AzSCI

Arizona's Science Assessment

2025
Technical Report

Submitted to the
Arizona Department of Education
October 2025

Pearson

FOREWORD

This technical report provides information on Arizona's Science Test (AzSCI), the statewide accountability assessment administered to Arizona students in grades 5, 8, and 11 each spring to assess their performance on the Arizona Science Standards. The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

The Arizona Department of Education (ADE) is a service organization committed to raising academic outcomes and empowering parents. The ADE Assessments agency is responsible for statewide assessment of students enrolled in Arizona public schools, working closely with educators in the development and administration of statewide assessments. All Arizona public schools, including district schools and charter schools, are required to properly administer state and federally mandated assessments.

Arizona Department of Education
1535 West Jefferson Street
Phoenix, AZ 85007
(602) 542-5057 (Assessments)
https://www.azed.gov/

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: INTRODUCTION

This technical report documents evidence of reliability and validity for the spring 2025 administration of Arizona's Science Test (AzSCI) in grades 5, 8, and 11 (Cohort 2026) to support test users in evaluating the intended purposes, uses, and interpretations of test scores. The evidence includes descriptions of the test design, development, and administration procedures; student test results; and psychometric analyses, including calibration, equating, and scaling to ensure comparable test results across different test forms and administrations. The report concludes with a synthesis of the evidence to construct the validity argument for the AzSCI assessments.

1.1. Assessment Overview

AzSCI is the statewide summative science achievement test administered each spring to Arizona students in grades 5, 8, and 11. It is a criterion-referenced assessment aligned with the Arizona Science Standards, designed to measure student progress toward achievement of the standards. Students do more than answer recall questions about science; they apply the practices, or behaviors, of scientists and engineers to investigate real-world phenomena and design solutions to problems. AzSCI is a grade band assessment in which students in grade 5 take the assessment based on the standards for grades 3–5, students in grade 8 take it based on the standards for grades 6–8, and students in grade 11 take it based on the standards for high school.

AzSCI is administered online as a computer-based assessment, with paper-accommodated forms available as needed. The needs of the student are also addressed through other supports, including accessibility features built into the online platform and accommodations such as assistive technology, a scribe, and/or sign language. Each AzSCI test form includes 50 operational items consisting of multiple-choice and technology-enhanced item types, administered in either independent or cluster item sets. Field test items are also embedded on each assessment that do not count toward students' scores.

Student performance is reported as an overall scale score and an accompanying performance level classification. The four performance levels are *Minimally Proficient*, *Partially Proficient*, *Proficient*, and *Highly Proficient*. Students performing at Levels 3 and 4 demonstrate proficiency in the assessed content. Student performance on reporting categories is reported as one of three levels of mastery: *Below Mastery*, *At/Near Mastery*, or *Above Mastery*. The primary intended score interpretation of the assessment is that AzSCI test scores provide reliable and valid information about important knowledge and skills that students are attaining in Physical Science, Life Science, and Earth and Space Science. While ultimate use of the test scores is determined by Arizona educators and other stakeholders, the primary intended uses of the AzSCI test scores include the following:

- Schools and districts use the AzSCI assessment and its results to monitor trends in student performance and design professional development for teachers.
- Teachers use the AzSCI assessment and its results to integrate assessment with their instructional planning.
- Parents/guardians use the AzSCI assessment and its results to get information about what their child knows and can do and their child's progress from year to year.

1.2. Background

The Arizona Science Standards were adopted by the State Board of Education in 2018. AzSCI replaced the previous Arizona science assessment known as Arizona's Instrument to Measure Standards Science (AIMS Science) aligned to the 2004 standards. The changes for AzSCI to accommodate the 2018 standards included measurement targets, test designs, item types, and test administration conditions. To support this effort, Pearson, WestEd, the Arizona Department of Education (ADE), and Arizona educators collaborated to develop item specifications and blueprints to guide the test development process.

A pilot test was conducted in 2020 to try out a small group of items aligned to the 2018 standards, evaluate psychometric characteristics of the items and item clusters, and collect data about student experiences during the test administration. Information collected from the pilot was used to develop items for the full standalone field test in spring 2021. Similar to the pilot, the purpose of the full standalone field test was to try out a large group of items aligned to the 2018 standards; evaluate psychometric characteristics of the items, different item types, and item clusters; and build an item bank for the first operational administration in spring 2022.

1.3. Testing Requirements

AzSCI is an accountability assessment developed to fulfill requirements of both state and federal laws (State Law ARS 15-741; Federal Law: 34 CFR 200.2 *Participation in Assessments*) that mandate all public-school students participate in the assessments that measure student achievement of grade-level content standards. Students with significant cognitive disabilities and whose Individualized Education Program (IEP) designates them as eligible for an alternate assessment, the Multi-State Alternate Assessment (MSAA), should not be administered the AzSCI assessment.

1.4. Educator Involvement

This section addresses the involvement of Arizona educators in test development as indicated by Standard 4.8 of the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Arizona educators were involved in many steps of the process, as shown in Table 1.1 that presents the major events regarding the development, administration, and reporting of the AzSCI assessment in 2024–2025.

Arizona educators have several opportunities to participate in meetings and provide feedback on assets developed for field testing. The content and bias and sensitivity community reviews enable community members, including past and present Arizona educators, to evaluate items for content, standards alignment, grade-level appropriateness, and bias and sensitivity. These meetings represent a continuation of stakeholder involvement in the development process. Arizona educators are also involved in the development of the performance level descriptors (PLDs) and the establishment of cut scores during the standard setting meetings as needed.

Table 1.1. Schedule of Major Events

Event	Date(s)
Educator Item Review Committee	June 10–13, 2024
Community and Parent Item Review Committee	June 20–21, 2024
Form Construction	Aug. 5–9, 2024
Administration Training	November 18, 2024 (live date)
Spring 2025 Administration Window	March 17 – April 11, 2025
Release of Electronic Score Reports and Student Data Files	May 22–23, 2025
Data Review Committee	June 16–18, 2025
Release of Paper Reports	June 12, 2025

Chapter 2: TEST DESIGN

This chapter provides information regarding test design as indicated by Standards 1.11, 4.0, 4.1, 4.12, 12.4, and 12.8 (AERA et al., 2014). This transparency into the intended construct and rationale behind the assessments supports the valid interpretations of the test scores and alignment with Arizona's standards-based educational system.

Accessibility was the foundation of the AzSCI test design to make sure all students have access to the content based on the Arizona Science Standards, which begins with rigorous curriculum, instructional resources, and training for teachers. Principles of Universal Design are adhered to throughout the item and test creation process to accommodate the needs and abilities of all learners. AzSCI is available to be administered in online settings including group, small group, or one-on-one settings. It is also available in appropriate accommodations including American Sign Language (ASL), Braille, large print, or Special Paper Version (SPV) in which the proctor transcribes a student's responses into an online form.

2.1. Academic Content Standards

The Arizona Science Standards, available online at https://www.azed.gov/standards-practices/k-12standards/standards-science, were written by a group of educators, content experts, and community members to reflect an increase in rigor of the standards. Guided by *A Framework for K–12 Science Education* (National Research Council, 2012) and *Working with Big Ideas of Science Education* (Harlen, 2015), the standards provide a vision and structure to prepare Arizona students to be scientifically literate and college and career ready, outlining what all students need to know, understand, and be able to do by the end of high school and reflecting the following shifts for science education:

- Organize the standards around 13 core ideas and develop learning progressions to build scientific literacy coherently and logically from kindergarten through high school.
- Connect the three dimensions of Core Ideas, Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs) to make sense of the natural world and understand how science and engineering are practiced and experienced.
- Focus on fewer, broader standards that allow for greater depth, more connections, deeper understanding, and more applications of content.

The Arizona Science Standards are organized around the Core Ideas in Physical Science, Life Science, and Earth and Space Science in addition to the SEPs and CCCs. The Core Ideas encompass the content that occurs at each grade and provide the background knowledge for students to develop sense-making around phenomena. They center around understanding the causes of phenomena in physical, life, and earth and space science; the principles, theories, and models that support that understanding; engineering and technological applications; and societal implications. Each standard embeds an SEP into a Core Idea and pairs the standard with one or more CCC.

The SEPs describe how scientists investigate and build models and theories of the natural world or how engineers design and build systems. They reflect science and engineering as they are practiced and experienced. There are eight practices:

1. Ask questions and define problems
2. Develop and use models
3. Plan and carry out investigations
4. Analyze and interpret data
5. Use mathematics and computational thinking
6. Construct explanations and design solutions
7. Engage in argument from evidence
8. Obtain, evaluate, and communicate information

CCCs cross boundaries between science disciplines and provide an organizational framework to connect knowledge from various disciplines into a coherent and scientifically based view of the world. They build bridges between science and other disciplines and connect the Core Ideas and SEPs throughout the fields of science and engineering. There are seven CCCs:

1. Patterns
2. Cause and effect
3. Structure and function
4. Systems and system models
5. Stability and change
6. Scale, proportion, and quantity
7. Energy and matter

2.2. Item Specifications

The AzSCI item specifications, available online at https://www.azed.gov/assessment/sci/, guide the item development process by defining the content limits and item types that can be used to assess each standard. During the item development and review processes, items are compared to the item specifications to ensure their alignment to the standard, grade-level appropriateness, and adherence to the content limits. The item specifications were developed in 2018 as part of Pearson and WestEd's evaluation of the new Arizona Science Standards to make suggestions to ADE that would guide item development and test design. Subsequent review and development of the specifications was an iterative process involving ADE, Pearson, and a committee of Arizona educators. By September 2019, the specifications were approved and continue to be updated each year as needed.

2.3. Test Blueprint

The test blueprint defines the content and structure of the assessment by domain, SEP, grade, and cognitive complexity and guides item selection during form construction based on the goal of testing every standard within a three-year window. To address this goal, the Pearson content team created a tracking spreadsheet for each grade that lists each standard and marks the standards selected for use each spring. This allows Pearson to identify which standards remain to be tested in future administrations to adhere to the three-year cycle. The AzSCI blueprints define the following information:

- A range for the number of items to be assessed from each content domain and SEP
- A range for the number of items based on item types and cognitive complexity
- A range for the number of items for each grade within a grade band
- The total number of points per item type

The Pearson content team submitted the initial blueprint drafts to ADE for review, with adjustments made as requested. In August 2020, an advisory committee of Arizona educators provided feedback on the draft, which was subsequently approved and used for item development. The blueprint was revised in 2021–2022 to better reflect the distribution of the standards; rather than allocating an equal percentage across Physical Science, Life Science, and Earth and Space Science, the standards coverage dictates the percentage across the domains. For example, a higher percentage of the test is dedicated to Physical Science that has a greater percentage of standards.

Table 2.1, Table 2.2, and Table 2.3 present a summary of the AzSCI blueprints by domain, SEP, and on- and off-grade standards for grades 5 and 8.

Table 2.1. AzSCI Blueprint Summary by Domain

Domain	Grade 5	Grade 8	Grade 11
Physical Science	40–48%	36–44%	32–40%
Life Science	28–36%	30–38%	34–42%
Earth and Space Science	20–28%	22–30%	22–30%

Table 2.2. AzSCI Blueprint Summary by SEP

Practice (and Categories)	Grade 5	Grade 8	Grade 11
Investigating (Asking Questions and Defining Problems, Planning and Carrying Out Investigations, Using Mathematic and Computational Thinking, and Analyzing and Interpreting Data)	20–42%	14–26%	16–26%
Sensemaking (Developing and Using Models and Constructing Explanations and Designing Solutions)	26–42%	40–60%	34–48%
Critiquing (Engaging in Argument from Evidence and Obtaining, Evaluating, and Communication of Information)	20–34%	18–30%	24–38%

Note. Assessment reporting categories for SEPs may vary.

Table 2.3. AzSCI Blueprint Summary for On- and Off-Grade Standards (Grades 5 and 8)

Grades	#Items (Goal)	%Items (Goal)	#Items (Range)	%Items (Range)
On-Grade Standards: Grades 5 and 8	30	60%	28–32	56–64%
Lower-Grade Standards: Grades 4 and 7	10	20%	8–12	16–24%
Lower-Grade Standards: Grades 3 and 6	10	20%	8–12	16–24%

2.4. Cognitive Complexity

The performance expectations for the Arizona Science Standards are written with high cognitive complexity, incorporating knowledge with practice while unifying concepts to develop scientific explanations. Assigning the cognitive load to AzSCI items requires the use of a model that accounts for how the dimensions interact, the degree of independence with which students apply the dimensions in exploring and explaining phenomena, and the dimensions' connection with the context of the problem presented for student interaction. As such, Arizona modified the Task Analysis Guide in Science (TAGS) models (Tekkumru-Kisa et al., 2015) to recognize that cognitive demand increases as the number of integrated dimensions increases. An item′s cognitive complexity is classified according to three levels: Doing Science Tasks, Guided Science Tasks, and Scripted Science Tasks. Table 2.4 identifies the AzSCI operational targets.

Table 2.4. AzSCI Cognitive Complexity Operational Targets

TAGS Level	Range (All Grades)
Doing Science Tasks: Students are required to DO science by using practices to DEVELOP an understanding of a scientific or engineering phenomenon. Students must develop a model, explanation, or argument from raw data or information. Students must be able to determine which data or information is appropriate and how to use it.	0–5%
Guided Science Tasks: Students use higher-level thinking to work through guided or scaffolded tasks. Students are told what information (model, data, etc.) to use or are provided with information and then required to develop the actual answer.	66–84%
Scripted Science Tasks: Students follow a script (defined actions or procedure) to complete a task.	16–28%

2.5. Test Structure

Table 2.5 summarizes the AzSCI test design for all grades. Each test form has 60 items (50 operational + 10 embedded field test). The 50 operational items on the base form are worth a total of 55 raw score points, while the field test items are not counted toward students' test scores. The assessment is administered in three units, each with 20 items. All items on the assessment are associated with a scientific phenomenon presented in a stimulus or series of stimuli. The items are part of one of two sets: (a) an independent set that includes at least two non-dependent items associated with one or more short stimuli or (b) an item cluster set that includes five items associated with longer, more complex stimuli.

The online platform allows for the use of a variety of technology-enhanced item types where students can apply critical thinking skills to demonstrate a deeper understanding of the three dimensions of the Arizona Science Standards. As such, the items may be multiple-choice (MC), technology-enhanced (TE), or two-part evidence-based selected response (EBSR). EBSRs may be two-part dependent (TPD) or two-part independent (TPI). MC, TE, and TPD items are worth 1 point, while TPI items are worth 2 points. TE interactions include bar graph, multiple select, inline choice, hot spot, graphic gap match, gap match, line graph, match, match table grid, and point graph. At least one item in each unit is a 2-point TPI item.

Table 2.5. AzSCI Test Design

Unit	#OP Items from Independent Sets	#OP Items from Cluster Sets	#FT Items
1	15 items (from at least five IN sets)	n/a	5 items (from 2 IN sets): MC: 0–3 items TE: 0–3 items 1 TPI or TPD item
2	n/a	15 items (from 3 CL sets)	5 items (from 1 CL set): MC: 0–3 items TE: 0–3 items 1 TPI or TPD item
3	n/a	20 items (from 4 CL sets)	n/a

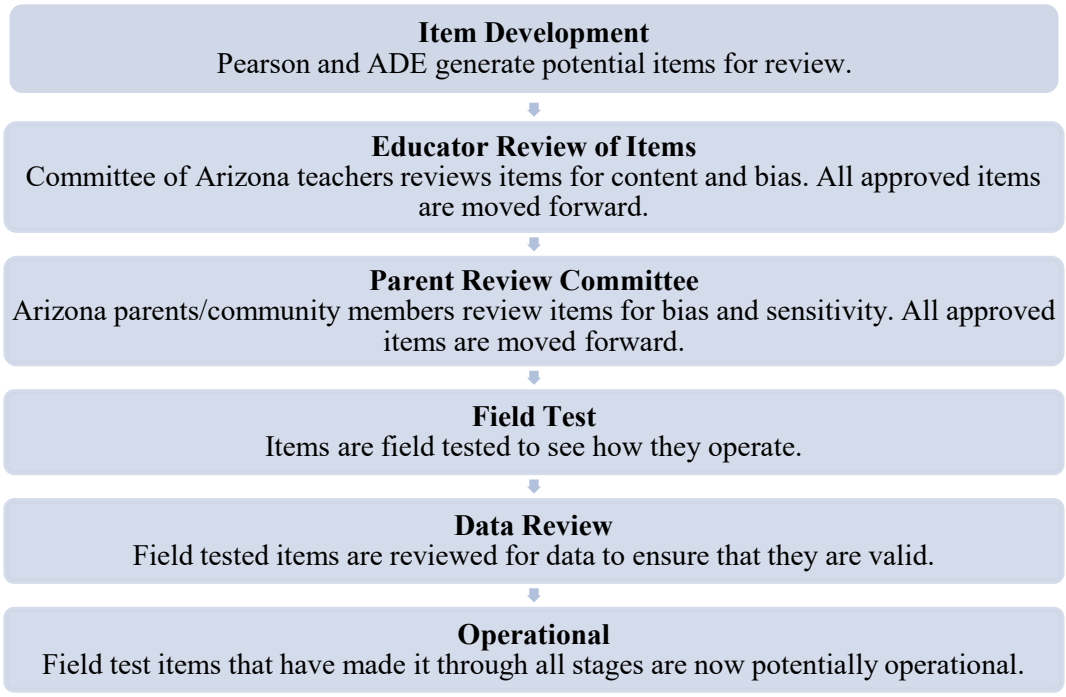
Unit	#OP Items from Independent Sets	#OP Items from Cluster Sets	#FT Items
Form as a Whole	15 items: MC: 3–8 items TE: 3–8 items TPD: 1–3 items TPI: 1–2 items	35 items: MC: 8–17 items TE: 8–17 items TPD: 3–4 items TPI: 3–4 items	10 items: MC: 0–6 items TE: 0–6 items TPD or TPI: 2 items

Note. OP = operational, FT = field test, IN = independent set, CL = cluster set, MC = multiple-choice, TE = technology-enhanced, TPD = two-part dependent, TPI = two-part independent, n/a = not applicable

Chapter 3: TEST DEVELOPMENT

This chapter addresses Standards 1.11, 3.2, 3.6, 4.0, 4.4, 4.6, 4.7, 4.8, 4.10, 4.12, 7.0, 7.2, 12.4, and 12.8 (AERA et al., 2014) regarding item development and test construction. ADE and Pearson worked together to construct the AzSCI tests based on the steps depicted in Figure 3.1.

Figure 3.1. Item Development Process



Items used to develop the operational test forms were drawn from the operational ready items in the item bank. Because the AzSCI test is set-based, accompanying stimuli were also needed for the items. Independent sets are associated with one or two brief stimuli, and cluster sets have several stimuli that are more detailed. The item development process is iterative, allowing for multiple opportunities for review of the items by various stakeholders including ADE and external item content and bias review participants. Newly developed items are then field tested during the spring administration, followed by a data analysis and data review process with Arizona stakeholders. Items that pass data review are added to the operational item bank.

This multistage development and review process provides ample opportunity to evaluate items for accessibility, appropriateness, and adherence to the principles of Universal Design. In this way, accessibility serves as a primary area of consideration throughout the item development process. This focus on accessibility is critical in developing an assessment that allows for the widest range of student participation as educators seek to provide access to the general education curriculum and foster higher expectations for students.

3.1. Content Development and Management Tool

The item pool and test development process are managed within Pearson's Assessment Banking and Building solutions for Interoperable assessments tool (ABBI) that acts as a content development and management tool, item bank, and publication system supporting both paper-pencil and online publication. The item development workflow is designed to move items and assets from inception through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes at each review and maintains previous versions of each item. As items travel through the review process, every version of each asset is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

ABBI 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. Forms are also built in ABBI. After items are used, ABBI stores the resulting statistics, including exposure statistics and classical and item response theory (IRT) statistics.

The item development process is predicated on a high level of interaction between test developers at Pearson and ADE, as well as with Arizona educators and stakeholders. Pearson's ABBI manages item content throughout the entire lifecycle of an item. It also manages item content beyond the operational life of the item, including items identified for use in sample tests or other training materials. ABBI provides on-demand reports of the content and item bank status. Each item is directed through a sequence of reviews and approvals by Pearson and ADE before it is identified for field test or operational administration.

3.2. Item Bank Analysis

Pearson conducted an item bank analysis at the start of the test development cycle to identify gaps that were then used to inform creation of an asset development plan to determine the priorities for new item development. For all items, item statistics and metadata were evaluated. The second step was to review all the additional items included in the item bank. Standards that were underrepresented in the item bank or represented by items with poorly performing statistics were identified as candidates for item development. Based on the gap analysis, Pearson's assessment specialists identified the following goals for development:

- Increase any standard with less than five items.
- Increase standards covered by independent items.
- Even out the number of item types.
- Increase standards covered in each domain under 60% of the total items (grades 5 and 8).
- Even out the number of items at all levels of PLDs.
- Increase the number of braillable and signable items.

3.3. Item Development

The initial phase of item development involved drafting science phenomena, a task led by Pearson and subsequently reviewed by ADE. The process continued with the creation of an outline detailing how these phenomena would be presented to students, with Pearson preparing the initial draft and ADE providing feedback. This collaborative approach was consistently applied to the development of items and stimuli; Pearson was responsible for authoring, after which the completed materials were submitted to ADE for approval.

An integral part of this process was a review of all assets by Pearson research librarians who verified accuracy and by Pearson copy editors who reviewed for clarity and correct use of grammar, punctuation, and spelling. All asset creators and reviewers at Pearson also applied the principles of Universal Design to meet the goal of maximizing accessibility and minimizing construct-irrelevant demands for all items. To meet these goals, text complexity was controlled, graphics were designed to be clear, and subject matter that might affect the student's performance was monitored. Pearson also paid close attention to respecting the diverse cultures of the American Indian tribes in Arizona, particularly to the presentation of topics related to animals.

All items aligned to the 2018 standards and SEPs, with some items also aligning to the CCCs. The compilation of items across item sets, both independent and cluster, support a multi-dimensional alignment.

3.4. Item Review

ADE review was the first of several external reviews of the newly developed items. Educators and community members also had opportunities to participate in review committees known as Item Review Committees (IRCs). The IRC Committee Review (i.e., the content and bias review) allowed educators to apply their familiarity with Arizona students and the Arizona Science Standards to provide feedback on the accuracy and appropriateness of the item and stimulus content. An IRC Community Review (i.e., the bias and sensitivity review) also allowed parents and other community stakeholders to review assets.

Committee members received training from Pearson content specialists and were provided with resources, including a checklist, to guide the review process. All feedback was recorded in ABBI. The overall goals for both committees were to confirm alignment to the standards, ensure that assets had no bias or sensitivity issues, and revise the assets as needed to be appropriate for Arizona students. An additional benefit of these interactions was that Pearson gained insight to help guide future item development.

ADE and Pearson engaged in a reconciliation process to review committee feedback. Pearson revised assets based on ADE guidance and made the newly edited versions available for ADE review. With ADE approval, the assets went through a final editorial review at Pearson to confirm that they met expectations.

3.5. Form Construction

Once the newly developed items were ready for field testing, the next step was to construct the test forms, beginning with selecting and positioning the items. Form construction involved Pearson and ADE content specialists and an ADE accessibility expert. In spring 2025, grade 5 had 15 field test forms with 10 embedded field test items per form (i.e., each form has the same 50 operational items but different field test items), grade 8 had 16 field test forms with 10 embedded field test items per form, and grade 11 had 14 field test forms, each with 10 embedded field test items.

3.5.1. Preparation for Item Selection

Parameters based on the test construction blueprint for each grade were loaded into ABBI by Pearson psychometricians and verified by Pearson assessment specialists. Different test map views were also configured based on the specific needs of various users, including Pearson assessment specialists, ADE and Pearson psychometricians, and Pearson publishing teams. Test maps for each stage were maintained throughout all steps of production. Pearson updated the test maps when any replacements or changes to items or item metadata were made.

Pearson psychometricians had previously loaded statistics from the previous administrations, and Pearson assessment specialists had updated the ABBI item status used to indicate eligibility for operational or field test selection based on the results from data review. Item statistics included, but were not limited to, classical difficulty (*p*-value) and IRT difficulty (Rasch), item discrimination (point-biserial correlation by total score), the Rasch model fit indices (infit), differential item functioning (DIF) flags as a measure of possible bias, and distractor analysis.

3.5.2. Item Selection and Positioning

The overriding goal in selecting items for the forms was adhering to the blueprint requirements. Additional criteria for item selection included item positioning and both content and statistical considerations. For each grade, a Pearson assessment specialist did an initial pull of operational items using the tools embedded in ABBI to verify blueprint alignment and acceptable statistics according to the test construction specifications. A different assessment specialist reviewed the form and provided feedback, identifying issues such as clueing. After issues were resolved, the Pearson and ADE psychometrician reviewed the form and provided feedback to meet the blueprint and other statistical requirements. This process repeated until the form met psychometric approval by ADE and the Pearson team.

The form was then reviewed by ADE content in collaboration with Pearson content to finalize forms. This was done virtually. The last stage of form construction was done in person in collaboration with content, psychometrics and the accommodation specialist.

Pearson selected field test items after the operational form was approved by ADE. Each form had a total of 10 field test slots, five for independent-set items and five for cluster-set items. Because cluster sets were developed with a total of 10 items, each set was tested on two forms. Similarly, independent sets, which were developed with a total of five items, were tested over two forms, with two items on one form and three items on another. ADE reviewed the field test selections, and Pearson revised as needed.

3.5.3. Sampling Plan

Each grade had one core online test form in the spring 2025 administration, with field test items embedded across multiple forms. All forms within a grade had the same operational items but different field test items. Grade 5 had 15 online forms, grade 8 had 16 online forms, and grade 11 had 14 online forms that were randomly assigned at a student level within a testing group, created by a district, by TestNav, Pearson's online test delivery platform. Each alternative form type such as the Special Paper Version (SPV), Braille, and ASL had only one form per grade.

3.6. Data Review

Following the spring 2025 administration, item analysis was conducted to generate statistics for the field tested items. Grade 5 had 144 field tested items, grade 8 had 149 field tested items, and grade 11 had 140 field tested items. Field tested items were flagged based on the criteria in Table 3.1. During data review, committee members reviewed the flagged items and their statistics to determine their eligibility for the operational item pool. Two different committees met for data review. One group focused solely on the items flagged for DIF, while another group reviewed the items flagged by the remaining statistics (e.g., item difficulty, point biserial, distractor analysis, and Rasch values). The DIF committee looks at the possibility of bias in each item flagged for DIF.

The meeting began with a training session that introduced the item review process, including an overview of the item statistics and how they should be used to evaluate items. Decisions about an item's quality cannot be made on statistics alone; the item itself and the content it measures should also be considered. Thus, the groups also reviewed the content of the items and how the items functioned according to the statistics before making a consensus decision about whether the item should be accepted or rejected for operational use. Revisions were recommended for the rejected items if applicable. Decisions for all items reviewed were documented in ABBI.

Table 3.1. Item Statistical Flagging Criteria

Statistic	Criterion	Possible Indication
P-value	< 0.2 or > 0.9	Very difficult or easy item
Point-biserial correlation	< 0.25	Poorly discriminating item
Distractor point-biserial correlation (MC only)	> 0.05	Possible miskey*
Omit rate	> 2%	Skipped item
Rasch difficulty	< -3 or > 3	Easy or difficult item
Item fit statistics	< 0.6 or > 1.4	Poor fit
Score point percentage (2-point items only)	< 1%**	Very few students got a certain score
Differential item functioning (DIF)	B, C	Item could be biased toward a certain student demographic group

*Possible miskey because the key should have a positive point-biserial correlation
**I.e., there should be at least 1% of students at each score point (2-point items only)

Table 3.2 presents the data review results based on the spring 2025 data. Committee members made these decisions based on the item content, using the item statistics to guide their discussion. Accepted items were added to the operational item pool for future use. Because the data review committee only reviewed the flagged field tested items, this table does not reflect the total number of field tested items because many did not have any statistical issues or they had fatal statistical issues (e.g., negative point-biserial) that removed them from the item pool.

Table 3.2. Data Review Results: Number of Flagged Field Tested Items

Grade	#Accepted	#Accepted w/Edits	#Rejected
5	40	0	28
8	25	0	48
11	32	0	57

3.7. Accommodated Forms

Each grade had one form of the paper-pencil Special Paper Version (SPV). The Pearson content team worked with ADE to produce paper-equivalent versions of the items used on the online test form. Upon approval of the item set, the Pearson publishing team worked with ADE to determine an approved paper-based test template for each grade. There were three rounds of review between ADE and Pearson before the document was approved to print. A final PDF printer proof was provided to ADE.

Upon approval of the paper-pencil form, Pearson began work on the Large Print and Braille forms. The Large Print forms are enlarged versions of the paper-pencil test forms. The publishing team enlarged the entire test book file to reach an 18-point font equivalent. The final Large Print printer proof file was posted for ADE's review and approval.

The Inkprint Braille version of the test was modified based on the Braille modification document to reflect any item omissions or modifications on the Student Braille Test Book. Pearson Braille Services reviewed all forms presented for Braille to determine if forms were well-suited for Braille testers. Any recommended modifications were reviewed in conjunction with ADE to arrive at final decisions. ADE then reviewed the Inkprint Test Book, the Student Braille Test Book proof, the Braille Test Administration Directions, and the Braille memo before production of the Braille material commenced.

Each grade and content area also had one form created for ASL testers. After approval by ADE of the online test form, Pearson ASL team began work for ASL translation. The Pearson ASL team created scripts to be used for filming of the ASL translation by professional ASL signers. Video sessions for ASL Filming were attended by the Pearson ASL team and Pearson content for any questions that arose during translation. ADE had final approval of any modifications necessary for successful ASL filming. All ASL videos and test forms were reviewed and approved by ADE before final production.

Chapter 4: TEST ADMINISTRATION

This chapter describes how the AzSCI assessments were administered, including the procedures used to ensure that the test administration was conducted in a secure and standardized manner, as indicated by Standards 1.10, 3.1, 3.9, 3.10, 4.2, 4.5, 4.15, 4.16, 4.21, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 7.0, and 7.8 (AERA et al., 2014). The AzSCI assessment is administered online via TestNav, Pearson's online testing platform that students use to access the assessment, with accommodated forms available as needed. PearsonAccess^{next} (PAN) is the student test management portal that test administrators use to manage student tests and registrations and order materials if needed.

District Test Coordinators (DTCs), School Test Coordinators (STCs), and Test Administrators (TAs) received online training and the supporting documents to ensure fidelity of implementation and the validity of the assessment results and to help prevent, detect, and respond to irregularities in academic testing and maintain testing integrity practices for technology-based assessments. For example, TAs were instructed to use the Test Administration Directions (TAD), as well as for the Special Paper Version (SPV) tests and entering student responses into TestNav.

When all TAs use the same well-defined administration procedures and are provided the same training, manuals, and supporting documents, administration is optimally standardized and poised to be fair to all students. DTCs were responsible for supporting the TAs in understanding and following the administration procedures. Comprehensive test coordinator training and materials targeted to their role and responsibility ensure that they are appropriately prepared to support the test administrators.

4.1. Test Units

The assessment for each grade was divided into three units to better manage the test administration, with a combined total of 60 items. Each test unit was estimated to take 60–90 minutes each. The AzSCI test was not timed. A test unit must be completed by the end of the regularly scheduled school day. Students taking the same test within the same school were not required to test on the same day, and students did not have to take Unit 1, Unit 2, and Unit 3 on the same day. It was recommended to take Unit 1 followed by Unit 2, then Unit 3, although this was not required. When two or three test units were scheduled the same day, a significant break was recommended as best practice between test units.

4.2. Administration Materials

Table 4.1 describes the materials provided to support the standardized administration of the AzSCI assessments and ensure fair testing for all students. The TAD and Test Coordinator Manual (TCM) were produced in collaboration with ADE. The Pearson program team drafted each manual using the previous year's version as a template. The manuals were then composed in desktop publishing software and sent for an editorial review. After a review of all comments and edits by the program team, the file was delivered for ADE review. There were multiple rounds of review between ADE and Pearson before the document was approved to print. ADE was provided with a final web-ready 508 compliant version in addition to the final printer's proof. Hard copies were not sent automatically to all participating schools, although a limited number were available for additional order during the additional order window. The materials were available on the ADE website at https://www.azed.gov/assessment/sci/.

Table 4.1. Administration Materials

Material	Description
Test Administration Directions (TAD)	Provides an overview of the AzSCI test administration, including the user roles in PAN and the test administration schedule, and directions about what to do before, during, and after testing.
Test Coordinator's Manual (TCM)	Indicates the responsibilities of the DTCs before, during, and after testing and explains the procedures for test administration. DTCs must review the TCM and the TAD well in advance of training STCs and TAs and before administering the tests. DTCs are responsible for ensuring the appropriate and correct administration of the AzSCI in all schools within the district or under the same charter.
PAN User's Guide	Explains how to navigate PAN and the tasks related to the AzSCI test administration.
Arizona Accommodation Manual	Lists the current accommodations, accessibility features, and tools available on Arizona's achievement assessments.

4.3. Administration Training

Mandatory test administration training was provided by ADE and Pearson and delivered through Pearson's online Arizona Learning Management System (AzLMS) that contained the training modules summarized in Table 4.2 that were required for DTCs, STCs, TAs, and other school staff involved in testing or test results.

The online training modules were available prior to the beginning of the testing window and throughout the testing window. The training modules addressed the specific responsibilities of the DTC, STCs, and TAs and provided important information from the three documents the testing staff are required to use (i.e., the TAD, TCM, and *PAN User's Guide*). These training modules are updated for each test administration in correspondence with the updates to the required documents. Each of the six modules requires approximately 30–45 minutes to complete. DTCs are required to ensure that all testing staff have viewed the applicable training modules and to track staff training completion. DTCs must obtain a score of 80% or higher on each module's final quiz to be certified to access the secure test administration materials. DTCs are allowed multiple attempts to obtain a score of 80% or higher on each module's final quiz.

Table 4.2. Administration Trainings

Training	Description
AzSCI Training for Test Coordinators	This training covered the AzSCI test administration for grades 5, 8, and 11, including an overview of the test administration, websites and resources, and responsibilities before, during, and after testing. This training module was required to be completed by DTCs and STCs.
Accommodations	This training covered the test accommodations. This was required for all DTCs and STCs but could be shared with staff members.
Achievement Test Administration Responsibilities	This training covered the test administration of AASA and AzSCI for all employees who administered, proctored, or were in contact with test materials. The purpose of this training was to provide guidance on consistent test administration across the state, increase the number of valid student tests, reduce test improprieties, and limit staff exposure to accusations of testing violations and discipline.
Test Security and Ethics	This training covered policies and practices to ensure the security and confidentiality of testing materials and the reliability and validity of test score interpretation. This training module was required for all employees who administered, proctored, or came in contact with testing materials or the test environment.

Training	Description
PearsonAccess^{next} (PAN)	This training covered PAN and was required for DTCs, STCs, and other testing staff who assisted with registering students or managing test sessions in PAN.
Technology Training	This training module covered technology requirements, TestNav information, and troubleshooting details for the online tests. It was required for all DTCs, STCs, and Technology Coordinators (TCs).

4.4. Sample Tests

In addition to the module training, TAs are instructed to become familiar with the online system by accessing sample items. Sample tests are available in TestNav year-round to help students become familiar with the AzSCI item types. The sample tests were created following Pearson's standard item and test development process, including item content and bias review by Arizona educators and community members. The sample tests reflect the AzSCI test specifications and blueprints, comprising 17 items in grade 5, 17 items in grade 8, and 16 items in grade 11. Because the sample tests do not include an item for each of the aligned Arizona Science Standards and do not provide scores for students, they should NOT be used to evaluate a student's performance level. Students access the test as a guest, so no personal information needs to be provided.

There is a sample test for each grade, and every eligible item type was represented. An accompanying scoring guide identified standard and TAGS levels. The portal and scoring guides are both available on the ADE website at https://www.azed.gov/assessment/sci.

4.5. Accommodations

Accommodations are specific practices and procedures that provide students with equitable access during the assessment. They are made to provide a student equal access to learning and equal opportunity to demonstrate what is known and are intended to reduce or even eliminate the effects of a student's disability. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. There should be a direct connection between a student's disability, special education need, or language need and the accommodation(s) provided to the student during educational activities, including assessment.

Students should receive the same accommodations for classroom instruction, classroom assessments, district assessments, and state assessments. No accommodations should be provided during assessments that are not also provided during instruction. However, not all accommodations appropriate for instruction are appropriate for use during a standardized state assessment. Table 4.3 presents the accommodations available to students while testing on Arizona assessments.

Table 4.3. Available Accommodations

Accommodation	Description
Adult Scribe	A student who requires one-on-one adult assistance during daily instruction may orally dictate or use gestures to indicate a selected response for multiple-choice items only while an adult enters this in the test. The adult may not ask or answer any questions during the session or influence student responses in any way.
American Sign Language (ASL)	ASL requires the use of a different test form that must be indicated in PearsonAccess^{next} (PAN). The ASL test form must be requested using the Additional Accommodations online request form.

Accommodation	Description
Braille test booklet	Braille tests must be requested using the Special Paper Version (SPV) test online request form. Requires adult transcription: An adult must transfer the student's responses exactly as written into the TestNav system.
Large print test booklet	Large Print tests must be requested using the Special Paper Version (SPV) test online request form. The 504 plan or IEP must clearly state the font size used for instruction and the type of materials teachers enlarge for the student. Requires adult transcription: An adult must transfer the student's responses exactly as written into the TestNav system.
Paper test booklet	A student who cannot access the computer for classroom work due to injury, illness, or vision impairments may need a paper test in lieu of taking the test with peers on the computer. Requires adult transcription: An adult must transfer the student's responses exactly as written into the TestNav system.
Sign test content	Any student who requires signing of content during daily instruction may have any of the content of writing, mathematics, and science signed.
Simplified test administration directions	The test administrator may provide verbal directions in simplified English for the scripted directions from the *Test Administration Directions* manual. This must take place in a setting that does not disturb other students.
Translated test administration directions	Exact oral translation, in the student's native language, of the scripted directions from the *Test Administration Directions* manual are permitted. No test content or directions embedded within the test may be translated.
Translation dictionary	During testing, English learner students may use the word-for-word published paper translation dictionary that is used regularly for classroom instruction. Students with a visual impairment may use an electronic dictionary with other features turned off.

4.6. Universal Test Administration Conditions

The following Universal Test Administration Conditions are testing situations and conditions that may be offered to any student to provide a comfortable and distraction-free testing environment. They do not require an accommodations request. While some of the items listed as Universal Test Administration Conditions might be included in an IEP or 504 plan as an accommodation, for achievement testing purposes these are not considered testing accommodations and are available to any student who needs them.

- Testing in a small group, 1:1, or in a separate location on campus or in a study carrel
- Being seated in a specific location within the testing room or at special furniture
- Having the test administered by a familiar test administrator
- Using a special pencil or pencil grip
- Using a place holder
- Read-aloud (text-to-speech or human reader) content of the ELA writing, mathematics, and science assessments
- Using devices that allow the student to see the test: glasses, contacts, magnification, CCTV, dome magnifiers, enlarged monitors, enlarged keyboard, and special lighting
- Using different contrast settings or color overlays
- Using devices that allow the student to hear the test directions: hearing aids, cochlear implants, and amplification
- Wearing noise buffers after the scripted directions from the *Test Administration Directions* manual have been read

- Signing the scripted directions from the *Test Administration Directions* manual
- Repeating the scripted directions from the *Test Administration Directions* manual
- Having assistance with logging into an online test
- Reading the test quietly to themselves as long as other students are not disrupted
- A phone or electronic device needed for medical care is permitted. The phone needs to stay close to the Test Administrator or proctor as well as the student and should be monitored to assure the device is only being used for medical purposes during testing
- Individual students may take a stretch break (1 or 2 minutes) during the test session (students may not talk, use electronic devices, go to lunch, or leave the testing room)
    - Paper test booklet and scratch paper must be collected
    - Students must sign out of TestNav without submitting the test. The test administrator will need to resume the student's test session using PAN.
- Students may use the restroom (only one student at a time)
    - The TA must collect the student's paper test booklet and scratch paper.
    - Students must sign out of TestNav without submitting the test. The test administrator will need to resume the student's test session using PAN.
- The use of scratch paper (plain, lined, or graph; school provided). Scratch paper must be securely shredded at the conclusion of testing
- Each testing session must be completed in the same school day in which it was started. The AASA and AzSCI are untimed. Do not start a test unit unless there is sufficient time to complete the test in the same school day.
- Students cannot leave for lunch during a test session. If necessary, lunch may be brought to the student. Test units should be scheduled in a way that provides the student more than adequate time to complete the test.

4.7. Universal Test Tools

The Universal Test Tools provided in Table 4.4 are available to all students taking the AzSCI assessment and cannot be disabled.

Table 4.4. Universal Test Tools

Universal Test Tool	Description
Alternate Mouse Pointer	There are six alternate mouse pointers available for students in TestNav. Alternate options include a medium, large, or extra-large sized white pointer, and extra-large sized black, green, or yellow pointer.
Answer Masking	Allows student to electronically cover and reveal individual answer choices.
Answer Eliminator	Allow students to cross out answer options for multiple-choice and multi-select items.
Area Boundaries	Allows student to click anywhere on the selected response text or button for multiple choice items.
Bookmark for Review	Allows student to mark an item for review so that it can be easily found later.
Contrast	Allows the student to change the background and text color based on need or preference. The Contrast setting will not change images or artwork. The options are white background with black text; cream background with black text; light blue background with black text; black background with white text; light magenta background with black text; blue background with yellow text; and pale green background with gray text.

Universal Test Tool	Description
Expand/Collapse Passage	Allows student to expand a passage for easier readability. Expanded passages can also be collapsed.
Highlighter	Allows student to highlight text in a passage or item.
Key Board Navigation	Key Board Navigation - Can be used in supported item types using the following keyboard commands: Move forward: Tab Move backwards: Shift and Tab Select buttons: Enter or Spacebar Navigate and select radio buttons: Arrow Up or Arrow Down Select and unselect boxes: Spacebar Increase / Decrease screen size: Ctrl + and Ctrl - Copy: Ctrl + C Paste: Ctrl + V
Line Reader	An adjustable box allows the student to focus on one line or a few lines at a time. The box can be adjusted to increase or decrease the number of lines shown. The Line Reader and Magnifier tools may be used simultaneously.
Magnifier	Allows the student to make part of the screen larger. When in use, the magnifier can be moved around the screen as needed.
Notes/Comments	Allows student to open an on-screen notepad and take notes or make comments. Notes carry over within a passage set. In non-passage items, notes are attached to the specific test item on which they are entered.
Pause and Restart	Students may sign out of TestNav. Before the student can resume testing, the Test Administrator will need to resume the student's session in TestNav.
Review Test	Allows student to review the test before submitting it.
Student Readiness Tool	Video lessons and interactive practice questions are embedded by grade and content area to provide students a hands-on experience with the tools, questions, and test support in the test environment.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech	Allows student to access text-to-speech for content of writing, mathematics, and science.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Allows student to enlarge the font and images in the test up to 500%. Undo zoom in and return the font and images in the test to original size. Zoom may be set in PAN up to 200% prior to testing, then the student may use the keys Ctrl + to Zoom up to 500% in TestNav.

4.8. Pearson Customer Support

To provide support to schools before, during, and after testing, Pearson operates and provides tiered technical support Monday through Friday from 7:00 a.m. to 7:00 p.m. CST. DTCs, STCs, and TAs can contact the customer support line with questions pertaining to the TestNav and PAN system and test administration procedures. The toll-free support number, e-mail address, and chat link are disseminated to the field through the AzSCI PAN system and related communications.

4.9. Test Security

All test coordinators, administrators, and proctors must be trained in proper test security procedures, must sign an Achievement Tests Staff Security Agreement form (as shown in Figure 4.1), and must adhere to test security procedures. Test materials should be secured prior to, and at the conclusion of, all testing sessions. Test Administrators and proctors may not assist students in answering test items and may not translate, reword, or explain any test content. No test content may ever be discussed before, during, or after test administration. It is unethical and shall be viewed as a violation of test security for any person to:

- Log into TestNav as a student unless assisting student with log in procedures
- Share their username/password for PAN
- 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 test item before, during, or after the test administration
- Allow students access to test content prior to testing
- Allow students to share information during the test administration
- Read any parts of the test to students, except as indicated in the TAD or as part of an approved accommodation
- Influence students' responses by making any kind of gestures (e.g., pointing to items or 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; this instruction may only be given before the students take the test
- Review students' responses
- Change students' answer choices
- Read or review students' scratch paper
- Participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures

Figure 4.1. Test Security Agreement

Achievement Tests (AASA, AzSCI, ACT Aspire, and ACT)
School Year 2024–2025 Staff Test Security Agreement

I acknowledge that all Achievement Tests are secure tests and agree to the following conditions of use to ensure the security of the test. For this document, Achievement Tests refers to AASA, AzSCI, ACT Aspire, and ACT.

1. I shall take necessary precautions to safeguard test materials.
    a. I shall sign an *Achievement Tests Staff Security Agreement* for School Year 2024-2025.
    b. Access to test materials, including online tests, is restricted. I shall not attempt to gain access to test materials beyond that which is granted to me by my school/district test coordinator, superintendent, or charter representative.
    c. If test materials are distributed to me, I shall keep them under lock and key except during actual test times. This includes any student data sheets or student information sheets provided to me.
    d. I shall not permit students to remove test material from the testing room.
    e. I shall not examine, read, or review the Achievement Tests.
        i. I shall not disclose, nor allow to be disclosed, the content of the test.
        ii. I shall not discuss any test item at any time.
        iii. I shall not examine, read, or review any student responses.
        iv. I shall not log into any student online test.
    f. I shall not erase or change any student responses or any marks (including stray marks) on a scorable test booklet or answer document.
    g. If test materials are distributed to me, I shall return all test materials to the school/district test coordinator immediately upon the completion of testing.
    h. I shall not use any test materials for instruction before or after test administration. I shall follow *Test Preparation and Administration Practices* approved by the State Board of Education.
    i. I shall not provide prohibited or inappropriate resources or content support to students during testing, including but not limited to graphic organizers, reference sheets, and calculators (except for tests and test sections where calculators are allowed).
2. I understand that the district superintendent or charter representative will develop, distribute, and enforce disciplinary procedures for the violation of test security by staff.

Individuals who will administer or proctor Achievement Tests for school year 2024-2025 must also agree to the following conditions to ensure the correct administration of the tests.

3. I shall participate in and complete training activities prior to administering the tests.
4. I shall review the Test Administration Directions prior to administering the test.
5. I shall follow all instructions in the Test Administration Directions, including reading the directions to students exactly as scripted.

By signing my name to this document, I am assuring my district/charter and the Arizona Department of Education that I will abide by the above conditions and that anyone I supervise, who will have access to the Achievement Tests or the testing environment, will also sign a Test Security Agreement.

Signed By:_____ Date: _____

Printed Name: _____

Title:_____ School: _____

Please return signed copy as per instructions from your school/district test coordinator.
Signed copies will be maintained by school/district administrators for 6 years.

In addition to test security procedures required of all educators involved in the testing process, TestNav has built-in security features for the test content and personal data that relies on multiple levels of protection, including restricted user access, encryption of data in transit and at rest, systems monitoring for abnormal behavior, application, server, and network security testing, and qualified, verified, and trusted support personnel.

Pearson uses Advanced Encryption Standard (AES) encryption for data at rest and Hypertext Transfer Protocol Secure (HTTPS) to provide encryption and data-in-motion security for online testing by creating a secure channel on the network with the Secure Socket Layer (SSL)/Transport Layer Security (TLS) protocols. Test content can only be viewed through a valid test registration and login, all of which are logged within the platform's audit trail system and cannot be deleted.

TestNav also locks down the student's desktop during testing to prevent students from accessing outside resources that could be used for cheating, such as email, instant messaging, or internet browsing. TestNav will stop students' tests if another background application attempts to interfere with or take focus away from the secure testing environment. These types of interruption cannot be blocked during testing and therefore could present additional opportunities for students to access unauthorized resources. However, TestNav also has a blocklist feature that prevents students from starting their test if certain applications that pose a threat to disrupt testing are running at the time TestNav is launched. In these situations, the student and/or proctor are prompted to shut down the offending application before attempting to start TestNav again.

Chapter 5: SCORING

All items (i.e., MC, TE, and EBSR items) on the AzSCI assessments were machine-scored, with an attemptedness rule that a student needed to answer at least one item in each unit. The machine-scored items on all forms (i.e., online, SPV, Braille, and ASL forms) were processed using TestNav, Pearson's online testing platform.

Machine scoring refers to the scoring of student responses to MC, TE, and EBSR items by TestNav based on pre-defined scoring rules and answer key. EBSR items are two-part items, in which each part is either MC or TE item. TE items require scoring models that go beyond traditional answer keys, defined during item development and encoded in the item's XML using rule-based logic. Scoring models may include multiple correct response patterns and partial credit criteria, depending on the item type. For example, common interaction types include choice interaction (students select one or more options), hot spot or text selection (students click or highlight designated regions of an image or text), or match interaction (students match items from two sets into correct pairs). Scoring rules for each interaction type are reviewed during item development and quality control processes.

To ensure the accuracy and reliability of machine scoring, Pearson Psychometrics conducted a key check and adjudication with sufficient data near the end of the test administration to verify that all scoring rules and keys were implemented correctly for each item prior to final scoring and reporting. A key check is the process of verifying the correct answer key for MC items. Adjudication is the process of verifying that student responses for TE items are scored appropriately, and the functionalities of the items perform correctly. For each TE item, the frequency distribution of responses scored correctly was created, along with the frequency distribution of responses scored as incorrect. Pearson Content Specialists analyzed for correctness every response string not checked in a previous administration in the frequency reports and indicate whether the response should be scored as correct. If discrepancies were identified during the key check or adjudication process, Pearson and ADE Content Specialists reviewed the flagged item(s) and worked to resolve the issue.

Chapter 6: REPORTING

The following AzSCI reports were available online in PAN at https://az.pearsonaccessnext.com. PDF versions of the reports and district-wide electronic student data files were available for downloading, May 23, 2025. District-level user roles were provided access to all school-level and district-level reports, including all Confidential Student Score Reports for students who tested in the district. School-level user roles were provided access to all school-level reports and all Confidential Student Score Reports for students who tested in the school. A Family Guide for interpreting reports was also available for download. Figure 6.1 and Figure 6.2 present sample reports.

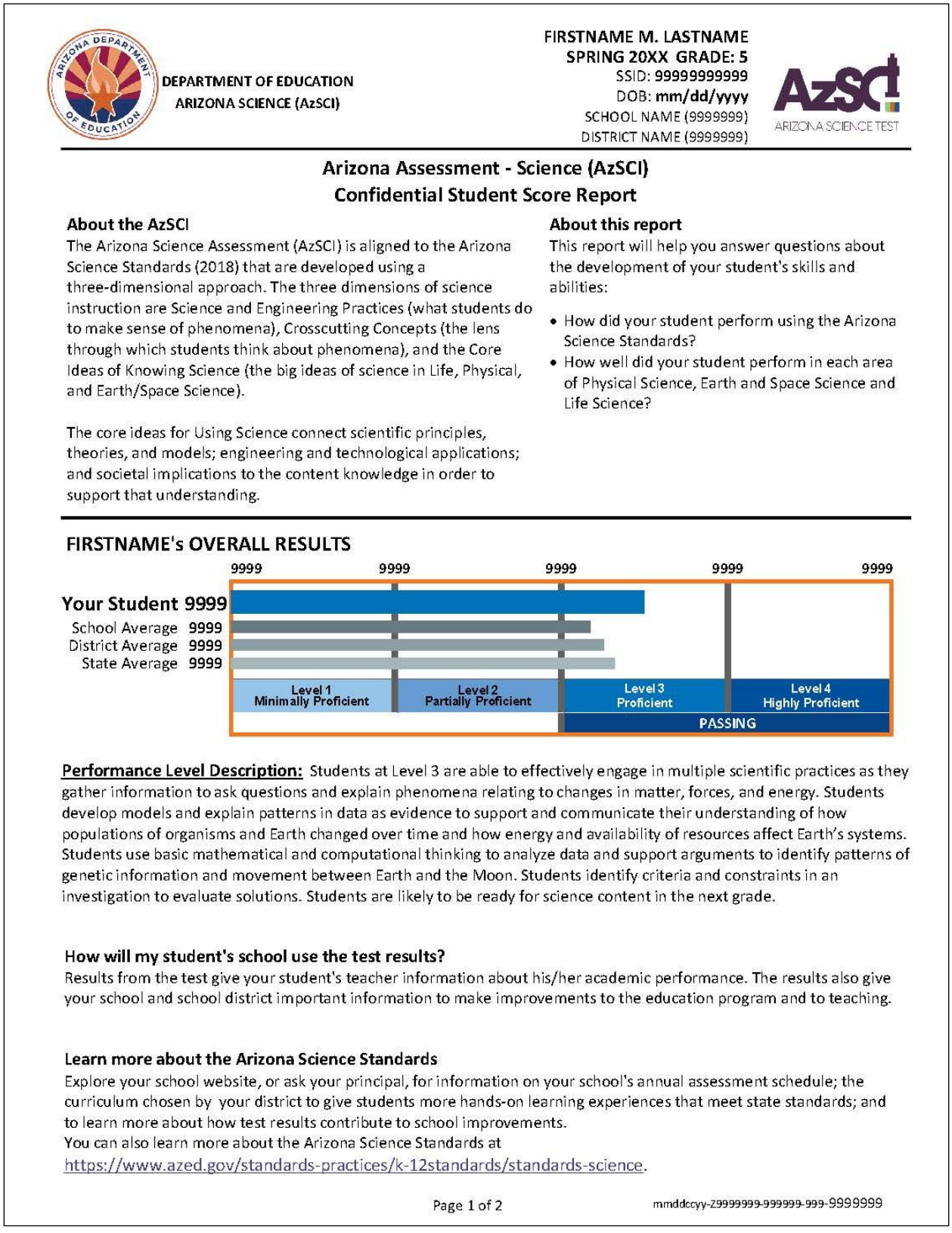
- District-level
    - Confidential Roster Report with Summary (school-level[1], student roster by grade)
    - District Confidential Roster Report with Summary (district-level, student roster by grade)
    - Student Data File
    - Summary Data File
- School-level
    - Confidential Student Score Report (individual student report)
    - Family Report Guide
    - Informe del Estudiante (individual student report in Spanish)
    - Confidential Roster Report with Summary (school-level, student roster by grade)
    - Summary Data File

AzSCI reports have been designed with the user's comprehension in mind. The goal of these reports is to deliver accurate assessment data and ensure that it is correctly interpreted and understood. Similar colors are used for groups of similar elements, such as performance levels, throughout the design to guide the user to compare like elements and avoid comparison of dissimilar elements.

All score report data are based on the total number of students whose tests have been scored. All score report data in PAN, except for individual students' score reports and Confidential Roster Reports, can be disaggregated into testing groups if they were set up by the school during the specified timeframe. The Confidential Student Score Report (individual student report) includes the average scale scores for the school, district, and state to allow for visual comparison. Two copies of the printed Confidential Student Score Report and Family Report Guide were also provided. Printed reports are packed by school and shipped to participating districts.

[1] Districts receive their own copy of the school-level Confidential Roster Report with Summary. For example, if a district has five schools, they will have a copy of all five rosters in one PDF file.

Figure 6.1. Sample Report—Confidential Student Score Report



DEPARTMENT OF EDUCATION
ARIZONA SCIENCE (AzSCI)

FIRSTNAME M. LASTNAME
SPRING 20XX GRADE: 5
SSID: 99999999999
DOB: mm/dd/yyyy
SCHOOL NAME (9999999)
DISTRICT NAME (9999999)

AzSCI ARIZONA SCIENCE TEST

Arizona Assessment - Science (AzSCI)
Confidential Student Score Report

About the AzSCI

The Arizona Science Assessment (AzSCI) is aligned to the Arizona Science Standards (2018) that are developed using a three-dimensional approach. The three dimensions of science instruction are Science and Engineering Practices (what students do to make sense of phenomena), Crosscutting Concepts (the lens through which students think about phenomena), and the Core Ideas of Knowing Science (the big ideas of science in Life, Physical, and Earth/Space Science).

The core ideas for Using Science connect scientific principles, theories, and models; engineering and technological applications; and societal implications to the content knowledge in order to support that understanding.

About this report

This report will help you answer questions about the development of your student's skills and abilities:

- How did your student perform using the Arizona Science Standards?
- How well did your student perform in each area of Physical Science, Earth and Space Science and Life Science?

FIRSTNAME's OVERALL RESULTS

Performance Level Description: Students at Level 3 are able to effectively engage in multiple scientific practices as they gather information to ask questions and explain phenomena relating to changes in matter, forces, and energy. Students develop models and explain patterns in data as evidence to support and communicate their understanding of how populations of organisms and Earth changed over time and how energy and availability of resources affect Earth's systems. Students use basic mathematical and computational thinking to analyze data and support arguments to identify patterns of genetic information and movement between Earth and the Moon. Students identify criteria and constraints in an investigation to evaluate solutions. Students are likely to be ready for science content in the next grade.

How will my student's school use the test results?

Results from the test give your student's teacher information about his/her academic performance. The results also give your school and school district important information to make improvements to the education program and to teaching.

Learn more about the Arizona Science Standards

Explore your school website, or ask your principal, for information on your school's annual assessment schedule; the curriculum chosen by your district to give students more hands-on learning experiences that meet state standards; and to learn more about how test results contribute to school improvements.
You can also learn more about the Arizona Science Standards at
https://www.azed.gov/standards-practices/k-12standards/standards-science.

Page 1 of 2

mmddccyy-Z9999999-999999-999-9999999

FIRSTNAME M. LASTNAME
SPRING 20XX GRADE: 5

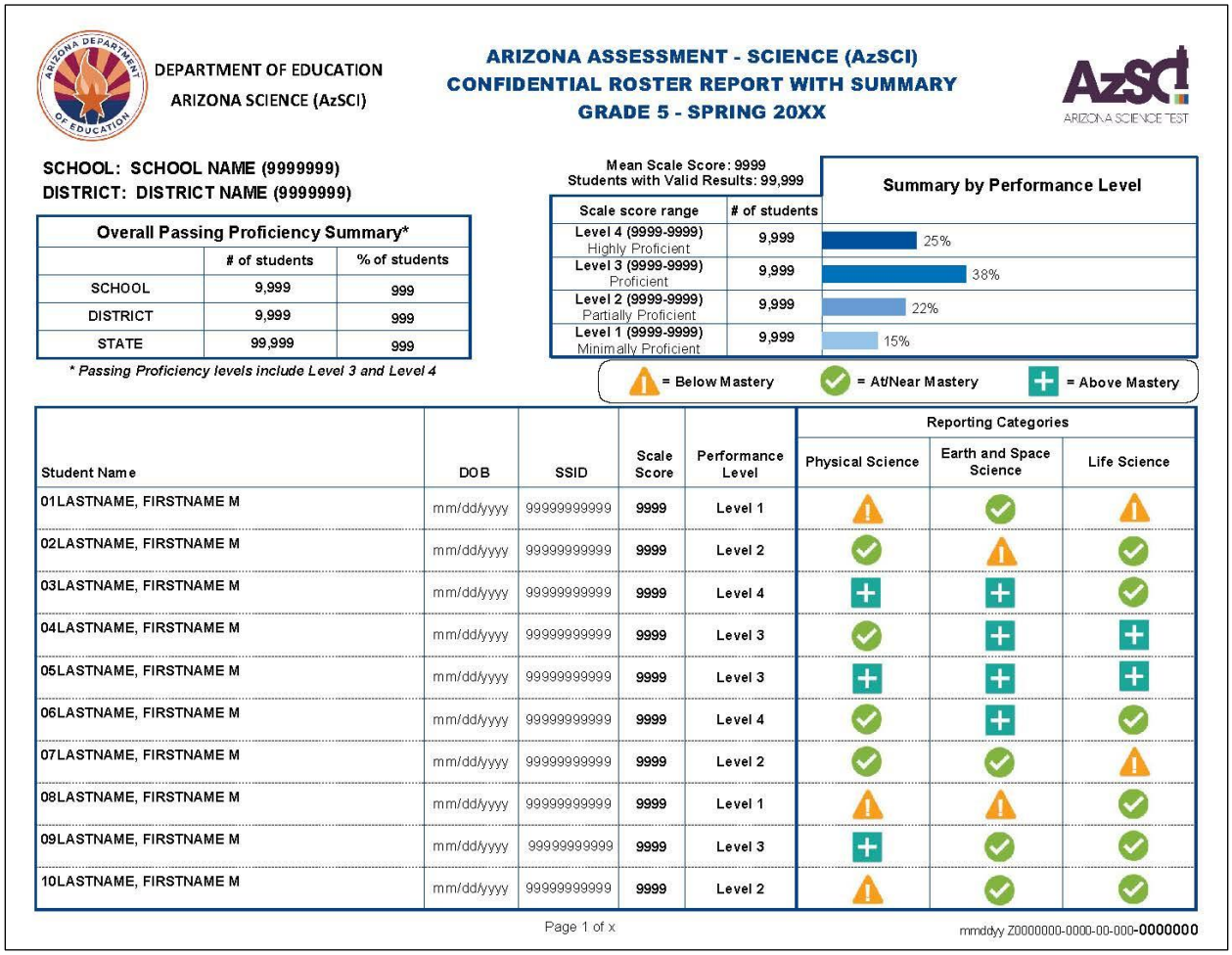
Legend: Reporting Categories
= Below Mastery = At/Near Mastery = Above Mastery

Science and Engineering Practices and Crosscutting Concepts Reporting Categories		PERFORMANCE
Physical Science:	Students performing at this level show an advanced understanding of the three-dimensions in Physical Science content, including: • All matter in the Universe is made of very small particles. • Objects can affect other objects at a distance. • Changing the movement of an object requires a net force to be acting on it. • The total amount of energy in a closed system is always the same but can be transferred from one energy store to another during an event.	
Earth and Space Science:	Students performing at this level show a good understanding of the three-dimensions in Earth and Space Science content, including: • The composition of the Earth and its atmosphere and the natural and human processes occurring within them shape the Earth's surface and its climate. • The Earth and our solar system are a very small part of one of many galaxies within the Universe.	
Life Science:	Students performing at this level likely need more support of the three-dimensions in Life Science content, including: • Organisms are organized on a cellular basis and have a finite life span. • Organisms require a supply of energy and materials for which they often depend on, or compete with, other organisms. • Genetic information is passed down from one generation of organisms to another. • The unity and diversity of organisms, living and extinct, is the result of evolution.	

For more information about AzSCI, go to https://www.azed.gov/assessment/sci.
If you require your child's report in an alternative format, please contact ADE's Assessment Section at Testing@azed.gov.

Page 2 of 2

Figure 6.2. Sample Report—Confidential Roster Report with Summary



DEPARTMENT OF EDUCATION
ARIZONA SCIENCE (AzSCI)

ARIZONA ASSESSMENT - SCIENCE (AzSCI)
CONFIDENTIAL ROSTER REPORT WITH SUMMARY
GRADE 5 - SPRING 20XX

AzSCI ARIZONA SCIENCE TEST

SCHOOL: SCHOOL NAME (9999999)
DISTRICT: DISTRICT NAME (9999999)

Overall Passing Proficiency Summary*		
	# of students	% of students
SCHOOL	9,999	999
DISTRICT	9,999	999
STATE	99,999	999

** Passing Proficiency levels include Level 3 and Level 4*

Mean Scale Score: 9999
Students with Valid Results: 99,999

Scale score range	# of students
Level 4 (9999-9999) Highly Proficient	9,999
Level 3 (9999-9999) Proficient	9,999
Level 2 (9999-9999) Partially Proficient	9,999
Level 1 (9999-9999) Minimally Proficient	9,999

Summary by Performance Level

= Below Mastery = At/Near Mastery = Above Mastery

Student Name	DOB	SSID	Scale Score	Performance Level	Reporting Categories: Physical Science	Earth and Space Science	Life Science
01LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 1			
02LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 2			
03LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 4			
04LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 3			
05LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 3			
06LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 4			
07LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 2			
08LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 1			
09LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 3			
10LASTNAME, FIRSTNAME M	mm/dd/yyyy	99999999999	9999	Level 2			

Page 1 of x

mmddyy Z0000000-0000-00-000-**0000000**

Chapter 7: CLASSICAL ITEM ANALYSIS

This chapter presents classical statistics for the data used for calibration, equating, and scaling of the spring 2025 AzSCI assessment as indicated by Standards 1.8, 1.10, 2.5, 2.19, 3.6, 4.14, and 7.4 (AERA et al., 2014). AzSCI only had one core online form with different embedded field test sets for each grade, with a total of 15 online forms for grade 5, 16 online forms for grade 8, and 14 online forms for grade 11.

7.1. Data

Classical item analysis was conducted based on the calibration samples as described in Section 8.1. Table 7.1 presents the demographic information of the students included in the calibration sample by gender, ethnicity (Hispanic or Non-Hispanic), race, and special education, English learner (EL), and low socioeconomic status (SES). Because only a few students took the accommodated forms, these students were not included in the item analysis. Students who did not complete the test were also excluded.

Table 7.1. Number of Students in the Calibration Sample by Subgroup

Subgroup	Grade 5	Grade 8	Grade 11
All	81,344	82,342	82,949
Male	41,264	41,798	41,771
Female	40,080	40,544	41,178
Hispanic	39,921	39,629	40,158
Non-Hispanic	41,423	42,713	42,791
American Indian	4,281	4,560	4,808
Asian	3,074	3,036	2,835
Black or African American	5,981	5,944	5,396
Multi-racial	5,125	4,908	4,441
Native Hawaiian or Other Pacific Islander	472	507	485
White	59,661	60,348	61,011
Missing	2,750	3,039	3,973
Special Ed.	12,192	9,872	7,334
English Learner (EL)	9,193	7,344	4,883
Low Socioeconomic Status (SES)	41,876	39,985	37,390

7.2. Descriptive Statistics

Table 7.2 presents descriptive statistics on total raw scores for the spring AzSCI assessment by grade, including the number of students included in the classical analysis, the number of operational items on the assessment, the maximum possible raw score, the mean raw score, the standard deviation (SD) of the raw score, and the minimum/maximum obtained raw score.

Table 7.2. Raw Score Descriptive Statistics

Grade	#Students	#Items	Max. Possible Raw Score	Mean Raw Score	SD Raw Score	Min. Raw Score	Max. Raw Score
5	81,344	50	55	25.44	11.41	0	55
8	82,342	50	55	23.20	10.32	0	55
11	82,949	50	55	19.96	9.62	0	54

7.3. Classical Item Analysis

Classical item analysis was conducted to show how the items performed for each grade-level assessment. Item difficulty is measured by the *p*-value bounded by 0.0 and 1.0 that indicates how easy or hard an item is for students. The *p*-value for 1-point items is based on the proportion of students who answered an item correctly and is derived by dividing the number of students who got the item correct by the total number of students who answered it. For multiple-point items, the *p*-value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item. A high *p*-value indicates that an item is easy (high proportion of students answered it correctly), whereas a low *p*-value indicates that an item is difficult. For example, a *p*-value of 0.79 indicates that 79% of students answered the item correctly. Easy and hard items are both necessary to include on an assessment to balance the test difficulty. The AzSCI assessment targets *p*-values in the range of 0.2 to 0.90.

Item discrimination is represented by the point-biserial correlation bounded by -1.0 and 1.0 that indicates how well an item discriminates, or distinguishes, between low-performing and high-performing students. The point-biserial correlation is based on the relationship between student performance on a specific item and performance on the entire test based on their test score. Students who do well on a test are expected to select the right answer to any given item, and students who do poorly are expected to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. An item with a high positive point-biserial correlation discriminates between low-performing and high-performing students better than an item with a point-biserial correlation near zero. A negative point-biserial correlation indicates that lower-performing students did better on that item than higher-performing students. The AzSCI assessment targets point-biserial correlations of 0.25 or higher.

Table 7.3 presents a summary of the classical item analysis, and Appendix A presents the statistics for each item. If the classical item statistics for the operational items were outside of the item selection criteria presented in Table 3.1, the items will be reviewed during test construction of the next testing cycle for possible replacement in future administrations.

Table 7.3. Classical Item Analysis Summary

Grade	#Items	Mean P-Value	Mean Point-Biserial
5	50	0.46	0.45
8	50	0.43	0.42
11	50	0.36	0.39

7.4. Distractor Analysis

Table 7.4 and Table 7.5 present the point-biserial correlations associated with a correct option and the incorrect options at various percentiles. As expected, the point-biserial correlation for a correct option was around 0.20 or higher for most items, whereas the point-biserial correlation for incorrect options was negative or very close to zero. The results show that students with higher proficiency tended to choose a correct option, and students with lower proficiency tended to choose an incorrect option. This indicates that the distractors appear to perform appropriately.

Table 7.4. Distractor Analysis Summary: Point-Biserial Correlations for Correct Options

Grade	#MC Items	Min.	P25	P50	P75	Max.
5	23	0.21	0.34	0.41	0.52	0.63
8	19	0.20	0.36	0.41	0.46	0.50
11	23	0.10	0.27	0.34	0.41	0.47

Note. Min. = minimum, P25 = 25th percentile, P50 = 50th percentile (median), P75 = 75th percentile, Max. = maximum

Table 7.5. Distractor Analysis Summary: Point-Biserial Correlations for Incorrect Options

Grade	#MC Items	Min.	P25	P50	P75	Max.
5	23	-0.36	-0.27	-0.19	-0.15	-0.01
8	19	-0.32	-0.24	-0.19	-0.15	0.20
11	23	-0.29	-0.21	-0.15	-0.08	0.08

Note. Min. = minimum, P25 = 25th percentile, P50 = 50th percentile (median), P75 = 75th percentile, Max. = maximum

A distractor analysis was also conducted for each multiple-choice item, as presented in Appendix A. The response distribution for an item across all possible choices (e.g., a correct option and distractors) was calculated. The point-biserial correlation and omit rate associated with each response option was calculated as well. Typically, a negative point-biserial correlation is sought for distractors because less-proficient students should be more likely to choose an incorrect option.

Chapter 8: CALIBRATION, EQUATING, AND SCALING

This chapter describes the calibration, equating, and scaling procedures that took place for the spring 2025 AzSCI assessment, addressing Standards 1.10, 5.1, 5.2, 5.3, 7.2, 7.4, and 12.9 (AERA et al., 2014).

8.1. Calibration Sample

To ensure valid calibration results, several data cleaning steps occurred upon receipt of raw data from the scanning and scoring processes. These steps allowed for calibration to be conducted on valid student responses. The cleaning process removed the following records from the calibration datasets for each grade level:

- Records with invalidated tests that are marked Do Not Report (DNR) in PearsonAccess^{next} (PAN)
- Records that indicate the student took an accommodated form
- Records with non-valid attempts noted by less than one response
- Duplicate records (e.g., students indicated as taking the test more than once)
- Records in which a student was enrolled in an exclusionary school list from ADE

8.2. Calibration Methods

Item response theory (IRT) models were used in the item calibration. All tests were calibrated separately by grade. If there was more than one operational form, all operational forms were calibrated concurrently. All calibration activities were replicated by two psychometricians independently as a quality control measure. The calibration results were also reviewed independently by a senior-level psychometrician at Pearson.

The Rasch model (Rasch, 1960) was used for 1-point items, and the partial-credit model (Masters, 1982) was used for multiple-point items for calibration. Parameter estimation for items was implemented using Winsteps 4.8.1.0 (Linacre, 2022b) that uses joint maximum likelihood estimation (JMLE), as described by Wright & Masters (1982).

The Rasch model estimates item difficulty and student ability on the same scale. Under the Rasch model, the probability that student j with ability θ answers item i with difficulty of b correctly is as follows:

$$P_i(\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

The partial-credit model is an extension of the Rasch model for items in which students may receive partial credit. Thus, the partial-credit model reduces to the Rasch model when items have only two response categories (i.e., 0 or 1). According to the partial-credit model, the probability that student j scores x on item i, which has a maximum possible score of m ($k = m+1$ possible response categories), is expressed as follows:

$$P_{ix}(\theta_j) = \frac{\exp \sum_{l=0}^{x} (\theta_j - D_{il})}{\sum_{k=0}^{m_i} [\exp \sum_{l=0}^{k} (\theta_j - D_{il})]}$$

where $x = 0, 1, \ldots, m_i$, D_{il} is a step difficulty for score l and by definition,

$$\sum_{l=0}^{0}(\theta_j - D_{il}) = 0$$

The step difficulty D_{il} can be decomposed such that

$$D_{il} = b_i + h_{il}$$

where b_i is an overall difficulty for item i, and h_{il} is a threshold for score l (Embretson & Reise, 2000; Linacre, 2022a). This parameterization allows b_i in the partial-credit model to be comparable to b_i in the Rasch model.

8.3. Calibration Results

All items converged during calibration using typical procedures for Winsteps software. Standard error of estimates for the Rasch difficulty measures indicated that the parameters were well-estimated. Table 8.1 presents a summary of the IRT statistics, and Appendix B presents the item-level IRT statistics resulting from the calibration of the spring AzSCI assessment.

Table 8.1. IRT Statistics Summary

Grade	#Items	Mean Rasch
5	50	0.15
8	50	-0.04
11	50	-0.03

An item-person map shows the distribution of item difficulty and the distribution of student ability in one graph, as they are on the same scale. This graph is useful for Rasch models to evaluate the extent to which the item difficulty and student ability distributions are aligned because they assume the probability of a correct answer is affected only by a student's ability and the item difficulty. Figure B.1 – Figure B.3 in Appendix B present the item difficulty distribution on the lefthand side and the student ability distribution on the right. Each marker in the item difficulty distribution is an item, and the item difficulty values are rounded with an increment of 0.20 before they are plotted. Horizontal dotted lines represent the three performance level cuts (*Partially Proficient*, *Proficient*, and *Highly Proficient*) for the total test.

In addition to the item-person map, two more graphs are presented to summarize the characteristics of each operational assessment. The test characteristic curve (TCC) shows an expected total raw score across different student abilities, while the CSEM curve presents an amount of standard error across different student abilities. The CSEM has an inverse relationship with the test information function (TIF) as follows:

$$SE(\theta) = \frac{1}{TI(\theta)}$$

where $SE(\theta)$ is the CSEM, and $TI(\theta)$ is the TIF (Embretson & Reise, 2000). Figure B.4 – Figure B.9 in Appendix B present the TCC, TIF, and CSEM curves.

8.4. Equating

The spring 2025 AzSCI tests were equated and placed on the operational AzSCI scale using a non-equivalent groups anchor item (NEAT) design. A set of anchor items was selected from the existing item bank. The anchor items were selected such that they contributed approximately 30% of the total score points and their content representation was as similar as possible to the blueprint. The location of all anchor items stayed within three positions from where they were in the previous year.

A fixed anchor parameter equating was implemented within Winsteps to place the tests on the operational reporting scale. This was implemented by constraining the parameter estimates in the existing item bank for the anchor items to equal the final parameter estimates obtained in the original AzSCI calibration analyses. The displacement statistic, which estimates the difference between the fixed parameter and the estimate had the item parameter not been constrained, was evaluated for each anchor item.

Items with a displacement statistic greater than 0.30 or less than -0.30 were reiteratively removed from the anchor set. The criterion of 0.30 has been used to flag displaced anchor items under a common item, non-equivalent group equating design for many state programs (Miller et al., 2004). If more than one anchor item was flagged, the item with the largest magnitude of displacement value was dropped from the anchor set. The displacement values of the remaining anchor items were then re-estimated by implementing the fixed anchor parameter equating with the remaining anchor items. This process was repeated until all the anchor items had displacement values of a magnitude smaller than 0.30 and greater than -0.30.

Table 8.2 presents the number of items for the initial anchor set of each grade and the number of items dropped from each initial anchor set.

Table 8.2. Summary of Anchor Items

Grade	#Items in Initial Anchor Set	#Items Dropped from Anchor
5	18	1
8	14	0
11	17	0

8.5. Scaling Methods

Scaling constants for the total score were determined such that the theta score, based on the total test, was transformed to have the reporting scale range from 1200 to 1500 across all grades. The scale scores for the *Partially Proficient* and *Proficient* cuts were fixed at 1300 and 1350, respectively, for each grade, and the *Highly Proficient* cut was allowed to freely vary. Thus, scaling constants were calculated by solving the following equations:

$$A \times \theta^{PartiallyProficient} + B = 1300, \text{ and}$$
$$A \times \theta^{Proficient} + B = 1350$$

where A and B are the scaling constants to transform the *Partially Proficient* and *Proficient* theta cuts to the 1300 and 1350 scale scores, respectively. The scaling constants were applied to a theta score to transform it to the reporting scale score. Appendix B presents the raw-to scale score conversion tables for each grade.

In addition to the total scale score, the scale score for each domain (i.e., Physical Science, Earth and Space Science, and Life Science) is reported individually. The scale scores for the domains are generated by including the items associated with each domain and using the item parameter estimates from the concurrent calibration across all domains. Scores associated with SEPs are not reported per the Technical Advisory Committee's (TAC's) recommendation (ADE, 2022).

8.6. IRT Assumptions

It is important to evaluate how the Rasch models fit the data because reported scale scores are derived from theta estimated under the IRT models. Three major assumptions are investigated: unidimensionality, local item independence, and item fit.

8.6.1. Unidimensionality

An assumption under the Rasch models is unidimensionality; that there is exactly one latent variable an instrument intends to measure (e.g., science proficiency). This is a more traditional and strict definition of the unidimensionality assumption. On the other hand, essential unidimensionality, in which there is one dominant latent variable with some minor latent variable(s), is a more practically applicable assumption (Stout, 1990).

Principal component analysis (PCA) is a statistical technique widely applied to investigate the dimensionality of data (Jackson, 1993; Velicer & Jackson, 1990). Many decision rules have been proposed to determine the number of dimensions using PCA results. Horn's (1965) parallel analysis is a Monte Carlo simulation technique used to determine the number of factors to retain from a PCA. Parallel analysis compares the observed eigenvalues from a correlation matrix to be analyzed with those obtained from uncorrelated normal variables (Ledesma & Valero-Mora, 2007). In other words, expected eigenvalues are obtained by simulating normal, random samples that "parallel" the observed data in terms of sample size and number of variables. Numerous studies have shown parallel analysis to be an effective and appropriate method to determine the number of factors underlying a construct (Glorfeld, 1995; Humphreys & Montanelli, 1975; Zwick & Velicer, 1986), including the least variability and sensitivity to different factors.

PCA was conducted for the operational form in each grade. Table 8.3 presents the first 10 eigenvalues from the PCA for each operational form, as well as the percentage of total variance explained by the first component (%Var). Reckase (1979) claimed that at least 20% of the total variance should be accounted for by the first principal component to obtain acceptable parameter estimates in a unidimensional model. Because the same blueprint was used to construct the operational forms, only one set of eigenvalues from the parallel analysis is presented. The graphical presentation of eigenvalues (i.e., scree plot) is presented in Figure B.10 – Figure B.12 in Appendix B. The PCA results with the parallel analysis criterion and Reckase's index show only one dominant dimension, which supports unidimensionality.

Table 8.3. Eigenvalues from PCA

Grade	1	2	3	4	5	6	7	8	9	10	%Var
5	16.24	1.54	1.16	1.07	1.00	0.96	0.94	0.92	0.90	0.89	32%
8	14.19	1.38	1.13	1.06	0.98	0.97	0.95	0.94	0.93	0.91	28%
11	12.81	1.45	1.29	1.18	1.14	1.09	1.00	0.98	0.92	0.91	26%

8.6.2. Local Item Independence

Local item independence is another assumption under the Rasch models that assumes any item pair is uncorrelated, conditioned on the latent trait an instrument is intended to measure (e.g., science proficiency). A violation of local item independence would impact parameter estimation under the Rasch models because JMLE performed by Winsteps (Linacre, 2022b) relies on uncorrelated item pairs. Winsteps produces raw score residual correlations for pairs of items on a test, which are analogous to Yen's Q3 statistics (Yen, 1984). For an item pair with a residual correlation greater than 0.70, only one item is needed on the test (Linacre, 2022a).

Table 8.4 summarizes the distribution of the residual correlations. Most residual correlations are slightly negative or slightly positive and none are greater than 0.70, which indicates that the local item independence assumption holds for the AzSCI tests.

Table 8.4. Q3 Statistics

Grade	#Item Pairs	Mean	SD	Min.	P10	P25	P50	P75	P90	Max.	#Items Exceeding 0.70
5	1,225	-0.02	0.03	-0.11	-0.05	-0.03	-0.02	-0.01	0.01	0.16	0
8	1,225	-0.02	0.02	-0.10	-0.04	-0.03	-0.02	-0.01	0.00	0.21	0
11	1,225	-0.02	0.03	-0.12	-0.05	-0.03	-0.02	-0.01	0.01	0.23	0

Note. SD = standard deviation, Min. = minimum, P10 = 10th percentile, P25 = 25th percentile, P50 = 50th percentile, P75 = 75th percentile, P90 = 90th percentile, Max. = maximum

8.6.3. Item Fit

Item fit was monitored using weighted mean-square (MNSQ) that indicates the degree of accuracy and predictability with which the data fit the model (Linacre, 2022b). In Winsteps and Rasch literature, weighted mean-square is also referred to as infit MNSQ. The infit MNSQ is sensitive to unexpected responses at or near the item's calibrated level. Items were flagged for misfit using a set of conservative criteria. For infit MNSQ, values less than 0.60 or greater than 1.40 were flagged in accordance with Wright and Linacre's (1994) recommendation.

Table 8.5 presents a summary of the item fit statistics, and Table B.1 – B.3 in Appendix B present the IRT statistics for each item. Items flagged by Winsteps' infit statistics are reviewed during test construction for possible replacement in future administrations.

Table 8.5. IRT Item Fit Summary Statistics

Grade	#Items	#Flagged Items by Infit	%Flagged
5	50	1	2
8	50	0	0
11	50	0	0

Chapter 9: TEST RESULTS

This chapter presents the test results of the spring 2025 AzSCI administration, addressing Standard 1.8, 2.11, 2.15, 3.1, 3.3, 3.6, 3.15, 5.3, 7.4, 12.17, and 12.18 (AERA et al., 2014).

Students receive an overall scale score, and student performance is reported as one of four performance levels: Level 1: *Minimally Proficient*, Level 2: *Partially Proficient*, Level 3: *Proficient*, and Level 4: *Highly Proficient*. Student performance on reporting categories is reported as one of three levels of mastery: *Below Mastery*, *At/Near Mastery*, or *Above Mastery*. Students who score *Below Mastery* demonstrate performance in the reporting category that is clearly below *Proficient*. Students who score *At/Near Mastery* demonstrate performance in the reporting category that is exactly at or immediately above/below *Proficient*. Students who score *Above Mastery* demonstrate performance in the reporting category that is clearly *Proficient* or higher.

The results, summarized below, are based on the population data contained within the final electronic data files (note that the data in this chapter are different from the calibration sample). The results in this section of the technical report may differ slightly from the final testing results presented on the ADE website due to small differences in the application of exclusion rules. Official results typically use more detailed school-level information than is used to conduct research analyses. Please note that the results in the following tables are presented as evidence of reliability and validity of the test scores and should not be used for state accountability purposes.

- Table 9.1 presents the test results for all students by grade, including the mean and standard deviation of the total scale scores and the percentage of students in the overall performance levels. Overall performance levels are determined based on students' total score on the assessment. Results from the last four years are included to show longitudinal performance, in which %Level 3 and 4 combined improved across years for grades 5 and 8 while it improved through 2024 and declined in 2025 for grade 11.
- Table 9.2 presents the percentage of students in each level of mastery by domain. %Level 2 and 3 combined is higher for Earth and Space Science for all grades compared to the other domains.
- Appendix C presents the test results by subgroup. Histograms of the scale score distribution for the total score are also presented.
- Table 9.3 presents the mean and standard deviation of the scale score and the performance level distribution by accommodation for students who used the available accommodations. These tables only include the accommodations captured in the student data file (i.e., accommodations used by students during the spring 2025 administration).
- Table 9.4 presents the frequency distribution statistics for total scale score by performance level. Results indicate that average scale scores increase when moving from lower to higher performance levels across all grades.

Table 9.1. Overall Test Results by Year

Grade	Year	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
5	2025	81,588	1332.29	44.91	27.7	38.5	24.8	9.1
	2024	81,086	1329.81	45.43	30.7	35.2	26.0	8.2
	2023	81,004	1329.97	44.81	29.7	35.9	25.9	8.4
	2022	80,889	1324.21	38.91	28.4	44.3	22.5	4.8
8	2025	82,414	1328.04	40.95	28.6	40.6	25.4	5.3
	2024	82,934	1326.87	40.86	26.3	46.4	21.7	5.5
	2023	85,600	1326.37	40.12	27.2	45.9	22.1	4.9
	2022	87,698	1322.08	37.69	29.8	46.2	20.2	3.8
11	2025	83,348	1319.66	37.46	36.1	44.6	16.0	3.3
	2024	82,097	1320.84	38.87	35.8	41.7	18.4	4.1
	2023	78,651	1321.78	38.17	29.0	49.3	18.9	2.8
	2022	76,418	1319.27	36.99	31.4	48.4	17.0	3.2

Note. SS = scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Table 9.2. Performance Distributions by Domain: Percent of Students at each Level of Mastery

Grade	Domain	N	%Level 1	%Level 2	%Level 3
5	Physical Science	81,588	52.6	22.2	25.2
	Earth and Space Science	81,588	46.0	39.2	14.8
	Life Science	81,588	51.6	25.6	22.8
8	Physical Science	82,414	58.5	23.6	17.8
	Earth and Space Science	82,414	47.2	37.1	15.7
	Life Science	82,414	47.9	30.4	21.8
11	Physical Science	83,348	59.8	26.5	13.7
	Earth and Space Science	83,348	57.0	32.5	10.5
	Life Science	83,348	66.9	23.1	10.0

Note. Level 1 = *Below Mastery*, Level 2 = *At or Around Mastery*, Level 3 = *Above Mastery*

Table 9.3. Test Results by Accommodation

Grade	Accommodation	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
5	Adult Transcription	36	1342.97	63.15	36.1	11.1	25.0	27.8
	American Sign Language	20	1283.15	38.98	85.0	10.0	0.0	5.0
	Assistive Technology	13	1294.08	23.40	69.2	23.1	7.7	0.0
	Braille Test Booklet	6	–	–	–	–	–	–
	Large Print Test Booklet	10	–	–	–	–	–	–
	Sign Test Content	2	–	–	–	–	–	–
	Simplified Directions	322	1294.14	29.50	65.5	30.4	3.4	0.6
	Translate Directions	101	1290.88	23.24	69.3	28.7	2.0	0.0
	Translation Dictionary	133	1292.95	24.89	64.7	33.1	2.3	0.0

Grade	Accommodation	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
8	Adult Transcription	1	–	–	–	–	–	–
	American Sign Language	19	1288.32	22.31	84.2	10.5	5.3	0.0
	Assistive Technology	11	1328.82	58.83	36.4	45.5	0.0	18.2
	Braille Test Booklet	6	–	–	–	–	–	–
	Large Print Test Booklet	6	–	–	–	–	–	–
	Sign Test Content	0	–	–	–	–	–	–
	Simplified Directions	179	1293.19	29.93	69.3	27.9	1.7	1.1
	Translate Directions	74	1290.50	24.21	77.0	21.6	1.4	0.0
	Translation Dictionary	112	1290.47	23.69	75.9	22.3	1.8	0.0
11	Adult Transcription	2	–	–	–	–	–	–
	American Sign Language	22	1293.73	20.47	54.5	45.5	0.0	0.0
	Assistive Technology	4	–	–	–	–	–	–
	Braille Test Booklet	4	–	–	–	–	–	–
	Large Print Test Booklet	7	–	–	–	–	–	–
	Sign Test Content	13	1289.08	19.97	61.5	38.5	0.0	0.0
	Simplified Directions	30	1292.27	22.73	70.0	30.0	0.0	0.0
	Translate Directions	222	1290.05	18.84	76.6	23.0	0.5	0.0
	Translation Dictionary	285	1291.55	20.01	74.4	24.6	0.7	0.4

Note. SS = scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*. Statistics for subgroups with less than 11 students are omitted in compliance with FERPA regulations.

Table 9.4. Scale Score Distribution by Performance Level

Grade	Performance Level	N	Average Scale Score	%	Cumulative %
5	Level 1	22,588	1280.66	27.7	27.7
	Level 2	31,382	1324.42	38.5	66.1
	Level 3	20,232	1370.44	24.8	90.9
	Level 4	7,386	1419.14	9.1	100.0
8	Level 1	23,608	1281.36	28.6	28.6
	Level 2	33,471	1323.58	40.6	69.3
	Level 3	20,931	1369.27	25.4	94.7
	Level 4	4,404	1416.18	5.3	100.0
11	Level 1	30,079	1283.82	36.1	36.1
	Level 2	37,155	1323.01	44.6	80.7
	Level 3	13,359	1370.09	16.0	96.7
	Level 4	2,755	1421.12	3.3	100.0

Note. SS = scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Chapter 10: RELIABILITY AND VALIDITY

This chapter provides evidence supporting the reliability and validity of scores on the spring 2025 AzSCI assessment, addressing Standards 1.8, 1.9, 1.21, 2.3, 2.7, 2.8, 2.11, 2.15, 2.19, 3.1, 3.3, 3.6, 3.15, and 7.4 (AERA et al., 2014).

10.1. Reliability

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) refer to reliability as the "consistency of scores across replications of a testing procedure" (p. 33). A reliable test produces stable scores, meaning that very similar score distributions would result if the test were administered repeatedly under similar conditions to the same students without memory or fatigue affecting the scores. The level of reliability/precision of scores has implications for validity in that the scores must be consistent and precise enough to be useful for intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. The range of certainty around the score should also be small enough to support educational decisions.

Reliability was evaluated based on the internal consistency for all tests. For test reliability, coefficient alpha, which is based on classical test theory (CTT), is a frequently used measure of internal consistency. Coefficient alpha (α) is computed as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right)$$

where *k* is the number of items, σ_X^2 is the variance of the total score, and σ_i^2 is the variance of item *i* (Crocker & Algina, 1986; Cronbach, 1951).

Typically, a test score is obtained from a single observation of performance and represents an estimate of the trait being measured. As an estimate, an observed test score contains some measurement error and does not perfectly reflect an individual's true score. The degree of measurement error in a test score can be estimated using a statistic called the standard error of measurement (SEM), which is calculated as follows:

$$SEM = \sigma_X \sqrt{1 - r}$$

where σ_X is a standard deviation of total score *X*, and *r* is a reliability coefficient, such as the coefficient alpha (Crocker & Algina, 1986).

Table 10.1 presents the coefficient alphas and SEMs (computed based on the calibration sample) for the total and domain scores. These results suggest that the AzSCI assessments produce reliable scores.

Table 10.1. Coefficient Alpha and SEM by Total and Domain Score

Grade	Domain	N	#Items	Coefficient Alpha	SEM
5	Total	81,344	50	0.92	3.30
	Physical Science	81,344	21	0.83	2.03
	Earth and Space Science	81,344	12	0.69	1.63
	Life Science	81,344	17	0.81	2.01
8	Total	82,342	50	0.90	3.24
	Physical Science	82,342	21	0.78	2.10
	Earth and Space Science	82,342	14	0.67	1.78
	Life Science	82,342	15	0.80	1.70
11	Total	82,949	50	0.88	3.27
	Physical Science	82,949	16	0.73	1.89
	Earth and Space Science	82,949	13	0.70	1.51
	Life Science	82,949	21	0.74	2.19

In contrast to the CTT-based SEM, an IRT-based SEM (i.e., CSEM) varies across an ability continuum. The CSEM should be lower around important performance level cuts (e.g., *Proficient*), which indicates higher measurement precision. The CSEM tends to be higher for the upper and lower ends of the ability continuum because there are usually fewer items that measure those difficulty levels. Figure B.4 – Figure B.9 in Appendix B present the TCC and CSEM curves of the assessments. As expected, the CSEMs around the performance level cuts were the lowest.

10.2. Differential Item Functioning

Because test scores can have many sources of variation, the test developers' task is to create assessments that measure the intended abilities and skills without introducing extraneous elements or construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores will reflect these unintended skills and knowledge, as well as what is purportedly assessed by the test. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). One of the factors that may render test scores biased is differing cultural and socioeconomic experiences.

Analysis of DIF is a statistical method to detect potential bias of an item. DIF is defined as a difference between groups (e.g., male and female) in the probability of answering an item correctly. DIF analyses are conditioned on the ability that the assessment is intended to measure (e.g., science proficiency). DIF is an indicator that the item might exhibit bias for one group over the other, not that it actually does. If DIF exists on an item, a committee composed of subject experts reviews the item to determine whether it actually shows bias.

The Mantel-Haenszel (MH) method (Holland & Thayer, 1988; Mantel & Haenszel, 1959) was used to investigate DIF on 1-point items. The MH method is frequently used and efficient in terms of statistical power (Clauser & Mazor, 1998). The Mantel-Haenszel chi-square statistic is computed as follows:

$$MH-\chi^2=\frac{(\sum_k F_k-\sum_k E(F_k))^2}{\sum_k Var(F_k)}$$

where F_k is the sum of scores for the focal group at the *k*th level of the matching variable (Zwick et al., 1993). The MH statistic is sensitive to *N* such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the MH delta (ΔMH) DIF statistic was computed, developed by the Educational Testing Service (ETS). To compute the ΔMH DIF, the MH alpha (the odds ratio) is first computed:

$$\sigma_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k}$$

where N_{r1k} is the number of correct responses in the reference group at ability level *k*, N_{f0k} is the number of incorrect responses in the focal group at ability level *k*, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level *k*, and N_{r0k} is the number of incorrect responses in the reference group at ability level *k*. The *ΔMH DIF* is computed as follows:

$$\Delta MH\ DIF = -2.35 ln(\alpha_{MH})$$

Positive values of *ΔMH DIF* indicate items that favor the focal group, whereas negative values indicate items that favor the reference group. The MH chi-square statistic and the *ΔMH DIF* were used in combination to identify both the operational and field test items that exhibit strong, weak, or no DIF for single-point items.

The standardized mean difference (SMD) is another DIF method applied to multiple-point items (Dorans & Schmitt, 1991; Zwick et al., 1993). The SMD is an effect size index of DIF that compares the mean scores of the reference and focal groups for an item, adjusting for the distribution of the reference and focal groups on the conditioned variable, which for the analyses is the raw score. The SMD is computed as follows:

$$SMD = \sum_{k} P_{F_k} (m_{F_k} - m_{R_k})$$

where P_{F_k} is the proportion of the focal group at the *k*th level of the matching variable, m_{F_k} is the mean score on the item for the focal group at the *k*th level of the matching variable, and m_{R_k} is the mean score on the item for the reference group at the *k*th level of the matching variable (Zwick et al., 1993). A negative SMD value indicates an item in which the focal group has a lower mean than the reference group, conditioned on the matching variable (e.g., science proficiency), whereas a positive SMD value indicates an item for which the reference group has a lower mean than the focal group, conditioned on the matching variable.

Table 10.2 presents the summary of DIF classification criteria for both the MH method and SMD. An alpha level of 0.05 was used for all MH and SMD statistics.

Table 10.2. DIF Flag Categories

Category	Description	MH Criterion	SMD Criterion
A	No DIF	$\|\Delta MH\ DIF\|$ is not significantly different from 0 ($p < 0.05$) or $\|\Delta MH\ DIF\| < 1.0$	MH chi-square not significantly different from 0 ($p < 0.05$) or $\|SMD\| < 0.17$
B	Weak DIF	$\|\Delta MH\ DIF\|$ is significantly different from 0 ($p < 0.05$) and $1.0 \leq \|\Delta MH\ DIF\| < 1.5$ or $\|\Delta MH\ DIF\|$ is significantly different from 0 but not from 1 and $\|\Delta MH\ DIF\| \geq 1.0$	MH chi-square significantly different from 0 ($p < 0.05$) and $0.17 < \|SMD\| \leq 0.25$
C	Strong DIF	$\|\Delta MH\ DIF\|$ is significantly different from 1 ($p < 0.05$) and $\|\Delta MH\ DIF\| \geq 1.5$	MH chi-square significantly higher than 0 ($p < 0.05$) and $\|SMD\| > 0.25$

DIF analysis was conducted for 10 different group pairs:

1. Female vs. Male
2. Hispanic vs. Non-Hispanic
3. American Indian vs. White
4. Asian vs. White
5. Black or African American vs. White
6. Native Hawaiian or Other Pacific Islander vs. White
7. Multi-racial vs. White
8. Students with Disability vs. Students without Disability
9. Economically Disadvantaged vs. Not Economically Disadvantaged
10. English Learner vs. English as a First Language

Table 10.3 presents the number of operational items exhibiting strong DIF between any two groups. Any items that display strong DIF are flagged for possible replacement in the future administration, as strong DIF is one of the holistic item replacement evaluation criteria used for item selection. DIF results with a sample size of less than 200 per group should not be considered statistically reliable (Clauser & Mazor, 1998; Mazor et al., 1992).

Table 10.3. Number of Items Exhibiting Strong DIF

Grade	Total #Items	#Items with Strong DIF
5	50	0
8	50	1
11	50	1

10.3. Correlations Among Domains

Correlations were examined between the total raw score and domain raw scores (Physical Science, Earth and Space Science, and Life Science). The data used to calculate the correlations were based on the calibration sample described in Chapter 8. Disattenuated correlations between were also computed, calculated based on the following formula:

$$r_{T_{xy}} = \frac{r_{xy}}{\sqrt{r_x r_y}}$$

where $r_{T_{xy}}$ is a corrected correlation for attenuation between scores *x* and *y*, r_{xy} is an observed correlation between the scores *x* and *y*, and r_x and r_y are reliabilities for *x* and *y*, respectively. Coefficient alphas (presented in Table 10.1) were used to calculate the corrected correlation coefficients for attenuation. The disattenuated correlations could be greater than 1.00.

Table 10.4 presents the test correlations and disattenuated correlations between the total raw score and the domain raw scores. The numbers in the lower diagonal of the table are the disattenuated correlations.

Table 10.4. Correlations and Disattenuated Correlations between Total and Domain Raw Scores

Grade	Score	Total	Physical Science	Earth and Space Science	Life Science
5	Total	1.00	0.94	0.86	0.93
	Physical Science	1.08	1.00	0.74	0.80
	Earth and Space Science	1.08	0.98	1.00	0.70
	Life Science	1.08	0.98	0.94	1.00
8	Total	1.00	0.93	0.87	0.91
	Physical Science	1.11	1.00	0.71	0.77
	Earth and Space Science	1.12	0.98	1.00	0.71
	Life Science	1.07	0.97	0.97	1.00
11	Total	1.00	0.90	0.87	0.92
	Physical Science	1.12	1.00	0.72	0.72
	Earth and Space Science	1.11	1.01	1.00	0.71
	Life Science	1.14	0.98	0.99	1.00

10.4. Validity Evidence

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 11). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for a particular purpose or use.

A validity argument should begin with clear statements regarding the purpose of a test and intended interpretations and uses of the test results. The purpose of the AzSCI tests is to assess the science proficiency of students based on the Arizona Science Standards. The objective of the proceeding sections is to highlight validity evidence for each aspect and to guide readers where to look for the evidence. Different aspects of validity evidence, which are in line with the *Standards* (AERA et al., 2014), are considered throughout this technical report. Providing validity evidence is an ongoing activity for any assessment as it matures.

10.4.1. Evidence Based on Test Content

Validity evidence based on test content refers to the extent to which a test is aligned with the construct the assessment is intended to measure (AERA et al., 2014). AzSCI measures a student's level of science proficiency based on the skills specified in the Arizona Science Standards. Thus, alignment of the AzSCI test to the standards is critical. A goal is testing every standard within a three-year window.

Item specifications and test blueprints are the core documents that ensure that the assessments are aligned to the Arizona Science Standards, as described in Chapter 2. The AzSCI specifications and blueprints were developed in an iterative process involving ADE, Pearson, and a committee of Arizona educators. The item specifications help define how the content in the Arizona Science Standards could be assessed given the proposed format of the AzSCI test. The test blueprint defines the standards to be assessed for each test form, the number of items per standard, the number of item types, the number of points per item type, and the total number of items and points per test form. For AzSCI, it was important to consider the relative weight of Physical Science, Life Science, and Earth and Space Science for each grade.

Once the item specifications and blueprints were established, item and test development took place. It was a rigorous and iterative process involving the Pearson content team and ADE to ensure that the AzSCI assessments meet the test blueprint and other content criteria and psychometric targets, as described in Chapter 3. Beyond the test blueprint, ADE and Pearson attempted to include items measuring different levels of rigor to cover the Arizona Science Standards as much as possible.

Alignment of test forms to the test blueprints is a thoughtful, careful task that involves collaboration among assessment specialists, psychometricians, and ADE. Developing test forms is challenging because test blueprints can be highly complex, specifying not only the range of items and points for each reporting category and standard, but also cross-cutting criteria such as distribution across item types, TAGS, writing genre, etc. In addition to meeting complex blueprint requirements, test developers worked to meet psychometric goals so that accommodated test forms measure equivalently across the range of student ability.

The results from the alignment study (Christopherson, 2023) were used to identify which standards required additional development, ensuring that all standards are assessed on a rotating basis within a three-year period, as described in Section 3.3.

10.4.2. Evidence Based on Response Processes

Evidence based on response processes refers to the cognitive process engaged in by students when answering test items, or the "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA et al., 2014, p. 15). A full standalone field test was administered in spring 2021 to try out a large group of items aligned to the 2018 standards, evaluate psychometric characteristics of the items and item clusters, and build an operational item bank. An online survey was prepared for test administrators to provide feedback about the student experience on the AzSCI field test administration. Results from this survey were analyzed by ADE and Pearson to improve the AzSCI assessment for future administrations. For more information about the full standalone field test, please refer to the spring 2021 AzSCI field test technical report (ADE, 2021).

As described in Chapter 3, all newly developed items also go through a rigorous item review process, including content, bias, and sensitivity committees with Arizona educators, parents, and community members. Reviewers evaluated the item for its alignment to the Arizona Science Standards, grade appropriateness, editorial completeness and accuracy, and the presence of any content that could be biased or sensitive in nature. Only the items accepted by the committees were considered eligible to be field tested.

10.4.3. Evidence Based on Internal Structure

Validity evidence based on internal structure refers to the extent to which an item or a component of a test ties to the assessment it is intended to measure (AERA et al., 2014, p. 16). AzSCI is designed to measure students' overall science proficiency based on the Arizona Science Standards composed of the Physical Science, Life Science, and Earth and Space Science domains. AzSCI items across all domains were calibrated concurrently under the unidimensional Rasch models (Masters, 1982; Rasch, 1960) as described in Chapter 8. To evaluate the unidimensionality assumption of the Rasch models, PCA was conducted for each operational form. The results of the PCA analysis with the parallel analysis criterion (Horn, 1965) and Reckase's index (1979) presented in Table 8.3 indicated there is one dominant dimension for science and the remaining components are non-significant.

Another assumption under the Rasch models is local item independence. The local item independence assumption is typically evaluated using Q3 statistics (Yen, 1984). Winsteps (Linacre, 2022b) produces raw score residual correlations for pairs of items on a test, which are analogous to the Q3 statistics. A distribution of the residual correlations by form, presented in Table 8.4, showed that most statistics are either slightly negative or slightly positive, which indicates the item independence assumption generally holds for AzSCI.

In addition to the total scale score, the scale score for each domain (i.e., Physical Science, Earth and Space Science, and Life Science) is reported individually. The scale scores for the domains are generated by including the items associated with each domain and using the item parameter estimates from the concurrent calibration across all domains. Details about scaling methods are described in Section 8.5. Correlations between the total score and domain score are presented in Table 10.4 and showed they are at least moderately, if not highly, correlated to each other, as expected.

A point-biserial correlation, as an indicator of interrelationship between an item and a construct that it is intended to measure, is calculated as a correlation between an item raw score and a total raw score. The point-biserial correlations should be higher than or equal to 0.25, as any item with a lower correlation is flagged during item selection. It is one of the psychometric criteria considered for item selection. The point-biserial correlation was calculated for distractors of multiple-choice items as well. Table 7.4 and Table 7.5 show that all the multiple-choice items have negative point-biserial correlations, except a few distractors with a slightly positive correlation close to zero. The results indicate that the distractors work as expected.

Differential item functioning (DIF) analysis is a statistical method to detect potential bias of an item for (or against) a manifest group (e.g., female). DIF is defined as a difference between groups (e.g., male and female) in the probability of getting an item correct, given the same level of ability within the construct that an assessment is intended to measure. Details on DIF analysis are presented in Section 10.2. Items showing strong DIF are flagged for possible replacement in future administrations.

10.4.4. Evidence Based on Performance Standards

Validity evidence concerning performance standards refers to the extent to which passing scores are aligned to performance standards (Kane, 1994). Performance level descriptors (PLDs) highlight the knowledge, skills, and processes students possess at different performance levels (Egan et al., 2012). The PLDs are the foundation of standard setting meetings. The PLDs for AzSCI, provided on the ADE website at https://www.azed.gov/assessment/sci/, were carefully developed by Pearson, reviewed by a group of Arizona educators in 2021, and approved for use in the standard setting conducted in June 2022 where the performance level cut scores for the AzSCI assessment were recommended by a group of educators using the Extended Modified (Yes/No) Angoff standard setting method. See Section 11.1 for more details on standard setting.

10.4.5. Evidence Based on Relations to Other Variables

Validity evidence concerning a relation to other variables refers to the extent to which test scores are related to other external measures (AERA et al., 2014, p. 16). Arizona's Academic Standards Assessment (AASA) is Arizona's statewide content-based achievement test for mathematics and English language arts (ELA). Because the AzSCI and AASA assessments are administered to all eligible Arizona students, scores on the tests are expected to be positively correlated.

Table 10.5 presents the correlation between AzSCI and AASA scale scores from the spring 2025 administration. AzSCI is highly correlated with both AASA ELA and mathematics, with the correlations ranging from 0.76 to 0.83. The correlation is higher with ELA than mathematics for both grades, which could be attributed to AzSCI including relatively high reading loads compared to mathematics. AASA is not administered to high school students.

Table 10.5. Correlation between AzSCI and AASA Scale Scores

	AASA ELA		AASA Mathematics	
Grade	N	Correlation	N	Correlation
5	77,754	0.83	77,934	0.78
8	78,077	0.80	78,383	0.76

Error! Not a valid bookmark self-reference. presents the correlation between the AzSCI and ACT scale scores from the spring 2025 administration for grade 11. AzSCI is highly correlated with ACT ELA, mathematics, and science, with the correlations ranging from 0.68 to 0.71. The correlation is higher with ELA than mathematics and science.

Table 10.6. Correlation between AzSCI and ACT Scale Scores

	ACT ELA		ACT Mathematics		ACT Science	
Grade	N	Correlation	N	Correlation	N	Correlation
11	79,022	0.71	81,447	0.68	81,324	0.70

10.4.6. Summary

Overall, the validity evidence supports the use of AzSCI scores. The PCA revealed unidimensionality of AzSCI, which supports the use of unidimensional Rasch models. The AzSCI scores were also positively correlated to the AASA ELA and mathematics scores for grades 5 and 8 and to the ACT ELA, mathematics, and science for grade 11. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment cycle. Additional evidence should and will be added to the AzSCI technical report in the future, as appropriate.

Chapter 11: CLASSIFICATION INTO PERFORMANCE LEVELS

Scores from the AzSCI tests are used to classify students into one of four performance levels: *Minimally Proficient, Partially Proficient, Proficient,* and *Highly Proficient*. This chapter provides information regarding classification of students into these four categories, including the consistency and accuracy with which students who took the spring 2025 AzSCI assessment were assigned to the performance levels, addressing Standards 1.8, 1.9, 2.13, 2.14, 2.16, 5.5, 5.21, 5.22, 5.23, and 7.4 (AERA et al., 2014).

11.1. Standard Setting

Arizona educators made recommendations for cut scores for each performance level on the AzSCI assessments during the standard setting workshop in June 2022 using the Extended Modified (Yes/No) Angoff procedure (Davis & Moyer, 2015; Plake et al., 2005). The cut scores were ultimately approved by the State Board of Education in July 2022. Documentation regarding the standard setting is provided in the standard setting report (Pearson, 2022).

Table 11.1 presents the final scale score ranges for the AzSCI performance levels, and Table 11.2 presents the scale score and associated CSEM at the performance level cuts. The performance level cuts were set to 1300 and 1350 for *Partially Proficient* and *Proficient*, respectively, whereas the cut score for *Highly Proficient* was allowed to freely vary for each grade. The CSEM is identical across all grades within each cut (i.e., 13 for *Partially Proficient*, 12 for *Proficient*, and 14 for *Highly Proficient*).

Table 11.1. Performance Level Cut Scores

Grade	*Minimally Proficient*	*Partially Proficient*	*Proficient*	*Highly Proficient*
5	1200–1299	1300–1349	1350–1394	1395–1500
8	1200–1299	1300–1349	1350–1398	1399–1500
11	1200–1299	1300–1349	1350–1401	1402–1500

Table 11.2. CSEM at Performance Level Cuts

	Partially Proficient Cut		*Proficient* Cut		*Highly Proficient* Cut	
Grade	Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
5	1300	13	1350	12	1395	14
8	1300	13	1350	12	1399	14
11	1300	13	1350	12	1402	14

11.2. Classification Consistency and Accuracy

Classification consistency is the agreement between students' performance level classification from two independent administrations of the same test (or two parallel forms of the test). Classification accuracy refers to the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston & Lewis, 1995).

In conjunction with internal consistency, classification consistency is an important type of reliability and is particularly relevant to high-stakes decisions, such as passing or not passing the AzSCI tests. As a form of reliability, classification consistency represents how reliably students can be classified into performance levels. For tests such as AzSCI, classification consistency is most important for students whose ability is near the *Proficient* cut score. Students whose ability is far above or far below the value established for *Proficient* are unlikely to be misclassified because repeated administration of the test will nearly always result in the same classification. Students whose true scores are close to the cut score are a more serious concern. These students' true scores will likely lie within the SEM of the cut score. For this reason, the measurement error at the cut scores should be considered when evaluating the classification consistency of a test.

Classification consistency and accuracy were estimated using the total scale score for the *Proficient* cut based on the procedures described by Livingston and Lewis (1995). Classification consistency is calculated as the proportion of students classified consistently between two parallel forms, and classification accuracy is calculated as the proportion of students classified the same between observed scores and true scores, as shown in Table 11.3.

Table 11.3. Classification Consistency and Accuracy for the *Proficient* Cut

Observed Performance	Expected Performance	Classification Outcome
Not Proficient	Not Proficient	Consistent Classification
Not Proficient	Proficient	Inconsistent Classification
Proficient	Not Proficient	Inconsistent Classification
Proficient	Proficient	Consistent Classification
Not Proficient	Not Proficient	Accurate Classification
Not Proficient	Proficient	False Negative
Proficient	Not Proficient	False Positive
Proficient	Proficient	Accurate Classification

Cohen's kappa (κ) coefficient (Cohen, 1960) is another way of expressing overall consistency. This statistic assesses the proportion of consistent classification expected beyond chance and is therefore most often lower than the unadjusted value of overall consistency. Cohen's kappa is calculated as follows:

$$\kappa = \frac{P - P_c}{1 - P_c}$$

where P_c is the probability of consistent classification by chance, and P is the probability of consistent classification (unadjusted by chance). Students can be misclassified in one of two ways. Students who are truly not *Proficient* but were classified as being *Proficient*, based on the assessment, are false positives. Similarly, students who are truly *Proficient* but were classified as being not *Proficient* are false negatives.

Table 11.4 presents the classification consistency and accuracy results, generated by BB-class (Brennan, 2004). These results are for classifying students into four performance levels using the total score on the assessment for students in the calibration sample. Included in the table are the sample size (N), classification consistency (Consistency), classification inconsistency (Inconsistency), probability of consistent classification by chance (Chance), Cohen's Kappa (κ), classification accuracy (Accuracy), false positive (False Positive), and false negative (False Negative). Inconsistency is defined as one minus Consistency.

Table 11.4. Classification Consistency and Accuracy Results

Grade	N	Consistency	Inconsistency	Chance	κ	Accuracy	False Positive	False Negative
5	81,344	0.73	0.27	0.29	0.62	0.81	0.10	0.09
8	82,342	0.72	0.28	0.31	0.59	0.80	0.11	0.09
11	82,949	0.71	0.29	0.36	0.55	0.79	0.12	0.08

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. AERA.

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Lawrence Erlbaum Associates.

Arizona Department of Education (ADE). (2021). *AzSCI 2021 field test technical report*. Pearson.

Arizona Department of Education (ADE). (2022, October). *Technical advisory committee meeting, October 12–13, 2022 – hybrid: Meeting notes*. Internal publication.

Brennan, R. L. (2004). BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy [computer software] (Version 1.0). University of Iowa.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. http://dx.doi.org/10.1177/001316446002000104

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 12*, 671–684.

Davis, L. L., & Moyer, E. L. (2015). *PARCC performance level setting technical report*. Partnership for Assessment of Readiness for College and Careers (PARCC).

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach.* ETS Research Report 91-47. Educational Testing Service.

Egan, K. A., Schneider, C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed work. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Erlbaum.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55,* 377–393.

Green, D. R. (1975, December). *Procedures for assessing bias in achievement tests.* Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Harlen, W. (Ed.). (2015). *Working with big ideas of science education*. InterAcademy Partnership (IAP). https://www.azed.gov/sites/default/files/2021/09/Working%20with%20Big%20Ideas%20of%20Science%20Education.pdf

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193–206.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*(8), 2204–2214.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions*, *17*, 133–159.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research, and Evaluation, 12,* 2.

Linacre, J. M. (2022a). Winsteps® Rasch measurement computer program user's guide, Version 4.8.1.0. Winsteps.com.

Linacre, J. M. (2022b). *Winsteps® (Version 4.8.1.0)* [Computer Software]. http://www.winsteps.com/

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443–451. https://doi.org/10.1177/0013164492052002020

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://www.azed.gov/sites/default/files/2021/09/Framework%20for%20K-12%20Science%20Education.pdf

Pearson. (2022, June). *Arizona Science (AzSCI) standard setting meeting*. Report prepared under contract with the Arizona Department of Education.

Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting multiple performance standards using the yes/no method: An alternative item mapping method*. Meeting of the National Council on Measurement in Education, Montreal, Canada.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danmarks Paedogogiske Institut.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230.

Stout, W. F. (1990). A new item response theory modelling approach and applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Tekkumru-Kisa, M., Stein, M. K., & Schunn, C. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science. *Journal of Research in Science Teaching, 52*(5), 659–685. https://www.lrdc.pitt.edu/schunn/research/papers/tekkumru-kisa-stein-schunn-2015.pdf

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1–28.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 26,* 44–66.

Appendix A: ITEM-LEVEL CTT STATISTICS

This appendix includes the following item-level CTT results:

- Table A.1 – Table A.3 present the item-level CTT statistics for each grade, including the item type, maximum number of points possible, number of students (N), *p*-value, and the point-biserial correlation between an item and total raw score.
- Table A.4 – Table A.6 present the item-level distractor analysis for the multiple-choice items, including the percentage of students who selected correct and incorrect response options, the point-biserial correlation associated with each option, and the overall omission rate for the item.

Table A.1. Item-Level CTT Statistics, Grade 5

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	81,344	0.55	0.21
2	MC	1	81,344	0.77	0.41
3	XI	1	81,344	0.49	0.31
4	XI	1	81,344	0.42	0.45
5	MC	1	81,344	0.51	0.29
6	MX	1	81,344	0.66	0.47
7	MC	1	81,344	0.21	0.36
8	MX	2	81,344	0.63	0.64
9	MC	1	81,344	0.34	0.34
10	MC	1	81,344	0.36	0.59
11	XI	1	81,344	0.71	0.47
12	XI	1	81,344	0.54	0.43
13	MC	1	81,344	0.43	0.52
14	MX	1	81,344	0.19	0.50
15	MX	2	81,344	0.43	0.55
16	MC	1	81,344	0.29	0.31
17	MC	1	81,344	0.33	0.39
18	MX	1	81,344	0.31	0.52
19	XI	1	81,344	0.32	0.31
20	MX	1	81,344	0.44	0.36
21	MX	1	81,344	0.18	0.34
22	MC	1	81,344	0.41	0.46
23	MC	1	81,344	0.38	0.26
24	MC	1	81,344	0.46	0.51
25	MX	1	81,344	0.39	0.61
26	MC	1	81,344	0.66	0.41
27	MX	1	81,344	0.55	0.51
28	MC	1	81,344	0.20	0.43
29	XI	1	81,344	0.40	0.37
30	MX	2	81,344	0.40	0.41
31	MC	1	81,344	0.48	0.50
32	MX	2	81,344	0.47	0.35
33	MX	2	81,344	0.51	0.51

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
34	MC	1	81,344	0.71	0.53
35	XI	1	81,344	0.60	0.53
36	MC	1	81,344	0.55	0.57
37	MX	1	81,344	0.33	0.53
38	MC	1	81,344	0.46	0.30
39	MC	1	81,344	0.51	0.50
40	MC	1	81,344	0.56	0.41
41	XI	1	81,344	0.35	0.48
42	MX	1	81,344	0.34	0.48
43	MX	1	81,344	0.59	0.51
44	MC	1	81,344	0.51	0.63
45	MC	1	81,344	0.43	0.40
46	MX	1	81,344	0.51	0.44
47	MC	1	81,344	0.64	0.58
48	MC	1	81,344	0.47	0.36
49	MC	1	81,344	0.48	0.52
50	MC	1	81,344	0.54	0.51

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table A.2. Item-Level CTT Statistics, Grade 8

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	82,342	0.78	0.37
2	XI	1	82,342	0.24	0.21
3	MX	1	82,342	0.32	0.41
4	MX	1	82,342	0.39	0.35
5	MX	1	82,342	0.38	0.40
6	MC	1	82,342	0.47	0.29
7	MX	1	82,342	0.30	0.50
8	MC	1	82,342	0.48	0.51
9	MC	1	82,342	0.78	0.37
10	MX	1	82,342	0.42	0.59
11	MX	2	82,342	0.37	0.56
12	MC	1	82,342	0.30	0.36
13	MC	1	82,342	0.54	0.50
14	MX	1	82,342	0.30	0.46
15	MC	1	82,342	0.27	0.34
16	MC	1	82,342	0.36	0.20
17	MC	1	82,342	0.63	0.43
18	XI	1	82,342	0.52	0.51
19	MC	1	82,342	0.41	0.43
20	MC	1	82,342	0.36	0.48
21	MX	1	82,342	0.71	0.36
22	XI	1	82,342	0.65	0.44
23	XI	1	82,342	0.57	0.26
24	MX	2	82,342	0.35	0.37

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
25	MC	1	82,342	0.71	0.48
26	MC	1	82,342	0.37	0.45
27	MX	2	82,342	0.45	0.58
28	XI	1	82,342	0.29	0.43
29	MC	1	82,342	0.49	0.36
30	MX	1	82,342	0.34	0.51
31	MC	1	82,342	0.52	0.45
32	MC	1	82,342	0.48	0.20
33	MX	2	82,342	0.22	0.25
34	MX	1	82,342	0.18	0.33
35	XI	1	82,342	0.24	0.38
36	MC	1	82,342	0.68	0.48
37	MC	1	82,342	0.39	0.45
38	MX	1	82,342	0.47	0.55
39	MX	1	82,342	0.27	0.44
40	MC	1	82,342	0.27	0.46
41	MC	1	82,342	0.53	0.38
42	MC	1	82,342	0.62	0.41
43	MC	1	82,342	0.42	0.49
44	XI	1	82,342	0.56	0.51
45	MX	2	82,342	0.42	0.44
46	MX	1	82,342	0.27	0.41
47	MX	1	82,342	0.20	0.33
48	XI	1	82,342	0.33	0.50
49	MC	1	82,342	0.63	0.46
50	MC	1	82,342	0.17	0.36

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table A.3. Item-Level CTT Statistics, Grade 11

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	82,949	0.50	0.24
2	MC	1	82,949	0.51	0.37
3	MC	1	82,949	0.30	0.24
4	XI	1	82,949	0.30	0.43
5	XI	1	82,949	0.23	0.35
6	MX	1	82,949	0.22	0.44
7	MC	1	82,949	0.45	0.34
8	MC	1	82,949	0.41	0.47
9	MX	1	82,949	0.18	0.40
10	MX	1	82,949	0.39	0.42
11	MC	1	82,949	0.30	0.45
12	MX	2	82,949	0.40	0.47
13	MC	1	82,949	0.24	0.29
14	XI	1	82,949	0.27	0.33
15	MX	1	82,949	0.19	0.39

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
16	MC	1	82,949	0.47	0.26
17	MC	1	82,949	0.47	0.43
18	MC	1	82,949	0.38	0.44
19	MC	1	82,949	0.29	0.34
20	MC	1	82,949	0.46	0.46
21	MC	1	82,949	0.37	0.35
22	MX	1	82,949	0.31	0.29
23	MC	1	82,949	0.55	0.39
24	MX	1	82,949	0.35	0.56
25	MX	2	82,949	0.40	0.55
26	MC	1	82,949	0.24	0.36
27	MX	1	82,949	0.32	0.47
28	MC	1	82,949	0.61	0.46
29	MC	1	82,949	0.29	0.52
30	MX	2	82,949	0.56	0.50
31	MC	1	82,949	0.50	0.37
32	MX	1	82,949	0.39	0.53
33	MC	1	82,949	0.62	0.38
34	MX	1	82,949	0.61	0.55
35	MX	2	82,949	0.23	0.21
36	MC	1	82,949	0.26	0.27
37	MC	1	82,949	0.27	0.33
38	MC	1	82,949	0.30	0.57
39	XI	1	82,949	0.58	0.49
40	MX	2	82,949	0.29	0.41
41	XI	1	82,949	0.17	0.41
42	MX	1	82,949	0.46	0.47
43	MC	1	82,949	0.30	0.29
44	MC	1	82,949	0.46	0.34
45	MC	1	82,949	0.27	0.19
46	MC	1	82,949	0.29	0.10
47	MC	1	82,949	0.34	0.41
48	MC	1	82,949	0.35	0.29
49	MX	1	82,949	0.32	0.39
50	MX	1	82,949	0.15	0.51

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table A.4. Distractor Analysis of Multiple-Choice Items, Grade 5

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%Omit
1	54.6	0.21	23.4	-0.13	5.7	-0.17	16.3	-0.03	0.0
2	76.6	0.41	10.2	-0.24	10.3	-0.23	3.0	-0.19	0.0
5	51.5	0.29	27.9	-0.03	14.6	-0.27	6.0	-0.15	0.0
9	33.8	0.34	19.9	-0.07	23.1	-0.13	23.2	-0.19	0.1
13	42.5	0.52	19.1	-0.18	16.8	-0.28	21.5	-0.20	0.1
16	29.4	0.31	24.7	-0.15	29.2	-0.08	16.6	-0.11	0.0
17	33.3	0.39	28.3	-0.13	23.2	-0.19	15.1	-0.12	0.0
22	40.9	0.46	16.1	-0.13	13.4	-0.27	29.5	-0.19	0.1
23	38.2	0.26	25.7	-0.02	23.0	-0.15	13.0	-0.14	0.1
24	45.9	0.51	21.1	-0.14	13.0	-0.29	19.9	-0.25	0.1
26	65.5	0.41	11.4	-0.16	16.6	-0.27	6.4	-0.18	0.1
31	47.7	0.50	14.5	-0.30	15.7	-0.20	22.1	-0.17	0.0
34	70.9	0.53	8.2	-0.27	11.6	-0.30	9.2	-0.25	0.0
36	54.7	0.57	9.1	-0.22	26.6	-0.30	9.6	-0.30	0.0
38	45.7	0.30	20.5	-0.01	13.2	-0.30	20.6	-0.11	0.0
39	50.7	0.50	8.9	-0.22	21.9	-0.27	18.5	-0.20	0.1
40	55.8	0.41	19.7	-0.21	10.5	-0.29	13.9	-0.09	0.0
44	51.3	0.63	9.9	-0.22	21.1	-0.36	17.6	-0.27	0.1
45	43.3	0.40	20.8	-0.14	21.5	-0.19	14.3	-0.18	0.1
47	64.1	0.58	6.7	-0.21	13.0	-0.30	16.1	-0.33	0.1
48	47.2	0.36	10.4	-0.15	27.3	-0.16	15.1	-0.18	0.1
49	48.2	0.52	26.0	-0.16	15.9	-0.32	10.0	-0.25	0.1
50	54.4	0.51	16.8	-0.28	18.9	-0.28	9.9	-0.15	0.1

Note. Pt. Bis. = Point-Biserial. The item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table A.5. Distractor Analysis of Multiple-Choice Items, Grade 8

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%Omit
1	78.0	0.37	9.1	-0.22	10.7	-0.24	2.3	-0.10	0.0
6	47.4	0.29	11.8	-0.14	26.3	-0.10	14.4	-0.15	0.0
9	77.9	0.37	4.9	-0.15	7.2	-0.19	10.0	-0.24	0.0
12	30.5	0.36	11.8	-0.26	22.0	-0.26	35.8	0.06	0.1
13	54.4	0.50	12.4	-0.15	19.9	-0.29	13.2	-0.24	0.1
15	26.7	0.34	22.2	-0.19	31.9	-0.08	19.2	-0.09	0.1
16	35.7	0.20	20.7	-0.19	26.1	-0.22	17.5	0.20	0.0
17	62.8	0.43	10.8	-0.23	14.1	-0.25	12.4	-0.15	0.0
19	41.4	0.43	29.8	-0.19	9.9	-0.22	19.0	-0.15	0.0
25	70.7	0.48	7.4	-0.22	8.3	-0.29	13.6	-0.24	0.0
29	49.4	0.36	12.3	-0.20	19.7	-0.27	18.5	-0.02	0.1
31	52.1	0.45	16.6	-0.16	15.6	-0.28	15.7	-0.17	0.0
32	47.6	0.20	28.5	0.02	11.9	-0.17	12.0	-0.17	0.0
36	68.2	0.48	15.4	-0.22	12.0	-0.32	4.4	-0.19	0.0
37	38.8	0.45	21.9	-0.28	24.4	-0.24	14.8	0.00	0.0
41	52.9	0.38	14.2	-0.29	16.4	-0.15	16.5	-0.10	0.0
42	61.9	0.41	19.0	-0.17	11.5	-0.27	7.5	-0.18	0.0
43	41.9	0.49	24.4	-0.25	25.4	-0.18	8.3	-0.20	0.0
49	62.8	0.46	8.0	-0.21	15.1	-0.30	14.1	-0.16	0.1

Note. Pt. Bis. = Point-Biserial. The item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table A.6. Distractor Analysis of Multiple-Choice Items, Grade 11

Item	Correct Option		Distractor 1		Distractor 2		Distractor 3		
Number	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%Omit
1	49.8	0.24	12.3	-0.16	9.4	-0.25	28.6	0.01	0.0
2	51.5	0.37	34.0	-0.26	6.5	-0.22	8.0	-0.04	0.0
3	30.0	0.24	10.6	-0.05	34.9	-0.09	24.5	-0.12	0.0
7	44.6	0.34	20.7	-0.25	25.6	-0.08	9.0	-0.12	0.1
8	41.1	0.47	17.9	-0.27	26.3	-0.20	14.7	-0.12	0.1
11	30.1	0.45	19.6	-0.08	25.7	-0.20	24.3	-0.20	0.2
13	24.0	0.29	21.9	-0.01	21.2	-0.22	32.8	-0.06	0.2
16	47.2	0.26	22.6	-0.07	21.3	-0.21	8.8	-0.03	0.0
17	46.6	0.43	18.9	-0.16	18.5	-0.29	16.1	-0.10	0.0
19	28.6	0.34	29.8	-0.17	29.0	-0.14	12.7	-0.05	0.0
20	45.5	0.46	15.7	-0.20	21.8	-0.17	17.0	-0.24	0.0
21	36.9	0.35	24.8	-0.13	20.2	-0.14	18.2	-0.14	0.1
23	54.8	0.39	15.9	-0.13	17.7	-0.25	11.5	-0.16	0.1
28	60.6	0.46	7.7	-0.18	22.2	-0.26	9.4	-0.24	0.1
31	49.9	0.37	35.8	-0.26	10.4	-0.16	3.9	-0.04	0.0
33	62.1	0.38	9.7	-0.16	14.4	-0.23	13.8	-0.16	0.0
36	25.8	0.27	19.6	-0.25	30.6	-0.07	23.9	0.03	0.1
43	30.5	0.29	18.9	-0.22	16.6	-0.14	34.0	0.02	0.1
44	46.0	0.34	16.7	-0.10	23.6	-0.22	13.7	-0.12	0.1
45	26.8	0.19	34.9	0.08	21.5	-0.15	16.8	-0.16	0.1
46	29.0	0.10	30.2	0.05	19.9	-0.22	20.8	0.04	0.1
47	34.2	0.41	13.9	-0.18	23.2	-0.20	28.6	-0.10	0.1
48	35.2	0.29	21.8	-0.13	28.2	-0.08	14.8	-0.13	0.1

Note. Pt. Bis. = Point-Biserial. The item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Appendix B: ITEM-LEVEL IRT STATISTICS

This appendix includes the following item-level IRT results:

- Table B.1 – Table B.3 present the IRT statistics, including item type, Rasch difficulty, standard error (SE) of Rasch, and infit values.
- Table B.4 – Table B.6 present the raw-to-scale score conversion tables.
- Figure B.1 – Figure B.3 present the item-person map for each post-equated operational form.
- Figure B.4 – Figure B.9 present the test characteristic curve (TCC), test information curve (TIF) and conditional standard error of measurement (CSEM) curve for each post-equated operational form.
- Figure B.10 – Figure B.12 present the scree plot from the principal component analysis (PCA) for each operational form. The scree plot shows only the first 10 components.

Table B.1. Item-Level IRT Statistics, Grade 5

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-0.2954	0.0078	1.24
2	MC	-1.4952	0.0089	0.94
3	XI	0.0064	0.0078	1.15
4	XI	0.3279	0.0079	1.00
5	MC	-0.1099	0.0078	1.17
6	MX	-0.8755	0.0081	0.91
7	MC	1.5978	0.0095	1.03
8	MX	-0.6669	0.0054	0.83
9	MC	0.7701	0.0083	1.11
10	MC	0.6418	0.0081	0.83
11	XI	-1.1822	0.0084	0.91
12	XI	-0.2648	0.0078	1.00
13	MC	0.3093	0.0079	0.91
14	MX	1.7318	0.0098	0.85
15	MX	0.2711	0.0056	1.04
16	MC	0.7303	0.0082	1.08
17	MC	0.8765	0.0084	1.06
18	MX	0.8590	0.0084	0.87
19	XI	0.8268	0.0083	1.14
20	MX	0.3458	0.0079	1.11
21	MX	1.8292	0.0100	1.00
22	MC	0.3908	0.0080	0.98
23	MC	0.5315	0.0081	1.20
24	MC	0.1396	0.0079	0.92
25	MX	0.4895	0.0080	0.82
26	MC	-0.8566	0.0081	0.99
27	MX	-0.3107	0.0078	0.91
28	MC	1.6687	0.0097	0.96
29	XI	0.4552	0.0080	1.08
30	MX	0.4686	0.0058	1.24

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
31	MC	0.0407	0.0078	0.94
32	MX	0.0558	0.0052	1.51
33	MX	-0.0823	0.0059	1.07
34	MC	-1.1577	0.0084	0.84
35	XI	-0.5881	0.0079	0.87
36	MC	-0.2999	0.0078	0.85
37	MX	0.8226	0.0083	0.90
38	MC	0.1464	0.0079	1.16
39	MC	-0.1021	0.0078	0.94
40	MC	-0.3580	0.0078	1.02
41	XI	0.6174	0.0081	0.93
42	MX	0.7440	0.0082	0.95
43	MX	-0.5064	0.0079	0.90
44	MC	-0.2616	0.0078	0.80
45	MC	0.5553	0.0081	1.10
46	MX	-0.0359	0.0078	1.00
47	MC	-0.7802	0.0080	0.82
48	MC	-0.0736	0.0078	1.09
49	MC	0.0328	0.0078	0.91
50	MC	-0.3268	0.0078	0.92

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced, SE = standard error, MNSQ Infit = mean-square infit. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table B.2. Item-Level IRT Statistics, Grade 8

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-1.9156	0.0090	0.95
2	XI	0.9737	0.0089	1.18
3	MX	0.1913	0.0079	0.95
4	MX	0.0280	0.0078	1.06
5	MX	0.1539	0.0079	1.02
6	MC	-0.2995	0.0077	1.14
7	MX	0.5911	0.0083	0.90
8	MC	-0.0821	0.0078	0.93
9	MC	-1.9099	0.0089	0.94
10	MX	0.0971	0.0079	0.85
11	MX	0.1175	0.0057	0.98
12	MC	0.3223	0.0080	1.00
13	MC	-0.6385	0.0077	0.91
14	MX	0.6221	0.0084	0.95
15	MC	0.7989	0.0086	1.04
16	MC	0.2895	0.0080	1.22
17	MC	-1.0555	0.0079	0.97
18	XI	-0.5120	0.0077	0.90
19	MC	-0.0025	0.0078	0.99
20	MC	0.2964	0.0080	0.94
21	MX	-1.4771	0.0083	1.00

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
22	XI	-1.1079	0.0079	0.93
23	XI	-0.9061	0.0078	1.16
24	MX	0.4662	0.0061	1.17
25	MC	-1.4742	0.0083	0.88
26	MC	0.2554	0.0080	0.98
27	MX	-0.0922	0.0054	0.98
28	XI	0.6730	0.0084	0.98
29	MC	-0.3992	0.0077	1.06
30	MX	0.3958	0.0081	0.91
31	MC	-0.5260	0.0077	0.96
32	MC	-0.3086	0.0077	1.22
33	MX	1.0126	0.0063	1.38
34	MX	1.4079	0.0098	1.04
35	XI	0.9679	0.0089	1.01
36	MC	-1.3368	0.0081	0.90
37	MC	0.1267	0.0079	0.97
38	MX	-0.2728	0.0077	0.86
39	MX	0.7752	0.0086	0.96
40	MC	0.7652	0.0086	0.94
41	MC	-0.5659	0.0077	1.03
42	MC	-1.0106	0.0078	0.98
43	MC	-0.0295	0.0078	0.93
44	XI	-0.7147	0.0077	0.89
45	MX	0.0242	0.0055	1.17
46	MX	0.8697	0.0087	1.02
47	MX	1.2195	0.0094	1.03
48	XI	0.6830	0.0084	0.99
49	MC	-1.0560	0.0079	0.94
50	MC	1.4837	0.0100	0.97

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced, MNSQ Infit = mean-square infit. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table B.3. Item-Level IRT Statistics, Grade 11

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-0.7371	0.0075	1.11
2	MC	-0.8173	0.0075	1.00
3	MC	0.2455	0.0082	1.13
4	XI	0.2629	0.0082	0.96
5	XI	0.6726	0.0089	1.02
6	MX	0.7615	0.0091	0.94
7	MC	-0.4966	0.0076	1.03
8	MC	-0.3256	0.0077	0.92
9	MX	1.0477	0.0098	0.95
10	MX	-0.2217	0.0077	0.97
11	MC	0.2387	0.0082	0.94
12	MX	-0.3874	0.0049	1.15

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
13	MC	0.5824	0.0087	1.06
14	XI	0.4180	0.0085	1.04
15	MX	0.8860	0.0094	0.94
16	MC	-0.5950	0.0075	1.12
17	MC	-0.5611	0.0075	0.96
18	MC	-0.1972	0.0077	0.94
19	MC	0.3869	0.0084	1.05
20	MC	-0.5656	0.0075	0.93
21	MC	-0.1512	0.0078	1.03
22	MX	0.0297	0.0080	1.05
23	MC	-0.9735	0.0075	0.98
24	MX	0.0270	0.0079	0.86
25	MX	-0.1772	0.0056	0.99
26	MC	0.5830	0.0087	1.01
27	MX	0.1270	0.0081	0.93
28	MC	-1.2483	0.0076	0.90
29	MC	0.3227	0.0083	0.87
30	MX	-1.0075	0.0052	0.98
31	MC	-0.6496	0.0075	1.01
32	MX	-0.2906	0.0077	0.86
33	MC	-1.3178	0.0076	0.97
34	MX	-1.2593	0.0076	0.81
35	MX	1.0333	0.0068	1.26
36	MC	0.4892	0.0086	1.09
37	MC	0.4482	0.0085	1.04
38	MC	0.2620	0.0082	0.83
39	XI	-1.1132	0.0075	0.88
40	MX	0.4048	0.0061	1.09
41	XI	1.1061	0.0099	0.95
42	MX	-0.5525	0.0075	0.92
43	MC	0.2193	0.0082	1.09
44	MC	-0.5610	0.0075	1.03
45	MC	0.4305	0.0085	1.17
46	MC	0.3026	0.0083	1.25
47	MC	0.0424	0.0080	0.99
48	MC	-0.0053	0.0079	1.10
49	MX	0.1545	0.0081	1.01
50	MX	1.2376	0.0103	0.86

Note. MC = multiple-choice, MX = multi-part, XI = technology-enhanced, MNSQ Infit = mean-square infit. Item number does not indicate item location on an operational test form, as field test items were embedded on the form but not included in the analysis.

Table B.4. Raw-to-Scale Score Conversion, Grade 5

Raw Score	Scale Score	CSEM	Performance Level
0	1200	60	1
1	1200	43	1
2	1200	31	1
3	1212	25	1
4	1225	22	1
5	1235	20	1
6	1244	19	1
7	1252	17	1
8	1259	17	1
9	1265	16	1
10	1271	15	1
11	1276	15	1
12	1281	14	1
13	1286	14	1
14	1290	14	1
15	1294	13	1
16	1298	13	1
17	1302	13	2
18	1306	13	2
19	1310	12	2
20	1314	12	2
21	1317	12	2
22	1321	12	2
23	1324	12	2
24	1328	12	2
25	1331	12	2
26	1334	12	2
27	1338	12	2
28	1341	12	2
29	1345	12	2
30	1348	12	2
31	1351	12	3
32	1355	12	3
33	1358	12	3
34	1362	12	3
35	1366	12	3
36	1369	13	3
37	1373	13	3
38	1377	13	3
39	1381	13	3
40	1385	13	3
41	1390	14	3
42	1394	14	3
43	1399	14	4
44	1404	15	4

Raw Score	Scale Score	CSEM	Performance Level
45	1409	15	4
46	1415	16	4
47	1422	17	4
48	1429	18	4
49	1436	19	4
50	1445	20	4
51	1456	22	4
52	1470	25	4
53	1488	31	4
54	1500	43	4
55	1500	60	4

Table B.5. Raw-to-Scale Score Conversion, Grade 8

Raw Score	Scale Score	CSEM	Performance Level
0	1200	59	1
1	1200	42	1
2	1200	30	1
3	1213	25	1
4	1227	22	1
5	1237	20	1
6	1246	19	1
7	1254	18	1
8	1261	17	1
9	1268	16	1
10	1274	15	1
11	1279	15	1
12	1284	14	1
13	1289	14	1
14	1294	14	1
15	1298	13	1
16	1302	13	2
17	1307	13	2
18	1311	13	2
19	1314	13	2
20	1318	13	2
21	1322	12	2
22	1326	12	2
23	1329	12	2
24	1333	12	2
25	1336	12	2
26	1340	12	2
27	1343	12	2
28	1347	12	2
29	1350	12	3
30	1354	12	3
31	1357	12	3

Raw Score	Scale Score	CSEM	Performance Level
32	1361	12	3
33	1365	12	3
34	1368	12	3
35	1372	12	3
36	1376	13	3
37	1379	13	3
38	1383	13	3
39	1387	13	3
40	1391	13	3
41	1396	14	3
42	1400	14	4
43	1405	14	4
44	1410	15	4
45	1415	15	4
46	1421	16	4
47	1427	16	4
48	1434	17	4
49	1442	18	4
50	1451	20	4
51	1461	22	4
52	1474	25	4
53	1492	30	4
54	1500	42	4
55	1500	59	4

Table B.6. Raw-to-Scale Score Conversion, Grade 11

Raw Score	Scale Score	CSEM	Performance Level
0	1200	60	1
1	1200	43	1
2	1203	30	1
3	1221	25	1
4	1234	22	1
5	1245	20	1
6	1254	19	1
7	1261	17	1
8	1268	16	1
9	1274	16	1
10	1280	15	1
11	1285	15	1
12	1290	14	1
13	1294	14	1
14	1299	13	1
15	1303	13	2
16	1307	13	2
17	1311	13	2
18	1315	13	2

Raw Score	Scale Score	CSEM	Performance Level
19	1319	12	2
20	1322	12	2
21	1326	12	2
22	1329	12	2
23	1333	12	2
24	1336	12	2
25	1339	12	2
26	1343	12	2
27	1346	12	2
28	1349	12	2
29	1353	12	3
30	1356	12	3
31	1360	12	3
32	1363	12	3
33	1367	12	3
34	1370	12	3
35	1374	12	3
36	1377	13	3
37	1381	13	3
38	1385	13	3
39	1389	13	3
40	1393	13	3
41	1398	14	3
42	1402	14	4
43	1407	14	4
44	1412	15	4
45	1417	15	4
46	1423	16	4
47	1430	17	4
48	1437	18	4
49	1444	19	4
50	1454	20	4
51	1464	22	4
52	1478	26	4
53	1496	31	4
54	1500	43	4
55	1500	60	4

Figure B.1. Item-Person Map, Grade 5

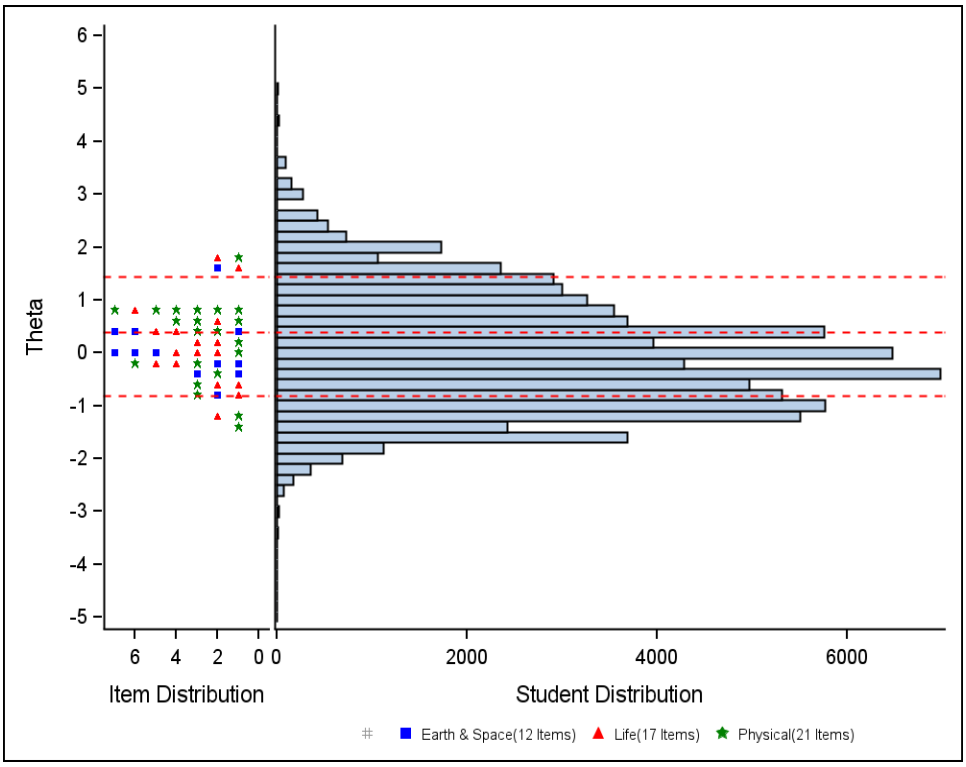


Figure B.2. Item-Person Map, Grade 8

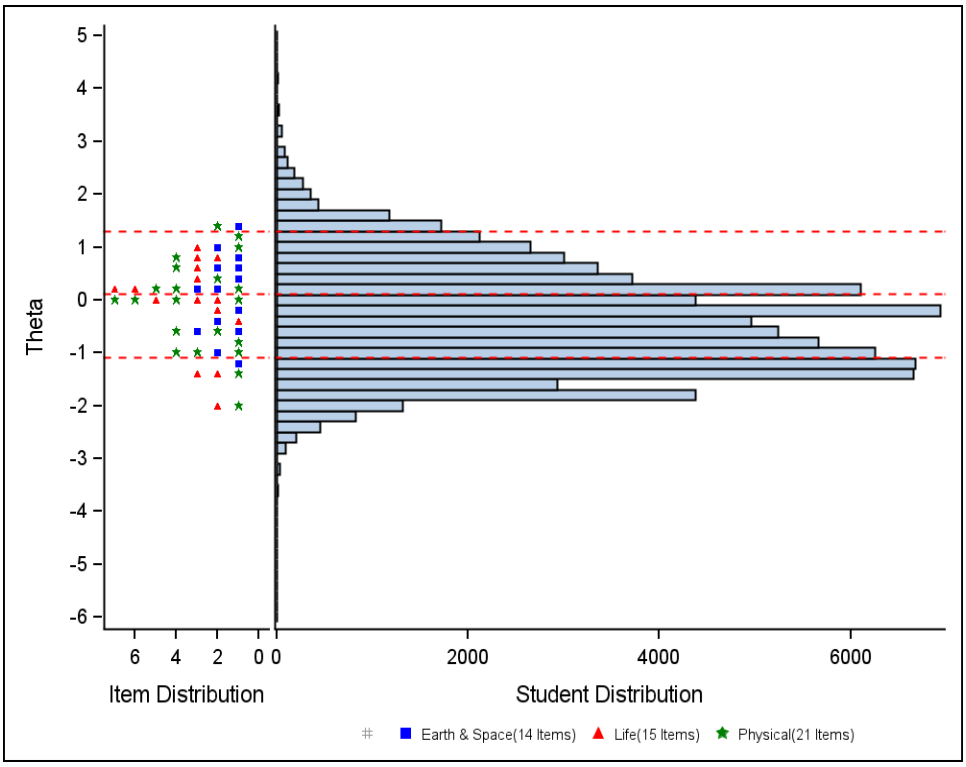


Figure B.3. Item-Person Map, Grade 11

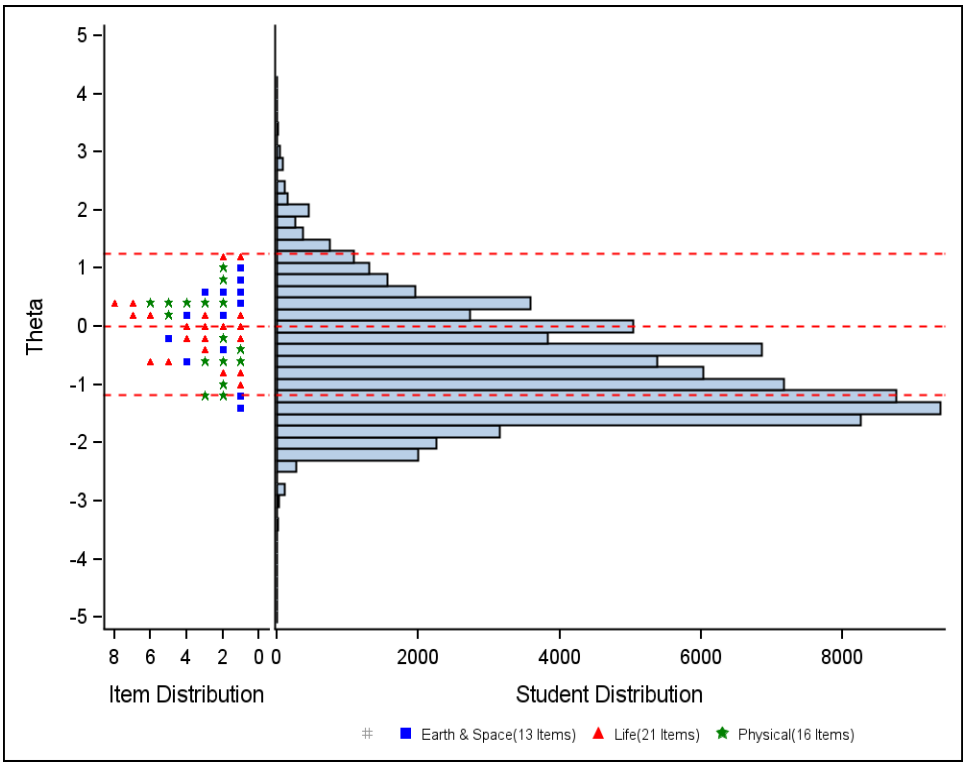


Figure B.4. TCC, Grade 5

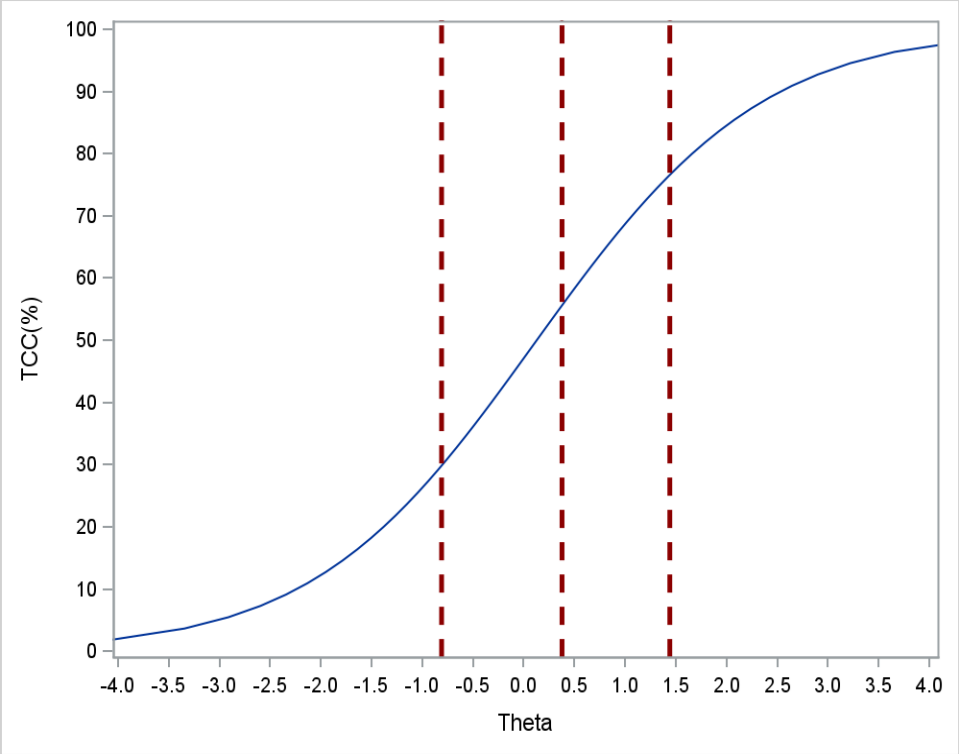


Figure B.5. TIF and CSEM, Grade 5

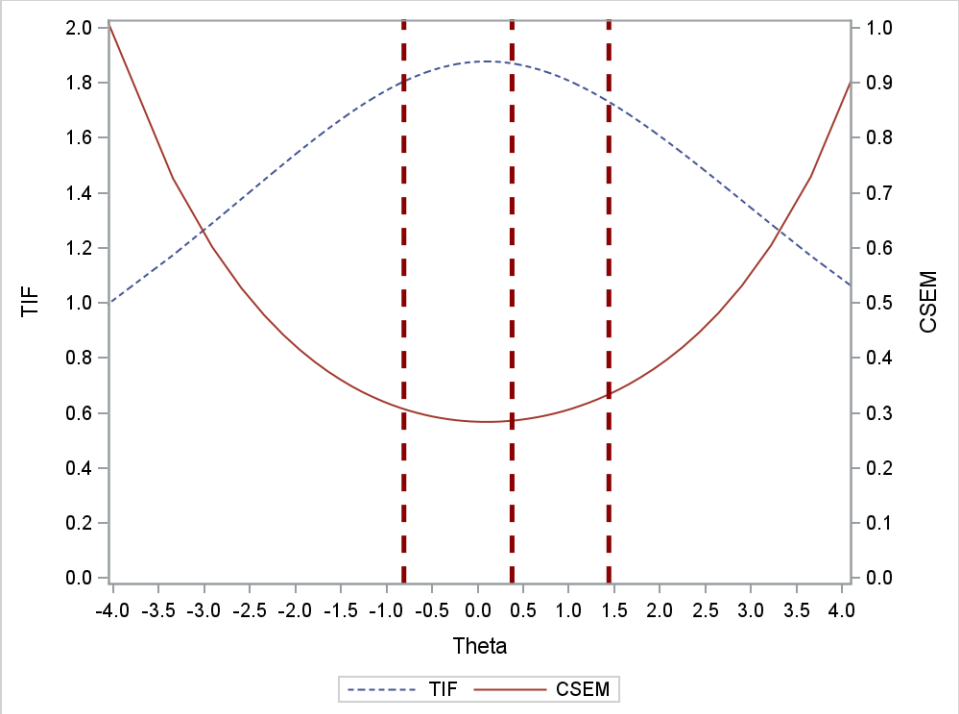


Figure B.6. TCC, Grade 8

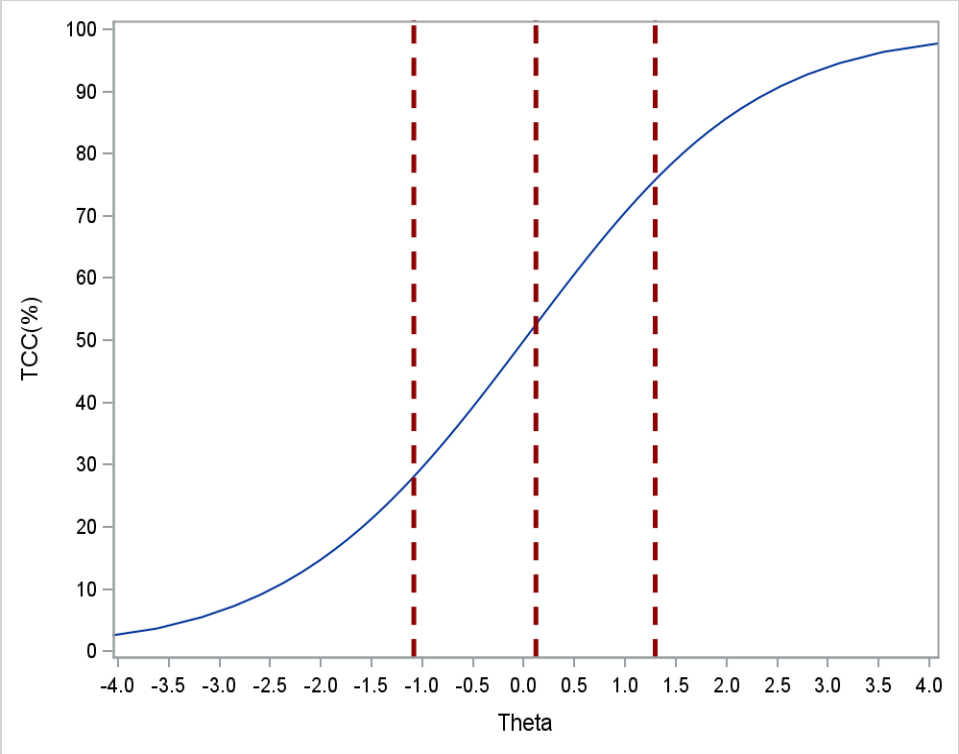


Figure B.7. TIF and CSEM, Grade 8

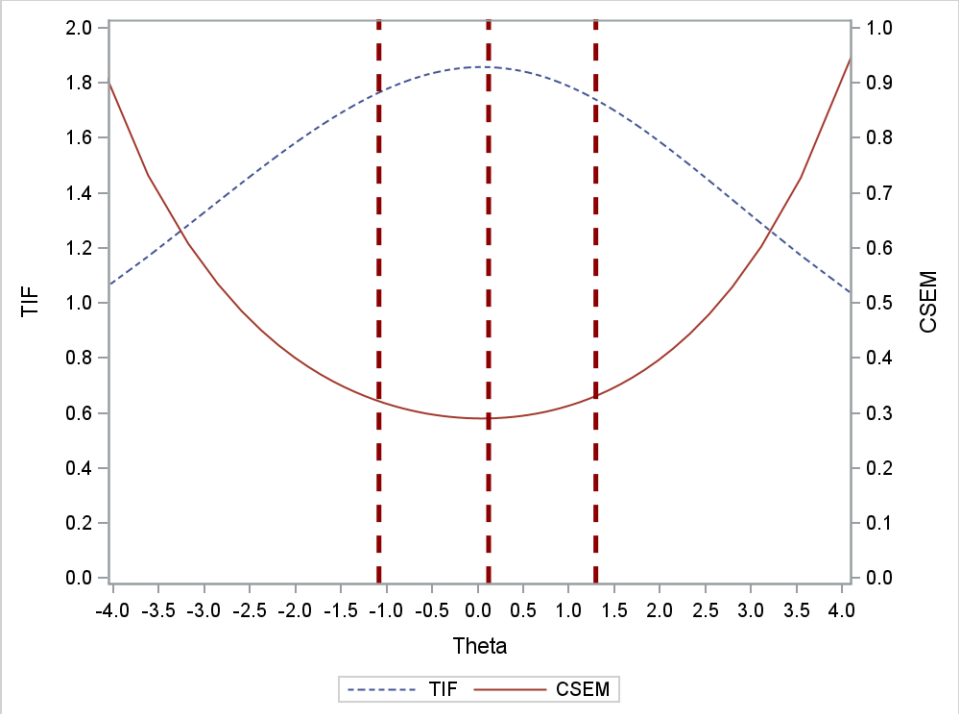


Figure B.8. TCC, Grade 11

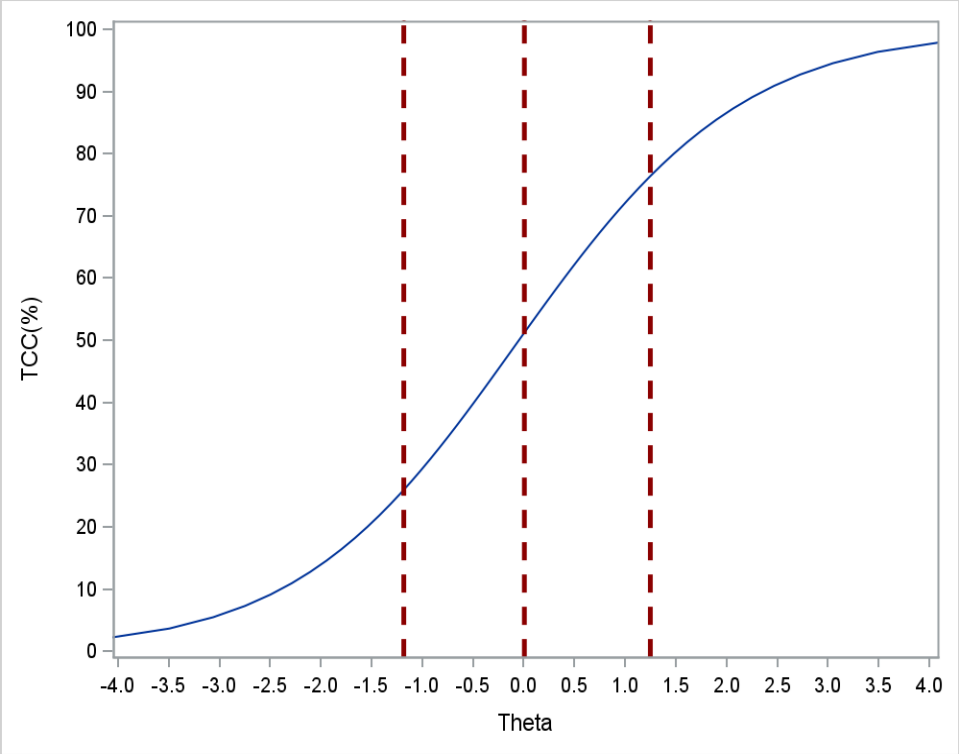


Figure B.9. TIF and CSEM, Grade 11

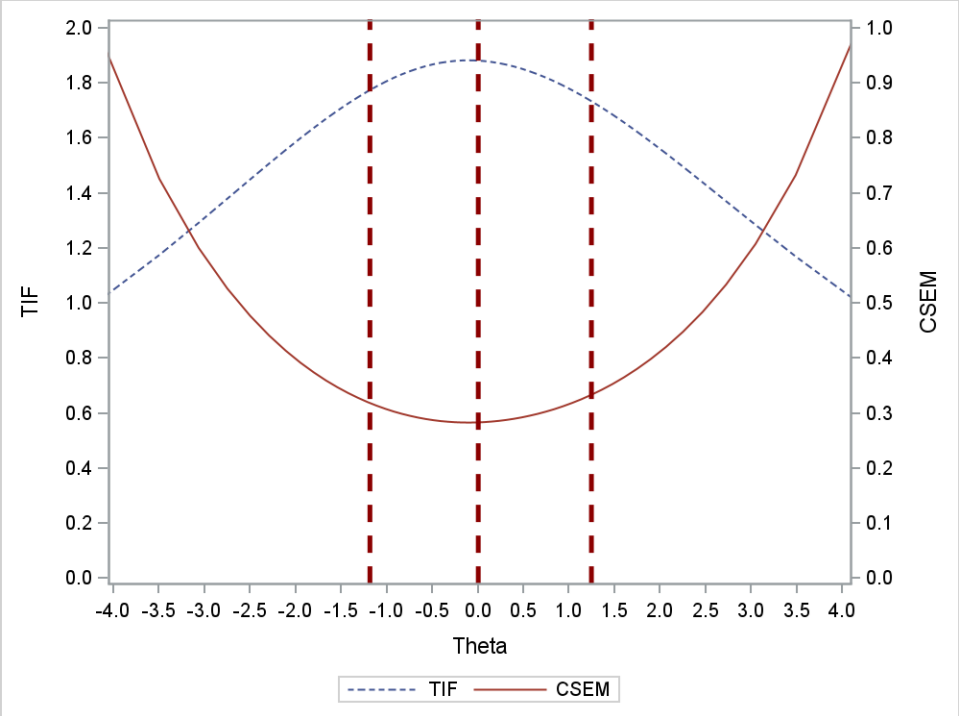


Figure B.10. Scree Plot, Grade 5

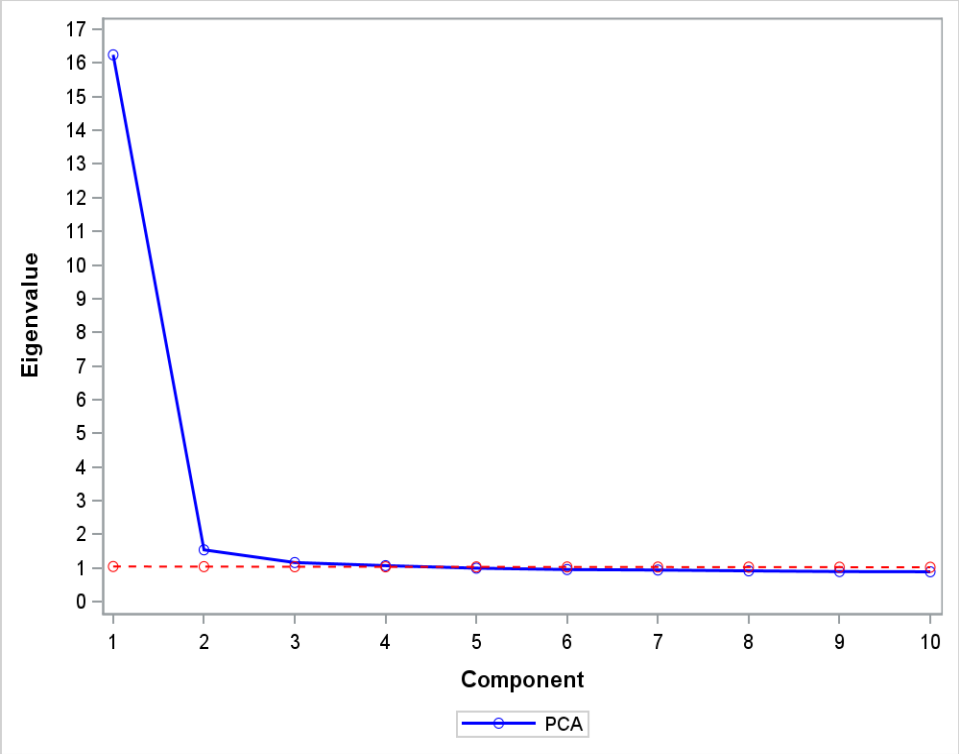


Figure B.11. Scree Plot, Grade 8

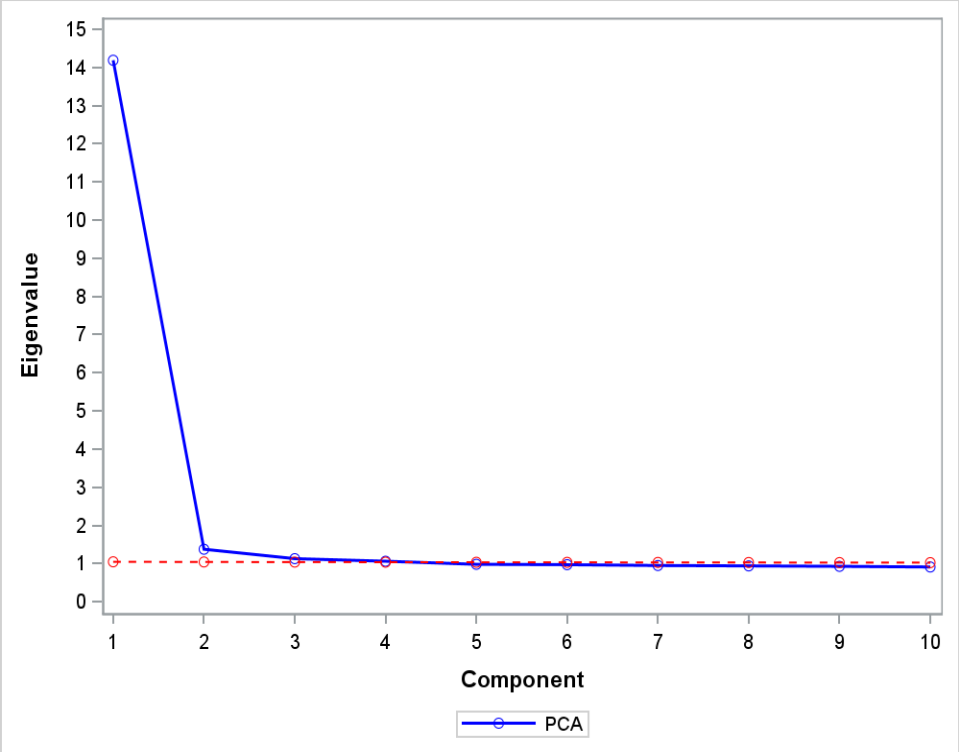
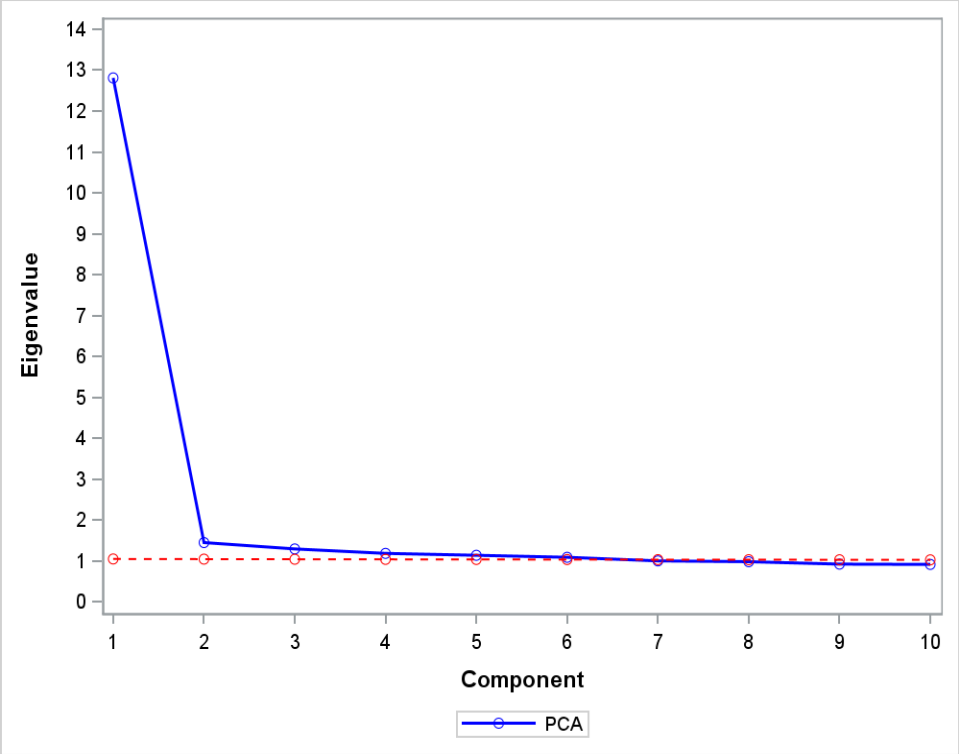


Figure B.12. Scree Plot, Grade 11



Appendix C: ADMINISTRATION RESULTS

This appendix presents the spring 2025 AzSCI results for all students and subgroups by gender, ethnicity (Hispanic or Non-Hispanic), race, and special education, English learner (EL), and low socioeconomic status (SES). Specifically:

- Table C.1 – Table C.3 present the overall results by subgroup, including the sample size, mean and standard deviation (SD) of the total combined scale score, and percentage of students at each overall performance level.
- Figure C.1 – Figure C.3 present histograms of the total scale score distribution.

Table C.1. Test Results by Subgroup, Grade 5

Subgroup	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
All	81,588	1332.29	44.91	27.7	38.5	24.8	9.1
Male	41,391	1335.62	46.53	26.4	36.3	26.6	10.7
Female	40,197	1328.86	42.92	29.0	40.7	23.0	7.3
Missing	0	–	–	–	–	–	–
Hispanic	39,976	1320.77	40.48	35.6	41.1	18.8	4.5
Non-Hispanic	41,612	1343.36	46.16	20.1	35.9	30.6	13.4
American Indian	4,349	1309.84	36.32	46.2	39.3	12.4	2.0
Asian	3,107	1358.56	47.59	12.9	28.5	36.8	21.8
Black or African American	6,080	1318.60	39.95	37.7	41.1	17.1	4.1
Multi-racial	5,303	1338.96	44.28	21.4	38.9	28.8	10.8
Native Hawaiian or Other Pacific Islander	485	1327.33	40.35	29.5	41.2	23.3	6.0
White	62,230	1333.37	44.83	26.7	38.6	25.5	9.3
Missing	34	1307.65	33.86	50.0	41.2	8.8	0.0
Special Education	13,207	1304.66	38.66	55.5	31.2	10.5	2.8
English Learner (EL)	9,404	1292.51	27.35	67.6	28.8	3.2	0.4
Low Socioeconomic Status (SES)	44,956	1319.54	40.07	36.8	40.9	18.0	4.2
Migrant	481	1306.51	36.19	53.0	33.5	12.1	1.5

Note. SS = total scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Table C.2. Test Results by Subgroup, Grade 8

Subgroup	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
All	82,414	1328.04	40.95	28.6	40.6	25.4	5.3
Male	41,815	1329.07	42.34	28.9	38.8	26.1	6.2
Female	40,599	1326.97	39.43	28.4	42.5	24.7	4.4
Missing	0	–	–	–	–	–	–
Hispanic	39,689	1317.58	37.10	36.8	42.4	18.4	2.4
Non-Hispanic	42,725	1337.75	41.95	21.0	38.9	31.9	8.1
American Indian	4,633	1309.89	33.65	44.0	42.6	12.2	1.2
Asian	3,055	1358.52	43.46	10.7	28.4	42.7	18.2
Black or African American	6,083	1315.16	35.87	38.0	44.3	15.6	2.0
Multi-racial	5,095	1334.83	40.57	22.3	40.7	30.2	6.7
Native Hawaiian or Other Pacific Islander	512	1322.87	36.82	28.1	49.6	18.9	3.3
White	62,979	1328.65	40.74	28.0	40.6	26.1	5.3
Missing	57	1311.65	35.64	49.1	33.3	15.8	1.8
Special Education	10,374	1299.58	32.61	59.3	32.1	7.5	1.1
English Learner (EL)	7,515	1289.42	24.26	72.3	25.7	2.0	0.1
Low Socioeconomic Status (SES)	43,113	1316.26	36.42	37.8	42.9	17.2	2.2
Migrant	504	1299.97	32.92	59.3	31.7	7.9	1.0

Note. SS = total scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Table C.3. Test Results by Subgroup, Grade 11

Subgroup	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
All	83,348	1319.66	37.46	36.1	44.6	16.0	3.3
Male	42,012	1321.37	40.58	37.3	40.4	17.7	4.6
Female	41,336	1317.91	33.91	34.9	48.9	14.3	2.0
Missing	0	–	–	–	–	–	–
Hispanic	40,371	1310.61	31.54	43.7	45.1	10.1	1.0
Non-Hispanic	42,977	1328.15	40.48	28.9	44.0	21.6	5.5
American Indian	4,877	1307.96	28.46	44.7	47.6	7.3	0.5
Asian	3,072	1350.58	46.91	16.5	35.9	31.9	15.7
Black or African American	5,713	1309.24	30.74	45.4	44.9	8.8	0.9
Multi-racial	4,817	1324.94	38.35	30.7	45.9	19.2	4.2
Native Hawaiian or Other Pacific Islander	502	1313.93	31.84	39.2	46.4	13.1	1.2
White	63,692	1319.69	37.20	35.9	44.6	16.4	3.1
Missing	675	1315.00	33.52	37.9	47.6	12.3	2.2
Special Education	8,209	1297.28	27.54	64.0	31.2	4.3	0.6
English Learner (EL)	5,179	1289.36	19.32	77.0	22.3	0.7	0.1
Low Socioeconomic Status (SES)	40,280	1310.67	32.06	44.1	44.6	10.2	1.2
Migrant	609	1298.67	25.33	61.4	35.0	3.6	0.0

Note. SS = total scale score, SD = standard deviation, Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Figure C.1. Total Scale Score Distribution, Grade 5

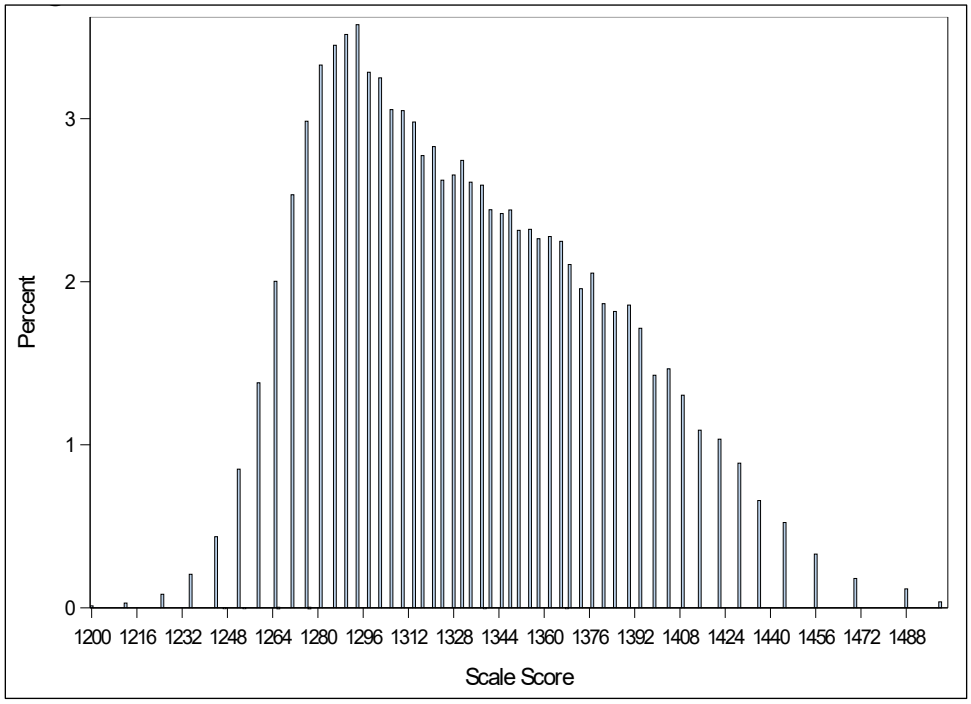


Figure C.2. Total Scale Score Distribution, Grade 8

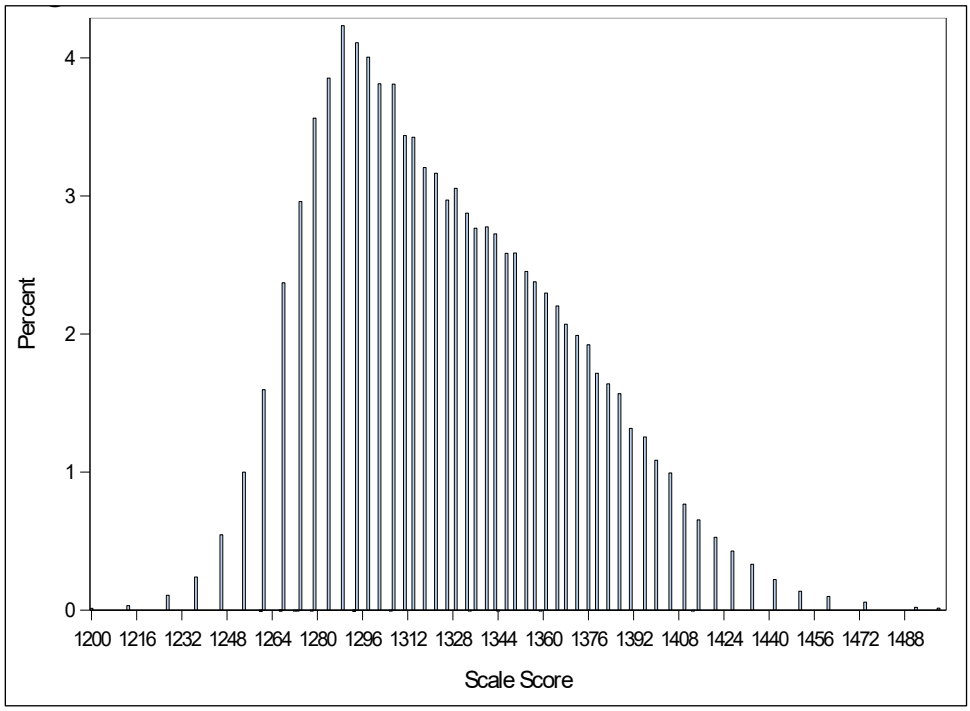


Figure C.3. Total Scale Score Distribution, Grade 11