are consistently presented.
- All tables and charts have titles and are consistently formatted.
- The number of the answer choice letters is approximately equal across the form.

- The answer key was checked by the initial reviewer and one additional independent reviewer.
- All stimuli have items associated with them.
- The topics of items, passages, or stimuli are not too similar to one another.
- There are no errors in spelling, grammar, or accuracy of graphics.
- The wording, layout, and appearance of the item matches how the item was field tested.
- There is gender and ethnic balance.
- The passage sets do not start with or end with a constructed-response item.
- Each item and the form are checked against the appropriate style guide.
- The directions are consistent across items and are accurate.
- All copyrighted materials have up-to-date permissions agreements.
- Word counts are within documented ranges.

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

5 TEST ADMINISTRATION

5.1 ELIGIBILITY

Arizona public school students in grades 3–8 and 10 were required to participate in Arizona's Statewide Achievement Assessment (AzM2) testing.[29] New for spring 2021, the requirements for testing were revised to meet ADE's approved Every Student Succeeds Act (ESSA) State Plan. All students in grades 3–8 were required to take the grade level ELA and mathematics tests. All high school students in grades 9–12 who belong to Cohort 2023 were required to take the grade 10 ELA and mathematics tests.

Students with significant cognitive disabilities whose current Individualized Education Plan (IEP) designated them as eligible for the alternate assessment, the Multi-State Alternate Assessment (MSAA), were excluded from AzM2.

5.2 ADMINISTRATION PROCEDURES

Key personnel involved with AzM2 administration include the district test coordinators (DTCs), school test coordinators (STCs), and test administrators (TAs) who proctor the test. For information about the roles and responsibilities of testing staff, refer to the following sections.

CAI's Secure Browser was required to access the computer-based AzM2 tests. The Secure Browser provided a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to desktop functionalities, such as the Internet and email. Other measures that protect the integrity and security of the online test are presented in Section 5.5 of this technical report.

Prior to each test administration, statewide DTC training sessions were conducted to provide information regarding both the paper-based testing (PBT) and computer-based testing (CBT) administrations. The training also provided an overview of the Test Delivery System (TDS), Online Reporting System (ORS), and the Test Information Distribution Engine (TIDE). Recorded training sessions and narrated training videos were posted online. The *Test Administrator Manual* (TAM) and Test Administration Directions were shipped to every testing district. Additionally, TAs were required to complete the online TA Certification Course before CBT administration.[30] DTCs and STCs were responsible for ensuring that all test administration personnel (for both PBT and CBT) were properly trained before the start of testing using the various resources.

Manuals and guides on test administrations are available on the AzM2 Portal.[31] The *Test Administrator User Guide* was designed to familiarize test administrators with the TDS and contains tips and screenshots throughout the text. The guide

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

[30] Standard 6.1: TAs should follow the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user carefully.

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

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

provides enough how-to information to enable TAs to access and navigate the TDS. The *User Guide* provides information on the following topics:

- Steps to take before accessing the system and logging in
- Navigating the TA Interface
- The Student Interface, used by students for CBT
- Training sites available for TAs and students
- Secure browsers and keyboard shortcut keys

The *AzM2 Test Coordinator's Manual* provides information about policies and procedures for AzM2 test coordinators. This manual is updated before each test administration and includes test administration policies and guidance for test coordinators before, during, and after the testing window.

The *AzM2 Test Administration Directions, Grade 10* and the *AzM2 Test Administration Directions, Grades 3–8* provide information about policies and procedures for the AzM2, both CBT and PBT versions. The *AzM2 Test Administration Directions*, which is updated before each test administration, includes test administration information, guidance, and directions.

The *AzM2 Test Administration Directions* provide easy-to-follow instructions for the online testing environment, such as creating online testing sessions, monitoring online sessions, verifying student information, assigning test accommodations, and starting and pausing test sessions.[32] Similar guidance is provided for the PBT environment, including instructions for the PBT session, monitoring sessions, verifying student information, and providing test accommodations. Additional instructions for administering tests to students using braille accommodated test booklets are provided in the *Supplemental Instructions for Braille* documents.

District and school personnel involved with AzM2 test administration played an important role in ensuring the validity of the assessment by maintaining both standardized administration conditions and test security.

DTCs were responsible for coordinating testing at the district level. They were ultimately accountable for ensuring that testing was conducted in accordance with the test security and other policies and procedures established by the Arizona Department of Education (ADE). They ensured that the TAs in each school were appropriately trained and aware of policies and procedures, and that they were trained to use the reporting system.

Districts may also identify STCs, who may assist in the identification and training of TAs. They may also create testing schedules and procedures for the school. If the school administers AzM2 online, the STCs may work with technology coordinators to ensure that the necessary secure browsers were installed, and any other technical issues were resolved. During the testing window, STCs must monitor testing progress, ensure that all students participate as appropriate, and handle testing incidents, as necessary.

TAs were responsible for reviewing necessary manuals and user guides to prepare the testing environment and ensuring that students did not have unapproved books, notes, or electronic devices available during testing. TAs were required to

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

administer AzM2 tests following the directions found in the *AzM2 Test Administration Directions.*[33] Any deviation in test administration must be reported by TAs to the STC, who reports it to the DTC. The DTC then reports it to the ADE.

TAs who administered computer-based AzM2 tests conducted a training test session using the AzM2 Sample Tests. TAs were required to pass a qualifying test before they were eligible to administer the AzM2 online.[34]

TAs had to ensure that only resources allowed for specific tests were available and no additional resources were used during the test. No calculators were permitted in AzM2 mathematics tests for grades 3–6. Scientific calculators were permitted in AzM2 Mathematics Part 1 for grades 7 and 8. Graphing calculators were permitted in AzM2 Mathematics Parts 1 and 2 for grade 10. Online calculators were provided as embedded tools within the appropriate CBT parts. Handheld calculators could be provided to students during the appropriate test sessions. Calculator guidance was provided in both the *AzM2 Test Coordinator's Manual* and the *AzM2 Test Administration Directions*. The online calculators were publicly available on the AzM2 Portal and securely available in a secure browser for students taking paper-pencil tests, if needed. Providing a calculator with prohibited functionality or in the incorrect test session was cause for test invalidation.

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

For the paper-pencil AzM2 tests, TAs had to ensure that students used No. 2 pencils to record their responses. STCs provided TAs with the materials needed to administer each test session. Secure materials were delivered or picked up immediately before the beginning of each test session. During mathematics testing and when responding to the writing prompt, students were permitted to use the scratch paper as a workspace. After testing, TAs needed to return the testing materials, including all scratch paper, to the STC.

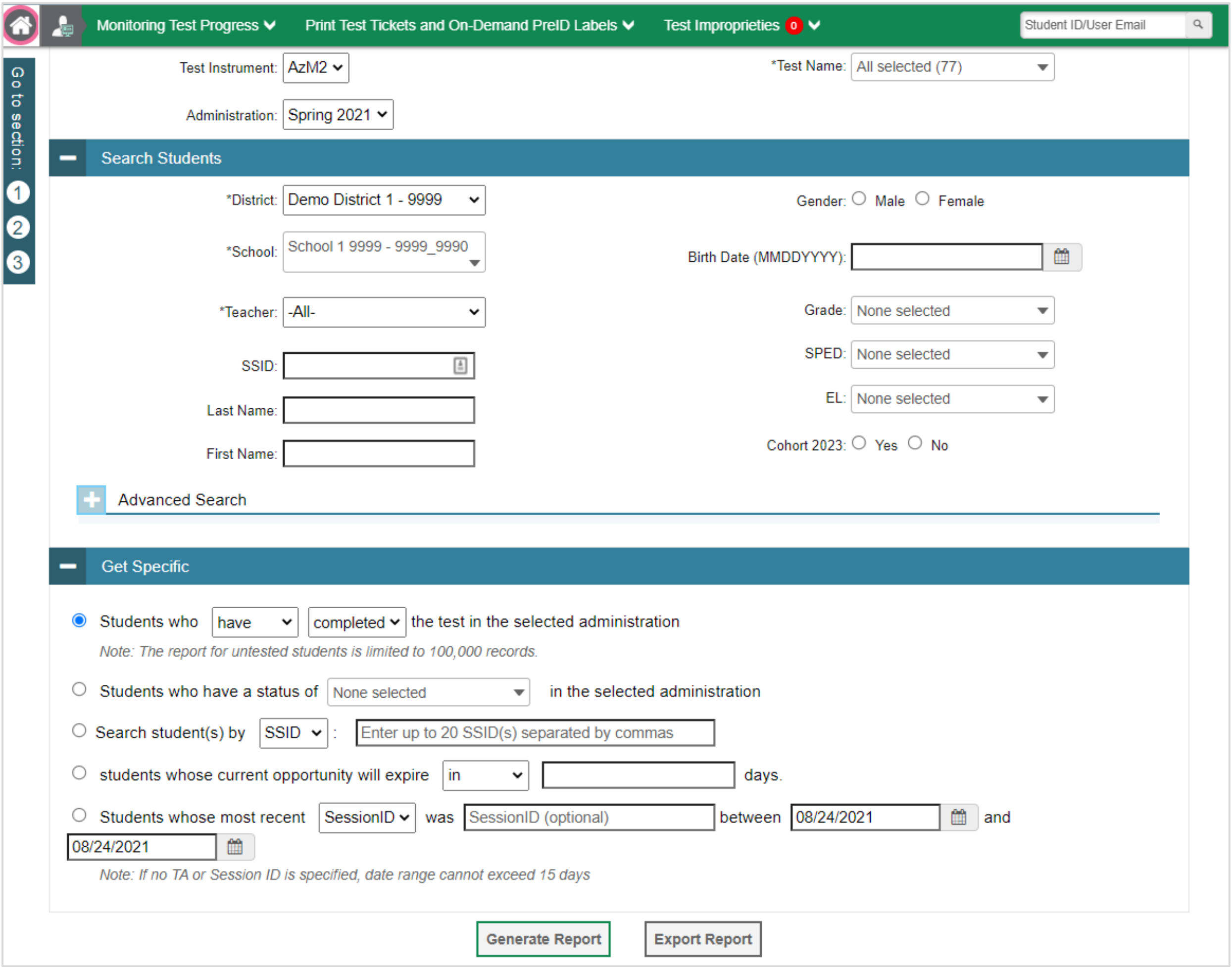
The STC and TAs worked together to determine the most appropriate testing option(s), testing environment, and the average time needed to complete each test. The appropriate protocols were established to maintain a quiet testing environment throughout the testing session. TAs also needed to ensure that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

[33] Standard 6.1: TAs should follow the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user carefully.

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

5.2.1 MANAGING TESTING

To help schools manage their test schedule, allocate testing resources, and prioritize testing, TIDE offered participation reports for online testers. Within TIDE, educators can generate up-to-the-minute reports showing students' test status. In addition, district-level users can monitor testing progress across schools.



Test Instrument: AzM2

*Test Name: All selected (77)

Administration: Spring 2021

Search Students

*District: Demo District 1 - 9999

Gender: ○ Male ○ Female

*School: School 1 9999 - 9999_9990

Birth Date (MMDDYYYY):

*Teacher: -All-

Grade: None selected

SSID:

SPED: None selected

Last Name:

EL: None selected

First Name:

Cohort 2023: ○ Yes ○ No

Advanced Search

Get Specific

● Students who have completed the test in the selected administration

Note: The report for untested students is limited to 100,000 records.

○ Students who have a status of None selected in the selected administration

○ Search student(s) by SSID : Enter up to 20 SSID(s) separated by commas

○ students whose current opportunity will expire in days.

○ Students whose most recent SessionID was SessionID (optional) between 08/24/2021 and 08/24/2021

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

5.3 TESTING CONDITIONS, TOOLS, AND ACCOMMODATIONS

This section summarizes the testing conditions, tools, and accommodations that are available to AzM2 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

TAs are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student to provide a more comfortable and distraction-free testing environment.[35] Universal test administration conditions are available for both PBT and CBT. Universal test administration conditions include:

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel
- Being seated in a specific location within the testing room or being seated at special furniture
- Having the test administered by a familiar TA
- Using a special pencil or pencil grip
- Using a placeholder
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT)
- Using devices that allow the student to hear the test directions: hearing aids and amplification
- Wearing noise buffers after the scripted directions have been read
- Signing the scripted directions
- Having the scripted directions repeated (at student request)
- Having questions about the scripted directions or the directions that students read on their own answered
- Reading the test quietly to himself/herself as long as other students are not disrupted
- Allowing extended time (Testing session must be competed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's IEP as an accommodation, for AzM2 testing purposes, these are not considered testing accommodations and are available to any student who needs them, not just to students with IEPs or Section 504 Plans.

5.3.2 UNIVERSAL TESTING TOOLS FOR COMPUTER-BASED TESTING

The AzM2 CBT platform offers numerous testing tools. All tools are available in the AzM2 Sample Tests, which are available to TAs and students before each test administration. TAs are encouraged to ensure that students who will participate in the computer-based AzM2 take the AzM2 Sample Tests and familiarize themselves with the available tools.

Exhibit 5.3.2.1 summarizes the universal test tools that are available to all students in all AzM2 tests; these features cannot be disabled by TAs.

Exhibit 5.3.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Click anywhere on the selected-response text or button for multiple-choice options.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen *Test Instructions and Help*.
Highlighter	Highlight text in a passage or item.

[35] Standard 3.4: Test takers should receive comparable treatment during the test administration and scoring process.
[35] Standard 4.5: If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.
[35] Standard 6.4: The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

Universal Test Tool	Description
Line Reader	This allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	This allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally throughout the session. In mathematics, comments are attached to a specific test item and available throughout the session.
Pause and Restart	This allows the session to be paused at any time and restarted and taken over a one-day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	This allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
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, paste) and basic text formatting tools (bold, underline, italics) are available 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 to the original size.

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

AzM2 testing requires specific subject-area tools or resources for certain portions of AzM2. 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.

Tool	Applicable Subject Area	Description of Tool
Calculator **Grades 7–8 (Part 1 only): Specific scientific calculators are acceptable.** **EOC (entire test): Specific graphing calculators are acceptable.**	Mathematics	CBT: Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted. PBT: Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.

5.3.4 ACCOMMODATIONS

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

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education (SPED) need, or language needs, and the accommodation(s) provided to the student during educational activities, including assessments. TAs are instructed to make accommodation decisions based on individual needs and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzM2 test that has not already been 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 AzM2 are generally limited to those listed in the *AzM2 Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section.[36] Arizona takes care to ensure that allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzM2. If a student's IEP calls for a testing accommodation that is not listed, TAs are instructed to contact the ADE for guidance.

Allowable accommodations are described on the following pages.[37]

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

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

5.3.4.1 ACCOMMODATIONS FOR STUDENTS WITH AN INJURY

Students with an injury, such as a broken hand or arm, which would make it difficult to participate in AzM2 may use, as appropriate, any of the universal test administration accommodations described in Exhibit 5.3.4.1.1. There are no specific CBT tools to support these accommodations.

Exhibit 5.3.4.1.1 Accommodations for Students with an Injury

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

5.3.4.2 ACCOMMODATIONS FOR EL, FEP, AND RFEP STUDENTS

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

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

Exhibit 5.3.4.2.1 summarizes accommodations that may be provided for EL, RFEP, and FEP students.

Exhibit 5.3.4.2.1 Allowable Accommodations for EL, RFEP, and FEP Students

Accommodation	Description of Use
Read-Aloud Test Content	CBT: Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the mathematics test. PBT: Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test upon student request.
Rest/Breaks	Provide students with rest breaks during testing sessions.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	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.

5.3.4.3 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 5.3.4.3.1, as designated in their IEP or Section 504 Plan.

Exhibit 5.3.4.3.1 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzM2 mathematics test.
Adult Transcription	If a student testing at a CBT school has an IEP indicating that they cannot enter their own responses on a computer, the school must order a Special Paper Version test for that student. An adult must transfer the student's responses exactly as provided orally or by gestures, into the paper-pencil booklet and then into the DEI, or directly into the DEI. If a student testing at a PBT school has an IEP indicating Adult Transcription, an adult must transfer the student's responses exactly as provided orally or by gestures into the paper-pencil booklet.
ASL and Closed Caption	In CBTs, this is available for the listening items on the reading ELA test.

Accommodation	Description of Use
Assistive Technology	This is the use of assistive technology for the writing response and/or other open-response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation requires Adult Transcription (refer to the appropriate entry in this table for rules on Adult Transcription).
Braille Test Booklet	Provide a paper braille test booklet. This accommodation requires Adult Transcription (refer to the appropriate entry in this table for rules on Adult Transcription).
Large Print Test Booklet	CBT: Either increase default zoom settings when a student participates in CBT or provide a PBT Large Print test booklet. PBT: Provide a Large Print test booklet. PBT Large Print Test booklets require Adult Transcription into the DEI. Refer to the appropriate entry in this table for rules on Adult Transcription.
Paper-Pencil Test Booklet	CBT: Student's IEP must indicate that the student cannot enter their own responses on the computer and requires a paper-pencil test or adult transcription. The school will provide a Special Paper Version booklet for the student. The student's responses must be transcribed into the paper-pencil booklet and then entered into the DEI or entered directly into the DEI. Refer to the appropriate entry in this table for rules on Adult Transcription.
Read-Aloud Test Content	CBT: Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the mathematics test. PBT: Read aloud, in English, any of the test content in the writing portion of the ELA test and the mathematics test.
Rest/Breaks	Provide students with rest breaks during testing sessions.
Sign Test Content	Sign any of the content of the writing portion of the ELA test. Sign any of the content of the mathematics test. Signing the content of the reading portion of the ELA test.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

5.4 SYSTEM SECURITY

5.4.1 SECURE SYSTEM DESIGN

CAI has developed a custom single sign-on application that is made available on Arizona's secure portal. This application is used to support access to CAI'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 CAI's portal, where they have access to CAI's secure applications, such as TIDE, the TDS, and the ORS. Nightly backups protect the data. The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful, or they will need to rerun the backup. The system can withstand failure of almost any component with little or no interruption of service.

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

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

5.4.2 SYSTEM SECURITY COMPONENTS

CAI has built-in security controls in all its data stores and transmissions.[38] Unique user identification is a requirement for all systems and interfaces. All of CAI's systems encrypt data at rest and in transit.

5.4.2.1 PHYSICAL SECURITY

AzM2 data resides on servers at Rackspace, CAI'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 CAI facilities and are accessed from CAI machines. CAI'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 CAI employees, have undergone rigorous background checks.

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

5.4.2.2 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.

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

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

5.4.2.3 SOFTWARE SECURITY

All of CAI'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.

CAI'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. CAI has worked with the ADE to maintain data security according to their specifications.

CAI 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, CAI runs automated functional tests of our TDS every morning, and logs from these runs are available for at least one week from the time of the run.

CAI psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior and provide a forensics analysis report. The forensics analysis report is described more completely in Section 5.6 on data forensics.

5.5 TEST SECURITY

Maintaining a secure test environment is critical to ensuring that scores represent what students know and can do. Because AzM2 was administered both as a PBT and a CBT assessment, test security procedures must guard against item exposure, cheating on the part of TAs or students, or other security problems for both testing modes.

The test security procedures involve the following:

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

TAs are trained on test security procedures, and both test security policies and procedures are clearly presented with the *AzM2 Test Administration Directions.*[39]

5.5.1 SECURITY OF TEST MATERIALS

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

The security of all test materials must be maintained before, during, and after test administration. Under no circumstances are students permitted to assist in preparing secure materials before testing or in organizing and returning materials after testing. After any administration, initial or make-up, secure materials (e.g., test booklets, test tickets, used scratch paper)

[39] Standard 6.7: Test users are responsible for protecting the security of test materials at all times.
Standard 7.9: If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.

are required to be returned immediately to the STC and placed in locked storage. Secure materials are never to be left unsecured and are not to remain in classrooms or be taken off the school's campus overnight. Secure materials are never to be destroyed (e.g., shredded, thrown in the trash), except for soiled documents. In addition, any monitoring software that would allow test content on student workstations to be viewed or recorded on another computer or device during testing needs to be turned off.

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

- capture images of any part of the test via any electronic device;
- duplicate in any way any part of the test;
- examine, read, or review the content of any portion of the test;
- disclose or allow to be disclosed the content of any portion of the test before, during, or after test administration;
- discuss any AzM2 test item before, during, or after test administration;
- allow students access to any test content before testing;
- provide any reference sheets to students during the mathematics test administration;
- allow students to share information during test administration;
- allow students to use scratch paper during the ELA reading test;
- read any parts of the test to students except as indicated in the Test Administration Directions or as part of an accommodation;
- influence students' responses by making any kind of gestures (for example, pointing to items, holding up fingers to signify item numbers or answer options) while students are taking the test;
- instruct students to go back and reread/redo responses after they have finished their test because this instruction may only be given before the students take the test;
- review students' responses;
- read or review students' scratch paper; or
- participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures.

Additional security violations for PBT include:

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

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

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

To access the computer-based AzM2 tests, a secure Internet browser is required. The Secure Browser provides a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to the desktop (Internet, email, and other files or programs installed on school machines). The Secure Browser did not display the IP address or other URL for the site. Users could not access other applications from within the 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 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 to exit the Secure Browser. The browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. Refer to the *Test Administrator User Guide* for further details.

Throughout the testing window, TAs were to report any test incidents (e.g., disruptive students, loss of Internet connectivity) to the STC 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. STCs notified district test coordinators of any test irregularities that were reported. DTCs were responsible for submitting requests for test invalidations to the ADE via CAI’s TIDE. The ADE made the final decision on whether to approve the requested test invalidation. DTCs 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 assessment administration can therefore cast doubt on the validity of the inferences based on those test scores. Multiple steps are taken to ensure that tests are administered properly, including providing clear test administration policies, effective TA training, and effective tools to identify possible irregularities in test administrations. With the introduction of remote test administration proctoring, the development and implementation of clear and precise training for TAs will be even more important. And whether tests are proctored locally in schools or remotely, monitoring test administrations for unusual activity will continue to be important.

For all online test administrations, quality assurance (QA) reports are generated during and after the testing windows. Many of these reports are geared toward ensuring the quality of test administrations, including item analysis reports which are used to ensure that items are performing as intended, blueprint match reports which ensure that the adaptive algorithm is performing as configured through simulations, and item exposure reports, which likewise indicate whether pool usage is consistent with the configuration of the adaptive algorithm during simulations. In addition, there are a suite of QA reports that are designed to assist in detecting irregularities in test administrations and which may indicate possible instances of cheating. These QA reports contain both individual-level and aggregate-level results. By aggregating unusual responses, we flag possible group-level testing anomalies. The aggregate-level categories include session, test administrator, and school.

Evidence evaluated includes

- changes in student performance across administrations (after the first operational administration);
- test-taking times;
- item response patterns using the person-fit index; and
- item response changes.

All analyses are performed on the completed tests at the student level and summarized for each aggregate unit, including testing session, TA, and school. The flagging criteria used for these analyses are configurable and can be changed by the user.

5.6.1 SMALL UNIT (SMALL GROUP) FLAGGING

For each aggregate unit, small groups are identified based on the number of tests included in the unit's analysis. Thus, a small unit identified in one analysis may not be a small unit in another analysis. For example, the number of tests used in the regression analysis could be lower than other analyses because the regression analysis is based on the merged students between two test administrations. The default small unit size is 5 or fewer students. For all analyses, the small unit is flagged if the percentage of flagged students is greater than 50% of students in the unit. The criteria of the small unit size and the flagging criteria for the analyses are configurable.

5.6.2 AGGREGATE UNIT FLAGGING (AGGREGATE UNIT SIZE > SMALL UNIT SIZE)

For the aggregate units with the unit size greater than the small unit size, the flagging criteria referenced to the state average and the standard deviation (SD) are computed in average and SD. The user can select which mean and SD to use, either weighted or unweighted, in the setting.

Aggregate units will receive a flag if their test-taking time is greater than 3 or smaller than -3 SDs of the state average. The number of standard deviations from the state mean used for flagging is configurable. The average and SD are computed based on the aggregate unit means, excluding the small units. For each aggregate unit, the state mean and SD are computed based on the aggregate level of analysis, such that the state mean and SD for evaluating student records is the simple average of all student records, while the state mean and standard deviation for sessions is the mean of all session means, the mean and SD for TAs is based on the mean of all TA means, and so on.

5.6.3 CHANGES IN STUDENT PERFORMANCE

When multiple testing occasions are available, the forensic analysis report predicts the expected level of achievement on the current test based on students' performance on the previous test administration. Fluctuation in individual student records does occur. For example, a student may have been ill during the previous test administration, causing the student to underperform, so that their current normal performance appears as an unusually large gain. However, such fluctuations are relatively rare.

Changes in students' test scores are examined between test administrations using a regression model. For multiple opportunity assessments that allow within-year comparisons, the most recent opportunity is regressed on previous performance (second most recent score), controlling for the number of days between two scores, to identify performance gains or losses that are substantially greater than might reasonably be expected.

For between-year comparisons, the scores between the current year and the previous year are evaluated. The most recent opportunity score in the current year (e.g., grade 4) will be regressed on the most recent score in the previous year performance (e.g., grade 3). Note that between-year comparisons are not available for the lowest grade tested (typically grade 3).

The score combinations in a regression analysis for within and between years are presented in Exhibit 5.6.3.1.

Exhibit 5.6.3.1 Score Comparisons Within and Between Years

Within-Year Comparison		**Between-Year Comparison**	
Y_{ti}	$Y_{(t-1)i}$	Y_{ti}	$Y_{(t-1)i}$
Most recent score	Send most recent score	Most recent score in current year	Most recent score in previous year
Opp1	→	Opp1	Opp1, Opp2, or Opp3
Opp2	Opp1	Opp2	Opp1, Opp2, or Opp3
Opp3	Opp2	Opp3	Opp1, Opp2, or Opp3

Within year: score comparison between opportunities
$Y_{ti} = a_0 + \beta_1 Y_{(t-1)i} + \beta_2 M_i + e_{ti}$
Y_{ti}: most recent opportunity score in current year for student i
$Y_{(t-1)i}$: second most recent opportunity score in current year for student i
M_i: difference in test end days between Y_t and Y_{t-1} for student i
e_{ti}: residual

Between year: score comparison
$Y_{ti} = a_0 + \beta_1 Y_{(t-1)i} + \beta_2 M_i + e_{ti}$
Y_{ti}: most recent score in current year for student i
$Y_{(t-1)i}$: most recent score in past year for student i
M_i: difference in test end days between Y_t and Y_{t-1} for student i
e_{tj}: residual

5.6.3.1 STUDENT-LEVEL FLAGGING

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as the observed score minus the predicted score in the regression model. 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 or less than -3.

The computation of the studentized t residuals is as follows:

Consider a simple regression model $Y = X\beta + e$.

The residuals can be expressed as $e = Y - \hat{Y} = Y - HY = (1 - H)Y$,

where $H = X(X'X)^{-1}X'$, called the hat matrix.

For linear models, the variance of the residual e_i for student i is, $Var(e_i) = \sigma^2(1 - h_{ii})$, and an estimate of the standard deviation (SD) of the residual is $SD(e_i) = s\sqrt{1 - h_{ii}}$.

The residuals can be standardized to better detect unusual observations. The ratio of the residual to its standard error, called *standardized residual*, is $e_{si} = \frac{e_i}{s\sqrt{1-h_{ii}}}$.

If the residual is standardized with an independent estimate of σ^2 , the result has a student's t distribution if the data satisfy the normality assumption. If we estimate σ^2 by s_i^2, the estimate of σ^2 obtained after deleting the i^{th} observation, the result is a studentized residual. Studentized t residuals can be computed as, $t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$, where i = student i, $s_{(i)}$ is the estimate of s after deleting the i^{th} observation.

5.6.3.1 AGGREGATE-LEVEL FLAGGING

The number of students with a large score gain or loss is aggregated for a testing session, TA, and school. Unusual changes in an aggregate performance between administrations within and/or between years is flagged based on either (1) t value of the average residual in an aggregate unit, or (2) the state average and the SD of residuals.

The size of the *t* value is determined both by the magnitude of the mean residual and the sample size in an aggregate unit, so that the same magnitude may be significant in larger groups, but not significant in smaller groups.

***t* Statistics**

For each aggregate unit, a critical t value is computed and flagged when $|t|$ is greater than 3,

$$t = \frac{\sum_{i=1}^{n} \hat{e}_i \,/n}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^{n} \sigma^2(1-h_{ii})}{n^2}}},$$

where s is the SD of residuals in an aggregate unit; n is the number of students in an aggregate unit (e.g., testing session, TA, school); σ^2 is the MSE from the regression; and $\hat{e}_i$ is the residual for the i^{th} student.

The variance of average residuals in the denominator is estimated in two components, conditioning on true residual e_i, $var\big(E(\hat{e}_i|e_i)\big) = s^2$ and $E\big(var(\hat{e}_i|e_i)\big) = \sigma^2(1-h_{ii})$. Following the law of total variance (Billingsley, 1995, p. 456),

$$var(\hat{e}_i) = var\big(E(\hat{e}_i|e_i)\big) + E\big(var(\hat{e}_i|e_i)\big) = s^2 + \sigma^2(1-h_{ii}), \text{ hence,}$$

$$var\left(\frac{\sum_{i=1}^{n} \hat{e}_i}{n}\right) = \frac{\sum_{i=1}^{n}(s^2+\sigma^2(1-h_{ii}))}{n^2} = \frac{s^2}{n} + \frac{\sum_{i=1}^{n}(\sigma^2(1-h_{ii}))}{n^2}.$$

State Mean and Standard Deviation

For each aggregate unit, aggregate unit means are also evaluated with respect to their deviation from the state mean. The state mean and standard deviation of residuals can be computed based on the average and SD. The user can select which mean and SD to use, either weighted or unweighted, in the setting. The state average and SD of residuals are computed based on the aggregate unit means such that the state mean and SD for sessions is the mean of all session means, the state mean and SD for TAs is based on the mean of all TA means, and so on. Small units are excluded in computing the unweighted state mean and SD of the aggregate means. An aggregate unit will be flagged if aggregate unit means of residuals is greater than 3 or smaller than -3 SDs of the state mean.

5.6.4 TEST-TAKING TIME

In the online environment, item response time is captured as the item page time (the length of time that each item page is presented) in milliseconds. Discrete items appear one item at a time on the screen, whereas stimulus-based items appear on the screen together. The page time is the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items. For each student, the total time taken to complete the test is computed by summing up the page time for all discrete items and stimulus-based items.

An example of unusual test-taking time would be a test record for a student who scores very well on the test, but with test-taking times on average far less than that required of students statewide. Such a pattern of short test-taking times and high scores might be expected if, for example, students already know or have been provided the answers to the questions. Conversely, if a TA helps students by "coaching" them to change their responses while taking the test, or leaves a test session open to manipulate student responses, the testing time could be much longer than expected.

The average and the SD of test-taking time are computed across all students for each test administration. Students receive a flag if their test-taking time is greater than 3 or smaller than -3 SDs of the state average. For aggregate level, group means are evaluated with respect to their deviation from the state mean testing time. The state average and SD are computed based on the aggregate unit test-taking time means, excluding the small units.

5.6.5 INCONSISTENT ITEM RESPONSE PATTERN (PERSON FIT)

In item response theory (IRT) models, person-fit indices are used to identify students whose patterns of responses to items are improbable given an IRT model. If a test has psychometric integrity, minimal irregularity will be evidenced in the item responses of the individual who responds to the items fairly and honestly.

In the IRT models used to score students' tests, the expectation that a student will respond correctly to an item depends both on the student's ability level and the difficulty of the item. Thus, high-ability students will have a higher probability of responding correctly to all items, but especially so as item difficulty increases. Sometimes, however, low-ability students answer difficult items correctly, perhaps through guessing. And sometimes, high-ability students respond incorrectly to easy items, perhaps through lack of attention. Generally, however, students' responses to test items are consistent with the scoring model.

For example, if a student is coached during a test administration or copies other students' responses, the student may provide correct responses to items at a higher probability than would be expected by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student.

The person-fit index is based on all item responses. An unlikely response to a few test items may not result in a flagged person-fit index. And 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 indicators of testing irregularities to determine whether cheating may be suspected.

Following Drasgow, Levine, and Williams (1985), the person-fit index l_z, is computed as

$$l_z(\theta) = \frac{l(\theta) - E\big(l(\theta)\big)}{Var\big(l(\theta)\big)},$$

where $l(\theta)$ is the log-likelihood of a vector of observed item scores for a given ability $l(\theta)$ with expected value $E\big(l(\theta)\big)$ and variance $Var\big(l(\theta)\big)$. The asymptotic distribution of l_z is a standard normal distribution (i.e., with an increasing number of administered items, i).

The asymptotic standard normal distribution of l_z only holds when the true person ability is known. In practice, the person ability is estimated from the same data that is used to compute l_z. The variance of l_z, can be considerably smaller than 1 when the true ability θ is replaced by its estimate $\hat{\theta}$ (Snijders, 2001). To remediate this, Snijders (2001) derived a family of asymptotically normal person fit statistics when $\hat{\theta}$ is used. When $\hat{\theta}$ is the maximum likelihood estimator, the fit statistic l_z^* is computed as l_z but with a corrected variance in the denominator. An alternative derivation for l_z^* is given by Lin, Jiang, & Rijmen (2021). The alternative method is presented here because it applies also to IRT models that are not unidimensional (e.g., the Rasch testlet model of Wang & Wilson, 2005).

The method of Lin et al. (2021) is based on a first-order Taylor series expansion of $Var\left(l(\hat{\theta}\,)\right)$, the denominator of l_z evaluated at the maximum likelihood estimate $\hat{\theta}$. A generalized person fit index can then be obtained as

$$l_{zg}^*(\hat{\theta}) = \frac{l(\hat{\theta}) - E\left(l(\hat{\theta})\right)}{\sqrt{Var\left(l(\hat{\theta})\right) - \frac{\left(h'(\hat{\theta})\right)^2}{I(\hat{\theta})}}}$$

where $h'(\hat{\theta}) = -\frac{dE\left(l(\hat{\theta})\right)}{d\hat{\theta}}$, which is the negative of the first derivative of the expected loglikelihood with respect to $\hat{\theta}$, and $I(\hat{\theta})$ is the test information evaluated at $\hat{\theta}$. The detailed calculations are given separately for unidimensional IRT models and the Rasch testlet model (Wang & Wilson, 2005).

5.6.5.1 UNIDIMENSIONAL MODELS

In the case of unidimensional IRT models, l_{zg}^* is equivalent to the l_z^* person fit index of Snijders (2001). We consider the general case where the test includes both binary items, modeled with a three-parameter logistic (3PL) model, and polytomous items, modeled with the generalized partial credit (GPC) model.

For a binary item *j*, define $p_j(\theta) = \Pr\left(Y_j = 1|\theta\right) = c_j + \frac{1-c_j}{1+Exp\left(-Da_j(\theta-b_j)\right)}$ and $q_j(\theta) = \Pr\left(Y_j = 0|\theta\right) = 1 - p_j(\theta)$. For a polytomous item *j* with possible scores *m* of $0,1,\cdots,M_j$, define $p_{jm}(\theta) = \Pr\left(Y_j = m|\theta\right) = \frac{Exp(\sum_{h=1}^{m} Da_j(\theta-b_{jh}))}{1+\sum_{k=1}^{M_j} Exp(\sum_{h=1}^{k} Da_j(\theta-b_{jh}))}$ with $p_{j0}(\theta) = \frac{1}{1+\sum_{k=1}^{M_j} Exp(\sum_{h=1}^{k} Da_j(\theta-b_{jh}))}$ and *D*=1.7; and define $p'_{jm}(\theta) = Da_j p_{jm}(\theta)\left[m - \sum_{k=1}^{M_j} kp_{jk}(\theta)\right]$. We then have, for a vector of observed item scores $(y_1,\ldots, y_j, \ldots, y_J)$

$$l(\hat{\theta}) = \sum_{j\in \mathrm{3PL}} \left(\log\left(\Pr\left(Y_j = y_j|\hat{\theta}\right)\right)\right) + \sum_{j\in \mathrm{GPC}} \left(\log\left(\Pr\left(Y_j = y_j|\hat{\theta}\right)\right)\right)$$

$$E\left(l(\hat{\theta})\right) = \sum_{j\in 3PL} \left(p_j(\hat{\theta})\log p_j(\hat{\theta}) + q_j(\hat{\theta})\log q_j(\hat{\theta})\right) + \sum_{j\in GPC}\sum_{m=0}^{M_j} p_{jm}(\hat{\theta})\log\left(p_{jm}(\hat{\theta})\right)$$

$$Var\left(l(\hat{\theta})\right) = \sum_{j\in 3PL} p_j(\hat{\theta})q_j(\hat{\theta})\left[\log\frac{p_j(\hat{\theta})}{q_j(\hat{\theta})}\right]^2 + \sum_{j\,\in GPC}\sum_{m=0}^{M_j}\left\{\left[\log\left(p_{jm}(\hat{\theta})\right) - \sum_{h=0}^{M_j}\left(p_{jh}(\hat{\theta})\log\left(p_{jh}(\hat{\theta})\right)\right)\right]^2 p_{jm}(\hat{\theta})\right\}$$

$$h'(\hat{\theta}) = -\sum_{j\in 3PL}\left(Da_j\frac{p_j(\hat{\theta})-c_j}{1-c_j}q_j(\hat{\theta})\log\left(\frac{p_j(\hat{\theta})}{q_j(\hat{\theta})}\right)\right) - \sum_{j\in GPC}\sum_{m=0}^{M_j}\{p'_{jm}(\hat{\theta})[\log p_{jm}(\hat{\theta})+1]\}$$

and information

$$I(\hat{\theta}) = \sum_{j\in\, 3PL} D^2a_j^2\frac{q_j(\hat{\theta})}{p_j(\hat{\theta})}\left(\frac{p_j(\hat{\theta})-c_j}{1-c_j}\right)^2 + \sum_{j\in\, GPC} D^2a_j^2\left\{\sum_{m=1}^{M_j} m^2p_{jm}(\hat{\theta}) - \left[\sum_{m=1}^{M_j} mp_{jm}(\hat{\theta})\right]^2\right\}$$

***t* Statistics**

Aggregate units are flagged with *t* smaller than -3, where

$$t = \frac{Average\ l_{zg}^*\ values}{\sqrt{s^2/n}},$$

where s^2 is the variance of l_{zg}^*values in the aggregate unit, and n is the number of students in the aggregate unit.

State Mean and Standard Deviation

Because the size of the t-statistic is determined both by the magnitude of the average l_{zg}^* values and the sample size, the same magnitude may be significant in larger groups, but not significant in smaller groups. Group means of l_{zg}^* values are evaluated with respect to their deviation from the state mean of l_{zg}^* values.

5.6.6 ITEM RESPONSE CHANGE

Students are allowed to revisit items as many times as they wish within a session and may even mark items to be revisited before completing the session. However, excessively high rates of response changes, especially those resulting in high rates of item score increases (i.e., response changes from wrong to right), may indicate irregularities in test administration. TAs could, for example, review students' responses and either coach them to modify their responses or keep the session active and change responses themselves.

To identify irregular patterns of response change, we examine the item score for the final response to each item and the penultimate response if one exists, and then count the number of instances in which the item score increases. Students with positive item score changes greater than 3 standard deviations (SDs) above the state mean are flagged, although the flagging value is configurable.

At the aggregate level, group means are evaluated with respect to their deviation from the state mean. Group means greater than 3 SDs of the state mean are flagged, although the flagging value is configurable. The summary of default flag indices used in forensic analysis is presented in Exhibit 5.6.6.1.

Exhibit 5.6.6.1 Summary of Default Flag Indices Used in Forensic Analysis

Testing Irregularity Index	Individual Test Flag	Aggregate Unit Flag
Within-Year: Changes in Test Scores	Flag if studentized ***t*** residual > 3 or studentized ***t*** residual < -3	Flag if aggregate unit average > state mean + 3SD or aggregate unit average < state mean – 3SD
Between-Years: Changes in Test Scores	Flag if studentized ***t*** residual > 3 or studentized ***t*** residual < -3	Flag if aggregate unit average > state mean + 3SD or aggregate unit average < state mean – 3SD
Test-Taking Time	Flag if total test time < state average – 3 SD or total test time > state average + 3 SD	
Person-Fit Index	Flag if lz value < -3.0	Aggregate unit average < state mean – 3SD
Item Response Change	Flag if positive score change > state average + 3 SD	
Small Unit Flag		

Note. Flagging criteria in the table are default values and can be configurable by a user.

6 REPORTING AND INTERPRETING AZM2 SCORES

A set of score reports that summarizes student performance in each grade and content area is provided for each administration. Score reports provide data on the performance of individual students and on the aggregated performance of students at various levels—such as state, districts, schools, and teachers. The test data are based on all students who participated in the Arizona's Statewide Achievement Assessment (AzM2) assessment for the 2019–2021 school year.

The score reports include reliable and valid information describing student progress toward mastery of the state content standards. Arizona provides individual student score reports that are shipped to the student's district for delivery to families. These reports detail student performance on overall tests and subscores. In addition, Arizona offers detailed individual- and aggregate-level data to educators via CAI's Online Reporting System (ORS), which provides score data for each AzM2 test, both online and paper-pencil. The ORS allows users to compare score data between individual students and the school, district, or overall state, and provides information about performance on subscore categories.

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 AzM2, both paper-pencil and online, are designed with stakeholders in mind—such as teachers, parents, and students, who are not technical measurement experts—and ensure that test results are used in ways that lead to valid inferences about student achievement and contribute to student learning.[40] For example, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy guides the reader to compare like elements and avoid comparison of dissimilar elements.

Sample reports are available at *https://AzM2portal.org*. The upcoming sections provide additional guidance for interpreting results.

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

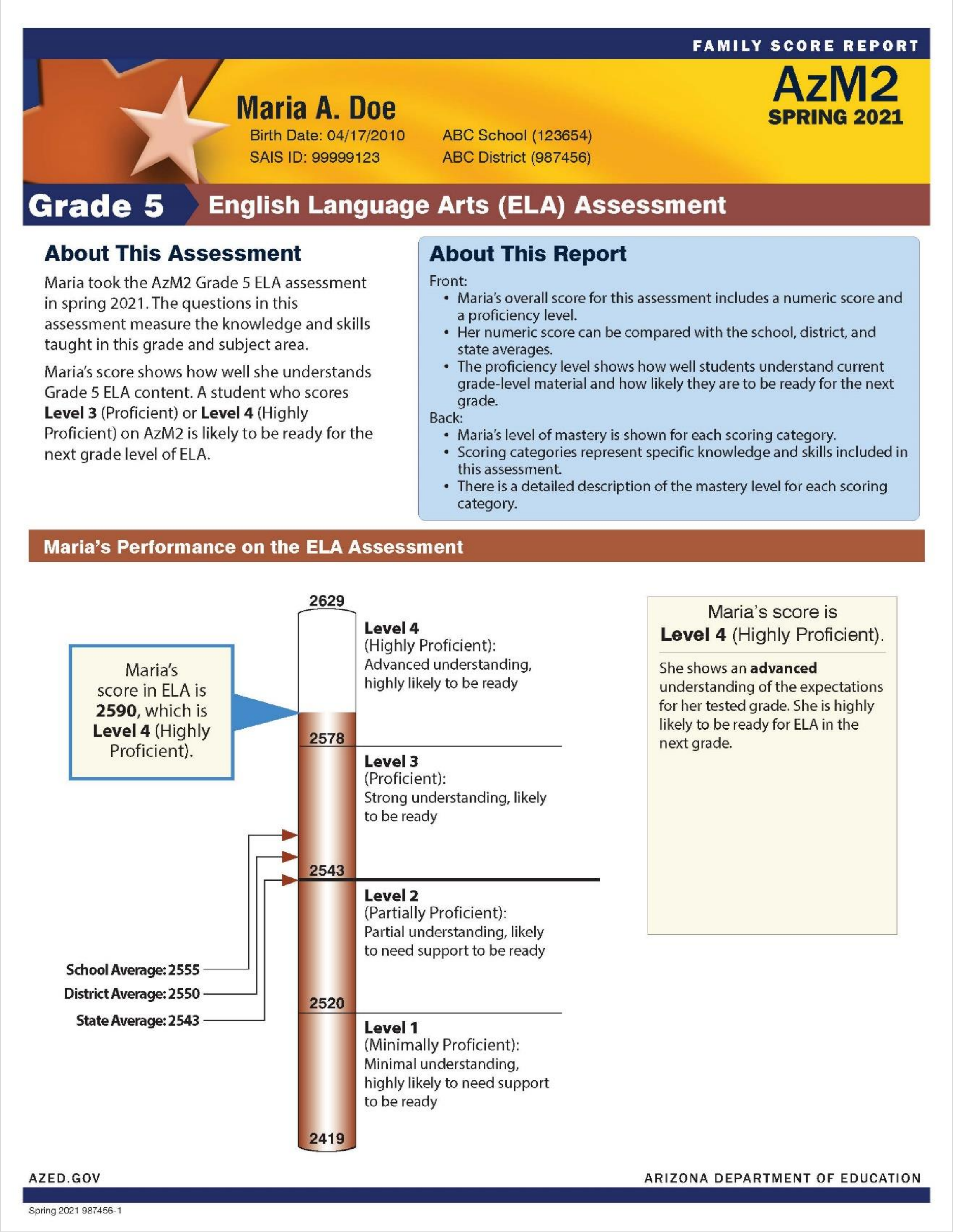
6.2 REPORTS PROVIDED

6.2.1 FAMILY REPORTS

Arizona provides full-color individual student reports (ISRs) to families of all AzM2 testers. Reports are designed to be useful to families, and include:

- full color to aid readers' interpretation of the data;
- scale scores and Performance-Level Descriptors;
- scoring category performance, including descriptions of what was assessed and what results mean for each scoring category to guide parents and students in their understanding of student scores:
    - A plus (+) symbol indicates that a student is performing above mastery in a particular scoring category.
    - A checkmark indicates that a student is performing at or near mastery within the scoring category.
    - The exclamation symbol indicates a student is performing below mastery in a scoring category.
- rubric scores for the writing portion of the English language arts (ELA) test, including descriptions of what those rubric scores mean; and
- school, district, and state average scores for comparative purposes.

In addition, beginning with the spring 2016 administration, the Arizona Department of Education (ADE) provided reports that included longitudinal data as presented at the bottom of the second page of the report. These data are designed to allow parents to track student achievement over time.



FAMILY SCORE REPORT

AzM2
SPRING 2021

Maria A. Doe

Birth Date: 04/17/2010
SAIS ID: 99999123

ABC School (123654)
ABC District (987456)

Grade 5 English Language Arts (ELA) Assessment

About This Assessment

Maria took the AzM2 Grade 5 ELA assessment in spring 2021. The questions in this assessment measure the knowledge and skills taught in this grade and subject area.

Maria's score shows how well she understands Grade 5 ELA content. A student who scores **Level 3** (Proficient) or **Level 4** (Highly Proficient) on AzM2 is likely to be ready for the next grade level of ELA.

About This Report

Front:

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

Back:

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

Maria's Performance on the ELA Assessment

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

2629

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

2578

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

2543

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

2520

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

2419

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

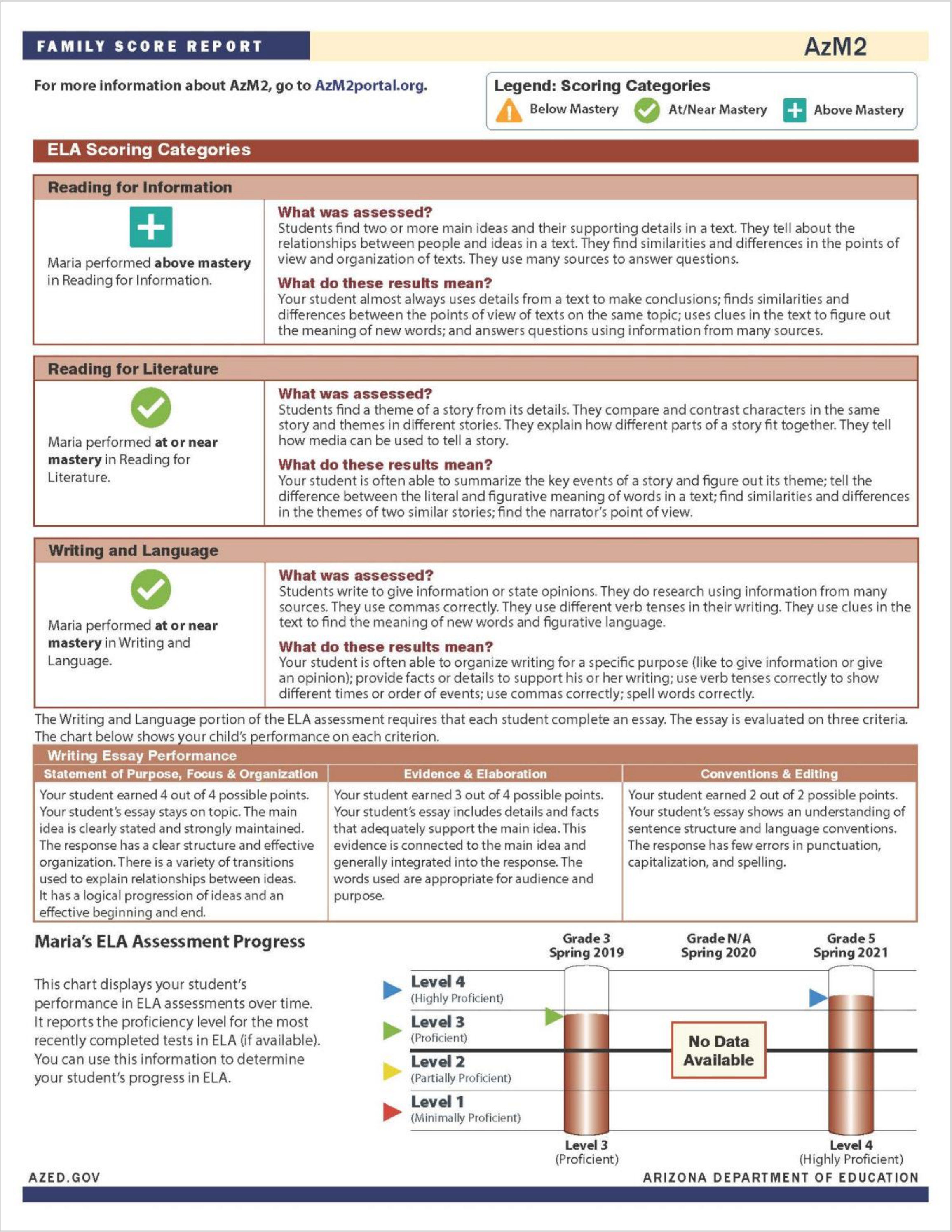
Maria's score is **Level 4** (Highly Proficient).

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

AZED.GOV

ARIZONA DEPARTMENT OF EDUCATION

Spring 2021 987456-1



FAMILY SCORE REPORT

AzM2

For more information about AzM2, go to AzM2portal.org.

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

ELA Scoring Categories

Reading for Information

+ Maria performed **above mastery** in Reading for Information.

What was assessed?
Students find two or more main ideas and their supporting details in a text. They tell about the relationships between people and ideas in a text. They find similarities and differences in the points of view and organization of texts. They use many sources to answer questions.

What do these results mean?
Your student almost always uses details from a text to make conclusions; finds similarities and differences between the points of view of texts on the same topic; uses clues in the text to figure out the meaning of new words; and answers questions using information from many sources.

Reading for Literature

✓ Maria performed **at or near mastery** in Reading for Literature.

What was assessed?
Students find a theme of a story from its details. They compare and contrast characters in the same story and themes in different stories. They explain how different parts of a story fit together. They tell how media can be used to tell a story.

What do these results mean?
Your student is often able to summarize the key events of a story and figure out its theme; tell the difference between the literal and figurative meaning of words in a text; find similarities and differences in the themes of two similar stories; find the narrator's point of view.

Writing and Language

✓ Maria performed **at or near mastery** in Writing and Language.

What was assessed?
Students write to give information or state opinions. They do research using information from many sources. They use commas correctly. They use different verb tenses in their writing. They use clues in the text to find the meaning of new words and figurative language.

What do these results mean?
Your student is often able to organize writing for a specific purpose (like to give information or give an opinion); provide facts or details to support his or her writing; use verb tenses correctly to show different times or order of events; use commas correctly; spell words correctly.

The Writing and Language portion of the ELA assessment requires that each student complete an essay. The essay is evaluated on three criteria. The chart below shows your child's performance on each criterion.

Writing Essay Performance

Statement of Purpose, Focus & Organization	Evidence & Elaboration	Conventions & Editing
Your student earned 4 out of 4 possible points. Your student's essay stays on topic. The main idea is clearly stated and strongly maintained. The response has a clear structure and effective organization. There is a variety of transitions used to explain relationships between ideas. It has a logical progression of ideas and an effective beginning and end.	Your student earned 3 out of 4 possible points. Your student's essay includes details and facts that adequately support the main idea. This evidence is connected to the main idea and generally integrated into the response. The words used are appropriate for audience and purpose.	Your student earned 2 out of 2 possible points. Your student's essay shows an understanding of sentence structure and language conventions. The response has few errors in punctuation, capitalization, and spelling.

Maria's ELA Assessment Progress

This chart displays your student's performance in ELA assessments over time. It reports the proficiency level for the most recently completed tests in ELA (if available). You can use this information to determine your student's progress in ELA.

AZED.GOV

ARIZONA DEPARTMENT OF EDUCATION

6.2.2 ONLINE REPORTING SYSTEM FOR EDUCATORS

AzM2 results are also reported using CAI's ORS, which is designed to support educators as they evaluate the needs of their students and reflect on their own curricula and practice. Navigation in the system mirrors the instructional decision-making process, meaning the user can intuitively navigate in any of the three dimensions inherent in the data, helping the user answer three kinds of questions:

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

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

The AzM2 online system produces the following online score reports: individual student reports, and aggregate reports at the teacher, school, district, and state level. The AzM2 online score reports are structured hierarchically. Upon selecting "Home" on the Welcome page, a user is taken to the Home Page Dashboard, which displays for all grades and content areas the number of students tested and the percentage of students passing by grade and content area. Users who have access to multiple districts or schools are first required to select a single district or school. Once an aggregate unit is selected in this instance, the summary table of student performance is displayed for the selected entity. For more detailed information for a subject and a grade, the user must select that subject and grade.

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

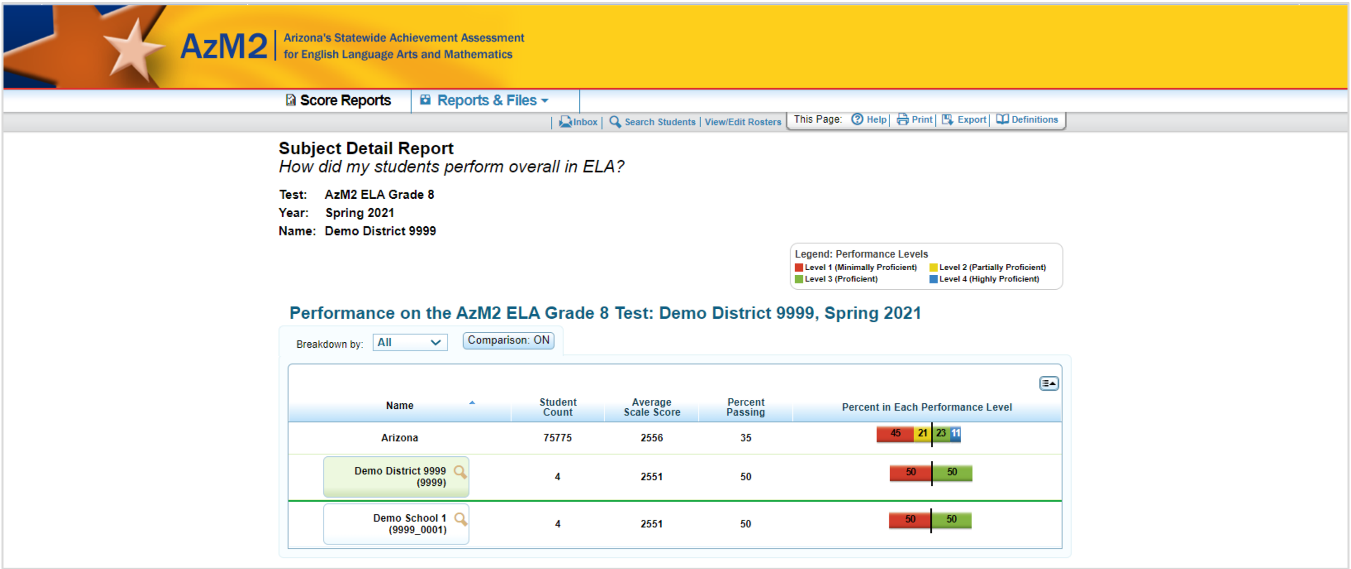
Exhibit 6.2.2.1 summarizes the types of online score reports available and the levels at which they can be viewed (e.g., student, roster, teacher, school, district).

Exhibit 6.2.2.1 AzM2 Online Score Report Summary

Type of Report Page	Level of Aggregation	Description
Home Page Dashboard	District, school, and teacher	Summary of performance and participation (Number Tested and Percentage Passing) across grades and subjects or course
Subject Detail	District	Average scale score, percentage passing, and percentage at each performance level for a district and each school within that district; ability to disaggregate data by subgroup
	School	Average scale score, percentage passing, and percentage at each performance level for a school and each teacher within that school; ability to disaggregate data by subgroup

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

6.2.2.1 SUBJECT DETAIL REPORTS


AzM2
Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics
Score Reports
Reports & Files
Inbox | Search Students | View/Edit Rosters
This Page: Help | Print | Export | Definitions

Subject Detail Report
How did my students perform overall in ELA?

Test: AzM2 ELA Grade 8
Year: Spring 2021
Name: Demo District 9999

Legend: Performance Levels
🟥 Level 1 (Minimally Proficient) 🟨 Level 2 (Partially Proficient)
🟩 Level 3 (Proficient) 🟦 Level 4 (Highly Proficient)

Performance on the AzM2 ELA Grade 8 Test: Demo District 9999, Spring 2021

Breakdown by: All
Comparison: ON

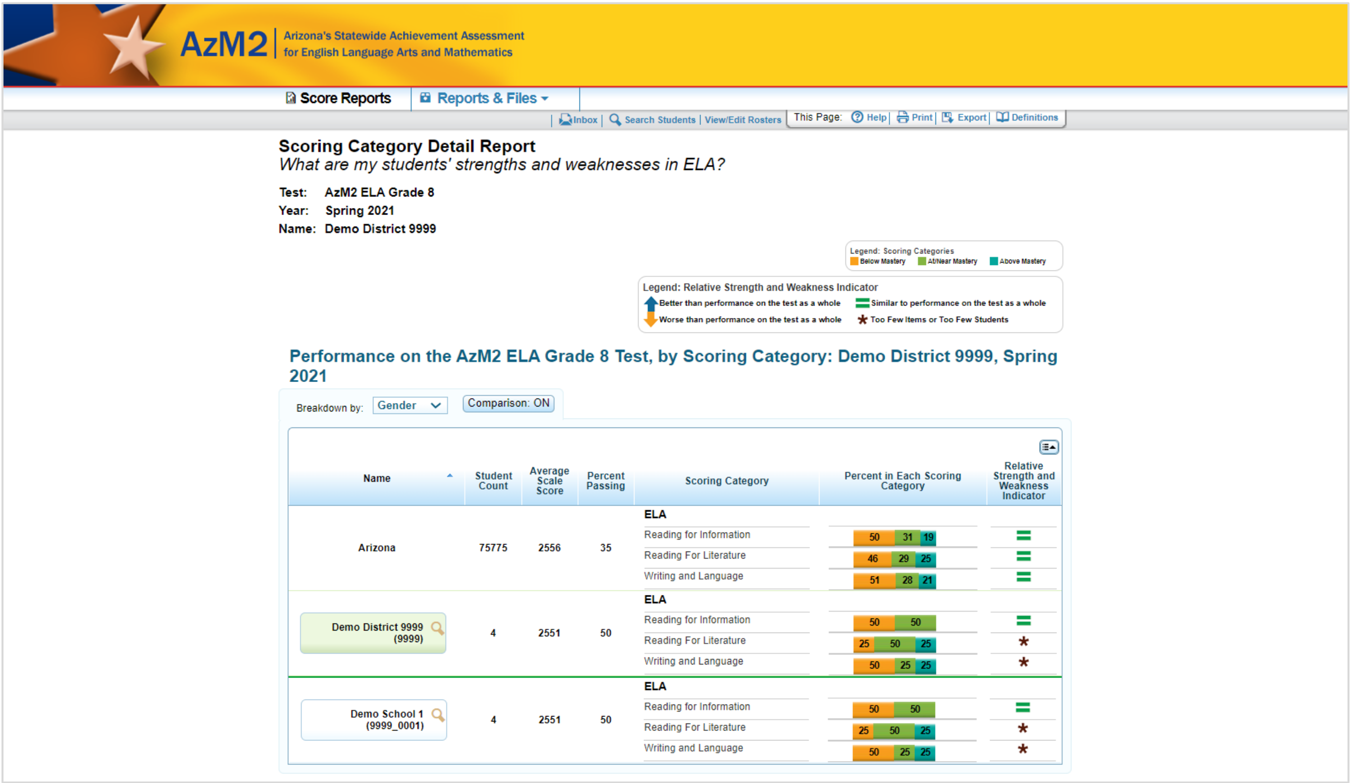
Name	Student Count	Average Scale Score	Percent Passing	Percent in Each Performance Level
Arizona	75775	2556	35	45 21 23 11
Demo District 9999 (9999)	4	2551	50	50 50
Demo School 1 (9999_0001)	4	2551	50	50 50

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

The Subject Detail Report page shows the following data:

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

6.2.2.2 SCORING CATEGORY DETAIL REPORTS


AzM2
Arizona's Statewide Achievement Assessment
for English Language Arts and Mathematics
Score Reports
Reports & Files
Inbox | Search Students | View/Edit Rosters
This Page: Help | Print | Export | Definitions

Scoring Category Detail Report
What are my students' strengths and weaknesses in ELA?

Test: AzM2 ELA Grade 8
Year: Spring 2021
Name: Demo District 9999

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

Performance on the AzM2 ELA Grade 8 Test, by Scoring Category: Demo District 9999, Spring 2021

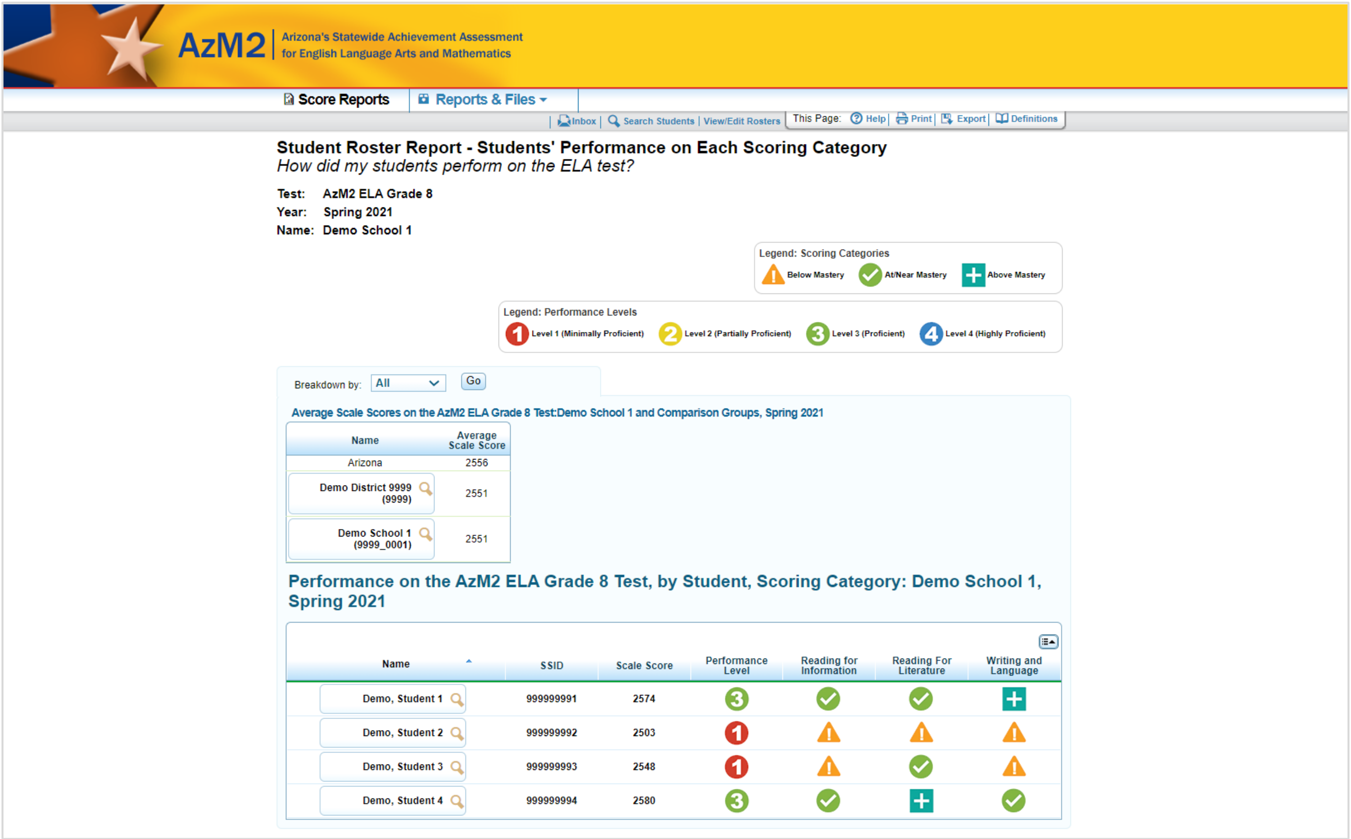
Breakdown by: Gender
Comparison: ON

Name	Student Count	Average Scale Score	Percent Passing	Scoring Category	Percent in Each Scoring Category	Relative Strength and Weakness Indicator
Arizona	75775	2556	35	**ELA**		
				Reading for Information	50 31 19	＝
				Reading For Literature	46 29 25	＝
				Writing and Language	51 28 21	＝
Demo District 9999 (9999)	4	2551	50	**ELA**		
				Reading for Information	50 50	＝
				Reading For Literature	25 50 25	*
				Writing and Language	50 25 25	*
Demo School 1 (9999_0001)	4	2551	50	**ELA**		
				Reading for Information	50 50	＝
				Reading For Literature	25 50 25	*
				Writing and Language	50 25 25	*

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

In addition to overall test scores, reporting category performance is reported as a strength and weakness indicator. The performance levels indicated on this report are relative to the test as a whole. Unlike performance levels provided at the subject level, these strengths and weaknesses do not imply proficiency. Instead, they show how the performance of a group of students is distributed across the scoring categories relative to their overall subject performance on a test. For example, a group of students may have performed very well in a subject but performed slightly lower in several scoring categories. Thus, the orange “down” sign for a scoring category does not imply a lack of proficiency. Instead, it simply communicates that these students’ performance on that scoring category was statistically lower than their performance across all other scoring categories put together. Although the students are doing well, an educator may want to focus instruction on these areas.

6.2.2.3 STUDENT ROSTER REPORTS



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

Test: AzM2 ELA Grade 8
Year: Spring 2021
Name: Demo School 1

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

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

Breakdown by: All Go

Average Scale Scores on the AzM2 ELA Grade 8 Test:Demo School 1 and Comparison Groups, Spring 2021

Name	Average Scale Score
Arizona	2556
Demo District 9999 (9999)	2551
Demo School 1 (9999_0001)	2551

Performance on the AzM2 ELA Grade 8 Test, by Student, Scoring Category: Demo School 1, Spring 2021

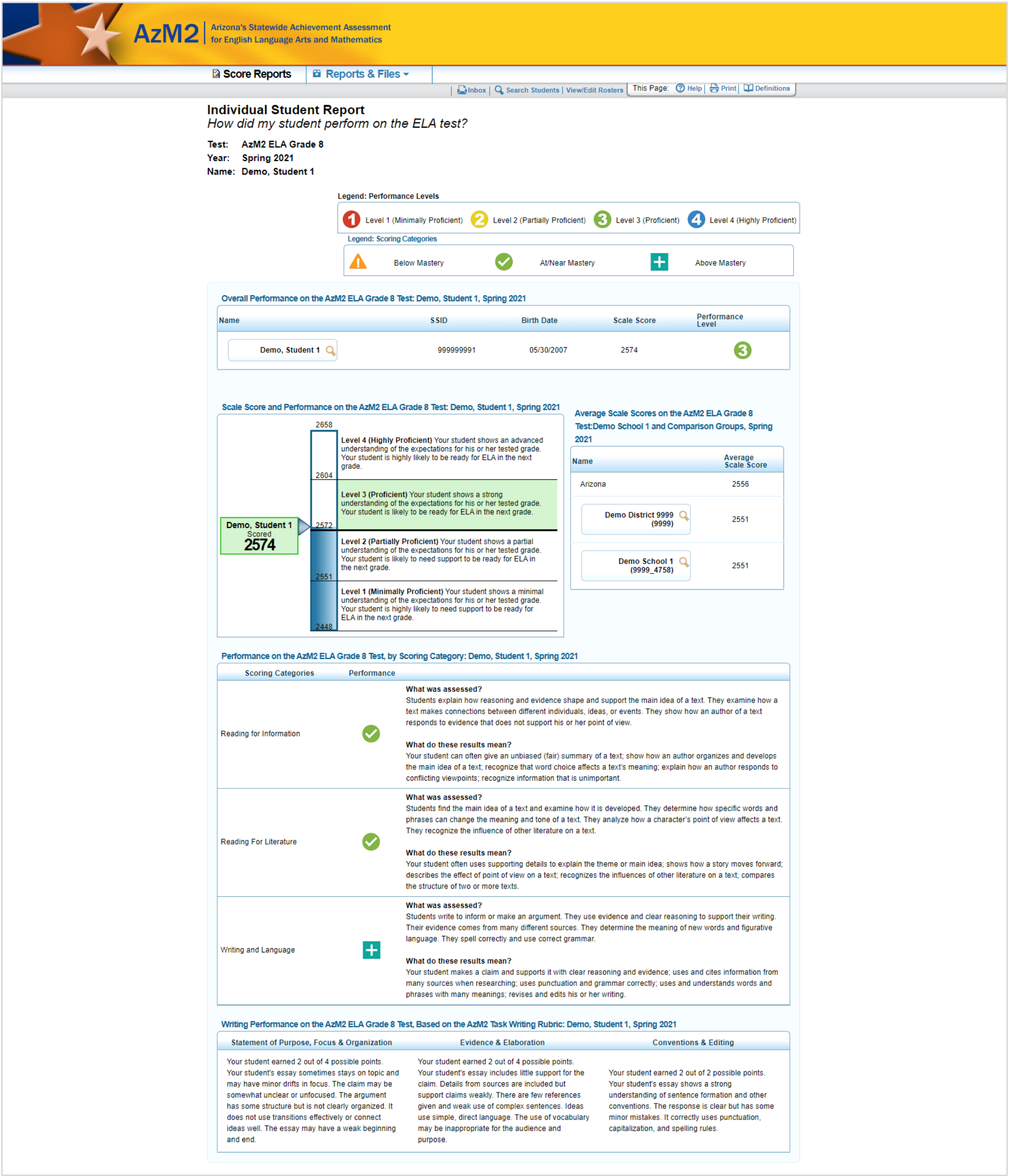
Name	SSID	Scale Score	Performance Level	Reading for Information	Reading For Literature	Writing and Language
Demo, Student 1	999999991	2574	③	✓	✓	➕
Demo, Student 2	999999992	2503	①	⚠	⚠	⚠
Demo, Student 3	999999993	2548	①	⚠	✓	⚠
Demo, Student 4	999999994	2580	③	✓	➕	✓

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

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

- **Scale Score.** This represents the score of each student who completed the test.
- **Performance level.** This represents levels of overall subject mastery with respect to the Arizona State Standards (4, representing Highly Proficient, to 1, representing Minimally Proficient).
- **Scoring Categories** This represents levels of scoring category mastery with respect to the Arizona State Standards, characterizing achievement at "above," "at or near," or "below" mastery on each scoring category.

6.2.2.4 INDIVIDUAL STUDENT REPORTS



Individual Student Report
How did my student perform on the ELA test?

Test: AzM2 ELA Grade 8
Year: Spring 2021
Name: Demo, Student 1

Legend: Performance Levels

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

Legend: Scoring Categories

⚠ Below Mastery ✓ At/Near Mastery + Above Mastery

Overall Performance on the AzM2 ELA Grade 8 Test: Demo, Student 1, Spring 2021

Name	SSID	Birth Date	Scale Score	Performance Level
Demo, Student 1	999999991	05/30/2007	2574	3

Scale Score and Performance on the AzM2 ELA Grade 8 Test: Demo, Student 1, Spring 2021

Demo, Student 1
Scored
2574

2658

Level 4 (Highly Proficient) Your student shows an advanced understanding of the expectations for his or her tested grade. Your student is highly likely to be ready for ELA in the next grade.

2604

Level 3 (Proficient) Your student shows a strong understanding of the expectations for his or her tested grade. Your student is likely to be ready for ELA in the next grade.

2572

Level 2 (Partially Proficient) Your student shows a partial understanding of the expectations for his or her tested grade. Your student is likely to need support to be ready for ELA in the next grade.

2551

Level 1 (Minimally Proficient) Your student shows a minimal understanding of the expectations for his or her tested grade. Your student is highly likely to need support to be ready for ELA in the next grade.

2448

Average Scale Scores on the AzM2 ELA Grade 8 Test:Demo School 1 and Comparison Groups, Spring 2021

Name	Average Scale Score
Arizona	2556
Demo District 9999 (9999)	2551
Demo School 1 (9999_4758)	2551

Performance on the AzM2 ELA Grade 8 Test, by Scoring Category: Demo, Student 1, Spring 2021

Scoring Categories	Performance	
Reading for Information	✓	**What was assessed?** Students explain how reasoning and evidence shape and support the main idea of a text. They examine how a text makes connections between different individuals, ideas, or events. They show how an author of a text responds to evidence that does not support his or her point of view. **What do these results mean?** Your student can often give an unbiased (fair) summary of a text; show how an author organizes and develops the main idea of a text; recognize that word choice affects a text's meaning; explain how an author responds to conflicting viewpoints; recognize information that is unimportant.
Reading For Literature	✓	**What was assessed?** Students find the main idea of a text and examine how it is developed. They determine how specific words and phrases can change the meaning and tone of a text. They analyze how a character's point of view affects a text. They recognize the influence of other literature on a text. **What do these results mean?** Your student often uses supporting details to explain the theme or main idea; shows how a story moves forward; describes the effect of point of view on a text; recognizes the influences of other literature on a text; compares the structure of two or more texts.
Writing and Language	+	**What was assessed?** Students write to inform or make an argument. They use evidence and clear reasoning to support their writing. Their evidence comes from many different sources. They determine the meaning of new words and figurative language. They spell correctly and use correct grammar. **What do these results mean?** Your student makes a claim and supports it with clear reasoning and evidence; uses and cites information from many sources when researching; uses punctuation and grammar correctly; uses and understands words and phrases with many meanings; revises and edits his or her writing.

Writing Performance on the AzM2 ELA Grade 8 Test, Based on the AzM2 Task Writing Rubric: Demo, Student 1, Spring 2021

Statement of Purpose, Focus & Organization	Evidence & Elaboration	Conventions & Editing
Your student earned 2 out of 4 possible points. Your student's essay sometimes stays on topic and may have minor drifts in focus. The claim may be somewhat unclear or unfocused. The argument has some structure but is not clearly organized. It does not use transitions effectively or connect ideas well. The essay may have a weak beginning and end.	Your student earned 2 out of 4 possible points. Your student's essay includes little support for the claim. Details from sources are included but support claims weakly. There are few references given and weak use of complex sentences. Ideas use simple, direct language. The use of vocabulary may be inappropriate for the audience and purpose.	Your student earned 2 out of 2 possible points. Your student's essay shows a strong understanding of sentence formation and other conventions. The response is clear but has some minor mistakes. It correctly uses punctuation, capitalization, and spelling rules.

Individual student reports (ISRs), which closely mirror the family reports, are also available through the ORS.

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 ORS and understanding paper family reports.[41] This section describes many of the measures presented in the paper and online score reports.

Performance levels represent levels of mastery with respect to the Arizona State Standards for a content-area assessment. Performance levels are labeled as Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance standards are the points on the achievement scale that differentiate performance levels. Three performance standards are used to classify students into one of the four performance levels. Performance standards were recommended by panels of Arizona educators following the first administration of AzM2 in 2015, and subsequently adopted by the Arizona State Board of Education. Panelists engaged in a rigorous, technically sound standard-setting process that is summarized in Section 7, Performance Standards, of this technical report and documented in detail in the 2015 standard-setting technical report, available from the ADE.

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

The student's performance in each content-area assessment is summarized in an overall test score referred to as a scale score. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale score is then used to determine how well students perform on each content-area assessment. Scale scores can be used to measure what 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 2,450 for grade 3 students in the 2017–2018 school year indicates the same level of achievement as an average scale score of 2,450 for grade 3 students in the 2018–2019 school year, even though the test may include a slightly different set of items.

As described in Section 8 on Scaling and Equating, for the ELA assessment, the scale score reported can range from 2,395 to 2,675. For the mathematics assessment, the scale score reported can range from 3,395 to 3,819. Overall scale scores for ELA and mathematics are mapped into four performance levels using three performance standards (i.e., cut scores). The AzM2 scale score ranges can be found in Exhibit 6.3.1.

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

Exhibit 6.3.1 AzM2 Scale Score Ranges

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
ELA				
Grade 3	2,395–2,496	2,497–2,508	2,509–2,540	2,541–2,605
Grade 4	2,400–2,509	2,510–2,522	2,523–2,558	2,559–2,610
Grade 5	2,419–2,519	2,520–2,542	2,543–2,577	2,578–2,629
Grade 6	2,431–2,531	2,532–2,552	2,553–2,596	2,597–2,641
Grade 7	2,438–2,542	2,543–2,560	2,561–2,599	2,600–2,648
Grade 8	2,448–2,550	2,551–2,571	2,572–2,603	2,604–2,658
Grade 10	2,458–2,566	2,567–2,580	2,581–2,605	2,606–2,668
Mathematics				
Grade 3	3,395–3,494	3,495–3,530	3,531–3,572	3,573–3,605
Grade 4	3,435–3,529	3,530–3,561	3,562–3,605	3,606–3,645
Grade 5	3,478–3,562	3,563–3,594	3,595–3,634	3,635–3,688
Grade 6	3,512–3,601	3,602–3,628	3,629–3,662	3,663–3,722
Grade 7	3,529–3,628	3,629–3,651	3,652–3,679	3,680–3,739
Grade 8	3,566–3,649	3,650–3,672	3,673–3,704	3,705–3,776
Grade 10	3,609–3,672	3,673–3,696	3,697–3,742	3,743–3,819

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

7 PERFORMANCE STANDARDS

In the summer of 2015, following the close of the first testing window, CAI convened panels of Arizona educators to recommend performance standards on each Arizona's Statewide Achievement Assessment (AzM2) assessment. Details of the panels, procedures, and outcomes are documented in the *Recommending AzM2 Performance Standards* technical report, which is available from the Arizona Department of Education (ADE).[42] 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 AzM2 is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Interpretation of the AzM2 test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the Arizona State Standards. The cut score establishing the Proficient level of performance is the most critical because it indicates that students are meeting grade-level expectations for achievement of the Arizona State Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standards for the AzM2 assessments are therefore central to the validity of test score interpretations.

Following the first operational administration of the AzM2 assessments in spring 2015, a standard-setting workshop was conducted to recommend to the Arizona State Board of Education a set of performance standards for reporting student achievement of the Arizona State Standards. The workshop consisted of a series of standardized and rigorous procedures that the Arizona educators serving as standard-setting panelists followed to recommend performance standards. The workshops employed the Bookmark procedure, a widely used method where standard-setting panelists used their expert knowledge of the Arizona State Standards and student achievement to map the Performance-Level Descriptors (PLDs) adopted by the Arizona State Board of Education to an ordered-item booklet (OIB) based on the first operational test form administered in spring 2015.

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant American College Testing (ACT) college-ready performance standard for the grade 11 English language arts (ELA) and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and geometry assessments. Panelists recommending performance standards for the grades 3–8 summative assessments were provided with the approximate location of relevant National Assessment of Educational Progress (NAEP) performance standards at grades 4 and 8, and the interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grades 3–8 and 11 assessments in ELA and mathematics to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous Arizona's Instrument to Measure Standards (AIMS) performance standards. Panelists were asked

[42] Standard 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
Standard 7.4: Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.

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

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

7.1.1 PERFORMANCE-LEVEL DESCRIPTORS

Student achievement on the AzM2 is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. 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 and the Arizona College and Career Readiness Standards. The AzM2 PLDs describe four levels of achievement:

1. Minimally Proficient
2. Partially Proficient
3. Proficient
4. Highly Proficient

Prior to convening the standard-setting workshops, CAI, in consultation with the ADE, drafted PLDs for each test that described the range of achievement encompassed by each performance level on the test. The PLDs were designed to be clear, concrete, and reflect Arizona's expectations for proficiency based on the Arizona State Standards. Following a cycle of revisions to the draft PLDs, the ADE invited Arizona educators to review PLDs for each of the assessments. Based on feedback from 166 educators, PLDs were further revised, and the resulting drafts were used by standard-setting panelists. ADE considered any need for clarification or revision that arose throughout the standard-setting process prior to publishing the final versions of the PLDs following the standard-setting workshop. AzM2 PLDs are available at www.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 (theta) and the percentage of students classified as meeting or exceeding each standard. Following the standard-setting workshop, panelist recommendations were submitted to the Arizona State Board of Education; the Board formally adopted the standards in August 2015.

Exhibit 7.2.1 Final Recommended Performance Standards for AzM2

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	0.60	39	1.80	5
5	-0.13	63	0.63	30	1.80	3
6	-0.12	61	0.58	34	2.03	4
7	-0.02	59	0.61	33	1.90	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	2.20	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 students classified at each performance level in the initial year of AzM2 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

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
		Mathematics		
3	27	31	27	15
4	29	29	32	10
5	29	31	27	13
6	38	30	21	11
7	48	22	18	13
8	43	24	20	13
Algebra I	45	23	23	9
Geometry	47	24	24	6
Algebra II	47	24	23	6

Exhibit 7.2.3 shows the percentage of students meeting the AzM2 proficient standard for each assessment in the base year of 2015 (meaning they are categorized as Proficient or Highly Proficient), and the approximate percentage of Arizona students that would be expected to meet the ACT college-ready standard, the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8, and the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2015 field test administration. As Exhibit 7.2.3 indicates, the performance standards recommended for AzM2 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 Percentages of Students Meeting AzM2 and Benchmark Proficient Standards

Grade/ Course	Percentage of Students Meeting Standard			
	AzM2 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

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

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

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

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (1PL; also known as the Rasch model) is used to calibrate AzM2 items that are scored either right or wrong, and takes the form

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

where b_i is the difficulty parameter for item i.

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

For items that have multiple, ordered-response categories (i.e., partial credit items), AzM2 items are calibrated using the Rasch family Masters' (1982) Partial Credit Model. Under Masters' Partial Credit Model, the probability of getting a score of x_i on item i given ability θ can be written as

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

with the constraint that $\sum_{k=0}^{0}(\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i. Item parameters for the assessments were calibrated following the spring administration in 2015, and vertical scales were established for reporting both English language arts (ELA) and mathematics. In addition, a series of linking studies were performed to allow the comparison of performance on the AzM2 to other state and national scales. A mode comparability study was also completed to examine possible effects of test administration mode. These studies were completed before establishing performance standards in summer 2015 and subsequent scoring and reporting of AzM2 results. AzM2 ELA is reported on a scale ranging from 2,395–2,675 across the grade-level and high school end-of-course (EOC) tests. AzM2 mathematics is reported on a scale ranging from 3,395–3,839 across grade-level and high school EOC (Algebra I, geometry, and Algebra II) tests.

8.1 ITEM RESPONSE THEORY PROCEDURES

The AzM2 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 AzM2 scales for scoring and reporting.[43] This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by Cambium Assessment, Inc. (CAI), the Arizona Department of Education (ADE), and HumRRO, which acts as a third-party quality assurance (QA) contractor.

Within AzM2, students can skip items in both the online and paper-pencil tests. While omitted items are scored as incorrect for purposes of ability estimation, all omitted responses are treated as not-administered for purposes of IRT analysis. All students who respond to at least one item within each test session are considered to have attempted a test. All attempted records are included in IRT analysis with the exclusion of students who had more than one record for the same test and records that are had been invalidated before scaling.

8.1.1 CALIBRATION OF AZM2 ITEM BANKS

Winsteps was used to estimate Rasch and Masters' Partial Credit Model item parameters for AzM2. Winsteps is publicly available software from Mesa Press. Winsteps employs a joint maximum likelihood approach toward estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters' (1982) Partial Credit Model, an extension of the one parameter Rasch model which allows for partial credit to be given on items, estimates the responses for polytomous items.

In spring 2015, operational items for each test were freely calibrated establishing the new AzM2 reference scales. Following the approval of final item parameter estimates for operational items, parameter estimates for the operational items were anchored to their new AzM2 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 AzM2 scale defined by the operational item parameters.

In spring 2021, pre-equated item parameters were used to score student test records for the ELA grades 3–8 and 10 assessments and for the mathematics grades 3–8 assessments. Post-equated algebra item parameters along with the pre-equated geometry item parameters were used to score student test records for the mathematics grade 10 assessment.

8.1.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

polytomously scored item if $m_i > 1$. Based on Masters' (1982) Partial Credit Model, the probability of getting a score of x_i on item i given ability θ can be written as

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

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

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

where b_i is the difficulty parameter for item i.

8.1.2.1 LIKELIHOOD FUNCTION

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

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

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

8.1.2.2 DERIVATIVES

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

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

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

The MLE of θ is found via the following iterative routine:

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

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

8.1.2.3 ESTIMATING ZERO AND PERFECT SCORES

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

8.2 ESTABLISHING A VERTICAL SCALE IN ELA AND MATHEMATICS

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade-level assessment (g) were embedded in the field-test slots of the assessment in the grade below (g − 1).[44] 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

Because the vertical scale essentially places each AzM2 assessment on the scale for the assessment in the grade above, we can best assure comparability of test scores between the grades by establishing the linkage using all available operational test items. Thus, to link the grade 4 assessments to the grade 5 scales, all operational items in the grade 5 assessment were made available for administration in the grade 4 embedded field-test (EFT) slots. The inclusion of all operational items in the vertical linking set ensures that the item set used to link to the target adjacent grade scale fully represents the measured construct in the target grade, allowing for valid inferences to be made with respect to student baseline performance for achievement in the subsequent grade level.

Because the AzM2 assessments of ELA in high school continue as EOC or grade-level measures of student achievement of the Arizona State Standards, each assessment can be linked to the grade above using all available operational items.

However, AzM2 assessments of high school mathematics are composed of a set of EOC tests that are not as consistently associated with grade-level instruction and which measure specific subsets of the content domain. For example, while mathematics coursework in high school follows a typical progression and it would therefore be possible to embed "grade 9" Algebra I EOC items in the grade 8 mathematics assessment, embed the "grade 10" geometry EOC items in the Algebra I EOC exam, and embed the "grade 11" Algebra II the geometry exam, the constructs measured across the four exams vary considerably and have implications for the interpretation of growth, or lack thereof, across assessments. For example, it is not clear what the expectation for growth should be in a vertical scale established by embedding geometry items in an Algebra I exam because geometry is not a focus of instruction in Algebra I courses. An alternative approach, and the one adopted by the ADE, was to link the grade 8 mathematics scale to both the Algebra I and geometry EOC scales because the grade 8 assessment includes items measuring both algebra and geometry. Because Algebra II builds on the knowledge and skills assessed in Algebra I, all Algebra II items were used to link the Algebra I assessments to the Algebra II scale.

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 IRT

[44] Standard 5.0: Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.
Standard 5.2: The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

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

8.2.2.1 FINAL LINKING SET

Exhibit 8.2.2.1.1 shows the number of items dropped and remaining in the final vertical linking set. To facilitate the development of a vertical scale that will be sensitive to student growth over time, we first evaluated the performance of vertical linking items between the grade levels in which they were administered to identify any items that were more difficult for students in the intended grade than they were for students in the lower grade. For mathematics, items that showed proportion correct scores lower in the intended grade than in the lower grade were dropped from the final vertical linking set. This resulted in dropping on average just over two items per linking set, with a maximum of six items dropped for the linkage between grade 6 and grade 7 mathematics assessments.

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

Exhibit 8.2.2.1.1 Number of Items Dropped and Remaining in the Final Vertical Linking Set

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

8.2.2.2 CHAIN LINKING

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

8.2.2.3 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 8.2.2.3.1 shows the vertical linking constants resulting from chain linking the within-grade scales and also from concurrently calibrating items from across grade levels. The linking constants are applied to their respective within-grade scale to place all item parameters on the grade 3 reference scale.

Exhibit 8.2.2.3.1 Vertical Linking Constants Resulting from Chain Linking Within-Grade Scales and Concurrent Calibration of Items Across Grades

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

Exhibit 8.2.2.3.2 shows the difference between linking constants between each of the grade levels assessed and can be examined more directly to assess the magnitude of gains across grade-level assessments.

Exhibit 8.2.2.3.2 Linking Constant Differences Between Each of the Grade-Level Scales

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

Relative gains are also represented graphically in Exhibit 8.2.2.3.3 and Exhibit 8.2.2.3.4 for ELA and mathematics, respectively, which plot the linking constants across grade-level assessments. As the linking constants indicate, for mathematics there is relatively large and steady growth across the grade-level and EOC assessments. For the ELA assessments, the cross-grade gains are more modest and tend to diminish in the higher grade levels.

Exhibit 8.2.2.3.3 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: ELA

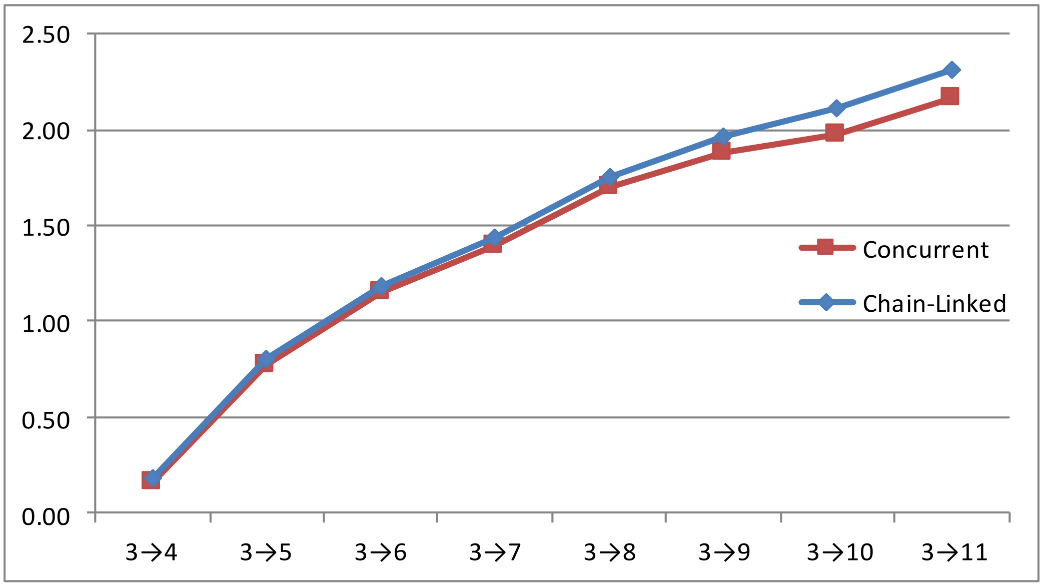
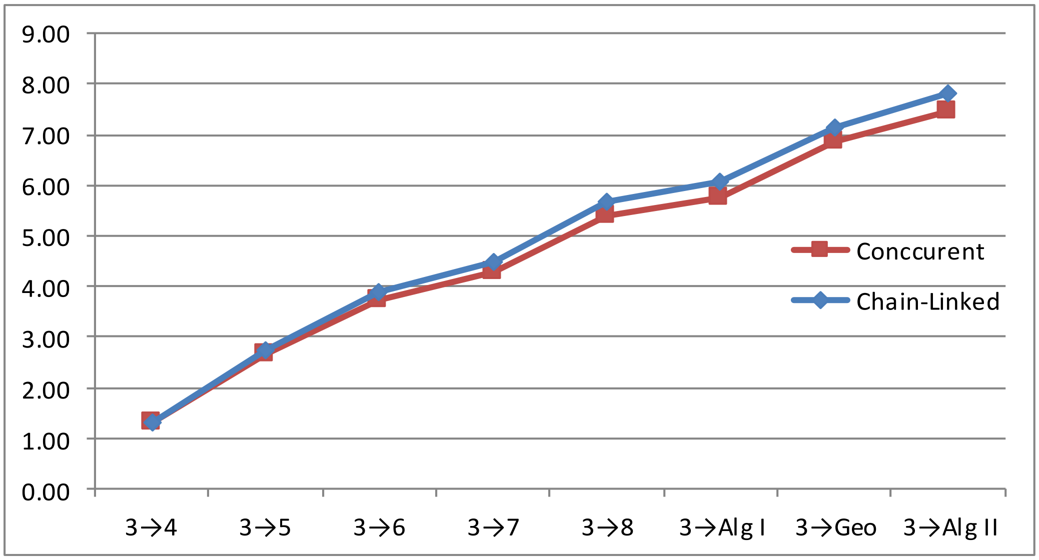


Exhibit 8.2.2.3.4 Vertical Linking Constants Estimated from Chain Linking and Concurrent Calibrations: Mathematics



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

The vertical linking constants also indicate much greater growth across grades and high school courses for mathematics than is observed for ELA. In mathematics, growth is on the order of about one standard deviation (SD) per year, except for grade 6 to grade 7, which showed just over a half SD gain. Similar one-half SD 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.

8.2.2.4 AZM2 2019 VERTICAL LINKING STUDY

It has been four years since the AzM2 vertical scales for mathematics and ELA were first established in 2015. As a part of an ongoing process in evaluating the stability of the vertical scales for AzM2, in spring 2019, the vertical linking study was repeated to evaluate results of the 2015 vertical linking study.

Both chain linking and concurrent calibration approaches were used to produce the 2019 vertical linking constants. The robustness of the vertical linking results between the chain linking and concurrent calibration methods was evaluated with respect to the convergence of the linking results across all grades per subject. Following the method used in 2015 to evaluate the performance of vertical linking items between the grade levels, the items showing higher proportion correct in the lower grade than in the grade above were removed from the linking sets. As expected, the 2019 linking constants produced by chain linking and concurrent calibration converged. The 2019 vertical linking constants resulting from chain linking and concurrent calibration in ELA and mathematics assessments are presented in Exhibit 8.2.2.4.1 and Exhibit 8.2.2.4.2.

Exhibit 8.2.2.4.1 Vertical Linking Constants Resulting from Chain Linking and Concurrent Calibration: ELA

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

Exhibit 8.2.2.4.2 Vertical Linking Constants Resulting from Chain Linking and Concurrent Calibration: Mathematics

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

Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain linking method preserves the within-grade measurement construct. For this reason, the vertical linking constants identified via chain linking were adopted as the AzM2 vertical scaling constants in 2015. Comparison of the chain linking results obtained in 2015 and 2019 is presented graphically in Exhibit 8.2.2.4.3 and Exhibit 8.2.2.4.4 for ELA and mathematics, respectively.

Exhibit 8.2.2.4.3 Comparison of 2015 and 2019 Vertical Linking Constants Estimated from Chain Linking Calibrations: ELA

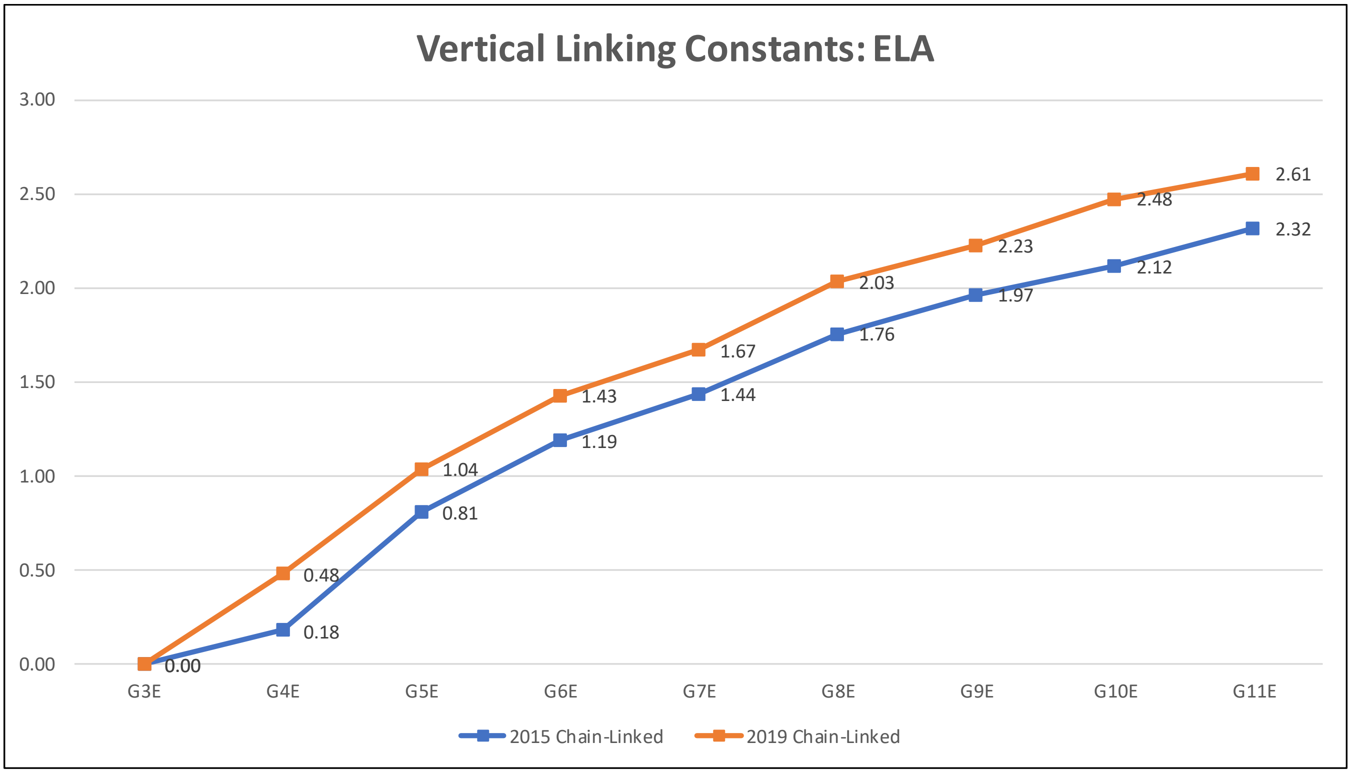
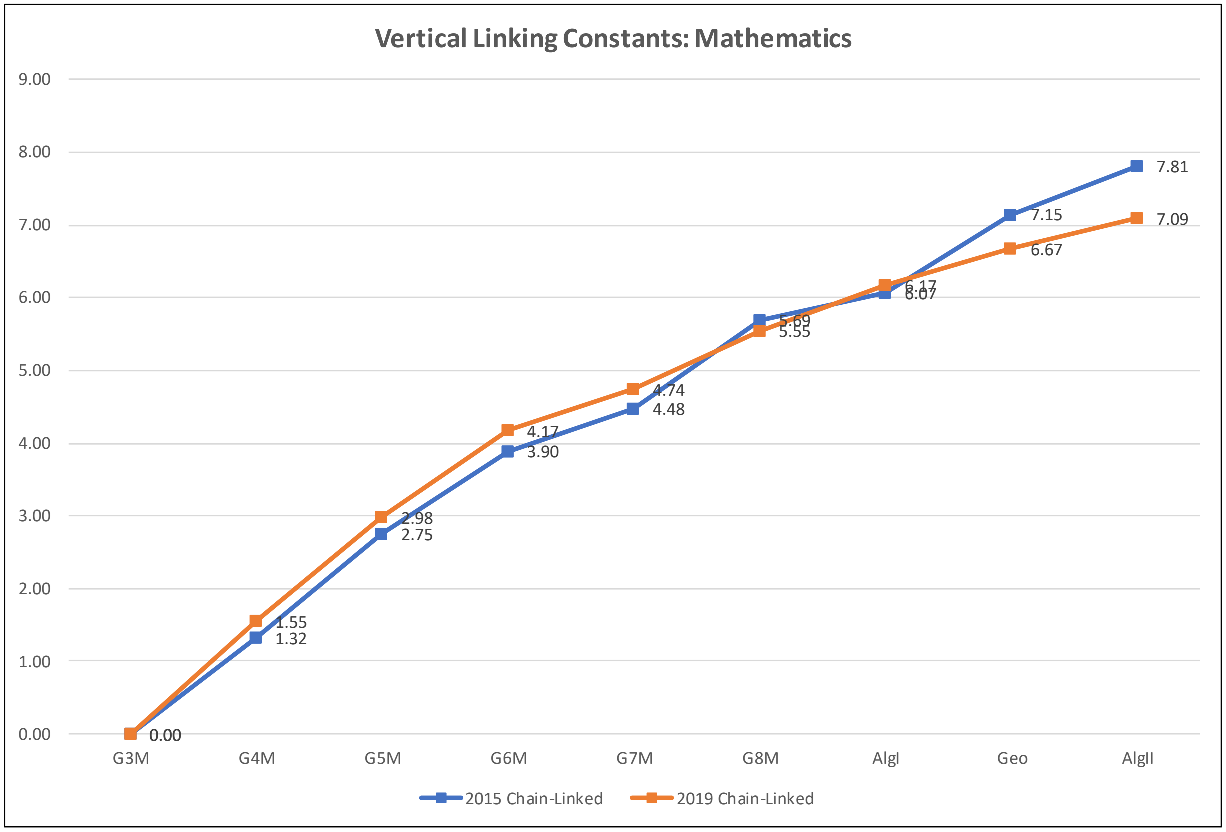


Exhibit 8.2.2.4.4 Comparison of 2015 and 2019 Vertical Linking Constants Estimated from Chain Linking Calibrations: Mathematics



Additionally, Exhibit 8.2.2.4.5 and Exhibit 8.2.2.4.6 show the comparison of the chain linking results obtained in 2015 and 2019 along with the standard error of the linking constants for ELA and mathematics, respectively. Similarity between the 2015 and 2019 vertical linking results is observed with respect to the difference between linking constants by grade. For ELA, although the vertical linking constants by grade in 2019 are uniformly higher than those in 2015, the difference between the 2015 and 2019 ELA linking constant for each grade is not larger than 0.4 logit. For mathematics, the vertical linking constants for grades 8, geometry, and Algebra II in 2019 are smaller than those in 2015, while the vertical linking constants for the other grades in 2019 are larger than those in 2015. The difference between the 2015 and 2019 mathematics linking constant for each grade is not larger than 0.5 logit, except for Algebra II, which is at 0.72 logit.

Exhibit 8.2.2.4.5 Vertical Linking Constants from 2015 and 2019: ELA

ELA	2015 Chain -Linked	2019 Chain-Linked	SE of 2019 Chain Linking Constant
G3E	0	0	NA
G4E	0.18	0.48	0.05
G5E	0.81	1.04	0.07
G6E	1.19	1.43	0.08
G7E	1.44	1.67	0.11
G8E	1.76	2.03	0.11
G9E	1.97	2.23	0.11
G10E	2.12	2.48	0.11
G11E	2.32	2.61	0.12

Exhibit 8.2.2.4.6 Vertical Linking Constants from 2015 and 2019: Mathematics

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

The vertical linking results are also similar between 2015 and 2019 in terms of the overall growth patterns across grades, as shown in Exhibit 8.2.2.4.7 and Exhibit 8.2.2.4.8. For each year, the vertical linking constants indicate much greater growth across grades and high school courses for mathematics than is observed for ELA. In mathematics for both years, growth is on the order of about one logit per year, with the exception of grade 6 to grade 7 and grade 8 to Algebra I. Gains in ELA are less pronounced, with somewhat larger gains in the elementary school years, with growth attenuating in the high school grades for both years.

Exhibit 8.2.2.4.7 Vertical Growth Between Grades for 2019: ELA

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

Exhibit 8.2.2.4.8 Vertical Growth Between Grades for 2019: Mathematics

Mathematics	# of Common Vertical Linking Items	Growth Between Grades	SE of Growth
G3M_G4M	43	1.55	0.04
G4M_G5M	43	1.43	0.03
G5M_G6M	41	1.19	0.04
G6M_G7M	26	0.57	0.02
G7M_G8M	43	0.81	0.06
G8M_AlgI	43	0.62	0.03
G8M_Geo	42	1.12	0.03
AlgI_AlgII	42	0.92	0.02

Similar vertical linking results across years suggest that the vertical linking scale established in the first year of test administration holds for subsequent years, which supports the monitoring and evaluation of student growth over time.

8.3 AZM2 REPORTING SCALE (SCALE SCORES)

The AzM2 assessments are reported on common scales within each subject (ELA and mathematics). The IRT vertical scale scores (SS) are formed by linking each grade-level assessment to the scale of the assessment in the grade level above. The vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate,[45]

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

where $a = 30, d = 2500$ for ELA tests, $a = 30,$ and $d = 3500$ for mathematics tests. $\theta_V = \theta + c$, where θ is the on-grade ability estimate and c is a vertical linking constant listed below for each of the tests, as described in the previous section. For reporting, the on-grade ability estimate is truncated at ± 3.5.

After transforming theta ability estimates to the vertical AzM2 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 AzM2 are mapped into four performance levels per grade/course. The performance-level designations are: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. The performance level is evaluated using the rounded scale score.

Exhibit 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	2,395–2,496	2,497–2,508	2,509–2,540	2,541–2,605
Grade 4	2,400–2,509	2,510–2,522	2,523–2,558	2,559–2,610
Grade 5	2,419–2,519	2,520–2,542	2,543–2,577	2,578–2,629
Grade 6	2,431–2,531	2,532–2,552	2,553–2,596	2,597–2,641
Grade 7	2,438–2,542	2,543–2,560	2,561–2,599	2,600–2,648
Grade 8	2,448–2,550	2,551–2,571	2,572–2,603	2,604–2,658
Grade 9	2,454–2,554	2,555–2,576	2,577–2,605	2,606–2,664
Grade 10	2,458–2566	2,567–2,580	2,581–2,605	2,606–2,668
Grade 11	2,465–2,568	2,569–2,584	2,585–2,607	2,608–2,675
Mathematics				
Grade 3	3,395–3,494	3,495–3,530	3,531–3,572	3,573–3,605
Grade 4	3,435–3,,529	3,530–3,561	3,562–3,605	3,606–3,645
Grade 5	3,478–3,562	3,563–3,594	3,595–3,634	3,635–3,688
Grade 6	3,512–3,601	3,602–3,628	3,629–3662	3,663–3,722

[45] Standard 5.2: The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
Grade 7	3,529–3,628	3,629–3,651	3,652–3,679	3,680–3,739
Grade 8	3,566–3,649	3,650–3,672	3,673–3,704	3,705–3,776
Algebra I	3,577–3,660	3,661–3,680	3,681–3,719	3,720–3,787
Geometry	3,609–3,672	3,673–3,696	3,697–3,742	3,743–3,819
Algebra II	3,629–3,689	3,690–3,710	3,711–3,750	3,751–3,839

8.4 LINKING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to reporting test scores for the spring 2015 and spring 2016 administrations of AzM2, CAI and ADE performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration.[46]

8.4.1 MODE LINKING

A matched samples design (Way, Davis, & Fitzpatrick, 2006) was used to investigate mode comparability. A covariate regression approach was implemented to construct equivalent groups of students taking the AzM2 assessments for both modes of test administration. For the spring 2015 mode investigation, the regression analysis identified for each student a predicted score on the paper-pencil AzM2 assessment from previous year achievement on Arizona's Instrument to Measure Standards (AIMS), covarying demographic variables that included gender, ethnicity, income level status, English learner (EL) status, and Individualized Education Plan (IEP) in the development of the prediction equation. A nearest neighbor search procedure was then applied to the predicted AzM2 scores to select the equivalent groups of students. This procedure resulted in the identification of two matched samples for each assessment to conduct the mode comparability study.

IRT parameter estimates were then calibrated independently for the matched online and paper-based testing (PBT) administration mode samples. The linking constant necessary to bring the matched sample paper-pencil item parameters on the matched sample online scale was then computed. Mean-mean linking was taken as the difference between the average item difficulty estimates from the matched-sample paper-pencil calibration and the average item difficulty estimates from the matched-sample online item parameter estimates.

Mode linking constants were estimated again following the spring 2016 administration of AzM2. Three approaches were used to identify matched samples for these analyses. In the first approach, 2014 AIMS paper-pencil test scores were used to predict student performance on the spring 2016 paper-pencil tests, with the resulting prediction model then used to identify a matched sample of online test takers. This approach allowed all available paper records to be included in the analysis but required constructing matched samples based on achievement scores estimated two years prior. To utilize a more recent and comparable test score, a second approach was used. In this approach, we identified students who were administered AzM2 on paper in 2015, but who participated online in spring 2016. We then identified a matched sample of students, based on AzM2 test scores, who took the paper-pencil version of AzM2 in both 2015 and 2016. For students at grade 3, there were no previous test scores with which to match student ability. We therefore used student performance on the multiple-choice items only on the spring 2016 AzM2 mathematics test to identify matched samples on the assumption that those items would be least susceptible to mode differences. To evaluate whether this approach yields results consistent with the other approaches, this approach was also applied to the grade 4 and grade 5 assessments.

[46] Standard 5.13: When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

Exhibit 8.4.1.1 presents the mode-linking constants for the ELA assessments resulting from the matched sample analysis conducted on the spring 2015 administration of AzM2 and the linking constants resulting from each of the matched sample approaches used following the spring 2016 administration. In the grades 4–8 assessments, whether the matched samples are based on spring 2014 AIMS or spring 2015 AzM2, the obtained mode-linking constants are generally small and equivalent across methods. For the high school EOC assessments, both approaches indicate that ELA assessments were somewhat more difficult online than on a paper-pencil form. The magnitude of those differences is greater when matching achievement based on 2014 AIMS than 2015 AzM2. We note that the R^2 for the prediction equation used to identify matched samples for ELA based on 2014 AIMS remained quite high (R^2 around 0.65) even for the high school assessments, although matching based on spring 2015 AzM2 achievement may nevertheless be more robust.

For grade 3 ELA, samples were matched based on student performance on the concurrently administered AzM2 mathematics multiple-choice (MC) items. To evaluate whether this approach yielded results consistent with the other two methods, we applied the same procedure in grades 4 and 5, where results indicated general convergence with the other methods, and indicating no effect for mode at grade 4 and a moderate mode effect at grade 5. When applied at grade 3, no mode effect was identified.

We note that any mode effect seems to interact with items, with some items easier when administered online, while others are more difficult. Thus, the mode effect is likely to be form-specific and vary across test administrations. And this seems to be the case when mode-linking constants are compared between the 2015 and 2016 administrations of AzM2. As shown in Exhibit 8.4.1.1, in spring 2015, mode effects were observed in grades 3, 4, and 8, but were more moderate at the other grades. In spring 2016, however, mode effects were absent or moderate in grades 3–8 but appeared in the high school EOC tests.

Exhibit 8.4.1.1 Mode-Linking Constants for AzM2 ELA Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3E	2015	0.13	−0.01	0.13	3.90
	2016—Mathematics MC Match	0.17	0.16	0.01	0.30
G4E	2015	−0.09	−0.19	0.11	3.30
	2016—2014 AIMS Match	0.21	0.19	0.02	0.60
	2016—2015 AzM2 Match	0.21	0.18	0.03	0.90
	2016—Mathematics MC Match	0.21	0.21	0.00	0.00
G5E	2015	0.04	−0.02	0.06	1.80
	2016—2014 AIMS Match	0.02	−0.02	0.04	1.20
	2016—2015 AzM2 Match	0.03	−0.02	0.05	1.50
	2016—Mathematics MC Match	0.04	−0.04	0.08	2.40
G6E	2015	0.07	−0.02	0.09	2.70
	2016—2014 AIMS Match	0.18	0.21	−0.03	−0.90
	2016—2015 AzM2 Match	0.20	0.16	0.04	1.20
G7E	2015	−0.08	−0.16	0.08	2.40
	2016—2014 AIMS Match	0.19	0.12	0.07	2.10
	2016—2015 AzM2 Match	0.12	0.05	0.07	2.10
G8E	2015	−0.04	−0.22	0.18	5.40
	2016—2014 AIMS Match	0.01	−0.01	0.02	0.60
	2016—2015 AzM2 Match	0.00	−0.05	0.05	1.50

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G9E	2015	0.13	0.09	0.04	1.20
	2016—2014 AIMS Match	0.07	−0.12	0.20	6.00
	2016—2015 AzM2 Match	0.08	−0.16	0.24	7.20
G10E	2015	−0.03	−0.10	0.07	2.10
	2016—2014 AIMS Match	0.10	−0.10	0.20	6.00
	2016—2015 AzM2 Match	0.09	−0.04	0.13	3.90
G11E	2015	0.12	0.15	−0.03	−0.90
	2016—2014 AIMS Match	0.16	−0.09	0.25	7.50
	2016—2015 AzM2 Match	0.14	−0.04	0.18	5.40

Exhibit 8.4.1.2 presents the mode-linking constants computed for the spring 2015 and spring 2016 administrations of the AzM2 mathematics assessments. As observed for ELA, in the grades 4–8 and Algebra I mathematics assessments, whether the spring 2016 matched samples were based on spring 2014 AIMS or spring 2015 AzM2, the obtained mode-linking constants are generally equivalent across methods. Effects of mode varied across grades, with the online form somewhat easier than a paper-pencil form at grade 4, somewhat more difficult at grade 7, and about the same at grades 5, 6, and 8. For the high school EOC assessments, both approaches indicate that mathematics assessments were somewhat more difficult online than on a paper-pencil form. As with ELA, the magnitude of those differences was greater when matching achievement based on 2014 AIMS than 2015 AzM2. In this case, we note that the R^2 for the prediction equation used to identify matched samples for mathematics based on 2014 AIMS remained quite a bit lower (R^2≈.40) for the high school assessments compared to the lower grades (R^2 ≈ .65), so that matching based on spring 2015 AzM2 achievement are likely more robust.

Exhibit 8.4.1.2 Mode-Linking Constants for AzM2 Mathematics Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3M	2015	−0.71	−0.77	0.06	1.80
	2016—Mathematics MC Match	−0.84	−0.57	−0.27	−8.10
G4M	2015	−0.40	−0.48	0.08	2.40
	2016—2014 AIMS Match	−0.43	−0.25	−0.17	−5.10
	2016—2015 AzM2 Match	−0.57	−0.43	−0.14	−4.20
	2016—Mathematics MC Match	−0.41	−0.24	−0.17	−5.10
G5M	2015	−0.09	−0.09	−0.01	−0.30
	2016—2014 AIMS Match	−0.06	−0.02	−0.04	−1.20
	2016—2015 AzM2 Match	−0.16	−0.12	−0.03	−0.90
	2016—Mathematics MC Match	−0.07	−0.06	0.00	0.00
G6M	2015	0.07	0.01	0.07	2.10
	2016—2014 AIMS Match	−0.01	0.04	−0.05	−1.50
	2016—2015 AzM2 Match	−0.09	−0.06	−0.03	−0.90
G7M	2015	0.15	0.07	0.08	2.40
	2016—2014 AIMS Match	0.18	0.07	0.11	3.30

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
	2016—2015 AzM2 Match	0.11	−0.03	0.14	4.20
G8M	2015	0.43	0.32	0.11	3.30
	2016—2014 AIMS Match	0.56	0.55	0.00	0.00
	2016—2015 AzM2 Match	0.47	0.47	0.01	0.30
Algebra I	2015	0.29	0.23	0.05	1.50
	2016—2014 AIMS Match	0.64	0.51	0.13	3.90
	2016—2015 AzM2 Match	0.72	0.57	0.15	4.50
Geometry	2015	1.12	0.99	0.13	3.90
	2016—2014 AIMS Match	1.34	1.15	0.20	6.00
	2016—2015 AzM2 Match	1.19	1.03	0.16	4.80
Algebra II	2015	1.45	1.36	0.09	2.70
	2016—2014 AIMS Match	1.45	1.17	0.28	8.40
	2016—2015 AzM2 Match	1.06	0.91	0.15	4.50

For grade 3 mathematics assessment, as with grade 3 ELA, samples were matched based on student performance on the mathematics multiple-choice items. Again, this approach was applied in grades 4 and 5 to evaluate it against the other two methods, where the results indicated general convergence, indicating that items administered online were somewhat easier at grade 4 and no mode effect at grade 5. When applied at grade 3, a relatively large effect for mode was identified, indicating that items administered online were easier than on a paper-pencil form.

As with ELA, the identified mode effects varied across test administrations. The advantage of online over paper-pencil identified in 2016 was not observed in 2015. Likewise, observed effects of mode at grade 7 and for Algebra I and Algebra II in 2016 were not as pronounced in 2015, while effects of mode observed at grade 8 in 2015 were not observed in 2016. Thus, as with ELA, the effect of mode appears to be form-specific and can be expected to vary across test administrations.

8.4.2 SCHOOL PERFORMANCE

In a separate approach to evaluating mode comparability, the ADE implemented an investigation based on the spring 2015 operational test administration statewide (Scott, 2015). In her study, Scott (2015) first identified which Arizona schools elected to administer AzM2 online and on paper-pencil forms and then examined the two samples of schools for any differences in performance on the spring 2014 PBT administration of AIMS. The rationale in selecting school-level analysis was based on schools having to choose only one of the two modes in which to assess all their students. This increased level of matching was appropriate because the mode used by the student was, and continues to be, a school-based decision, rather than student based. Having found no difference in mean 2014 performance between the two groups, there would be no expectation for performance differences on AzM2 except as a function of test administration mode. Following the spring 2015 administration of AzM2, ADE examined the performance of schools participating online and on paper-pencil forms, and again found performance on the AzM2 to be comparable between the two sets of schools.

8.5 LINKING THE AZM2 TO OTHER SCALES FOR PERFORMANCE COMPARISON

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

To facilitate comparisons of Arizona achievement to other national and international benchmarks, several external linking sets were embedded in the 2015 AzM2 field-test slots. Arizona identified the locations of performance standards of other assessments systems on the AzM2 scale; this information was used to inform panelists recommending performance standards for the AzM2.[47] The location of performance standards from the following assessments were identified on the AzM2 scale:

- Smarter Balanced, by linking to CAI's Independent College and Career Readiness (ICCR) item bank on the Smarter Balanced scale
- PISA, by embedding PISA items in the grade 10 ELA, Algebra I, and geometry EOC assessments
- Historical Arizona performance by embedding AIMS items to link to the AIMS scale
- Utah's SAGE via common items in the operational test form

After the calibration of the AzM2 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 AzM2 scale defined by the operational item parameters. All external linking items had two sets of item parameters: (a) external scale, and (b) AzM2 scale. To identify the location of external scale performance standards on the AzM2 scale, CAI identified the linking constants necessary to transform item parameters from the external reference scale to the AzM2 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 AzM2 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 AzM2 scale.

In the context of standard setting, this procedure enabled the ADE to identify a location in the AzM2 ordered-item booklet (OIB) that represented a level of difficulty similar to a particular level in the external scale. For example, after finding the linking constant necessary to put the Smarter Balanced item parameters on the AzM2 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 AMERICAN COLLEGE TESTING COLLEGE-READY CUT ON AZM2

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the American College Testing (ACT) college-ready cuts was identified on the AzM2 scale and provided to panelists during performance standards workshops in 2015. In order to identify the location of the ACT college-ready cuts for the grade 11 ELA and Algebra II AzM2 EOC 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 AzM2 to map the ACT college-ready benchmark on the AzM2 scale(s). To examine the relationships between the AzM2 and ACT assessments directly, the ADE

[47] Standard 5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

obtained the ACT test scores for Arizona students graduating from high school in spring 2016. The direct linking study using the AzM2 and ACT data is summarized in this section.

Although AzM2 is offered as a series of EOC tests in high school, most students take the Algebra II assessment at grade 11, so the focus of this investigation will be on the grade 11 ELA and Algebra II AzM2 assessments administered in spring 2015. From among the full set of spring 2015 grade 11 ELA and Algebra II test takers, there are 58,888 (93%) and 32,945 (56%) grade 11 students, respectively. These records represent the target sample for the analyses reported in this study.

Because many students did not take the ACT and the two subgroups differed systematically across demographic and achievement variables, the imputing approach is often employed to handle missing data in the analysis of the relationship between the AzM2 scores and subsequent performance on the ACT. However, previous studies for Minnesota and Ohio showed that imputing or deleting the missing records did not impact the linkage identified between their graduation tests and the ACT test. For this study, we instead divided the complete sample of merged records into model building and cross-validation samples of equal size. The cross-validation sample allows for better estimation model fit. Because the model is built using a sample independent from that used to evaluate model fit, estimates of model fit exclude sample dependent idiosyncrasies that would be reflected as model overfit in the model development sample.

ELA: Test takers with missing ACT or AzM2 scale scores were removed from the merged dataset. The ACT reading scale score for the remaining 25,977 students were regressed onto the applicable grade 11 ELA scale score and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2, was identified as the best model to predict ACT reading from prior performance on the AzM2 ELA test:

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

where

$\hat{Y}$ = ACT Reading Scale Score
X1 = AzM2 ELA Scale Score
X2 = Female–Male Contrast
X3 = American Indian–White Contrast
X4 = Multi-Ethnic Contrast
X5 = Asian Contrast
X6 = Hispanic-White Contrast
X7 = African American–White Contrast
X8 = Native Hawaiian–White Contrast
X9 = Free and Reduced-Price Lunch Contrast
X10 = EL Contrast

The overall model was statistically significant (*F* (10, 20388) = 1704.70, p < .0001; adjusted R^2 = 0.46). Application of this regression model indicates that an AzM2 ELA scale score of 2,585 is associated with the ACT reading college-ready cut score of 22.

Mathematics: The records with missing ACT or AzM2 scale scores were excluded from the analysis. Then, the ACT mathematics scale scores for the remaining 13,777 students were regressed onto the applicable AzM2 Algebra II test and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE, and largest adjusted R^2, was identified as the best model to predict ACT mathematics scores from prior performance on the AzM2 Algebra II test:

$$\hat{Y} = -305.7 + 0.08*X1 - 0.55*X2 - 1.55*X3 - 0.48*X4 - 0.44*X5 - 1.44*X6 - 1.41*X7 - 0.83*X8 - 1.22*X9 - 1.57*X10$$

where

$\hat{Y}$ = ACT Mathematics Scale Score
X1 = AzM2 Mathematics Scale Score
X2 = Female–Male Contrast
X3 = American Indian–White Contrast
X4 = Multi-Ethnic Contrast
X5 = Asian Contrast
X6 = Hispanic–White Contrast
X7 = African American–White Contrast
X8 = Native Hawaiian–White Contrast
X9 = Free and Reduced-Price Lunch Contrast
X10 = EL Contrast

The overall model was statistically significant (*F* (10, 13768) =1764.13, p < .0001; adjusted R^2 = 0.51). Application of this regression model indicates that an AzM2 mathematics score of 3,727 is associated with the ACT mathematics college-ready cut score of 22.

The validation set approach is a type of resampling method that estimates a model error rate by holding out a subset of the data from the fitting process (the testing dataset). The model is then built using the other set of observations (the training dataset). Then the model result is applied on the testing dataset in which we can then calculate the error. In summary, this general idea allows for the model to not overfit. In this study, the training dataset contained 50% randomly selected merged records and the testing dataset had the other 50% of students. The multiple regression built by the training set yielded the same AzM2 cut scores (ELA 2,585, mathematics 3,727) as the ones from the full data model. Then the predictive model was applied to the testing set. The root mean square error (RMSE) was calculated as the square root of the average squared errors found between the actual ACT score point and the model fitted values. Furthermore, we repeated this sampling and model fitting process 100 times to determine how the RMSE varied across random samples. For ELA, the average RMSE was 5.03 and the standard deviation (SD) of the RMSE was 0.02 across the 100 replications. For mathematics, the average RMSE was 2.79 and the SD was 0.02. The SD of the RMSE was very small, indicating that the sample selected for the modeling has no significant impact on the model fitting.

In addition, the equipercentile equating method was used to verify the linking between ACT and AzM2 test scores. The AzM2 scale score associated with the ACT cut score 22 is 2,585.72 for ELA and 3,727.46 for mathematics. These cut scores are consistent with those identified using regression models.

9 CONSTRUCTED-RESPONSE SCORING

The Arizona's Statewide Achievement Assessment (AzM2) assessments in English language arts (ELA) and mathematics utilize a variety of item types to assess students' mastery of the Arizona State Standards. The Arizona Department of Education (ADE) leverages Cambium Assessment, Inc.'s (CAI) item-scoring technology to machine-score student responses to most items, including traditional selected-response (multiple-choice) item types and machine-scored constructed-response (MSCR) items types. The MSCR item types are designed to capture and score a variety of response types, such as graphing, drawing, or arranging graphic regions, selecting or rearranging sentences or phrases within passages, or entering equations or words, allowing AzM2 items to assess a wide range of student knowledge and skills. In most cases, MSCR items that are developed for online administration are adapted for paper-pencil and responses are captured in a format that allows machine scoring.

In addition, some constructed-response items are scored by human raters; these items are referred to as "handscored." To support machine scoring of each essay response, in 2016, a sample of essay responses was handscored through verification, and those responses and scores were used to develop the statistical scoring models used to score the remaining responses. The statistical scoring models developed in spring 2016 will be used to score all essay responses in future test administrations. In addition, mathematics assessments administered on paper-pencil forms included a small number of items that were scored by human raters. Generally, these were items that required students to produce an equation. The reading components of the ELA assessments, both online and paper-pencil, and the mathematics assessments administered online are machine scored in their entirety.

CAI partners with Measurement, Inc. (MI), to fulfill all handscoring requirements. CAI provides the automated electronic scoring, and MI provides all handscoring for the AzM2 tests. This section describes the process for configuring and validating machine rubrics and the process for handscoring, including rules, descriptions of scorer training and systems used, and mechanisms for ensuring the reliability and validity of item scores.

9.1 MACHINE SCORING

9.1.1 EXPLICIT RUBRICS

As part of the item-development process for machine-scored item types which are scored with explicit rubrics, a rubric validation process was enacted to verify that rubrics are implemented as intended and responses are scored correctly. This procedure is typically conducted following the initial administration of items, usually when the item is field tested, and allows test developers to review the intent of the rubric versus the actual behavior. Actual student responses were reviewed by test development experts, along with resulting item scores, to ensure that the rubrics functioned as intended and awarded credit appropriately. Where necessary, test developers modified machine rubrics to address insufficiencies, automatically rescoring student responses for the item, and repeating the process to finalize and approve the machine-scored rubrics. Test developers reviewed a strategic sample of responses, including responses where high-achieving students scored poorly on the item and lower-achieving students scored well on the item. They also reviewed randomly selected responses from the population.

9.1.2 ESSAY AUTOSCORING

As part of the spring 2021 administration of AzM2, students in each grade were administered one of two writing tasks (one informational/explanatory, and the other, either opinion [grades 3–5] or argumentative [grades 6–8 and 10]) that had been calibrated during the spring 2016 administration. This section describes the processes performed to calibrate these, and the rest of the available writing prompts completed during the spring 2016 administration. As part of the spring 2016 administration of AzM2, students in each grade were administered one of two writing tasks (one informational/explanatory, and the other, either opinion [grades 3–5] or argumentative [grades 6–8 and 10]) in the writing component of each of the ELA online assessments.

Two approaches were used to develop the statistical models used to score the essay responses. For ICCR writing tasks administered online in the Florida field test (grades 8–10), ADE adopted the scoring models generated from student responses in the Florida field-test administration. Because the scoring models are based on semantic and syntactic features of the text that discriminate high- versus low-scoring essays as determined by human raters, the models are highly generalizable.

For the grades where scoring models did not already exist (grades 3–7), an alternative approach was employed that allowed for autoscoring to be implemented as part of the spring 2016 essay scoring. Because the ELA window is split into separate writing and reading assessment windows, with the online writing window closing several weeks before close of the reading test administration, the dual window afforded an opportunity to build and implement the statistical scoring models in time to meet spring reporting timelines.

To facilitate development of the scoring models, MI conducted rangefinding, where possible, based on student responses from the Florida assessment. The rangefinding process is designed to calibrate a sample of responses for scorer training, qualification, and monitoring. Responses exemplifying each score point are identified and annotated for scorer training. Additional responses are identified for use in qualifying readers for scoring and for establishing validity sets that are used to monitor reader performance. Thus, for grades 4–7, which were included in the Florida field test, rangefinding activities to support AzM2 rubric scoring were completed before the opening of the AzM2 assessment window.

For the grades 3 and 11 assessments, which had not been previously administered, MI pulled a sample of essay responses following the first week of the testing window with which to conduct rangefinding activities. The development of training materials and training of raters followed immediately so that handscoring could begin by the end of the fourth week of the testing window.

At the end of the second week of testing, CAI drew a random sample of 2,000 responses to each of the writing tasks administered at grades 3–7 for use in building the statistical scoring models. Those responses were routed to MI for handscoring. Each response was double scored, with any discrepancies routed for resolution scoring.

As handscoring activities were completed for each writing task, and scores were uploaded to CAI, work began to develop statistical scoring models for each rubric element, and to deploy those models to the Test Delivery System (TDS) to score all remaining essay responses.[48]

To develop the scoring models, the random sample of 2,000 responses was divided into a model building sample of 1,500 responses and a cross-validation sample of 500 responses. Model performance was evaluated on the cross-validation sample to ensure that model fit indices were not based on the model building sample, which may inflate fit indicators.

The statistical scoring models also yield an indicator of score confidence based on (1) responses with unusual features, and (2) responses scoring near rubric thresholds. For each model, a confidence threshold defined as two standard deviations

[48] Standard 4.19: When automated algorithms are to be used to score complex test taker responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

below the mean confidence value for the responses in the cross-validation sample was identified. Any scored response with the lowest 15% confidence index were automatically routed to MI for verification scoring.

The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and "learned" from a training set. During training, these features are related to human scores through a statistical model. The resulting estimates complete a prediction equation that predicts how a human would score a response with the measured features. Statistical rubrics are, effectively, proxy measures. Although they can directly measure some aspects of writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not make direct measures of argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Further, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may *predict* whether a response to a factual item is correct—higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the "reason" that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

As noted, the engine employs a "training set," a set of essay responses scored with maximally valid scores, which we obtain by having all responses double-scored by expert scorers and a thorough adjudication process for adjacent or discrepant scores. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Approximately 1,500 essay responses were selected at random from the set of scored essay responses to serve as the training set.

For each dimension in the rubric, the system estimates an appropriate statistical model relating the measures to the score assigned by humans. This model, along with its final parameter estimates, is used to generate a predicted or "proxy" score.

In addition to the training set, we draw an independent random sample of responses for cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are handscored, and agreement between human- and machine-assigned scores is examined. The cross-validation process ensures that the rubric generalizes across all responses and that the statistical model identified during training does not capitalize on peculiarities in the training set.

Exhibit 9.1.2.1 presents agreement indicators for the two initial human raters, and between the resolved human and statistical rubric score, for the two writing prompts randomly assigned in each grade in the spring 2021 administration.[49] Please refer to the *2016 AzM2 Technical Report*, available at www.azed.gov, for the values for the complete list of prompts**.** Indicators include percentage exact agreement, Pearson's correlation, a quadratic weighted kappa statistic, and the standardized mean difference between the scores. Although absolute values for evaluating statistics have been advanced (Condon, 2013; Wei & Higgins, 2013), the focus of these comparisons is degradation of agreement when moving from human–human agreement to machine–human agreement. Agreement between human raters is an indicator of how reliably the responses can be scored by human raters. Because the statistical rubrics attempt to reproduce human–assigned scores, evaluation of machine–human agreement is with respect to observed human–human agreement. Responses with poor human–human agreement will not be reliably scored by either humans or machines. For the training and validation sets of the prompts administered in spring 2021, Exhibit 9.1.2.2 presents the correlations among the dimension scores.

[49] Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

Exhibit 9.1.2.1 Summary of Human and Machine Scores for Spring 2021 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N of Human	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted K	SMD	% Exact	Pearson r	Weighted κ	SMD
3	13022	Conventions	2	2,092	1.49	1.62	0.68	0.63	0.65	0.70	0.52	0.02	0.71	0.76	0.60	0.20
		Elaboration	4		2.06	2.02	0.72	0.60	0.61	0.57	0.47	0.00	0.63	0.68	0.51	0.06
		Organization	4		2.14	2.08	0.74	0.60	0.67	0.61	0.53	0.03	0.64	0.68	0.52	0.09
3	13025	Conventions	2	2,093	1.46	1.52	0.71	0.66	0.59	0.68	0.49	0.01	0.66	0.73	0.58	0.09
		Elaboration	4		2.03	2.01	0.75	0.66	0.61	0.59	0.48	0.01	0.71	0.73	0.61	0.02
		Organization	4		2.05	1.99	0.80	0.74	0.64	0.59	0.51	0.01	0.68	0.65	0.56	0.07
4	13120	Conventions	2	2,091	1.20	1.15	0.68	0.64	0.63	0.66	0.53	0.04	0.64	0.69	0.54	0.07
		Elaboration	4		1.31	1.26	0.49	0.46	0.52	0.76	0.48	0.02	0.57	0.82	0.56	0.09
		Organization	4		1.46	1.45	0.55	0.52	0.61	0.76	0.57	0.04	0.59	0.79	0.59	0.02
4	13119	Conventions	2	2,094	1.29	1.32	0.64	0.63	0.60	0.68	0.52	0.06	0.67	0.73	0.59	0.04
		Elaboration	4		1.38	1.33	0.53	0.50	0.47	0.71	0.42	0.04	0.59	0.79	0.56	0.10
		Organization	4		1.53	1.51	0.60	0.53	0.59	0.70	0.51	0.03	0.65	0.77	0.60	0.03
5	13247	Conventions	2	2,097	1.45	1.48	0.66	0.62	0.69	0.74	0.60	0.04	0.71	0.76	0.62	0.04
		Elaboration	4		1.78	1.81	0.62	0.59	0.56	0.65	0.47	0.05	0.65	0.74	0.57	0.06
		Organization	4		1.94	1.92	0.65	0.61	0.65	0.69	0.54	0.02	0.69	0.77	0.61	0.03
5	13246	Conventions	2	2,093	1.46	1.49	0.61	0.62	0.63	0.73	0.56	0.10	0.71	0.78	0.65	0.06
		Elaboration	4		1.61	1.59	0.55	0.51	0.55	0.69	0.48	0.07	0.61	0.78	0.58	0.03
		Organization	4		1.83	1.81	0.66	0.56	0.61	0.67	0.51	0.00	0.62	0.71	0.53	0.03
6	13307	Conventions	2	2,095	1.46	1.49	0.66	0.64	0.64	0.68	0.53	0.03	0.69	0.74	0.60	0.05
		Elaboration	4		1.60	1.57	0.67	0.61	0.62	0.66	0.52	0.00	0.67	0.74	0.59	0.05
		Organization	4		1.84	1.79	0.73	0.63	0.63	0.64	0.52	0.02	0.68	0.70	0.57	0.06
6	13306	Conventions	2	2,097	1.59	1.63	0.59	0.59	0.56	0.70	0.47	0.07	0.63	0.76	0.55	0.08
		Elaboration	4		1.70	1.64	0.64	0.56	0.55	0.65	0.46	0.01	0.62	0.73	0.55	0.09
		Organization	4		1.91	1.88	0.71	0.63	0.61	0.66	0.51	0.05	0.63	0.69	0.53	0.05
7	13401	Conventions	2	2,084	1.67	1.71	0.50	0.52	0.63	0.81	0.59	0.05	0.68	0.83	0.63	0.07
		Elaboration	4		1.84	1.86	0.54	0.50	0.58	0.72	0.50	0.01	0.67	0.82	0.62	0.03
		Organization	4		2.01	2.00	0.55	0.42	0.63	0.74	0.53	0.04	0.66	0.83	0.58	0.01
7	13406	Conventions	2	2,090	1.45	1.51	0.62	0.60	0.58	0.70	0.50	0.03	0.72	0.78	0.65	0.10
		Elaboration	4		1.76	1.77	0.54	0.52	0.60	0.74	0.54	0.03	0.61	0.79	0.57	0.03
		Organization	4		1.92	1.92	0.52	0.45	0.54	0.70	0.45	0.04	0.66	0.84	0.61	0.00
8	13454	Conventions	2	2,677	1.55	1.59	0.63	0.61	0.69	0.79	0.63	0.03	0.72	0.80	0.65	0.06
		Elaboration	4		1.93	1.96	0.71	0.68	0.75	0.78	0.69	0.02	0.73	0.75	0.63	0.05
		Organization	4		2.06	2.04	0.76	0.72	0.76	0.75	0.68	0.01	0.76	0.76	0.67	0.02

Grade	ITS ID	Dimensions	Score Point	N of Human	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted K	SMD	% Exact	Pearson r	Weighted κ	SMD
8	13439	Conventions	2	2,719	1.62	1.70	0.58	0.53	0.62	0.78	0.55	0.02	0.73	0.84	0.66	0.12
		Elaboration	4		2.11	2.08	0.71	0.64	0.75	0.74	0.65	0.01	0.72	0.75	0.62	0.05
		Organization	4		2.21	2.20	0.81	0.75	0.72	0.69	0.62	0.05	0.79	0.75	0.69	0.01
10	13638	Conventions	2	2,580	1.60	1.68	0.57	0.52	0.58	0.71	0.51	0.05	0.59	0.76	0.52	0.15
		Elaboration	4		2.02	2.01	0.69	0.63	0.65	0.69	0.56	0.00	0.71	0.77	0.63	0.02
		Organization	4		2.10	2.12	0.73	0.68	0.69	0.67	0.58	0.00	0.73	0.74	0.64	0.02
10	13637	Conventions	2	1,417	1.59	1.65	0.58	0.54	0.58	0.69	0.49	0.06	0.60	0.77	0.53	0.09
		Elaboration	4		1.92	1.90	0.68	0.64	0.70	0.75	0.62	0.02	0.73	0.77	0.65	0.05
		Organization	4		2.06	2.08	0.72	0.64	0.74	0.76	0.66	0.03	0.75	0.78	0.67	0.01

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

Exhibit 9.1.2.2 Summary of Dimension Intercorrelations for Spring 2021 Writing Prompts

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

9.1.3 MACHINE-IDENTIFIED CONDITION CODES

Verifications with Machine-Identified Condition Codes

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

The available condition codes include:

- NO_RESPONSE: No non-blank characters are detected in the response.
- NOT_ENOUGH_DATA: Student response is fewer than 11 words.
- PROMPT_COPY_MATCH: Student response is substantially copied from the passage or item prompt (flagged when more than 50% of response text matches the prompt or when the response includes more than 70% sequential match with prompt).
- DUPLICATE_TEXT: Student response is substantially comprised of repeated text copied over and over (flagged when ratio of duplicate text is more than 70% of total response).
- NONSPECIFIC: Essay scoring engine predicts the assignment of a condition code.

Responses receiving the NO_RESPONSE condition code are considered not attempted and do not receive a score. All other condition codes imply an attempt and receive the lowest possible dimension score for purposes of ability estimation.

All responses assigned the NONSPECIFIC condition code for human verification:

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

For score reporting, NO_RESPONSE will be reported as Blank. All other condition codes will be reported as non-scorable responses (i.e., NS). Please note the responses receiving machine-assigned condition codes should not be routed for human verification with exception of NONSPECIFIC. Exhibit 9.1.3.1 presents percentages of the machine-assigned condition codes for spring 2017 administrations and Exhibit 9.1.3.2 presents percentages of the machine-assigned condition codes for spring 2018 administrations. Exhibit 9.1.3.3 presents percentages of the machine-assigned condition codes for spring 2019 administrations. Exhibit 9.1.3.4 presents the percentages of the machine-assigned condition codes for spring 2021 administrations.

Exhibit 9.1.3.1 Frequency of Machine-Assigned Condition Codes for Spring 2017 Writing Prompts

Machine-Assigned Condition Code		Percentage of Condition Code						
		PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension		ALL	ALL	ALL	ALL	C	E	O
G3E	13023	9	0	0	1	0	0	0
	13026	13	0	0	1	0	0	0
G4E	13094	26	0	0	1	0	0	0
	13095	9	0	0	1	0	0	0
G5E	13236	7	0	0	0	0	0	0
	13239	10	0	0	0	0	0	0
G6E	13304	9	0	0	0	0	0	0
	13308	6	0	0	0	0	0	0
G7E	13402	12	0	0	0	0	0	0
	13403	3	0	0	0	0	0	0
G8E	13437	7	0	0	0	2	0	2
	13452	4	0	0	0	0	0	0
G9E	13557	4	0	0	0	1	3	3
	13566	4	0	0	0	0	0	0
G10E	13639	4	0	0	0	0	6	6
	13640	1	0	0	0	0	3	0
G11E	13722	1	0	0	0	0	0	0
	13724	2	0	0	1	0	0	0

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

Exhibit 9.1.3.2 Frequency of Machine-Assigned Condition Codes for Spring 2018 Writing Prompts

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

Machine-Assigned Condition Code		Percentage of Condition Code						
		PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension		ALL	ALL	ALL	ALL	C	E	O
	13565	2	0	0	0	1	0	0
G10E	13635	2	0	0	0	0	0	0
	13636	5	0	0	0	0	0	0
G11E	13723	1	0	0	0	0	0	0
	13725	3	0	0	1	0	0	0

Exhibit 9.1.3.3 Frequency of Machine-Assigned Condition Codes for Spring 2019 Writing Prompts

Machine-Assigned Condition Code		Percentage of Condition Code						
		PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension		ALL	ALL	ALL	ALL	C	E	O
G3E	13022	9	0	0	1	0	0	0
	13025	12	0	0	1	0	0	0
G4E	13119	7	0	0	1	0	0	0
	13120	5	0	0	1	0	0	0
G5E	13246	8	0	0	0	0	0	0
	13247	3	0	0	0	0	0	0
G6E	13306	9	0	0	0	0	0	0
	13307	4	0	0	0	0	0	0
G7E	13401	4	0	0	0	0	0	0
	13406	9	0	0	0	0	0	0
G8E	13439	5	0	0	0	1	1	0
	13454	4	0	0	0	0	0	0
G9E	13555	2	0	0	0	0	0	0
	13556	4	0	0	0	1	2	1
G10E	13637	1	0	0	0	0	0	0
	13638	2	0	0	0	0	0	1
G11E	13720	4	0	0	1	0	0	0
	13721	1	0	0	0	0	0	0

Exhibit 9.1.3.4 Frequency of Machine-Assigned Condition Codes for Spring 2021 Writing Prompts

Machine-Assigned Condition Code		Percentage of Condition Code						
		PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension		ALL	ALL	ALL	ALL	C	E	O
G3E	13022	9	0	0	3	0	0	0
	13025	16	0	0	3	0	0	0
G4E	13119	8	0	0	1	0	0	0
	13120	6	0	0	1	0	0	0

Machine-Assigned Condition Code		Percentage of Condition Code						
		PROMPT COPY MATCH	DUPLICATE TEXT	NO RESPONSE	NOT ENOUGH DATA	NONSPECIFIC		
Dimension		ALL	ALL	ALL	ALL	C	E	O
G5E	13246	10	0	0	1	0	0	0
	13247	3	0	0	1	0	0	0
G6E	13306	10	0	0	0	0	0	0
	13307	5	0	0	0	0	0	0
G7E	13401	4	0	0	0	0	0	0
	13406	11	0	0	0	0	0	0
G8E	13439	5	0	0	0	3	3	3
	13454	4	0	0	0	0	0	0
G10E	13637	4	0	0	0	0	0	0
	13638	2	0	0	0	1	1	1

9.2 HANDSCORING

Handscoring of online essay responses for statistical model building and all essay responses from paper-based testing (PBT) administrations were routed to MI for scoring. As noted in Section 9.1, the sample of essay responses selected for statistical model building was independently scored by two readers. Any response assigned discrepant scores was routed for resolution scoring by a scoring trainer. In addition, all essay responses captured from PBT administrations were handscored, with 10% of all paper responses receiving a second reading (Reader 2) to monitor and maintain sufficient inter-rater reliability, as discussed in the following sections. For ELA handscoring, where scores from Reader 1 and Reader 2 were not in adjacent agreement, the response was sent for resolution scoring by a team leader or scoring director. The final item score was based on the resolution score, when present, or else on the initial read. For mathematics handscoring, where scores from Reader 1 and Reader 2 were not in exact agreement, the response was sent for resolution scoring by a team leader or scoring director. The final item score for mathematics was based on the resolution score, when present, or else on the initial read.

In spring 2021, all the essays were autoscored, and the essay responses with the low confidence index were routed to MI for human verification. The final essay score was the human verification score when present.

9.2.1 HANDSCORING PROCESS

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

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

- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
- Accommodating PBT scores when images are of insufficient quality
- Outputting data easily into MI's score reporting applications

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

When logging on to VSC Score, scorers were presented with a scoring set, which is the images-scoring equivalent of a physical packet of student responses. The scoring set was generated by randomly selecting student responses from the pool of non-scored student responses. The resultant set of responses was checked out to the scorer. The images they received had no demographic information on them. The scorer did not know the name, sex, school, or location of the student whose item was being scored. The scorer evaluated the first response, entered the score by clicking the appropriate values on the scoring toolbar, and clicked the Submit button. For multi-page responses, a scorer had to view each page of the response before a score was entered. Once the Submit button was clicked, the system recorded the score and the next response in the scoring set appeared for the scorer to score and submit. This process continued until all responses in the set had been scored.

When a scorer had a question about a response, he or she transferred the image (along with a virtual note including the question and/or comments) from the current scoring set to a review set assigned to a team leader or the scoring director. The team leader or scoring director submitted the appropriate score or returned the response to the scorer with comments. This procedure was used whenever a scorer had scoring concerns or found nonscorable responses (NSR) or responses requiring condition codes. Previously, condition codes were assigned to student responses by scoring leadership per Arizona specifications, such as a code noting that the response was left blank, the response was undecipherable or illegible, the response was made in non-English, and so on. Condition codes other than "blank" were then recoded to the lowest score for each dimension for ability estimation. Because the statistical scoring engine cannot assign condition codes, all non-blank responses were assigned a rubric score directly, with responses that would otherwise have received a non-blank condition code being assigned the lowest score point for each dimension.

After scoring all the responses in a set, the scorer reviewed all the responses and modified the scores before committing them to the system. Once the scores had been committed, the set was checked in and responses were routed to other scorers as necessary. Regardless of the specific requirements, however, student responses were not marked as complete until the requisite number of independent scorers had scored the response.

VSC prioritized the available responses in the queue to make sure that the newer responses were placed toward the back of the queue.

9.2.2 HANDSCORING QUALITY CONTROL

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students. The requirements for double scorings are defined to VSC at setup time. MI assumed a double-blind scoring rate of 10 for both the essays and mathematics constructed-response items.

9.2.3 HANDSCORING RELIABILITY AND VALIDITY

MI uses a two-pronged approach to provide the scoring teams for AzM2. First, the scoring leadership recruits qualified, experienced scorers who have successfully scored large-scale assessments for MI, and therefore have experience understanding MI's scoring approach. To ensure reliable and valid handscores, MI puts scoring directors, team leaders, and scorers through a rigorous screening and training process.[50]

Scoring directors, team leaders, and scorers are hired for AzM2 based on experience and performance. Potential new scorers are given a comprehensive content screening for reading and mathematics. This screening is used to identify potential scorers' aptitude for content area and grade level, as well as their reading comprehension and deductive reasoning skills, which are directly related to what they may be scoring. In addition to writing an extemporaneous essay, new hires are required to read a passage and answer questions pertaining to that passage, proofread a sample essay for writing conventions, and solve a series of mathematics problems. The results determine grade and content-area placement if a scorer is to be offered a position on a project. New scorers are selected based on their scores on MI's content screening assessment given for language arts and mathematics projects, the quality of their interview, their work history, and the references provided. The actual qualification for the scorers occurs at the end of training. In addition, the scorers are provided with ongoing validation that they are providing the state with consistency in their scoring using validation sets that are incorporated into the ongoing live scoring.

The Arizona training materials provided for the initial operational ELA scoring were scoring guides comprised of anchor responses, as well as training, qualifying, and recalibration sets approved for use by the state. These training materials were selected as a result of the approval of existing documentation from CAI's Item Tracking System (ITS), which is the repository for all item attributes, including scoring rubrics. New items, approved from the previous year's field test, will be incorporated based on the materials used during the field-test scoring. All materials and selected sets were submitted to Arizona for approval.[51]

MI's scoring directors ensured that ELA scoring guides had detailed annotations to explain how the scoring criteria are to be applied to each response's specific features and why the response should be assigned a particular score. The approach was to focus on the precise scoring rationale, which helped scorers define the lines between score points. All scoring guides and other training materials were presented to Arizona for review and approval before 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 before their use. The training and qualifying sets contained examples of responses from all score points arranged in random score-point order. MI created an appropriate number of training sets and qualifying sets based on the complexity of the item. Essay questions were more complex than single-point mathematics items. The sets were designed to help the scorers learn to apply the criteria illustrated in the scoring guide, ensure that the scorers become familiar with the process of scoring student responses, and assess the scorers' understanding of the scoring criteria before they can begin live scoring. MI worked with Arizona to finalize the number of training and qualifying sets for each item and determine the appropriate qualifying percentage. All

[50] Standard 4.20: The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

[51] Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

scoring decisions and supplemental responses were submitted more than one month before the start of scoring for review and approval by the State.

MI's scoring directors trained both new and experienced scorers within the scoring rooms, giving detailed explanations of all training materials.

MI's online training interface allowed observers from ADE to witness training in real time. Using TurboMeeting software, observers were able to see the responses being trained and discussed as each training set progressed. Observers were also allowed to hear the training through the software's audio function. In addition to observing the training of leadership virtually, representatives from Arizona also traveled to individual scoring sites to observe training in person. This allowed Arizona to observe MI's training techniques and interact with project leadership. The State was able to provide additional guidance on scoring rationale during the training process. These observations allowed MI to further ensure reliability in the handscoring efforts.

Recruited staff followed established training methodologies to ensure the reliability and validity of scores. Scorers were trained as a group, not individually, and all scorers (whether experienced or not) were required to train on all the scoring sets and, at the end of training, pass the qualifying sets with acceptable scores to prove that they were able to understand and apply the criteria. Unless a scorer was trained and qualified for a project successfully, he or she was not permitted to score any student responses.

Each member of MI's scoring staff was required to qualify for the scoring of student responses based on standards established by Arizona. Each staff member was also expected to maintain a consistent level of scoring quality throughout the scoring effort or he or she was released from the project. MI continually monitored performance to guarantee scoring accuracy.

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

Mathematics handscoring was monitored in the same way as essay scoring, with consistent read-behind and validation sets incorporated into the daily scoring schedule to ensure that scorers were providing accurate scoring on a consistent basis.

9.2.4 MACHINE-SCORING VERIFICATION

In addition to the regular ELA handscoring activities, MI also provided a percentage of second readings on machine-scored items. These read-behind scores were used to help ensure consistency and reliability with the ELA machine scoring. Responses requiring read-behinds were generated and sent to MI, where the most experienced scorers, team leaders, and scoring directors provided a second read verification. This process utilized blind scoring, with the scorer unaware of the first score provided by machine. Where scores from Reader 1 (machine) and Reader 2 (human) were in exact agreement or adjacent, the final item score was based on the initial machine read. Where scores from Reader 1 (machine) and Reader 2 (human) were not in exact agreement or adjacent, the final item score was based on the second human read.

10 QUALITY ASSURANCE PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of Arizona's Statewide Achievement Assessment (AzM2) test development, administration, and scoring and reporting of results. This section describes QA procedures associated with the following:

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

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

10.1 QUALITY ASSURANCE IN TEST CONSTRUCTION

Section 4.6 details the form construction process. Each form is built to exactly match the detailed test blueprint and the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the Depth of Knowledge (DOK) with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test. The statistical targets ensure that students will receive scores of similar precision, regardless of which form of the test they receive.

The form construction process is managed through CAI's Form Builder software, which automates important form construction activities to ensure development of equated test forms. Form Builder interfaces with CAI's Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves, and standard error of measurement curves (SEMCs) as test developers build a test map. This helps our content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

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

When submitting test forms for review by the Arizona Department of Education (ADE), Cambium Assessment, Inc. (CAI) produced a form evaluation workbook that includes an evaluation summary checklist, as well as summary statistics and test characteristic graphs.

All bookmaps (test maps), key files, and conversion tables were produced directly from Form Builder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents are generated directly from information maintained in ITS. The information stored in ITS is rigorously reviewed by multiple skilled reviewers to protect against errors. Automated production of these critical files (such as key files) virtually eliminates opportunities for errors.

10.2 QUALITY ASSURANCE IN PAPER-DELIVERED TEST PRODUCTION

Camera-ready documents are prepared after the test items have been selected, composed in forms, and reviewed per the ADE's specifications.

Paper-pencil tests go through a traditional production process. The test booklet production process starts with the creation of test maps (also referred to as bookmaps). The test map is built in the ITS and initiates the production of printed test forms. The process includes the following five steps:

1. The 1×1s (test items printed one per page) are generated based on the test map.
2. Blackline 1 is drafted and reviewed internally.
3. Blackline 1 is delivered to the Department for review and approval.
4. Should any changes be requested in the blackline 1 review, blackline 2 forms are produced, reviewed, and delivered to the ADE.
5. The documents are taken to blueline (camera-ready copy).

Step 1 is entirely automated within ITS. ITS houses destination templates that define the format of the 1×1s and automatically generates these documents based on the test map. At this stage, items are proofread by internal editorial and test development staff and the ADE. Additionally, they are reviewed to verify that all edits from previous rounds of review have been correctly implemented. Any changes required at this stage are entered directly into ITS to ensure consistency across all item uses.

Blackline 1 is a semi-automated process. With the appropriate destination template defined and 1×1 approval, ITS generates a Quark-readable document in the specified format. Through this integration, items are automatically styled with fonts, graphics, spacing, and other formatting specifications outlined in the ADE's style guide. Our production staff may adjust page layout, including instructions, borders, and other elements, to meet the ADE's guidelines. At this stage, reviewers check the document layout and formatting. Should any egregious errors be found in the content of an item, changes must be entered into ITS and the item must be re-exported to ensure consistent item use across all test forms. Changes to blackline 1 require a second blackline proof. Changes to subsequent blackline proofs require sign-off by senior management and the ADE.

The final QA step before printing is the blueline, or camera-ready copy, review stage. During this step, CAI and the ADE's staff review proofs from the print vendor, verifying that the file to be printed matches the previously approved blackline proof. At CAI, in addition to reviews by test development and forms production staff, two members of the technical team—who have not seen the items previously—independently take the tests. This process forces a close look at the items and gives a final opportunity to verify the keys.

During the production and review process, test book blacklines are accompanied by answer document blacklines, which are produced by MI. Answer documents reflect the demographic fields required by the ADE, as well as fields for pre-code labels and the scannable marks required for accurate data collection. The item sequence is based on test maps and corresponds directly with test books.

All blacklines in CAI's production queue are controlled by an electronic version-control server system that ensures that only the current version is immediately available to our production staff, preventing version-control errors. Like CAI's ITS, which controls and tracks all changes to items, this production system maintains historical records (including all older versions), which senior production staff can access if necessary. Each blackline after blackline 1 and the blueline (camera-ready copy) is automatically compared with the immediately preceding version using a PDF comparison tool that highlights all changes. This step has proved useful for identifying unintended changes made during the revision process. Such changes are difficult to detect because they can appear anywhere in a document and may be subtle. The PDF comparison tool highlights these changes so differences between versions can be mapped to an intended revision. All materials delivered will go through this process, ensuring that the ADE will receive error-free materials for review and that any changes requested by the ADE are implemented promptly and accurately.

At each of the review stages, proofs will be accompanied by proof tickets that identify the document being reviewed, its review stage, the scheduled and actual delivery dates, and the return date. Sign-off by the ADE is required at each stage before proceeding with subsequent steps.

10.3 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

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

1. Content for online delivery shares some processes with paper-pencil versions, but also requires additional, unique steps.
2. CAI's online Test Delivery System (TDS) must deliver the content reliably (and, with the right tools, the accommodations, layouts, etc.).

10.3.1 PRODUCTION OF CONTENT

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

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

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. Forms are finalized using the process described in Section 4.6, and final forms are approved in our Form Builder software.
3. Complete test packages are created with our test packager, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
4. Forms are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the field.
5. The final system is deployed to a staging environment accessible to ADE for user acceptance testing (UAT) and final review.

10.3.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

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

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

1. Content can be revised and re-approved for web display.
2. Online style sheets can change to revise the layout of all items on the test.

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

10.3.3 APPROVAL OF FINAL FORMS

Section 4.6 describes our process for constructing operational test forms, including the approval of test forms by the ADE. The forms are built in Form Builder (a component of ITS), and upon approval, they are ready for preliminary publication.

10.3.4 PACKAGING

The test packaging system performs two simultaneous roles in the preparation of computer-based products: It compiles the form definitions and other information about how the test is to be administered (e.g., where any embedded field-test items might be inserted) and pulls together the content packaged during web approval.

The test packager assigns form identifiers to each form, evaluates the form against the blueprint, and performs a quality check against the content. The content quality check includes checks to see that every asset (e.g., graphics) referenced in the item is included in the package, confirms that the item has not changed since it was web approved, and ensures that the items have received all the approvals necessary for publication.

10.3.5 PLATFORM REVIEW

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

10.3.6 USER ACCEPTANCE TESTING AND FINAL REVIEW

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

10.3.7 FUNCTIONALITY AND CONFIGURATION

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

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

1. Testing on the system itself to ensure function, performance, and capacity
2. Capacity planning
3. Continuous monitoring

CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the testing window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI 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 are captured for each assessed student—data about how long it takes to load, view, or respond to an item. All this information is logged, as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

10.4 QUALITY ASSURANCE IN DOCUMENT PROCESSING

10.4.1 SCANNING ACCURACY

When test documents are returned to be scored, they must be scanned first. When they were scanned, a quality control sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) was created so that all possible responses and all demographic grids were verified, including various typical errors that required editing via MI’s Data Inspection, Correction, and Entry (DICE) application program. This structured method of scan testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner, editing process (validation and data correction), data transfer to the project database, and scoring were all accurate according to the reporting rules provided by ADE.

10.4.2 QUALITY ASSURANCE IN EDITING AND DATA INPUT

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

- Score key verification
- Post analysis of item keys
- Response analyses to determine score frequency distribution by item verification of bank values of item statistics
- Live data checks to verify that data/results conformed to approved specifications comprehensive software test plan

- Double data entry correction process to verify student response and demographic information report data verification
- Reviewed and proofread all electronic and printed report deliverables

MI utilized a double data correction process to achieve the highest level of quality and accuracy in both Arizona CBT and PBT assessment student data. Data correction operators used their sophisticated 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 corrections.

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

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

MI developed verification routines as part of their standard data validation to detect duplicate student tests in the assessment, whether in a single Local Educational Agency (LEA) or across LEAs, and student moves between schools. MI staff then worked closely with the ADE to resolve these discrepancies through processes called Barcode Processing and Tested Roster. These processes and the business rules governing them are described in a set of requirements developed in conjunction with the ADE.

10.5 QUALITY ASSURANCE IN DATA PREPARATION

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

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

Data pass directly from the QM to the Database of Record (DOR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator is the tool that is used to pull data from the DOR for delivery to ADE and their QA contractor. CAI psychometricians ensure that data in the extract files match the DOR before delivery to the ADE.

10.6 QUALITY ASSURANCE IN TEST FORM EQUATING

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

To ensure security, all data is shared using ADE's Secure File Transfer Protocol site.

Prior to operational work, CAI produces simulated datasets for testing software and analysis procedures and shares it with the ADE and the QA contactor. All parties complete a dry run of calibration and post-equating activities and compare results. The practice runs serve two functions:

1. To verify accuracy of program code and procedures.
2. To evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.

Following the completion of these activities and the resolution of questions that arise, analysis specifications are finalized.

10.7 QUALITY ASSURANCE IN SCORING AND REPORTING

10.7.1 QUALITY ASSURANCE IN HANDSCORING

10.7.1.1 DOUBLE SCORING RATES, AGREEMENT RATES, VALIDITY SETS, AND ONGOING READ-BEHINDS

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students.

MI's Virtual Scoring Center (VSC) software, described in Section 9.2.1, provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read-behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses if they require additional monitoring.

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

We monitor their scoring intensively to ensure that all responses are scored accurately. If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, he or she is given interactive feedback and mentoring on the responses that have been scored incorrectly and is expected to change the scores. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

If a scorer's interrater agreement rate falls below the expected standard, the scorer will be re-trained. Should the scorer still be unable to score reliably, the scorer is assigned to another, non-Arizona-related project or dismissed.

In addition to using validity responses (also known as calibration or anchor responses) as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be pulled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following Arizona's review and approval. MI periodically administers validity sets to each of MI's scorers working on the scoring effort. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the State.

With the VSC program, the way in which the student responses are presented prevents scorers from having any knowledge about which responses are being single- or double-read, or which responses are validity set responses. A performance threshold of 75 is set to specify validity agreement standards as well as the frequency and total number of validity responses evaluated by each scorer based on client specifications.

10.7.1.2 HANDSCORING QA MONITORING REPORTS

MI generates detailed scorer status reports for each scoring project utilizing a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Arizona. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated real-time reports that show both daily and cumulative (project-to-date) data are available. These reports are available to Arizona 24 hours a day via a secure website. Project leadership reviews these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

Scorers are released when they are unable to demonstrate the ability to score responses according to the criteria and standards established by MI and Arizona and perform to the level of client expectation. Should Arizona request that certain responses be rescored, we are prepared to do so, if necessary. The reporting system can produce a list of all the responses a selected scorer has scored. In these situations, all responses scored by a scorer during the time frame in question can be identified, reset, and released back into the scoring pool. The aberrant scorer's scores are deleted, and the responses are redistributed to other qualified scorers for rescoring.

10.7.1.3 MONITORING BY THE ARIZONA DEPARTMENT OF EDUCATION

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

10.7.1.4 IDENTIFYING, EVALUATING, AND INFORMING THE STATE ON ALERT PAPERS

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the test taker or those around him or her. We also flag potential security breaches identified during scoring. For possible dangerous situations, scoring project management and staff employ a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

This process is also used to notify Arizona of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow up. The ADE has processes in place to communicate the presence of and information contained within the alert paper to student's school official.

10.7.2 TEST SCORING

CAI verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the State. The ability of each of these simulated students is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they provide a check of the full range of item responses and test scores in fixed-form tests, as well. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the Online Reporting System (ORS), we merged item response data with the demographic information taken from previous year assessment data. If current year enrollment data are available by the time simulated data files are created, we verify online reporting using current-year testing information. By populating the simulated data files with real school information, it is possible to verify that specific school types and special districts are being handled properly in the reporting system.

Specifications for generating simulated data files are included in the Analysis Specifications document submitted to and approved by the ADE each year. Although the ADE does not currently provide immediate reporting, review of all simulated data is scheduled to be completed before the opening of the test administration, so that the integrity of item administration, data capture, item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the testing window, a series of QA reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational testing window.

An additional set of forensic analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, TA, and school level that may indicate cheating. The QA reports can be generated on any desired schedule. Item analysis reports are evaluated frequently at the opening of the testing window to ensure that items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit 10.7.2.1 presents an overview of the QA reports.

Exhibit 10.7.2.1 Overview of Quality Assurance Reports

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

10.7.2.1 ITEM ANALYSIS REPORT

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

Item p-Value. For dichotomous items, the proportion of students selecting each response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for dichotomous items and polyserial correlations for polytomous items are computed. CAI psychometrics staff members evaluate all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between observed and expected score given the student's ability is computed for each item. The residuals for each are averaged across all students, and the average residual is used to flag an item. The item fit statistic is computed as follows:

Let X_{ij} be the variable for the response of student j to item i, and $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ be the probability that student j gets a score of x_{ij} to item i given his or her ability estimate $\hat{\theta}_j$. $P\left(X_{ij} = x_{ij}\middle|\hat{\theta}_j\right)$ is calculated using Rasch model

$$P\left(X_{ij} = x_{ij}\middle|\hat{\theta}_j\right) = \frac{\exp\left(\hat{\theta}_j - b_i\right)}{1 + \exp\left(\hat{\theta}_j - b_i\right)},$$

where b_i is the difficulty parameter of item i. If item i is a polytomously scored item, $P\left(X_{ij} = x_{ij}\middle|\hat{\theta}_j\right)$ is calculated using the Master's Partial Credit Model,

$$P\left(X_{ij} = x_{ij}\middle|\hat{\theta}_j\right) = \frac{\exp\sum_{k=0}^{x_{ij}}(\hat{\theta}_j - b_{ki})}{\sum_{l=0}^{m_i}\exp\sum_{k=0}^{l}\left(\hat{\theta}_j - b_{ki}\right)}.$$

The expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as

$$E(X_{ij}|\hat{\theta}_j) = \sum_{x_{ij}=0}^{m_i} x_{ij}P(X_{ij} = x_{ij}|\hat{\theta}_j).$$

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

$$\delta_{ij} = x_{ij} - E(X_{ij}|\hat{\theta}_j).$$

The statistic δ_{ij} is aggregated across all n students for item i,

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

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

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > .96$$

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

10.7.2.2 FORENSIC ANALYSIS

Another component in the suite of QA reports is geared toward detecting testing irregularities that may indicate possible cheating. The forensic analysis components of the QA reports are described in detail in Section 5.6. Evidence evaluated includes changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and were determined in partnership with ADE. Analyses are performed at the student level and summarized for each aggregate unit, including testing session, TA, and school.

10.7.3 REPORTING

Scores for online assessments are assigned by automated systems in real time. For the 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 IRT parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Once both online and handscoring items have passed through their validity and quality checks, the handscored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the QM system, where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies Arizona-specific scoring rules and assigns scores from the calibrated items, including calculating performance-level indicators, subscale scores, and other features, which then pass automatically to the reporting system and DOR. The scoring system is tested extensively before deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

After passing through the series of validation checks in the QM system, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QM checks and are uploaded to the DOR are they passed to the ORS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all the QM system's validation checks and ADE's independent data verification checks.

11 REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014*). Standards for educational and psychological testing*. Washington, DC: Author.

AzMERIT Testing Conditions, Tools and Accommodations Guidance Manual. Arizona Department of Education (2017, February). Retrieved from: https://cms.azed.gov/home/GetDocumentFile?id=5836103eaadebe14087eb770

Bentler, P.M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin, 107*(2), 238–46.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Camilli, G., & Shepard, L. (1994). Methods for identifying biased test items. California: Sage Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. doi:10.1080/10705510701301834

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Concon, W. (2013). Large-scale assessment, locally-developed measured, and automated scoring of essays: Fishing for the red herrings? *Assessing Writing, 18*(1), 100–108.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Estrada S., Burnham C., Feld J. K., Bergan J. R., & Bergan J. R. (2015). *Can Local Assessment Data Be Successfully Used as Part of an Arizona A-F Accountability System?* Leawood, KS: Assessment Technology Incorporated (ATI). Retrieved from: https://azsbe.az.gov/sites/default/files/media/ATI-Feasibility.pdf

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement. 13*, 253–264.

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

Karkee, T., Lewis, D. M., Hoskins, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lin, Z., Jiang, T., & Rijmen, F. *Person Fit z-statistics for Rasch Testlet Model*. Manuscript in preparation.

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement. 5*(1), 95–110.

Livingston, S.A., & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement, 32*(2), 179–197.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *8*(4), 453-461.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443–452.

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.). *Handbook of Structural Equation Modeling* (pp. 380–392). New York: Guilford Press.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115–135.

Scott, L. (2015). *Analysis of Mode Comparability of AzMERIT's Online and Paper Administrations for Spring 2015.* Arizona Department of Education, Recommending AzMERIT Performance Standards (pp. I-28–I-40). Retrieved from http://www.azed.gov/assessment/files/2014/11/spring-2015-azmerit-standard-setting_091415-full-report.pdf.

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

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

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

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 11*, 265–276.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments: Synthesis Report* (No. 44). Minneapolis, MN: National Center on Educational Outcomes.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149.

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.* Presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wei, Y., & Higgins, J. P. (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. [Research Support, Non-U.S. Government]. *Stat Med, 32*(7), 1191–1205.

Wesolowsky, G.O. (2000). Detecting Excessive Similarity in Answers on Multiple Choice Exams. *Journal of Applied Statistics, 27*, 909–921.