AzSCI

Arizona Science Assessment

2022
Technical Report

Submitted to the
Arizona Department of Education
November 2022

Pearson

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

This technical report documents the design, development, administration, technical processes, and results of the Spring 2022 administration of Arizona’s Science Test (AzSCI) to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. 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).

Chapter 1: INTRODUCTION

1.1. Assessment Overview

Arizona's Science Test (AzSCI) is the statewide achievement test for Arizona students in science in Grades 5, 8, and 11. It is a criterion-referenced assessment designed to measure student progress toward achievement of the Arizona Science Standards adopted by the State Board of Education in 2018. 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. It is a computer-based assessment, allowing for the use of a variety of innovative item types where students can apply critical thinking skills to demonstrate a deeper understanding of the three dimensions of the Arizona Science 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.

The 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 design include measurement targets, test designs, item types, and test administration conditions. To support this effort, Pearson, in early collaboration with WestEd, worked with the Arizona Department of Education (ADE), with input from Arizona educators, to develop item specifications and blueprints to guide the item and 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 field test 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.

1.1. Educator Involvement

This section addresses the involvement of Arizona educators in test development as indicated by Standard 4.6 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 Spring 2022 AzSCI assessment.

Arizona educators had several opportunities to participate in meetings and provide feedback on assets developed for field testing in Spring 2022. Because of the number of assets developed (160 items per grade, plus associated stimuli for each set of items), there were two weeks of content and bias review, also known as item review in June 2021. Dividing this work over two weeks allowed more educators to participate. There was also a bias and sensitivity community review in August 2021 that enabled community members, including past and present Arizona educators, to evaluate items. Both meetings represent a continuation of stakeholder involvement in the development process. In previous years, for example, Arizona educators were involved in reviewing the AzSCI performance level descriptors (PLDs).

The culmination of educator involvement in the test development cycle was a standard setting that occurred in June 2022. Arizona educators later reconvened in July 2022 for data review. The meetings in 2021 were held virtually. Those in 2022 were in Scottsdale, Arizona, with all participants in person.

Table 1.1. Schedule of Major Events

Event	Date(s)
Item Review Committee–Committee Review	June 21–25, 2021
Technical Advisory Committee (TAC)	July 28, 2021
Item Review Committee–Community Review	August 10–11, 2021
Administration Training	November 30, 2021 – April 15, 2022
Additional Order Window for Test Materials	February 28 – April 7, 2022
Technical Advisory Committee (TAC)	March 9, 2022
Spring 2022 AzSCI Test Administration Window	March 21 – April 15, 2022
Standard Setting	June 21–23, 2022
Data Review	July 20–22, 2022
Release of Electronic Score Reports and Student Data Files	August 3, 2022
Release of Paper Reports	August 16, 2022

Chapter 2: TEST DESIGN

This chapter provides information regarding test design as indicated by Standards 1.1, 1.11, 4.0, 4.1, 4.2, 4.12, 7.0, 7.2, 12.4, and 12.8 (AERA et al., 2014).

2.1. Arizona Science Standards

In October 2018, ADE adopted a new version of the Arizona Science Standards that were written by a group of educators, content experts, and community members and reflect an increase in rigor when compared to the previous version 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 standards around 13 core ideas and develop learning progressions to coherently and logically build scientific literacy from kindergarten through high school.
- Connect 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 three dimensions of 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 provides 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.

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 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

The standards are presented for each grade from kindergarten through high school. Each standard embeds an SEP into a Core Idea. The standards document then pairs the standard with one or more CCC. The complete set of standards can be accessed on the ADE website at https://www.azed.gov/standards-practices/k-12standards/standards-science.

2.2. Item Specifications

In Spring 2018, Pearson and its partner WestEd undertook a comprehensive and systematic evaluation of the new Arizona Science Standards to make suggestions to ADE that would guide item development and test design. One suggestion was the creation of item specifications, or detailed documents publicly available to all Arizona educators that provide an analysis of each standard, specifying content limits and identifying the item types that can be used to assess the standard. Item writers also use these specifications while developing items to make the best use of the available item types.

This document was envisioned as a companion to existing documents, such as the Arizona Science Standards. One important distinction was that the item specifications would identify content limits that would aid in item development and content review. The subsequent development of an item specifications document was an iterative process involving ADE, Pearson, and a committee of Arizona educators. By September 2019, the specifications were approved and shortly thereafter published for public access on the ADE website at https://www.azed.gov/assessment/sci/.

2.3. Test Blueprint

The test blueprint, in concert with the item specifications, defines the content and structure of the test and guides item selection. At each grade band, blueprint guidance is provided by domain, SEP, grade, and cognitive complexity. Item selection for forms is guided by 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. The standards selected for use in Spring 2022 were then marked. Using spreadsheets allows Pearson to quickly identify which standards remain to be tested in future administrations to ensure that standards will be assessed in a 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 items based on item types
- A range for the number of items based on cognitive complexity
- A range for the number of items for each grade within a grade band
- The total number of points per item type

An iterative process was used to develop the test blueprint. Pearson's assessment specialists drafted an initial blueprint that was submitted to ADE for review, and adjustments were made as requested. In August 2020, an advisory committee of Arizona educators provided feedback on the draft. The blueprint plan was subsequently approved and used by the Pearson AzSCI content team for item development.

The blueprint was revised before the Spring 2022 administration to better reflect the distribution of the standards. For example, rather than allocating a higher percentage of the test to Physical Science, which has a greater percentage of standards, the blueprint allocated an equal percentage across Physical Science, Life Science, and Earth and Space Science. At ADE's request, changes were also made to the reporting categories for SEPs because the investigating category did not include the SEP Analyzing and Interpreting Data (DATA). As a result, the investigating category contained fewer than 10 items after the adjustments were made to domain percentages. The SEP DATA category was moved into the investigating category to gather adequate data.

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

Table 2.1. AzSCI Blueprint by Domain

Domain	Grade 5	Grade 8	High School
SEPs and CCCs in Physical Sciences	40–48%	36–44%	32–40%
SEPs and CCCs in Life Sciences	28–36%	30–38%	34–42%
SEPs and CCCs in Earth and Space Sciences	20–28%	22–30%	22–30%

Table 2.2. AzSCI Blueprint by SEP

Practice (and Categories)	Grade 5	Grade 8	High School
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 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%

The performance expectations for the Arizona Science Standards are written with high levels of cognitive complexity, incorporating knowledge with practice and identifying and using unifying concepts to develop scientific explanations. Appropriately assigning the cognitive load to AzSCI items requires 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) for the AzSCI assessment to more accurately 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 operational targets for AzSCI.

Table 2.4. AzSCI Blueprint for Cognitive Complexity Operational Targets

Task Analysis Guide in Science (TAGS) Level	Percent 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%

2.4. Test Designs

Each AzSCI form has 60 items (50 operational + 10 field test). The base form has 50 operational items worth a total of 55 points. For grade bands 3–5 and 6–8, a total of 14 forms are used to embed field test items, with 10 field test items per form. For high school, a total of 12 forms are used to embed field test items, with 10 field test items per form. All items on the AzSCI assessment are based on a specific scientific phenomenon presented in a stimulus or series of stimuli. The items are part of one of two sets: an independent set or an item cluster set.

An independent item set has at least two non-dependent items associated to one or more short stimuli, whereas an item cluster set has five items associated to longer, more complex stimuli. In both types of sets, 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, whereas TPI items are worth 2 points. Interactions classified as TEs include bar graph, multiple select, inline choice, hot spot, graphic gap match, gap match, line graph, match, match table grid, and point graph.

Table 2.5 summarizes the AzSCI test design. Items in the independent and cluster sets are divided across two forms for field test purposes. All grade bands are administered in two units, each with 30 items. At least one item in each unit is a 2-point TPI item.

Table 2.5. AzSCI Test Design for Grades 3–5, 6–8, and 11

Unit	OP Items from Independent Sets	OP Items from Cluster Sets	FT Items from Independent and Cluster Sets
1	15 items (from at least five independent sets)	10 items (from two cluster sets)	5 items (from two independent sets) • 0–3 MC items • 0–3 TE items • 1 TPI or TPD item
2	n/a	25 items (from five cluster sets)	5 items (from one cluster set) • 0–3 MC items • 0–3 TE items • 1 TPI or TPD item
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

Note. OP = operational, FT = field test, MC = multiple-choice, TE = technology-enhanced, TPD = two-part dependent, TPI = two-part independent.

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.

Spring 2022 was the first operational administration for AzSCI. Item selection was based on data from a standalone field test administered in Spring 2021. The Spring 2022 forms were also used to embed 10 field test items in each form. Pearson developed 160 items per grade (480 items total) for the Spring 2022 administration. 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 overriding goal in selecting items for the forms was adhering to the blueprint requirements. Additional criteria for item selection included item positioning, content considerations, and statistical considerations.

3.1. Content Development and Management Tool

The item pool and content 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, classical item statistics, 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 staff before it is identified for field test or operational administration.

3.2. Form Construction Process

ADE and Pearson worked collaboratively to construct the AzSCI tests based on the following steps:

1. Asset development plan
2. Item development
3. Item review

4. Preparation for item selection
5. Item selection and positioning
6. Field test verification
7. Sampling plan

3.2.1. Asset Development Plan

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. A gap analysis process was used 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 that has less than five items.
- Increase coverage within the Earth and Space Science domain (Grade 11).
- Increase investigating SEP group.
- Increase standards covered by independent items.
- Increase graphing items.
- Increase "D" level TAGS coverage (i.e., Doing Science Tasks) (Grade 11).
- Even out the number of item types.
- Increase standards covered in each domain under 60% of the total items (Grades 5 and 8).

3.2.2. Item Development

Item development was guided by the item specifications. The first step was drafting the science phenomena. Pearson took the lead on this work, followed by a review by ADE. The next step was providing an outline describing how the phenomena would be presented to students. Again, Pearson did the initial work, and ADE provided feedback. The same type of collaboration was used for developing the items and stimuli; authoring responsibilities started with Pearson, with the completed sets going to ADE for approval.

Throughout all steps, Pearson responded to ADE feedback, revised, and resubmitted for approval as needed. An integral part of this process was a review of all assets by Pearson research librarians who verified accuracy and by Pearson copyeditors who reviewed for clarity and correct use of grammar, punctuation, and spelling. All asset creators and reviewers at Pearson also apply 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.

3.2.3. Item Review

ADE pre-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., 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., bias and sensitivity review) also allowed parents and other community stakeholders to review assets. 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.

Prior to beginning review, committee members received training from Pearson assessment specialists. They were also provided resources, including a checklist, to guide the review process. All feedback was recorded in ABBI.

3.2.4. Field Test Candidate Finalization

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.

3.2.5. Preparation for Item Selection

Test construction took place in ABBI. 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 Spring 2021 standalone field test, 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.

3.2.6. Item Selection and Positioning

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. Table 3.1 presents the acceptable item-level statistics for item selection. Acceptable test-level statistics include an average *p*-value between 0.4 and 0.5 and p-value distribution being reasonable. Because Spring 2022 was the first operational administration, there was no specific target regarding test characteristic curves (TCCs).

Table 3.1. Acceptable Item Statistics for Item Selection

Statistic	Criterion
P-value	>0.2 or < 0.9
Point-biserial correlation	≥ 0.25
Distractor point-biserial correlation (MC only)	≤ 0.05
Omit rate	≤ 2%
Rasch difficulty	≥ -3 or ≤ 3
Item fit statistics	≥ 0.6 or ≤ 1.4
Score point percentage (2-point items only)	≥ 1%*
Differential item functioning (DIF)	A or B for one group

*I.e., there should be at least 1% of students at each score point (2-point items only)

A different assessment specialist reviewed the form and provided feedback, identifying issues such as clueing. After issues were resolved, a Pearson psychometrician reviewed the form and provided feedback based on statistical considerations. This process repeated until the form met psychometric approval. The form was then provided to ADE for review, and revisions were made based on ADE feedback. This process continued until ADE gave approval.

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.

3.2.7. Sampling Plan

Grades 5 and 8 had 14 forms, and Grade 11 had 12 forms. All forms within a grade had the same operational items but different field test items. The test forms were randomly assigned at a student level within a testing group, created by a district, by TestNav, Pearson’s online test delivery platform. Only one paper-pencil version was available per grade.

3.3. Data Review

Field tested items were flagged based on the criteria in Table 3.2. During data review, committee members reviewed the flagged items and their item statistics to determine whether the field tested items were eligible for the operational item pool. One committee group focused solely on the items flagged for DIF, while another group reviewed the items flagged by the remaining statistics (i.e., all statistics in Table 3.2 except for DIF). The DIF group was formed by educators who had different cultural backgrounds and/or knew students in special populations such as students with disabilities.

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 the quality of an item cannot be made on statistics alone; the item itself and the content it measures should also be taken into consideration. Thus, the committee 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. Table 3.3 presents the data review results based on the Spring 2022 data. Accepted items were added to the operational item pool for future use.

Table 3.2. 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 3.3. Data Review Results: Number of Field Tested Items

Grade	#Accepted	#Accepted w/Edits	#Rejected
5	64	11	1
8	64	7	–
11	70	4	1

3.4. Special Paper Versions

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. ADE 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.

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).

Students in Grades 5, 8, and 11 (Cohort 2023) participate in the spring administration of the AzSCI test. Students with significant cognitive disabilities and whose Individualized Education Program (IEP) designates them as eligible for an alternate assessment, Multi-State Alternate Assessment (MSAA) and MSAA Science are excluded from the AzSCI assessment.

AzSCI is a computer-based test. Test administrators were instructed to use the *Test Administration Directions* (TAD) manual for the online administration of AzSCI, as well as for the Special Paper Version (SPV) tests and entering student responses into TestNav, Pearson's online testing platform that students use to access the assessment. PearsonAccess^{next} (PAN) is the student test management portal that test administrators use to manage student tests and registrations and to order materials if needed.

The assessment for each grade was divided into two 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. A test unit must be completed prior to starting the next test unit. Test Unit 1 must be completed prior to starting Test Unit 2. If both test units are administered on the same day, there must be a significant break between each test unit. ADE requires that a test unit be submitted within the day that it is started. Any test that is not complete at the end of the testing day is marked complete and submitted for scoring by Pearson.

4.1. Manuals

The *Test Administration Directions* (TAD) and *Test Coordinator Manual* (TCM) were produced in collaboration with ADE. The Pearson program team drafted the original manuscript using the previous year's TAD or TCM as a template for design, layout, and content. The document was 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 three 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, although a limited number of each could be ordered during the additional order window.

Test administrators were also provided a *PAN User's Guide* and the *Arizona Accommodation Manual* that lists the current accommodations, accessibility features, and tools available on Arizona's achievement assessments. The *PAN User's Guide* was posted in PAN, and the *Arizona Accommodation Manual* was posted on the ADE website.

4.2. Administration Training

Mandatory test administration training was provided by ADE and Pearson and delivered through Pearson's Training Management System (TMS) online at https://azachieve.tms.pearson.com/. The TMS contained three training modules as summarized in Table 4.1 that were required for District Test Coordinators, School Test Coordinators, Test Administrators, and other school staff involved in testing or test results.

Table 4.1. Administration Trainings

Training	Description
AzSCI Test Administration	This training covered the Spring 2022 AzSCI test administration for Grades 5, 8, and 11, including an overview of the test administration, websites and resources, PearsonAccess^{next} (PAN) information, and responsibilities before, during, and after testing.
Accommodations	This training covered the test accommodations. This was required for all District Test Coordinators 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 was 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.

4.3. Sample Tests

Sample Tests are available in TestNav year-round to help students become familiar with the item types on the AzSCI assessments. 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 AzSCI Sample Tests reflect the AzSCI test specifications and blueprints and had 15 items on each test. The Sample Tests do not include an item for each of the aligned Arizona Science Standards and do not provide scores for students. As such, 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. Every eligible item type was represented. An accompanying scoring guide identified standard and TAGS levels. The portal and scoring guides are both available on ADE website at https://www.azed.gov/assessment/sci.

4.4. 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 4.2 presents the accommodations available to students while testing on Arizona assessments, including AzSCI.

Table 4.2. Available Arizona Accommodations

Available 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).
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 response 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 response 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 response exactly as written into the TestNav system.
Math window	All students in Grades 3–8 and 11 may use their math window during testing.
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, 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.

Table 4.3 presents the number of students who used the available accommodations. This table only includes the accommodations captured in the student data file (i.e., accommodations used by students during the Spring 2022 administration).

Table 4.3. Frequency of Accommodations Used

Accommodation	Grade 5	Grade 8	Grade 11
Adult Transcription	14	3	205
Assistive Technology	5	7	20
Sign Test Content	22	30	141
Simplified Directions	404	349	161
Translate Directions	223	170	107
Translation Dictionary	243	214	203

4.5. 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, testing one-on-one, testing in a separate location on campus or in a study carrel
- Being seated in a specific location within the testing room or being seated at special furniture
- Having the test administered by a familiar test administrator
- Using a special pencil or pencil grip
- Using a place holder
- 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, and special lighting
- Using different contrast settings or color overlays
- Using devices that allow the student to hear the test directions: hearing aids 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 during the break)

  - 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 Test Administrator 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. Test units should be scheduled in a way that provides the student more than adequate time to complete the test.

4.6. 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	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	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; and blue background with yellow text.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Highlighter	Highlight text in a passage or item.
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.

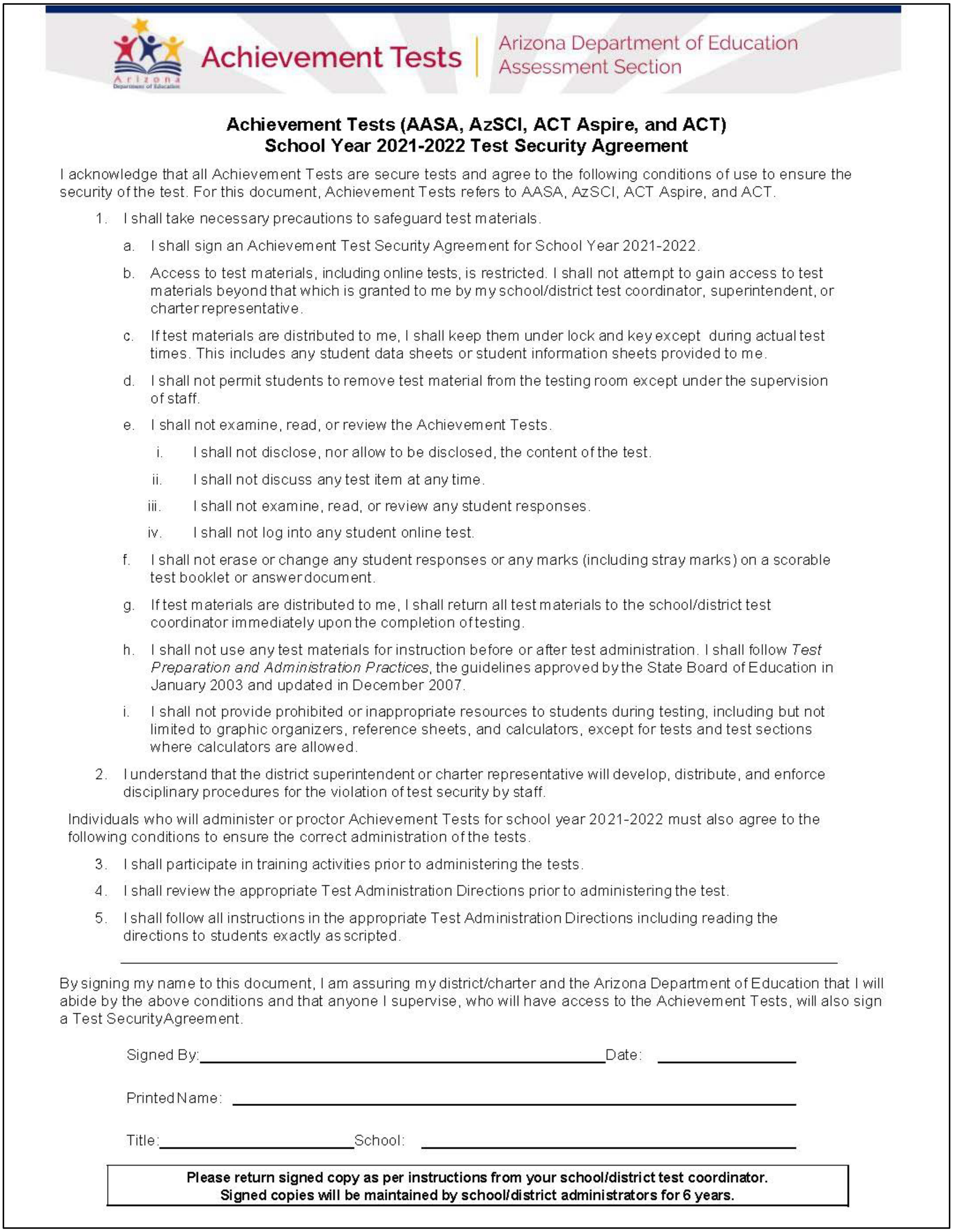
Universal Test Tool	Description
Pause and Restart	Students may sign out of TestNav. Before the student can resume testing, the Test Administrator will need to resume the student's session in TestNav.
Review Test	Allows student to review the test before submitting it.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech	Text-to-Speech for content of writing, mathematics, and science.
Tutorial	Learn and practice using TestNav tools and responding to each item type.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test up to 200%. Undo zoom in and return the font and images in the test to original size.

4.7. 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 | Arizona Department of Education
Assessment Section

Achievement Tests (AASA, AzSCI, ACT Aspire, and ACT)
School Year 2021-2022 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 Test Security Agreement for School Year 2021-2022.
   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 except under the supervision of staff.
   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*, the guidelines approved by the State Board of Education in January 2003 and updated in December 2007.
   i. I shall not provide prohibited or inappropriate resources 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 2021-2022 must also agree to the following conditions to ensure the correct administration of the tests.

3. I shall participate in training activities prior to administering the tests.
4. I shall review the appropriate Test Administration Directions prior to administering the test.
5. I shall follow all instructions in the appropriate 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, will also sign a Test SecurityAgreement.

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 and Reporting

5.1. Scoring

All items on the AzSCI assessments were machine-scored with maximum likelihood estimation (MLE) scoring, with an attemptedness rule that a student needed to answer at least one item in each unit. (See Section 2.4 for information on item types.)

Students received a scale score, and student performance was 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 was 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 was clearly below *Proficient*. Students who score *At/Near Mastery* demonstrate performance in the reporting category that was exactly at or immediately above/below *Proficient*. Students who score *Above Mastery* demonstrate performance in the reporting category that was clearly *Proficient* or higher.

5.2. Reporting

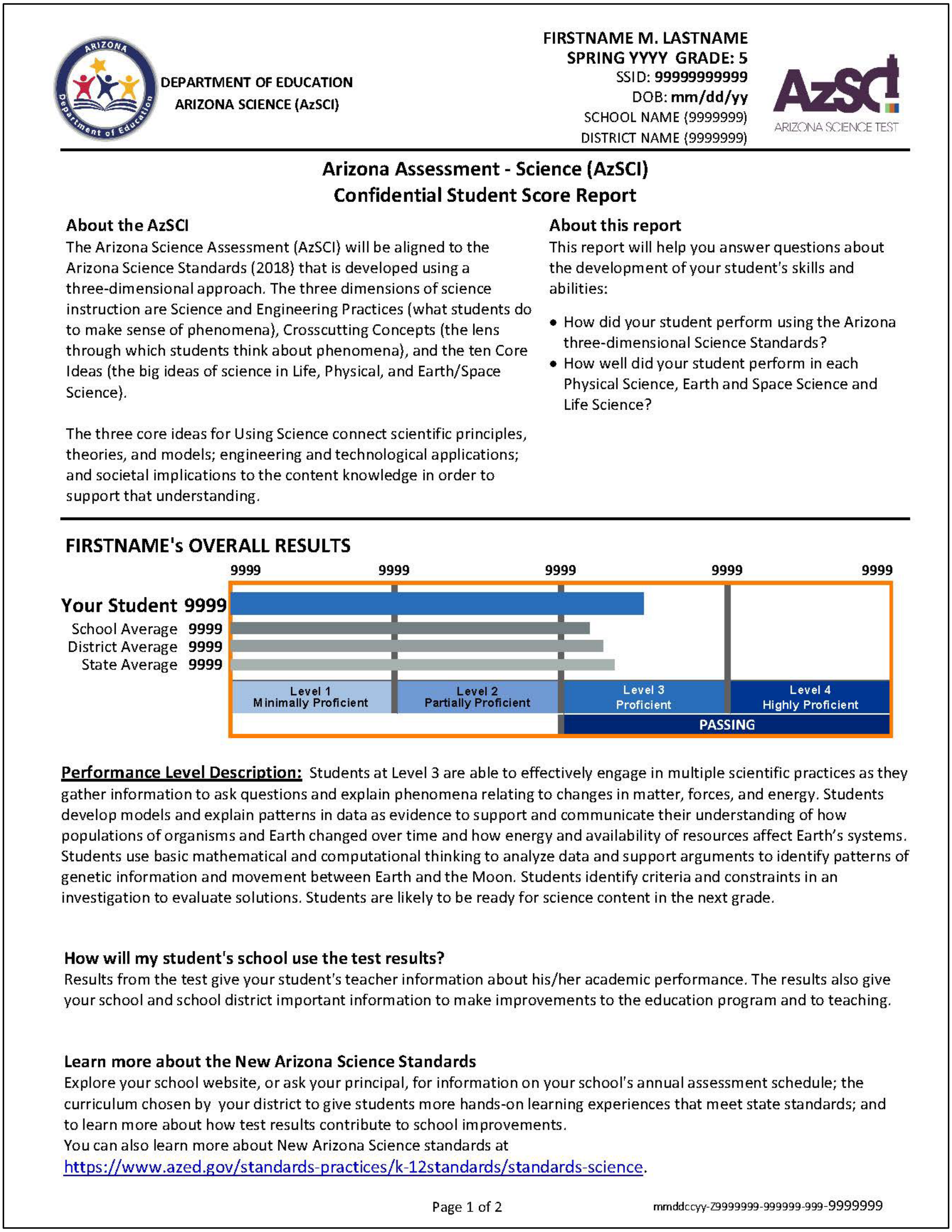
The following AzSCI reports were available online in PAN at https://az.pearsonaccessnext.com. PDF versions of the reports and district-wide electronic student data files were also available for downloading. District-level user roles provided access to all school-level and district-level reports, including all Confidential Student Score Reports for students who tested in the district. School-level user roles provided access to all school-level reports and all Confidential Student Score Reports for students who tested in the school. Figure 5.1 and Figure 5.2 present sample reports.

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

AzSCI reports have been designed with the user's comprehension in mind. The goal of these reports is not only to deliver accurate assessment data, but to ensure it is correctly interpreted and understood by the audience. To this end, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy guides the reader to compare like elements and avoid comparison of dissimilar elements. 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, 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.

Figure 5.1. Sample Report—Confidential Student Score Report



DEPARTMENT OF EDUCATION
ARIZONA SCIENCE (AzSCI)

FIRSTNAME M. LASTNAME
SPRING YYYY GRADE: 5
SSID: **99999999999**
DOB: **mm/dd/yy**
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) will be aligned to the Arizona Science Standards (2018) that is 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 ten Core Ideas (the big ideas of science in Life, Physical, and Earth/Space Science).

The three 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 three-dimensional Science Standards?
- How well did your student perform in each Physical Science, Earth and Space Science and Life Science?

FIRSTNAME's OVERALL RESULTS

9999 9999 9999 9999 9999

Your Student 9999
School Average **9999**
District Average **9999**
State Average **9999**

Level 1 Minimally Proficient
Level 2 Partially Proficient
Level 3 Proficient
Level 4 Highly Proficient
PASSING

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?

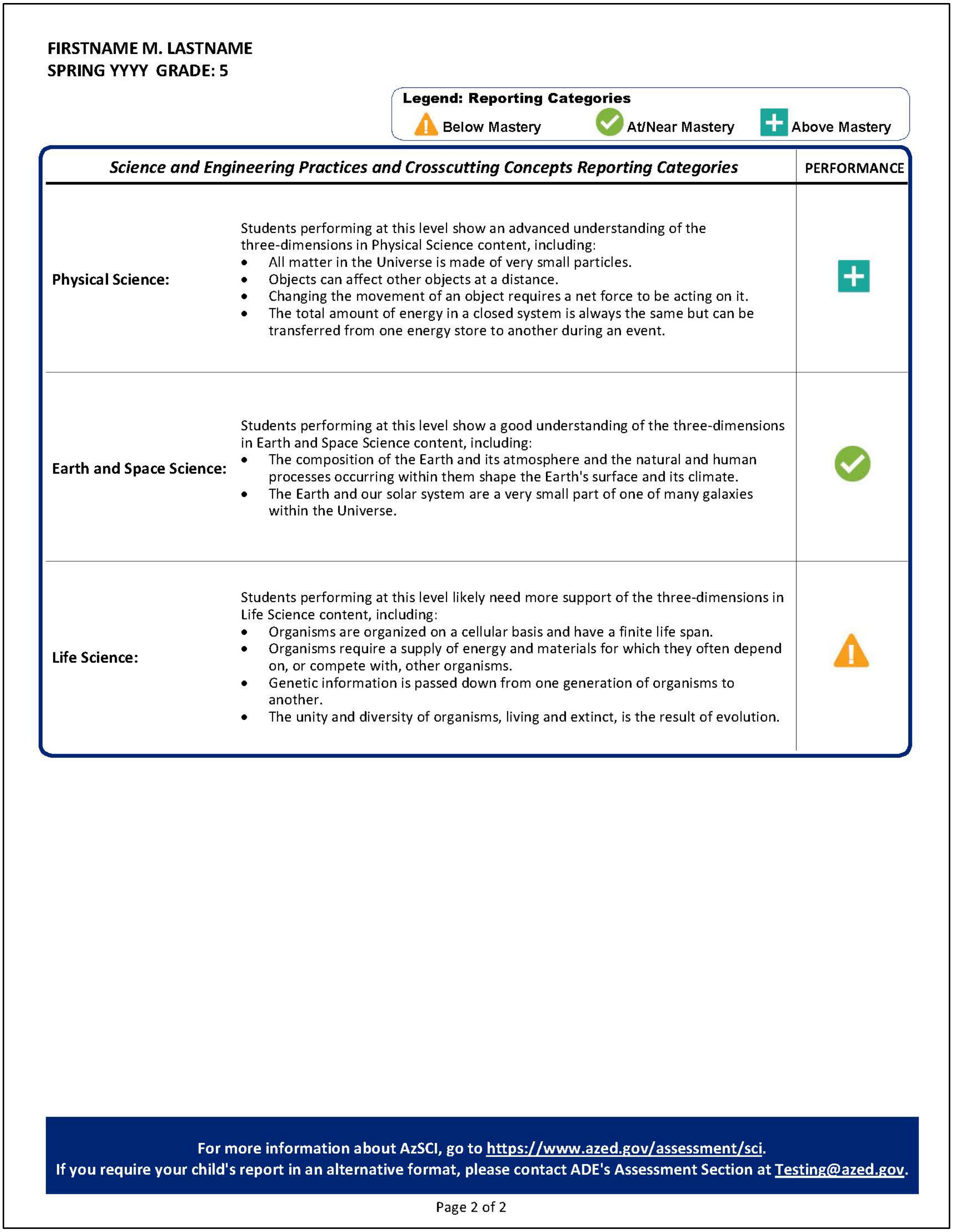
Results from the test give your student's teacher information about his/her academic performance. The results also give your school and school district important information to make improvements to the education program and to teaching.

Learn more about the New 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 New 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 YYYY GRADE: 5

Legend: Reporting Categories

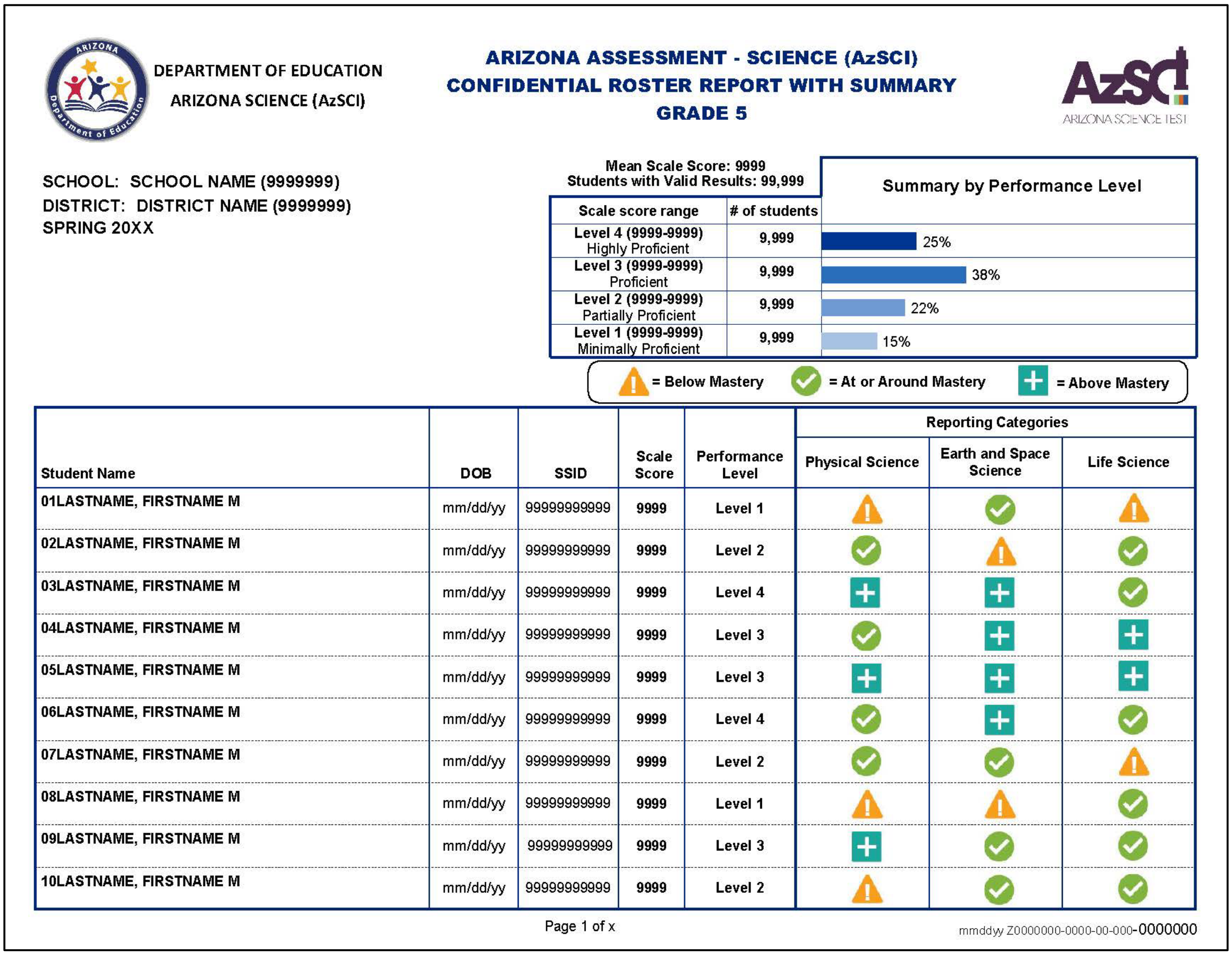
⚠ Below Mastery ✓ At/Near Mastery ➕ Above Mastery

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 5.2. Sample Report—Confidential Roster Report with Summary



DEPARTMENT OF EDUCATION
ARIZONA SCIENCE (AzSCI)

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

SCHOOL: SCHOOL NAME (9999999)
DISTRICT: DISTRICT NAME (9999999)
SPRING 20XX

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

Summary by Performance Level

Scale score range	# of students	
Level 4 (9999-9999) Highly Proficient	9,999	25%
Level 3 (9999-9999) Proficient	9,999	38%
Level 2 (9999-9999) Partially Proficient	9,999	22%
Level 1 (9999-9999) Minimally Proficient	9,999	15%

⚠ = Below Mastery ✓ = At or Around 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/yy	99999999999	9999	Level 1	⚠	✓	⚠
02LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 2	✓	⚠	✓
03LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 4	+	+	✓
04LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 3	✓	+	+
05LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 3	+	+	+
06LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 4	✓	+	✓
07LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 2	✓	✓	⚠
08LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 1	⚠	⚠	✓
09LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 3	+	✓	✓
10LASTNAME, FIRSTNAME M	mm/dd/yy	99999999999	9999	Level 2	⚠	✓	✓

Page 1 of x

mmddyy Z0000000-0000-00-000-0000000

Chapter 6: Classical Item Analysis

This chapter presents classical statistics for the data used for calibration and scaling for the Spring 2022 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).

6.1. Data

The classical item analysis was conducted based on the calibration samples as described in Section 7.1. Table 6.1 presents demographic information of the students included in the calibration sample for the Spring 2022 AzSCI assessments. 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 6.1. Frequency of Students by Subgroup: Number of Students by Grade

Subgroup	Grade 5	Grade 8	Grade 11
All	80,700	87,631	76,161
Male	40,761	44,587	38,150
Female	39,939	43,044	38,011
Hispanic	37,872	42,023	35,041
Non-Hispanic	42,828	45,608	41,120
American Indian	4,287	4,909	4,202
Asian	2,783	2,874	2,591
Black or African American	5,616	5,900	4,733
Multi-racial	4,565	4,457	3,371
Native Hawaiian or Other Pacific Islander	543	485	393
White	61,955	67,780	59,259
Missing	951	1,226	1,612
Special Ed.	10,362	9,276	6,723
EL	6,893	7,401	4,131
Low SES	32,783	33,219	26,631

6.2. Descriptive Statistics

Table 6.2 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 6.2. Classical Test Analysis Statistics

Grade	#Students	#Items	Max. Possible Raw Score	Mean Raw Score	SD Raw Score	Min. Raw Score	Max. Raw Score
5	80,700	50	55	24.85	10.00	0	55
8	87,631	50	55	21.40	9.81	0	54
11	76,161	50	55	19.68	9.54	0	55

6.3. 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. 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 0.9.

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 6.3 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 as 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 6.3. Classical Item Analysis Summary

Grade	#Items	Mean *P*-Value	Mean Point-Biserial
5	50	0.45	0.40
8	50	0.39	0.39
11	50	0.35	0.39

6.4. Distractor Analysis

Table 6.4 and Table 6.5 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 0.15 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.

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 incorrect options (i.e., distractors) because less-proficient students should be more likely to choose an incorrect option.

Table 6.4. Distractor Analysis Summary: Point-Biserial Correlations for Correct Options

Grade	#MC Items	Min.	P25	P50	P75	Max.
5	25	0.20	0.32	0.40	0.47	0.54
8	24	0.17	0.31	0.39	0.42	0.52
11	19	0.21	0.28	0.30	0.34	0.45

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

Table 6.5. Distractor Analysis Summary: Point-Biserial Correlations for Incorrect Options

Grade	#MC Items	Min.	P25	P50	P75	Max.
5	25	-0.33	-0.25	-0.19	-0.12	0.06
8	24	-0.31	-0.22	-0.17	-0.12	0.09
11	19	-0.31	-0.20	-0.14	-0.07	0.04

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

Chapter 7: Calibration and Scaling

This chapter describes the calibration and scaling procedures that took place for the Spring 2022 AzSCI assessment and summarizes the results, addressing Standards 1.10, 5.1, 5.2, 5.3, 7.2, 7.4, and 12.9 (AERA et al., 2014).

7.1. 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$^{\text{next}}$ (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., score sheets were double-scanned or students indicated as taking the test more than once)
- Records in which a student was enrolled in an exclusionary school list from ADE

7.2. Calibration Methods

Item response theory (IRT) models were used in the item calibration. Because Spring 2022 was the first operational AzSCI administration, free calibration was performed to establish a base scale. 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). Winsteps 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.

7.3. Calibration Results

All items for the AzSCI tests 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 7.1 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. Because this was the first operational administration of the AzSCI assessments, free calibration was used to estimate item parameters. When free calibration is performed, the average Rasch value is set to zero. Thus, the mean Rasch difficulty for all grades is zero. When fixed anchor calibration methods are used for future administrations, the mean Rasch should vary from zero.

Table 7.1. IRT Statistics Summary

Grade	#Items	Mean Rasch
5	50	0
8	50	0
11	50	0

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 particularly 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.2, and Figure B.3 in Appendix B present the item difficulty distribution on the lefthand side and the student ability distribution on the righthand side. 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 (i.e., *Partially Proficient*, *Proficient*, and *Highly Proficient*, respectively) for the total test.

In addition to the item-person map, two more graphs are presented to summarize the characteristics of each operational assessment in Figure B.4 – Figure B.9. The test characteristic curve (TCC) shows an expected total raw score across different student abilities, whereas 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(θ)* is the CSEM, and *TI(θ)* is the TIF (Embretson & Reise, 2000). Because the CSEM can be interpreted on the ability scale, the CSEM curve is presented over the TIF curve in this technical report.

7.4. Scaling Methods

Although student proficiency in science is estimated unidimensionally for the AzSCI assessment, subscores are reported at the domain level. 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:

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

where *A* and *B* are the scaling constants to transform the *Partially Proficient* and *Proficient* theta cuts to 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.

7.5. IRT Assumptions

It is important to evaluate how the Rasch models applied for AzSCI fit the data because reported scale scores are derived from theta estimated under the IRT models. Three major assumptions are investigated: (1) unidimensionality, (2) local item independence, and (3) item fit.

7.5.1. Unidimensionality

An assumption under the Rasch models is unidimensionality, that there is exactly one latent variable (e.g., science proficiency) that an instrument intends to measure. 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 the results of PCA. 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 extracted 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 7.2 presents the first 10 eigenvalues from the PCA for each operational form. 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 for each grade in Figure B.10, Figure B.11, and Figure B.12 in Appendix B. The PCA results with the parallel analysis criterion show only one dominant dimension, which supports unidimensionality.

Table 7.2. Eigenvalues from PCA

Grade	1	2	3	4	5	6	7	8	9	10
5	12.90	1.49	1.22	1.20	1.05	1.01	0.96	0.95	0.92	0.92
8	12.64	1.48	1.11	1.09	1.02	1.01	1.00	0.93	0.92	0.91
11	12.44	1.38	1.26	1.13	1.04	1.01	0.99	0.97	0.93	0.91

7.5.2. 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 (e.g., science proficiency) an instrument is intended to measure. 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).

As shown in Table 7.3 that summarizes the distribution of the residual correlations, most residual correlations are slightly negative or slightly positive, and none are greater than 0.70. The results of the residual correlations indicate that the local item independence assumption holds for the AzSCI tests.

Table 7.3. 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	0.24	0
8	1,225	-0.02	0.02	-0.09	-0.04	-0.03	-0.02	-0.01	0.00	0.13	0
11	1,225	-0.02	0.02	-0.08	-0.04	-0.03	-0.02	-0.01	0.01	0.09	0

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

7.5.3. 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 7.4 presents a summary of the item fit statistics, and Appendix B presents the statistics for each item. Items flagged by Winsteps' infit statistics will be reviewed during test construction for possible replacement in future administrations.

Table 7.4. IRT Item Fit Summary Statistics

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

Chapter 8: Test Results

This chapter contains information about the results of the administration of the Spring 2022 AzSCI assessment, 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).

Results presented in this chapter 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. 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 8.1 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 the performance levels for the total score. Table 8.2 presents the percentage of students in each level of mastery by domain, and Table 8.3 presents the mean and standard deviation of the scale score and the performance level distribution by accommodation. Appendix C presents the test results for each grade by subgroup. Histograms of the scale score distribution for the total score are also presented by grade in Appendix C.

Table 8.1. Overall Test Results

Grade	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
5	80,889	1324.21	38.91	28	44	23	5
8	87,698	1322.08	37.69	30	46	20	4
11	76,418	1319.27	36.99	31	48	17	3

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

Table 8.2. Performance Distributions by Domain: Percent of Students at each Level of Mastery

Grade	Domain	N	%Level 1	%Level 2	%Level 3
5	Physical Science	80,889	57	31	11
	Earth and Space Science	80,889	49	41	10
	Life Science	80,889	55	27	18
8	Physical Science	87,698	60	28	12
	Earth and Space Science	87,698	59	28	13
	Life Science	87,698	57	29	14
11	Physical Science	76,418	61	26	13
	Earth and Space Science	76,418	50	40	9
	Life Science	76,418	70	17	13

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

Table 8.3. Test Results by Accommodation

Grade	Accommodation	N	SS Mean	SS SD	%Level 1	%Level 2	%Level 3	%Level 4
5	Adult Transcription	14	1301.14	39.46	57	21	21	0
	Assistive Technology	5	*	*	*	*	*	*
	Sign Test Content	22	1276.77	29.62	77	18	5	0
	Simplified Directions	404	1292.96	30.14	64	30	6	0
	Translate Directions	223	1293.11	30.37	61	32	5	1
	Translation Dictionary	243	1294.20	30.29	60	34	6	1
8	Adult Transcription	3	*	*	*	*	*	*
	Assistive Technology	7	*	*	*	*	*	*
	Sign Test Content	30	1274.47	17.33	90	10	0	0
	Simplified Directions	349	1291.43	23.72	67	30	3	0
	Translate Directions	170	1290.08	22.56	69	31	0	1
	Translation Dictionary	214	1289.92	21.94	70	29	0	0
11	Adult Transcription	205	1343.46	41.24	16	39	36	9
	Assistive Technology	20	1289.55	31.29	60	40	0	0
	Sign Test Content	141	1289.55	24.95	72	26	1	1
	Simplified Directions	161	1286.68	21.80	76	24	0	0
	Translate Directions	107	1285.50	17.83	78	22	0	0
	Translation Dictionary	203	1288.85	20.36	71	29	0	0

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 and replaced with an asterisk (*).

Table 8.4 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 8.4. Scale Score Distribution by Performance Level

Grade	Performance Level	N	Average Scale Score	%	Cumulative %
5	Level 1	22,981	1279.21	28.41	28.41
	Level 2	35,806	1322.36	44.27	72.68
	Level 3	18,203	1366.48	22.50	95.18
	Level 4	3,899	1409.11	4.82	100.00
8	Level 1	26,096	1281.13	29.76	29.76
	Level 2	40,539	1321.13	46.23	75.98
	Level 3	17,748	1367.44	20.24	96.22
	Level 4	3,315	1413.38	3.78	100.00
11	Level 1	23,985	1281.48	31.39	31.39
	Level 2	36,990	1319.99	48.40	79.79
	Level 3	12,986	1368.02	16.99	96.78
	Level 4	2,457	1419.53	3.22	100.00

Note. Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

Chapter 9: Reliability and Validity

This chapter provides evidence supporting the reliability and validity of scores on the Spring 2022 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).

9.1. 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 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 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 9.1 presents the coefficient alphas and SEMs (computed based on the calibration sample) for the total and domain scores. The test-level and domain-level reliability coefficient alpha results suggest that the AzSCI assessments produce reliable scores.

Table 9.1. Coefficient Alpha and SEM by Total and Domain Score

Grade	Domain	N	#Items	Coefficient Alpha	SEM
5	Total	80,700	50	0.89	3.31
	Physical Science	80,700	21	0.74	2.14
	Earth and Space Science	80,700	12	0.62	1.69
	Life Science	80,700	17	0.80	1.87
8	Total	87,631	50	0.89	3.28
	Physical Science	87,631	19	0.77	1.97
	Earth and Space Science	87,631	16	0.72	1.90
	Life Science	87,631	15	0.71	1.78
11	Total	76,161	50	0.88	3.28
	Physical Science	76,161	18	0.74	1.91
	Earth and Space Science	76,161	11	0.64	1.43
	Life Science	76,161	21	0.75	2.26

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.

9.2. 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.

Two types of DIF, namely uniform DIF and non-uniform DIF, are typically investigated. Uniform DIF means that, given the ability, the probability of getting an item correct is always higher for one subgroup than the other across the full range of the ability continuum. In other words, the direction of DIF remains the same on the entire ability continuum. Non-uniform DIF occurs when the direction of DIF changes at some point within the ability continuum. To date, many DIF detection methods have been proposed. For the AzSCI assessment, two uniform DIF methods are used.

The Mantel-Haenszel (MH) method (Holland & Thayer, 1988; Mantel & Haenszel, 1959) was used to investigate DIF on one-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 statistic (ΔMH) was computed. Educational Testing Service (ETS) first developed the ΔMH DIF statistic. 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:

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

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 9.2 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 9.2. DIF Flag Categories

Category	Description	MH Criterion	SMD Criterion
A	No DIF	MH chi-square 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 $\|SMD\| \leq 0.17$
B	Weak DIF	MH chi-square significantly different from 0 ($p < 0.05$) and $1.0 \leq \|\Delta MH\ DIF\| < 1.5$	MH chi-square significantly different from 0 ($p < 0.05$) and $0.17 < \|SMD\| \leq 0.25$
C	Strong DIF	MH chi-square significantly higher than 1 ($p < 0.05$) and $\|\Delta MH\ DIF\| \geq 1.5$	MH chi-square significantly different from 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 9.3 presents the number of operational items exhibiting strong DIF between any two groups. The items displaying 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 9.3. Number of Items Exhibiting Strong DIF

Grade	#Items	#Items with Strong DIF
5	50	–
8	50	–
11	50	1

9.3. Correlations Among Domains

Correlations were examined between the total raw score and the 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 7. Table 9.4 presents the test correlations and disattenuated correlations between the total raw score and the domain raw score. The numbers in the lower diagonal of the table are the disattenuated correlations, which were 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 9.1, were used to calculate the corrected correlation coefficients for attenuation. The disattenuated correlations could be greater than 1.00.

Table 9.4. 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	0.92	0.84	0.91
	Physical Science	1.13	1.00	0.68	0.74
	Earth and Space Science	1.13	1.00	1.00	0.68
	Life Science	1.08	0.96	0.97	1.00
8	Total	1.00	0.91	0.89	0.87
	Physical Science	1.10	1.00	0.71	0.70
	Earth and Space Science	1.11	0.95	1.00	0.68
	Life Science	1.09	0.95	0.95	1.00
11	Total	1.00	0.91	0.84	0.93
	Physical Science	1.13	1.00	0.69	0.73
	Earth and Space Science	1.12	1.00	1.00	0.68
	Life Science	1.14	0.98	0.98	1.00

9.4. 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(s) 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 interested 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.

9.4.1. 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.

Item specifications and test blueprints are the core documents that ensure that the assessments are aligned to the Arizona Science Standards. 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. In the case of AzSCI, it was important to consider the relative weight of Physical Science, Life Science, and Earth and Space Science for each grade. Details about the development of the documents are presented in Chapter 2.

Once the item specifications and blueprints were established, item development took place. It was a rigorous and iterative process involving the Pearson content team and ADE staff, as described in Chapter 3. Arizona educators, parents, and community members also participated in the content, bias, and sensitivity committees to evaluate the newly developed items. Reviewers were asked to evaluate the item for its alignment, 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 appropriate to be field tested on the assessment.

The test development process described in Chapter 3 ensures that the AzSCI assessments meet the test blueprint and other content criteria and psychometric targets. Beyond the test blueprint, ADE staff and Pearson attempted to include items measuring different levels of rigor to cover the Arizona Science Standards as much as possible.

9.4.2. 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 presented in Chapter 3, all newly developed items for the AzSCI assessment also go through a rigorous item review process, including content, bias, and sensitivity committees. During the review process, a group of educators are trained to evaluate whether the items are aligned to the Arizona Science Standards and to assess important knowledge or skills identified by the standards and item specifications. The items deemed to be acceptable by the committees are eligible to be field tested on the AzSCI test.

9.4.3. 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 indicated in Chapter 7. 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 (Horn, 1965) criterion indicated that 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 7.3, showed that most statistics are either slightly negative or slightly positive, which indicates that the item independence assumption generally holds for the AzSCI tests.

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 7.4. Correlations between the total score and domain score are presented in Table 9.4, showing that they are at least moderately correlated to each other, if not highly correlated, 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 6.4 and Table 6.5 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 9.2. Items showing strong DIF are flagged for possible replacement in future administrations.

9.4.4. 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 experienced educators using the Extended Modified (Yes/No) Angoff standard setting method. See Section 10.1 for more details on standard setting.

9.4.5. 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. Because the AzSCI and AASA assessments are administered to all eligible Arizona students, scores on the tests are expected to be positively correlated. Table 9.5 presents the correlation between AzSCI and AASA scale scores from the Spring 2022 administration. AzSCI is highly correlated with both AASA ELA and Mathematics, with the correlations ranging from 0.74 to 0.82. 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, so there are no results for Grade 11.

Table 9.5. Correlation between AzSCI and AASA Scale Scores

	AASA ELA		AASA Mathematics	
Grade	N	Correlation	N	Correlation
5	74,736	0.82	75,061	0.74
8	80,240	0.78	80,769	0.76
11	–	–	–	–

9.4.6. Summary

Overall, the validity evidence provided above 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. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Additional evidence should and will be added to the AzSCI technical report in the future as appropriate.

Chapter 10: CLASSIFICATION INTO PERFORMANCE LEVELS

This chapter provides information regarding classification of students into performance levels for the Spring 2022 AzSCI assessments, 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).

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 section of the technical report provides information regarding classification of students into these four categories. Arizona educators made recommendations for cut scores for each performance level during the standard setting workshop in June 2022. Analyses were conducted to examine the consistency and accuracy with which students who took the Spring 2022 AzSCI assessment were assigned to the performance levels.

10.1. Standard Setting

Standard setting for the AzSCI tests was conducted 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. All technical documentation regarding the standard setting will be provided in the standard setting technical report when it is available.

Table 10.1 presents the final scale score ranges for the AzSCI performance levels, and Table 10.2 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., 12 for *Partially Proficient* and *Proficient* and 14 for *Highly Proficient*).

Table 10.1. 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 10.2. 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	12	1350	12	1395	14
8	1300	12	1350	12	1399	14
11	1300	12	1350	12	1402	14

10.2. 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 the AzSCI assessments, 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 in the diagonal in Table 10.3 (i.e., students classified consistently between two parallel forms, listed in bold). Similarly, classification accuracy is calculated as the proportion of students in the diagonal in Table 10.4 (i.e., students classified the same between observed scores and true scores, listed in bold).

Table 10.3. Classification Consistency for the *Proficient* Cut

		Expected Performance on Parallel Form	
		Not Proficient	Proficient
Observed Performance on Actual Form	Not Proficient	**Consistent Classification**	Inconsistent Classification
	Proficient	Inconsistent Classification	**Consistent Classification**

Table 10.4. Classification Accuracy for the *Proficient* Cut

		Expected Performance on Test	
		Not Proficient	Proficient
Observed Performance on Test	Not Proficient	**Accurate Classification**	False Negative
	Proficient	False Positive	**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 on the AzSCI tests. 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 10.5 presents the classification consistency and accuracy results for the Spring 2022 AzSCI assessment, 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 for each grade 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 10.5. Classification Consistency and Accuracy Results

Grade	N	Consistency	Inconsistency	Chance	κ	Accuracy	False Positive	False Negative
5	80,700	0.72	0.28	0.33	0.58	0.80	0.11	0.09
8	87,631	0.72	0.28	0.34	0.57	0.80	0.11	0.09
11	76,161	0.70	0.30	0.36	0.53	0.79	0.13	0.09

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.

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

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.

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	80,700	0.55	0.34
2	XI	1	80,700	0.63	0.38
3	MC	1	80,700	0.52	0.36
4	XI	1	80,700	0.35	0.44
5	MX	2	80,700	0.33	0.33
6	XI	1	80,700	0.32	0.26
7	MC	1	80,700	0.40	0.20
8	MC	1	80,700	0.47	0.29
9	MX	2	80,700	0.60	0.64
10	XI	1	80,700	0.34	0.39
11	MC	1	80,700	0.54	0.41
12	MC	1	80,700	0.48	0.41
13	XI	1	80,700	0.34	0.39
14	MC	1	80,700	0.60	0.47
15	MX	1	80,700	0.47	0.60
16	MC	1	80,700	0.28	0.29
17	MC	1	80,700	0.31	0.29
18	MX	1	80,700	0.29	0.47
19	MX	1	80,700	0.38	0.32
20	XI	1	80,700	0.29	0.31
21	MC	1	80,700	0.36	0.27
22	MC	1	80,700	0.62	0.41
23	MX	2	80,700	0.42	0.33
24	XI	1	80,700	0.22	0.24
25	MC	1	80,700	0.48	0.33
26	MX	1	80,700	0.53	0.40
27	MC	1	80,700	0.46	0.32
28	MC	1	80,700	0.66	0.53
29	MC	1	80,700	0.50	0.47
30	MC	1	80,700	0.57	0.52
31	XI	1	80,700	0.33	0.29
32	MC	1	80,700	0.81	0.49

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
33	XI	1	80,700	0.24	0.33
34	MX	1	80,700	0.69	0.56
35	MC	1	80,700	0.64	0.52
36	XI	1	80,700	0.69	0.37
37	XI	1	80,700	0.33	0.43
38	MC	1	80,700	0.67	0.54
39	XI	1	80,700	0.47	0.50
40	MX	1	80,700	0.28	0.43
41	XI	1	80,700	0.40	0.36
42	MC	1	80,700	0.62	0.38
43	MX	2	80,700	0.50	0.51
44	MC	1	80,700	0.53	0.40
45	MX	2	80,700	0.37	0.43
46	MC	1	80,700	0.45	0.23
47	MC	1	80,700	0.33	0.44
48	MC	1	80,700	0.35	0.38
49	MX	1	80,700	0.22	0.43
50	MC	1	80,700	0.46	0.47

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	87,631	0.53	0.31
2	MC	1	87,631	0.56	0.37
3	MX	1	87,631	0.26	0.41
4	MC	1	87,631	0.44	0.40
5	MC	1	87,631	0.77	0.40
6	MX	1	87,631	0.38	0.37
7	MX	2	87,631	0.51	0.56
8	XI	1	87,631	0.37	0.38
9	MC	1	87,631	0.62	0.52
10	MX	1	87,631	0.29	0.51
11	MC	1	87,631	0.48	0.30
12	MX	1	87,631	0.41	0.45
13	MX	2	87,631	0.28	0.32
14	XI	1	87,631	0.20	0.37
15	XI	1	87,631	0.32	0.43
16	MX	1	87,631	0.20	0.29
17	MC	1	87,631	0.53	0.47
18	MC	1	87,631	0.31	0.39
19	MC	1	87,631	0.50	0.41
20	MC	1	87,631	0.31	0.32
21	MC	1	87,631	0.67	0.48
22	XI	1	87,631	0.35	0.44
23	MC	1	87,631	0.46	0.35

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
24	MX	2	87,631	0.47	0.56
25	XI	1	87,631	0.24	0.40
26	MX	2	87,631	0.45	0.56
27	XI	1	87,631	0.26	0.53
28	MX	1	87,631	0.35	0.34
29	MC	1	87,631	0.38	0.41
30	MC	1	87,631	0.43	0.37
31	MX	1	87,631	0.20	0.33
32	MC	1	87,631	0.50	0.42
33	MC	1	87,631	0.34	0.17
34	MC	1	87,631	0.42	0.30
35	MC	1	87,631	0.49	0.47
36	XI	1	87,631	0.18	0.28
37	XI	1	87,631	0.49	0.52
38	MX	1	87,631	0.33	0.36
39	MC	1	87,631	0.47	0.30
40	MX	1	87,631	0.28	0.41
41	MC	1	87,631	0.46	0.50
42	MC	1	87,631	0.28	0.30
43	MX	1	87,631	0.49	0.51
44	MC	1	87,631	0.46	0.23
45	MX	2	87,631	0.23	0.23
46	XI	1	87,631	0.39	0.51
47	MC	1	87,631	0.49	0.38
48	XI	1	87,631	0.20	0.28
49	MC	1	87,631	0.36	0.42
50	MX	1	87,631	0.14	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	76,161	0.51	0.28
2	MX	2	76,161	0.48	0.45
3	MC	1	76,161	0.44	0.44
4	MX	1	76,161	0.33	0.58
5	MC	1	76,161	0.46	0.34
6	MC	1	76,161	0.41	0.45
7	MC	1	76,161	0.54	0.44
8	XI	1	76,161	0.29	0.48
9	XI	1	76,161	0.20	0.33
10	MX	1	76,161	0.17	0.42
11	MX	1	76,161	0.19	0.38
12	MX	1	76,161	0.18	0.47
13	MC	1	76,161	0.44	0.34
14	MC	1	76,161	0.41	0.31

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
15	MC	1	76,161	0.37	0.27
16	XI	1	76,161	0.79	0.29
17	MX	2	76,161	0.33	0.44
18	MC	1	76,161	0.44	0.29
19	MC	1	76,161	0.38	0.30
20	XI	1	76,161	0.20	0.37
21	MC	1	76,161	0.29	0.25
22	MC	1	76,161	0.45	0.29
23	MX	2	76,161	0.32	0.49
24	MX	1	76,161	0.45	0.48
25	MX	1	76,161	0.26	0.42
26	XI	1	76,161	0.15	0.20
27	MX	1	76,161	0.40	0.32
28	XI	1	76,161	0.40	0.28
29	MX	1	76,161	0.29	0.50
30	MC	1	76,161	0.43	0.36
31	MC	1	76,161	0.39	0.29
32	MX	2	76,161	0.37	0.62
33	MX	1	76,161	0.49	0.58
34	MX	1	76,161	0.39	0.44
35	MX	1	76,161	0.46	0.45
36	XI	1	76,161	0.25	0.45
37	MX	1	76,161	0.34	0.40
38	MX	2	76,161	0.43	0.43
39	XI	1	76,161	0.29	0.52
40	MC	1	76,161	0.43	0.30
41	MC	1	76,161	0.27	0.26
42	MX	1	76,161	0.13	0.38
43	MC	1	76,161	0.23	0.21
44	XI	1	76,161	0.52	0.42
45	MX	1	76,161	0.15	0.31
46	XI	1	76,161	0.18	0.48
47	MC	1	76,161	0.30	0.30
48	XI	1	76,161	0.48	0.39
49	MX	1	76,161	0.24	0.48
50	MC	1	76,161	0.39	0.30

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	55.4	0.34	10.1	-0.18	25.6	-0.12	8.9	-0.22	0.10
3	52.0	0.36	24.9	-0.16	9.4	-0.10	13.6	-0.23	0.10
7	39.6	0.20	9.5	-0.11	23.4	-0.20	27.4	0.06	0.20
8	47.1	0.29	28.2	-0.01	15.8	-0.26	8.7	-0.15	0.20
11	54.1	0.41	9.2	-0.20	20.1	-0.31	16.4	-0.04	0.20
12	48.3	0.41	16.1	-0.25	11.5	-0.31	23.9	-0.03	0.20
14	59.9	0.47	17.1	-0.20	13.3	-0.27	9.5	-0.21	0.20
16	27.9	0.29	31.0	-0.12	26.9	-0.12	13.9	-0.05	0.30
17	30.5	0.29	20.0	-0.13	29.4	-0.07	19.8	-0.13	0.30
21	36.2	0.27	28.2	0.06	25.7	-0.25	9.6	-0.17	0.30
22	61.9	0.41	13.1	-0.18	9.0	-0.28	15.7	-0.15	0.30
25	48.4	0.33	13.8	-0.09	20.4	-0.24	17.0	-0.09	0.40
27	45.7	0.32	11.8	-0.12	29.0	-0.17	13.5	-0.13	0.10
28	66.1	0.53	7.3	-0.22	16.4	-0.27	10.1	-0.30	0.10
29	49.6	0.47	14.6	-0.19	17.3	-0.25	18.4	-0.18	0.10
30	57.2	0.52	16.7	-0.29	15.8	-0.29	10.2	-0.13	0.10
32	80.9	0.49	5.9	-0.22	8.2	-0.31	4.9	-0.25	0.10
35	63.7	0.52	13.1	-0.31	18.3	-0.25	4.8	-0.21	0.10
38	67.0	0.54	11.6	-0.31	11.6	-0.25	9.5	-0.24	0.20
42	61.9	0.38	14.9	-0.20	13.5	-0.10	9.5	-0.25	0.20
44	53.3	0.40	20.4	-0.06	16.4	-0.33	9.6	-0.17	0.30
46	45.4	0.23	16.9	0.04	23.2	-0.23	14.3	-0.08	0.30
47	32.9	0.44	25.1	-0.04	18.3	-0.29	23.4	-0.18	0.30
48	35.1	0.38	16.8	-0.07	16.8	-0.23	31.0	-0.14	0.30
50	45.7	0.47	21.0	-0.27	20.7	-0.22	12.3	-0.10	0.30

Note. 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	52.6	0.31	28.9	-0.09	10.7	-0.23	7.8	-0.16	0.10
2	56.2	0.37	22.8	-0.13	10.1	-0.23	10.9	-0.18	0.00
4	44.1	0.40	18.9	-0.20	27.2	-0.17	9.7	-0.14	0.10
5	76.8	0.40	8.3	-0.23	11.7	-0.24	3.1	-0.15	0.10
9	62.2	0.52	11.0	-0.21	13.6	-0.28	13.0	-0.26	0.20
11	47.5	0.30	12.2	-0.16	26.2	-0.12	13.9	-0.12	0.30
17	53.3	0.47	9.2	-0.17	22.3	-0.30	14.8	-0.16	0.40
18	30.9	0.39	15.5	-0.16	23.8	-0.02	29.4	-0.24	0.40
19	50.0	0.41	17.6	-0.20	18.0	-0.21	14.0	-0.13	0.40
20	31.1	0.32	20.1	-0.03	22.6	-0.19	25.8	-0.12	0.40
21	66.5	0.48	9.6	-0.23	9.7	-0.28	13.9	-0.22	0.40
23	45.9	0.35	15.2	-0.22	20.5	-0.19	18.0	-0.04	0.40
29	37.6	0.41	15.2	-0.16	33.4	-0.20	13.6	-0.14	0.10
30	42.8	0.37	13.9	-0.21	26.1	-0.15	17.1	-0.12	0.10
32	50.0	0.42	15.2	-0.17	18.9	-0.28	15.8	-0.09	0.10
33	34.1	0.17	18.9	-0.02	28.5	-0.23	18.3	0.09	0.10
34	41.5	0.30	27.5	0.04	16.4	-0.26	14.5	-0.20	0.10
35	49.2	0.47	21.4	-0.12	14.7	-0.31	14.6	-0.22	0.10
39	46.9	0.30	10.9	-0.15	29.2	-0.14	12.9	-0.10	0.30
41	45.7	0.50	16.7	-0.18	18.8	-0.29	18.5	-0.17	0.30
42	28.0	0.30	22.7	-0.13	26.3	-0.01	22.7	-0.17	0.30
44	45.6	0.23	22.5	0.03	16.4	-0.26	15.2	-0.07	0.30
47	49.2	0.38	17.4	-0.19	17.7	-0.20	15.2	-0.11	0.40
49	35.6	0.42	30.3	-0.17	15.0	-0.21	18.8	-0.11	0.30

Note. 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 Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%Omit
1	51.4	0.28	7.6	-0.16	18.6	-0.27	22.3	0.02	0.10
3	43.5	0.44	17.9	-0.14	19.7	-0.20	18.8	-0.22	0.10
5	46.2	0.34	18.6	-0.24	27.2	-0.10	7.9	-0.11	0.10
6	41.2	0.45	16.5	-0.26	25.8	-0.20	16.4	-0.10	0.10
7	54.4	0.44	10.2	-0.19	23.0	-0.21	12.3	-0.22	0.10
13	43.8	0.34	12.8	-0.17	20.9	-0.26	22.3	-0.01	0.30
14	40.5	0.31	22.5	0.03	17.9	-0.26	18.9	-0.15	0.30
15	37.3	0.27	14.6	-0.10	16.7	-0.20	31.2	-0.04	0.30
18	44.4	0.29	10.6	-0.15	18.6	-0.20	26.1	-0.04	0.30
19	37.9	0.30	22.0	-0.09	23.5	-0.20	16.3	-0.07	0.30
21	29.2	0.25	17.2	-0.19	15.4	-0.12	37.8	0.01	0.40
22	45.4	0.29	18.3	-0.10	21.9	-0.16	14.0	-0.11	0.40
30	42.7	0.36	32.3	-0.17	16.5	-0.21	8.4	-0.08	0.10
31	39.2	0.29	25.2	-0.04	17.9	-0.26	17.6	-0.06	0.10
40	43.3	0.30	14.0	-0.07	22.0	-0.31	20.6	0.01	0.20
41	26.5	0.26	19.0	-0.23	30.7	-0.04	23.6	-0.01	0.20
43	22.9	0.21	31.1	-0.15	33.8	0.04	12.0	-0.11	0.20
47	29.5	0.30	26.0	-0.16	32.2	-0.04	11.9	-0.13	0.20
50	39.3	0.30	16.2	-0.08	33.4	-0.11	10.9	-0.21	0.30

Note. 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) 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	-0.5062	0.0077	1.05
2	XI	-0.8770	0.0079	1.00
3	MC	-0.3433	0.0077	1.03
4	XI	0.5064	0.0081	0.96
5	MX	0.7179	0.0061	1.19
6	XI	0.6267	0.0082	1.13
7	MC	0.2569	0.0078	1.19
8	MC	-0.1099	0.0077	1.11
9	MX	-0.6669	0.0052	0.82
10	MC	0.5210	0.0081	1.01
11	MC	-0.4415	0.0077	0.99
12	MC	-0.1666	0.0077	0.99
13	MC	0.5652	0.0081	1.01
14	MC	-0.7203	0.0078	0.92
15	MX	-0.0984	0.0077	0.81
16	MC	0.8765	0.0085	1.07
17	MC	0.7303	0.0083	1.08
18	MX	0.8361	0.0084	0.90
19	MX	0.3458	0.0079	1.07
20	XI	0.8268	0.0084	1.07
21	MC	0.4283	0.0080	1.11
22	MC	-0.8200	0.0078	0.97
23	MX	0.2985	0.0064	1.14
24	XI	1.2792	0.0092	1.10
25	MC	-0.1713	0.0077	1.07
26	MX	-0.3651	0.0077	0.99
27	MC	-0.0423	0.0077	1.07
28	MC	-1.0331	0.0080	0.85
29	MC	-0.2291	0.0077	0.93
30	MC	-0.5907	0.0077	0.88

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
31	XI	0.6003	0.0081	1.09
32	MC	-1.9131	0.0094	0.83
33	MC	1.1500	0.0089	1.02
34	MX	-1.1836	0.0082	0.80
35	MC	-0.9093	0.0079	0.86
36	XI	-1.1644	0.0081	1.00
37	MC	0.5867	0.0081	0.96
38	MC	-1.0822	0.0081	0.84
39	XI	-0.0900	0.0077	0.91
40	MX	0.8875	0.0085	0.95
41	XI	0.2616	0.0079	1.04
42	MC	-0.8180	0.0078	1.00
43	MX	-0.2609	0.0055	1.01
44	MC	-0.4066	0.0077	1.00
45	MX	0.4249	0.0058	1.10
46	MC	-0.0258	0.0077	1.16
47	MC	0.5998	0.0081	0.95
48	MC	0.4835	0.0080	1.01
49	MX	1.2681	0.0092	0.92
50	MC	-0.0422	0.0077	0.94

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

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-0.6977	0.0073	1.07
2	MC	-0.8695	0.0074	1.00
3	MX	0.6427	0.0083	0.97
4	MC	-0.2972	0.0074	1.00
5	MC	-1.9554	0.0085	0.91
6	MX	0.0280	0.0076	1.03
7	MX	-0.6259	0.0052	0.93
8	MC	0.0392	0.0076	1.01
9	MC	-1.1566	0.0075	0.85
10	MX	0.5035	0.0081	0.89
11	MC	-0.4564	0.0074	1.09
12	MX	-0.1281	0.0075	0.95
13	MX	0.7179	0.0061	1.19
14	XI	1.0447	0.0091	0.98
15	MC	0.3445	0.0079	0.96
16	MX	1.0773	0.0091	1.04
17	MC	-0.7319	0.0073	0.92
18	MC	0.3758	0.0079	0.99
19	MC	-0.5746	0.0073	0.98
20	MC	0.3638	0.0079	1.06
21	MC	-1.3728	0.0077	0.88

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
22	MC	0.1733	0.0077	0.96
23	MC	-0.3792	0.0074	1.05
24	MX	-0.4483	0.0051	0.96
25	XI	0.8131	0.0086	0.97
26	MX	-0.3135	0.0053	0.93
27	XI	0.6512	0.0083	0.86
28	MX	0.1358	0.0077	1.06
29	MC	0.0218	0.0076	0.98
30	MC	-0.2306	0.0074	1.02
31	MX	1.0276	0.0090	1.03
32	MC	-0.5733	0.0073	0.98
33	MC	0.2032	0.0077	1.21
34	MC	-0.1676	0.0075	1.09
35	MC	-0.5356	0.0073	0.93
36	XI	1.1930	0.0094	1.06
37	XI	-0.5050	0.0073	0.87
38	MX	0.2588	0.0078	1.03
39	MC	-0.4275	0.0074	1.09
40	MX	0.5697	0.0082	0.98
41	MC	-0.3720	0.0074	0.90
42	MC	0.5415	0.0082	1.07
43	MX	-0.5366	0.0073	0.88
44	MC	-0.3687	0.0074	1.15
45	MX	0.8293	0.0061	1.35
46	XI	-0.0490	0.0075	0.89
47	MC	-0.5397	0.0073	1.01
48	MC	1.0512	0.0091	1.06
49	MC	0.1285	0.0077	0.98
50	MX	1.5772	0.0104	0.96

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

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-0.8179	0.0078	1.07
2	MX	-0.7161	0.0048	1.14
3	MC	-0.4466	0.0079	0.94
4	MX	0.0531	0.0083	0.82
5	MC	-0.5746	0.0078	1.02
6	MC	-0.3390	0.0080	0.94
7	MC	-0.9577	0.0078	0.92
8	XI	0.2703	0.0086	0.91
9	XI	0.8594	0.0097	1.03
10	MX	1.0640	0.0102	0.94
11	MX	0.9482	0.0099	0.98
12	MX	0.9979	0.0101	0.89

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
13	MC	-0.4614	0.0079	1.04
14	MC	-0.3011	0.0080	1.07
15	MC	-0.1439	0.0081	1.10
16	XI	-2.2336	0.0093	0.98
17	MX	0.1268	0.0061	1.06
18	MC	-0.4884	0.0079	1.08
19	MC	-0.1774	0.0081	1.07
20	XI	0.9016	0.0098	0.99
21	MC	0.2847	0.0086	1.12
22	MC	-0.5356	0.0079	1.07
23	MX	0.0576	0.0057	1.04
24	MX	-0.5011	0.0079	0.90
25	MX	0.4773	0.0089	0.96
26	MC	1.2294	0.0107	1.10
27	MX	-0.2643	0.0080	1.05
28	XI	-0.2685	0.0080	1.10
29	MX	0.2973	0.0086	0.89
30	MC	-0.4074	0.0079	1.02
31	MC	-0.2374	0.0080	1.09
32	MX	-0.1402	0.0057	0.85
33	MX	-0.6909	0.0078	0.81
34	MX	-0.2391	0.0080	0.95
35	MX	-0.5458	0.0079	0.93
36	XI	0.5081	0.0090	0.93
37	MX	0.0328	0.0083	0.98
38	MX	-0.3941	0.0057	1.08
39	XI	0.2740	0.0086	0.87
40	MC	-0.4358	0.0079	1.07
41	MC	0.4398	0.0089	1.10
42	MX	1.4751	0.0115	0.95
43	MC	0.6678	0.0093	1.13
44	XI	-0.8548	0.0078	0.94
45	MX	1.2846	0.0109	1.03
46	MC	0.9850	0.0100	0.89
47	MC	0.2643	0.0086	1.07
48	XI	-0.6767	0.0078	0.97
49	MX	0.5957	0.0092	0.91
50	MC	-0.2452	0.0080	1.07

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 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	1204	25	1
4	1217	22	1
5	1228	20	1
6	1237	19	1
7	1245	18	1
8	1252	17	1
9	1258	16	1
10	1264	15	1
11	1269	15	1
12	1274	14	1
13	1279	14	1
14	1284	14	1
15	1288	13	1
16	1292	13	1
17	1296	13	1
18	1300	13	2
19	1304	13	2
20	1308	12	2
21	1311	12	2
22	1315	12	2
23	1318	12	2
24	1322	12	2
25	1325	12	2
26	1329	12	2
27	1332	12	2
28	1336	12	2
29	1339	12	2
30	1343	12	2
31	1346	12	2
32	1350	12	3
33	1354	12	3
34	1357	12	3
35	1361	13	3
36	1365	13	3
37	1369	13	3
38	1373	13	3
39	1377	13	3
40	1381	13	3
41	1385	14	3
42	1390	14	3
43	1395	14	4
44	1400	15	4

Raw Score	Scale Score	CSEM	Performance Level
45	1405	15	4
46	1411	16	4
47	1417	17	4
48	1424	18	4
49	1432	19	4
50	1441	20	4
51	1452	22	4
52	1465	25	4
53	1484	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	1218	25	1
4	1231	22	1
5	1241	20	1
6	1250	18	1
7	1258	17	1
8	1265	16	1
9	1271	16	1
10	1277	15	1
11	1282	15	1
12	1287	14	1
13	1291	14	1
14	1296	13	1
15	1300	13	2
16	1304	13	2
17	1308	13	2
18	1312	13	2
19	1316	12	2
20	1319	12	2
21	1323	12	2
22	1326	12	2
23	1330	12	2
24	1333	12	2
25	1337	12	2
26	1340	12	2
27	1343	12	2
28	1347	12	2
29	1350	12	3
30	1353	12	3
31	1357	12	3

Raw Score	Scale Score	CSEM	Performance Level
32	1360	12	3
33	1364	12	3
34	1367	12	3
35	1371	12	3
36	1374	12	3
37	1378	13	3
38	1382	13	3
39	1386	13	3
40	1390	13	3
41	1394	13	3
42	1399	14	4
43	1404	14	4
44	1409	15	4
45	1414	15	4
46	1419	16	4
47	1426	16	4
48	1433	17	4
49	1440	18	4
50	1449	20	4
51	1460	22	4
52	1473	25	4
53	1491	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	1204	31	1
3	1222	25	1
4	1235	22	1
5	1246	20	1
6	1255	19	1
7	1262	17	1
8	1269	16	1
9	1275	16	1
10	1281	15	1
11	1286	15	1
12	1291	14	1
13	1296	14	1
14	1300	13	2
15	1304	13	2
16	1308	13	2
17	1312	13	2
18	1316	12	2

Raw Score	Scale Score	CSEM	Performance Level
19	1319	12	2
20	1323	12	2
21	1326	12	2
22	1330	12	2
23	1333	12	2
24	1337	12	2
25	1340	12	2
26	1343	12	2
27	1347	12	2
28	1350	12	3
29	1353	12	3
30	1357	12	3
31	1360	12	3
32	1363	12	3
33	1367	12	3
34	1370	12	3
35	1374	12	3
36	1378	12	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	1429	17	4
48	1436	17	4
49	1444	19	4
50	1453	20	4
51	1463	22	4
52	1477	25	4
53	1495	31	4
54	1500	43	4
55	1500	60	4

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

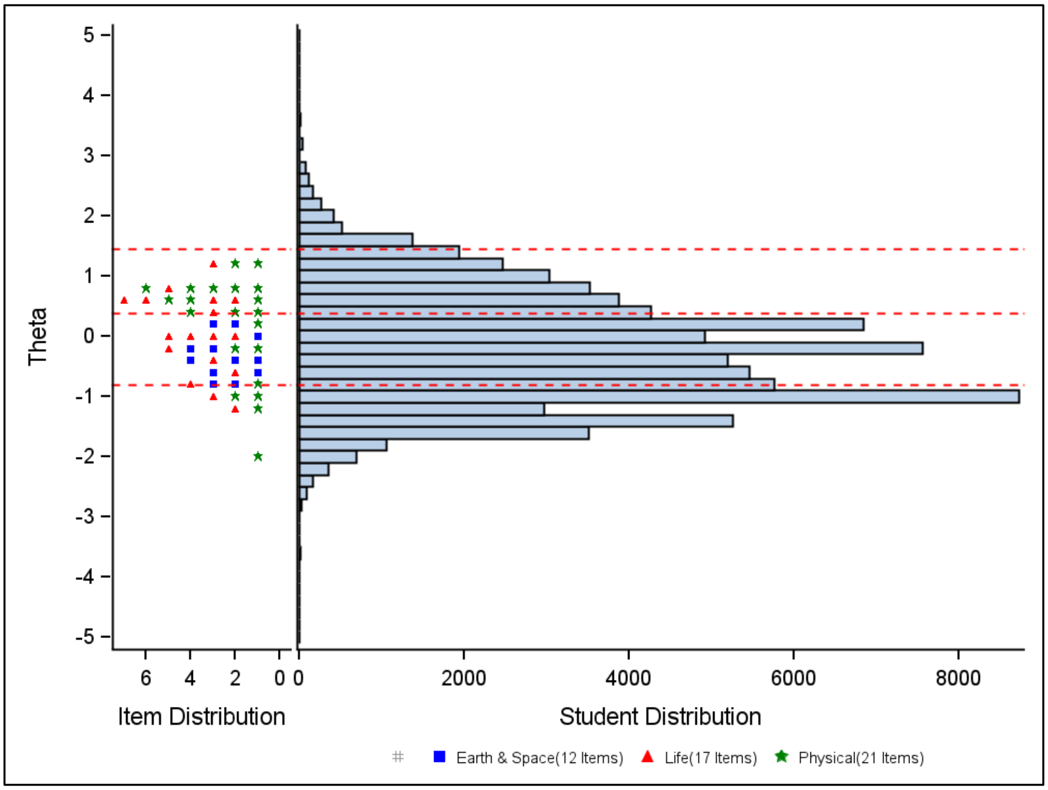


Figure B.2. Item-Person Map, Grade 8

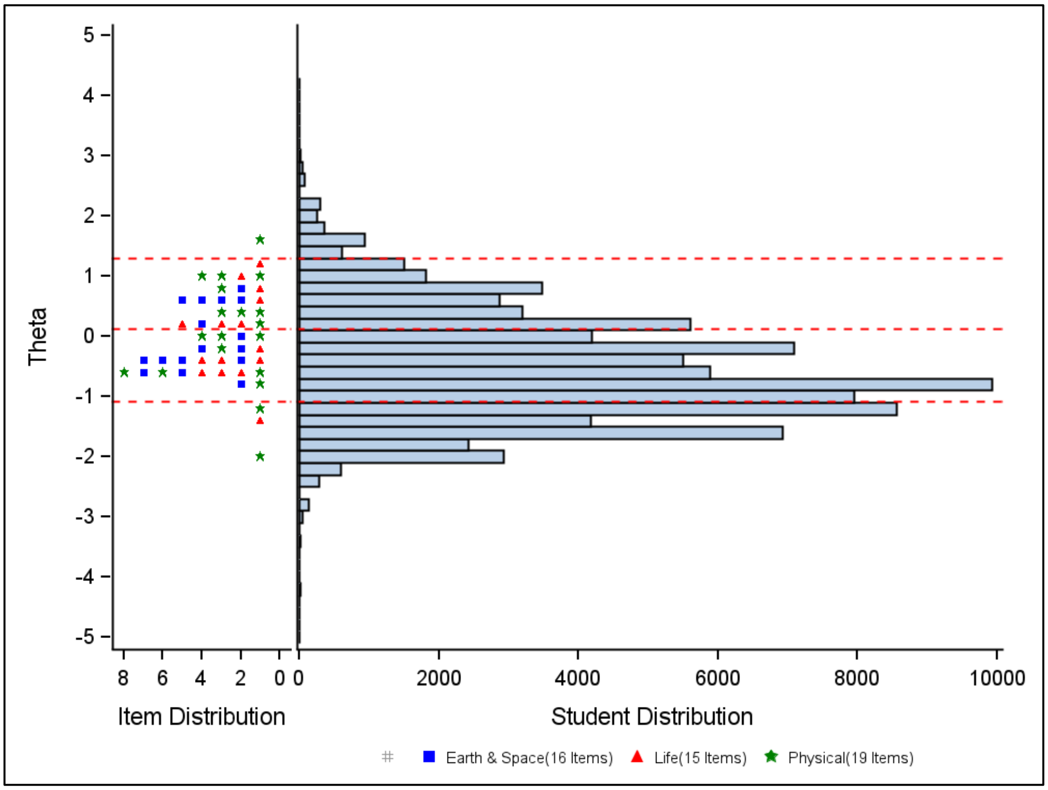


Figure B.3. Item-Person Map, Grade 11

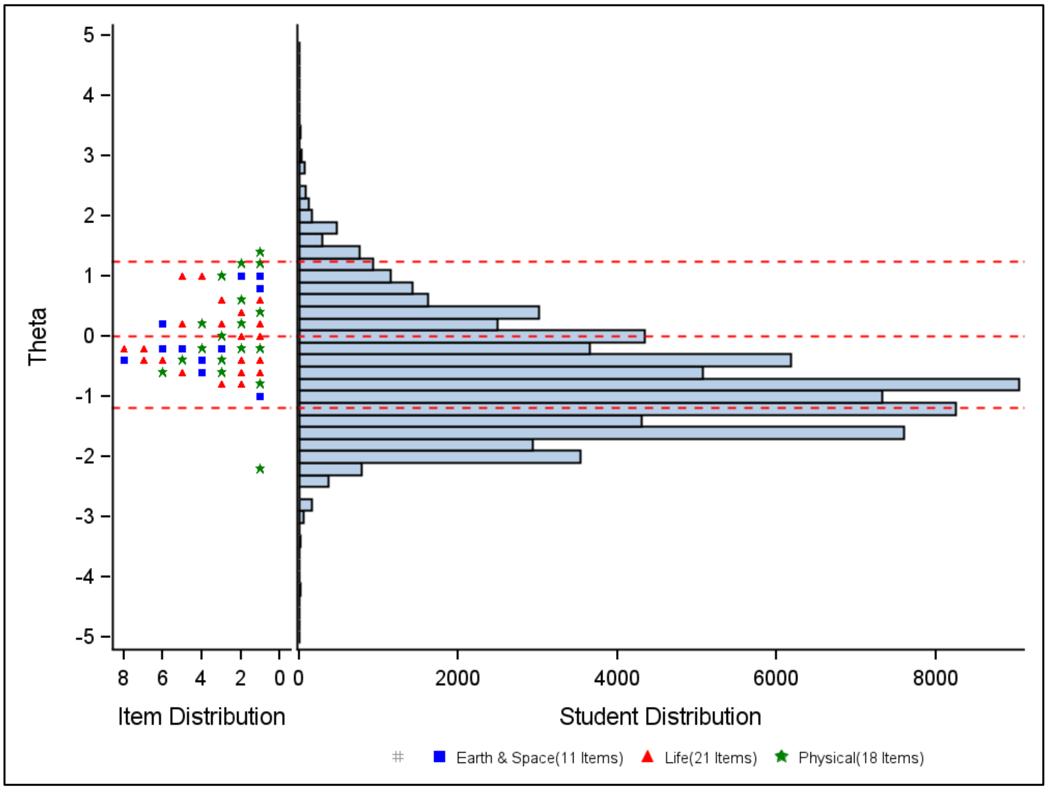


Figure B.4. TCC, Grade 5

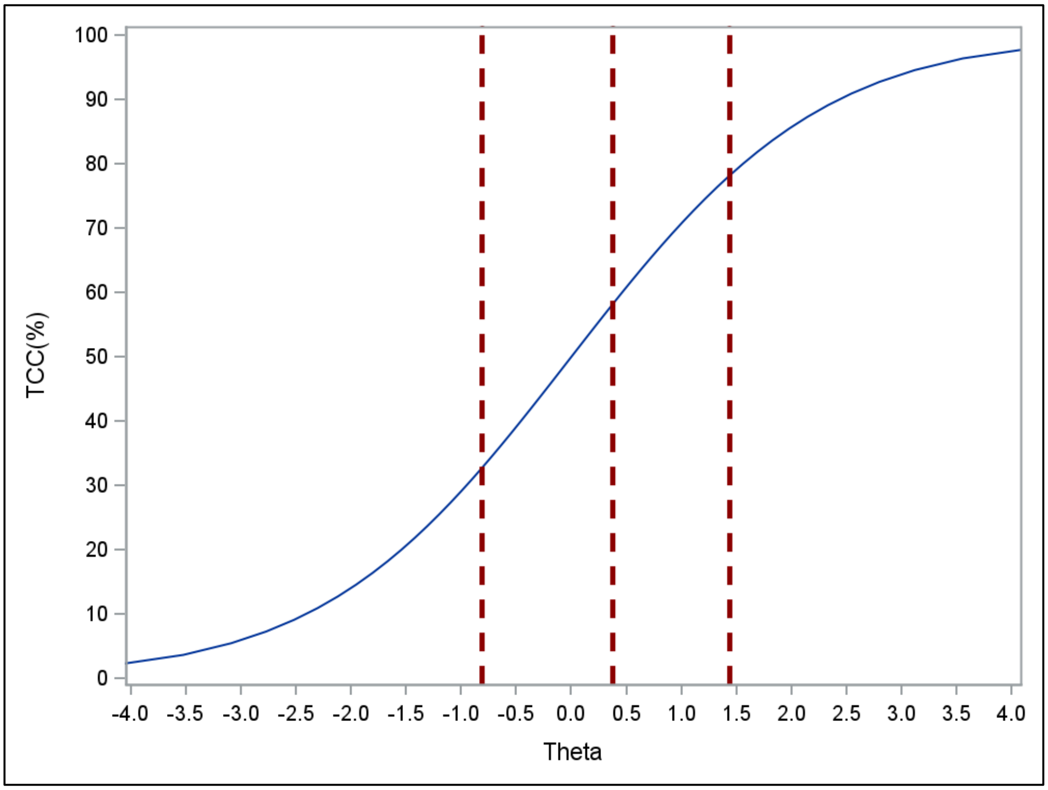


Figure B.5. CSEM, Grade 5

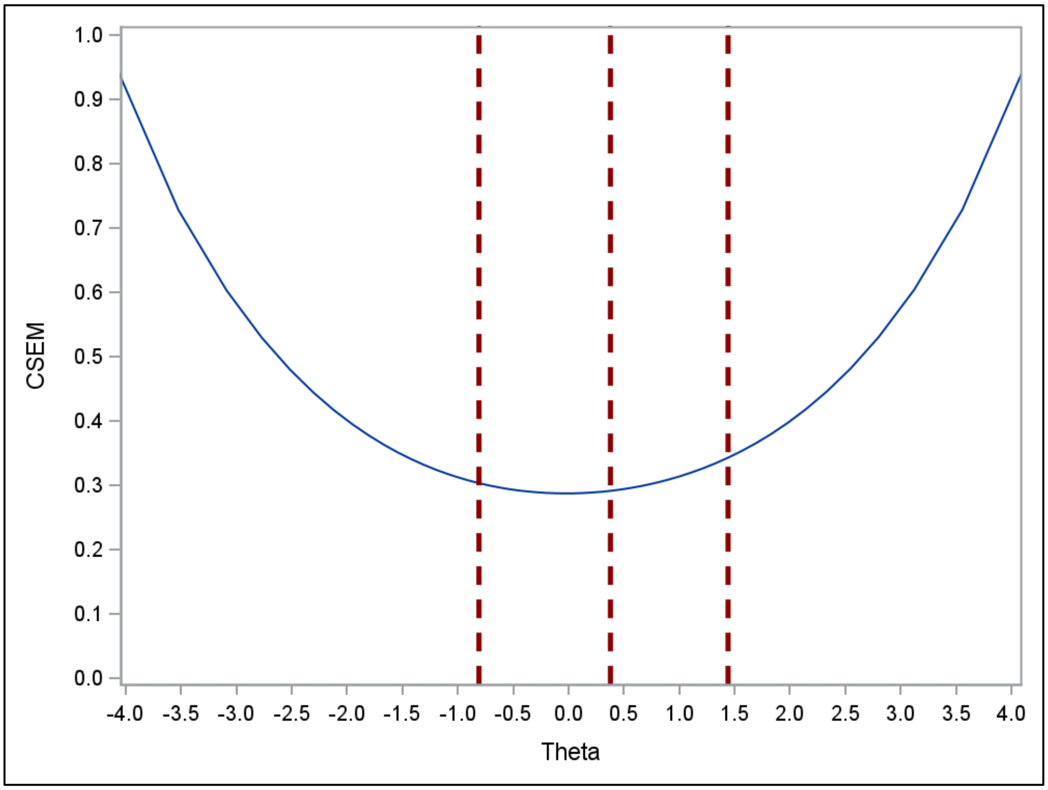


Figure B.6. TCC, Grade 8

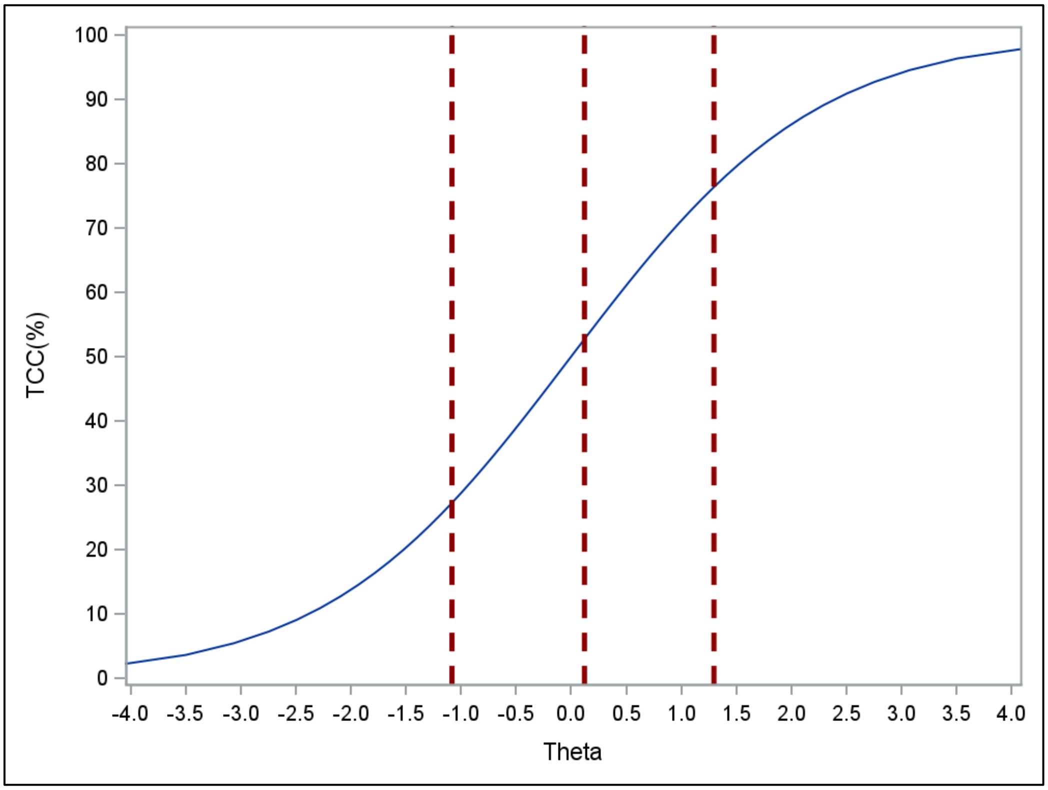


Figure B.7. CSEM, Grade 8

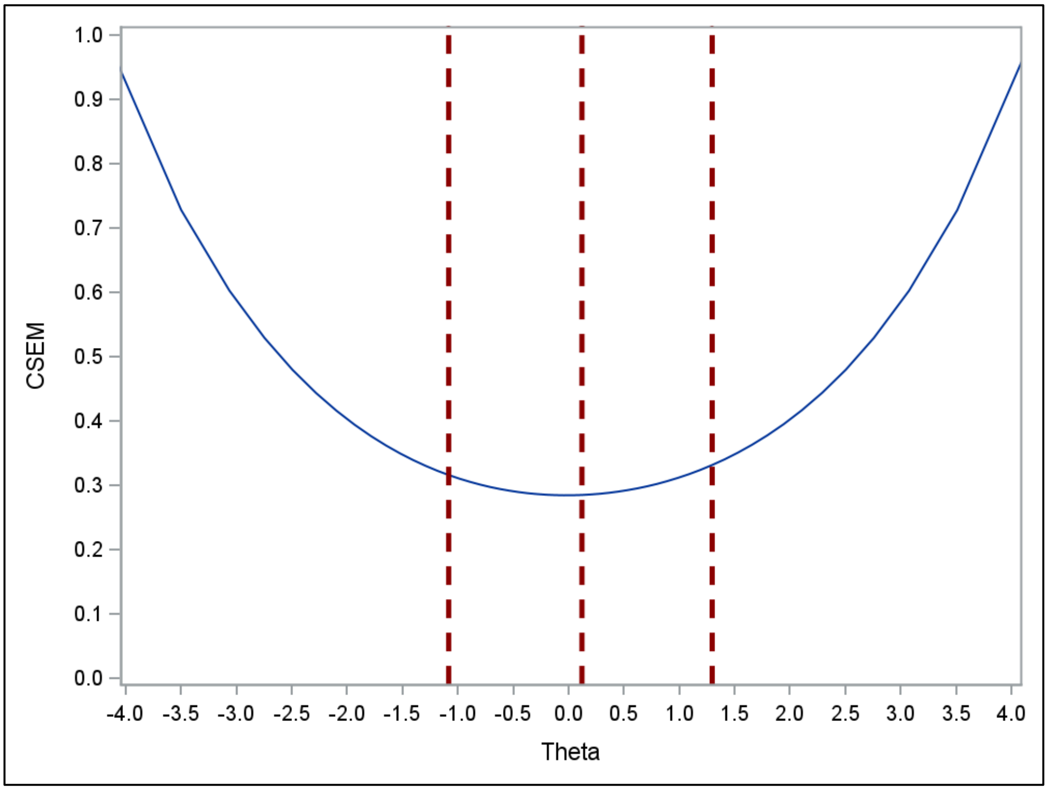


Figure B.8. TCC, Grade 11

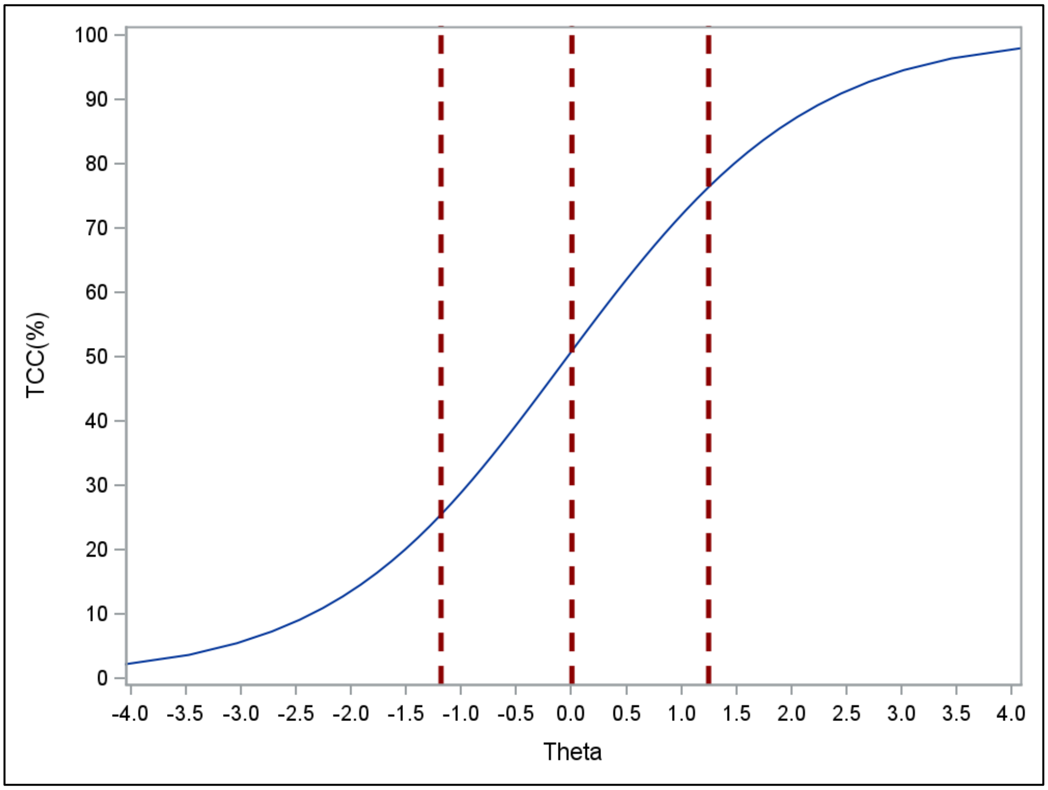


Figure B.9. CSEM, Grade 11

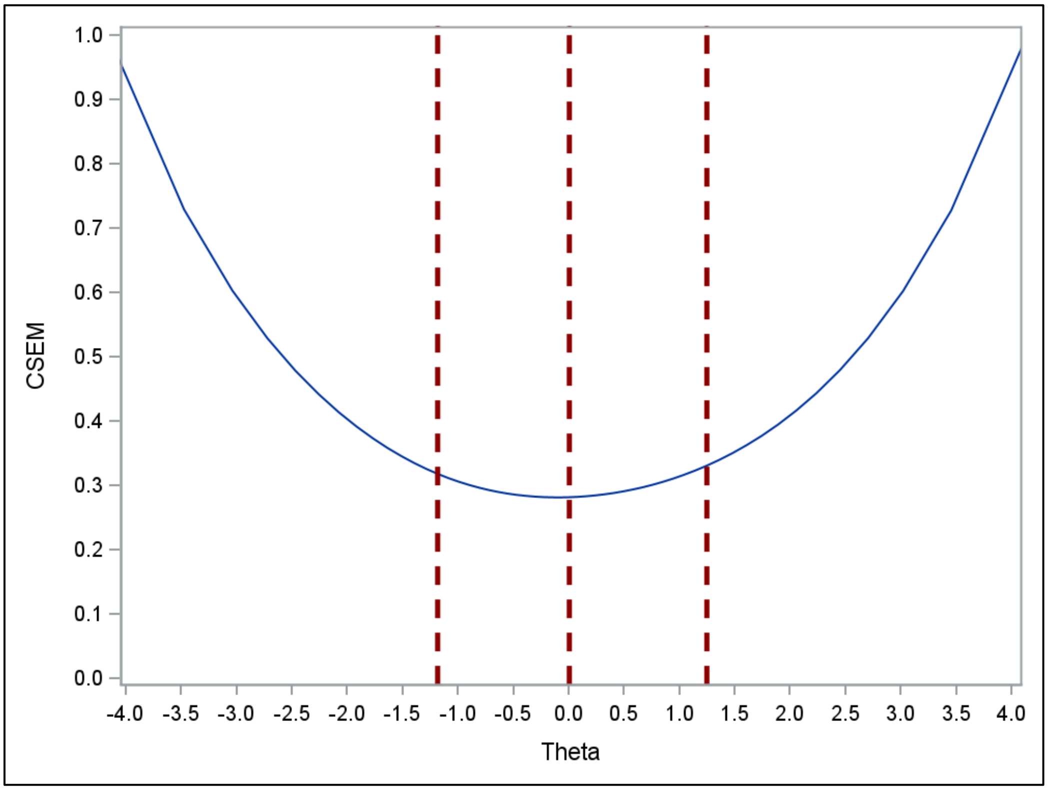


Figure B.10. Scree Plot, Grade 5

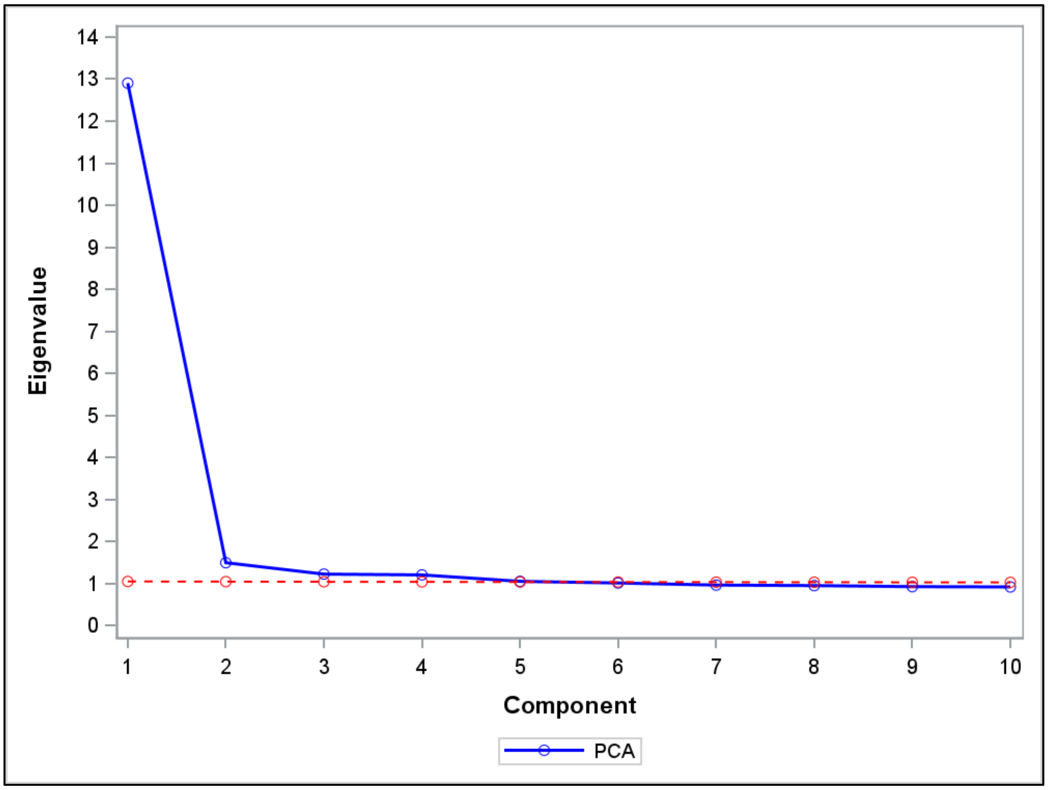


Figure B.11. Scree Plot, Grade 8

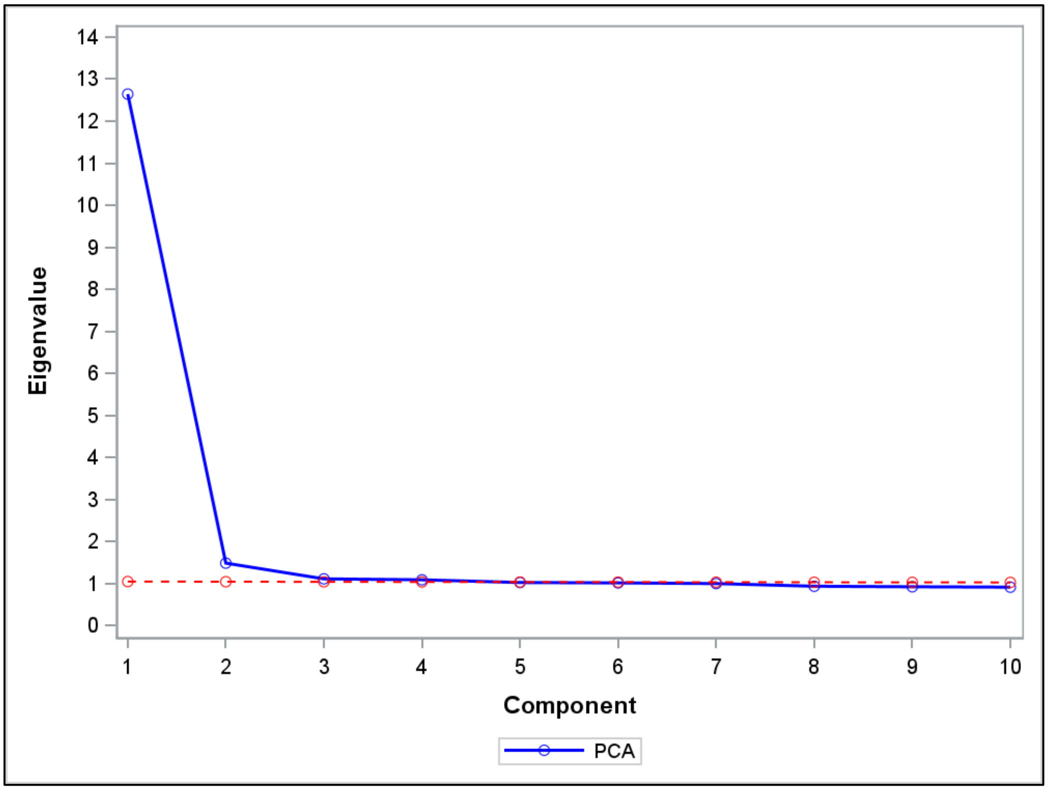
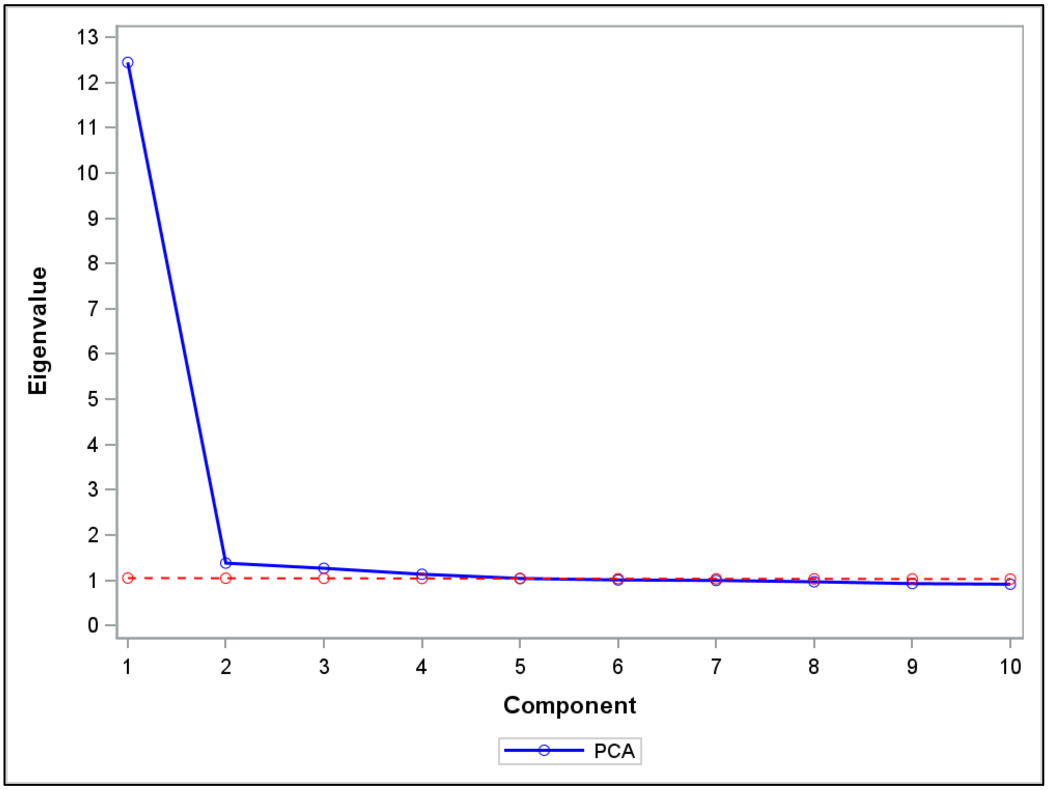


Figure B.12. Scree Plot, Grade 11



Appendix C: ADMINISTRATION RESULTS

This appendix presents the Spring 2022 AzSCI results for all students and subgroups. 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 performance level overall.
- 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	80,889	1324.21	38.91	28	44	23	5
Male	40,806	1325.55	40.19	28	42	24	6
Female	39,983	1322.88	37.51	29	46	21	4
Missing	100	1311.99	33.38	38	48	14	0
Hispanic	37,900	1314.39	35.23	36	46	16	2
Non-Hispanic	42,889	1332.93	39.92	21	43	29	7
American Indian	4,423	1303.74	31.97	49	41	9	1
Asian	2,841	1345.41	39.52	12	41	35	12
Black or African American	5,683	1309.26	34.28	42	45	12	1
Multi-racial	4,758	1329.96	38.83	22	46	26	6
Native Hawaiian or Other Pacific Islander	490	1319.85	38.11	33	46	18	3
White	62,594	1325.67	38.77	27	44	24	5
Missing	100	1311.99	33.38	38	48	14	0
Special Ed.	11,414	1299.32	34.53	57	33	9	1
EL	6,698	1289.65	24.74	69	29	2	0
Low SES	31,427	1311.26	34.47	40	45	14	2

Note. 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	87,698	1322.08	37.69	30	46	20	4
Male	44,554	1322.63	39.10	31	44	21	4
Female	43,072	1321.53	36.16	29	48	20	3
Missing	72	1319.03	38.75	38	38	21	4
Hispanic	39,676	1312.09	33.38	38	47	13	2
Non-Hispanic	47,951	1330.36	39.02	23	45	26	6
American Indian	4,981	1306.47	30.55	44	46	9	1
Asian	2,833	1349.70	41.31	12	36	39	13
Black or African American	5,923	1309.12	32.27	42	46	11	1
Multi-racial	4,854	1326.01	38.00	25	47	23	4
Native Hawaiian or Other Pacific Islander	518	1315.36	33.58	34	49	15	2
White	68,486	1322.99	37.57	29	47	21	4
Missing	103	1312.47	36.43	47	34	17	3
Special Ed.	10,134	1296.01	28.89	62	32	5	1
EL	6,924	1288.51	21.07	71	28	1	0
Low SES	32,200	1310.08	32.63	41	46	12	1

Note. 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	76,418	1319.27	36.99	31	48	17	3
Male	38,270	1320.25	39.64	33	45	18	4
Female	38,075	1318.31	34.12	30	52	16	2
Missing	73	1302.59	25.03	48	47	5	0
Hispanic	33,475	1309.47	31.29	40	49	10	1
Non-Hispanic	42,869	1326.94	39.24	25	48	22	5
American Indian	4,251	1306.91	28.84	41	50	8	1
Asian	2,654	1347.73	44.68	13	41	33	13
Black or African American	4,853	1307.10	30.14	42	48	9	1
Multi-racial	3,671	1323.53	37.40	27	49	20	3
Native Hawaiian or Other Pacific Islander	406	1313.81	35.62	37	48	13	2
White	60,479	1319.66	36.82	31	49	17	3
Missing	104	1308.28	29.28	40	48	12	0
Special Ed.	7,137	1294.88	26.41	62	34	3	1
EL	3,680	1288.08	18.64	72	28	0	0
Low SES	24,511	1308.93	30.75	40	49	10	1

Note. Level 1 = *Minimally Proficient*, Level 2 = *Partially Proficient*, Level 3 = *Proficient*, Level 4 = *Highly Proficient*

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

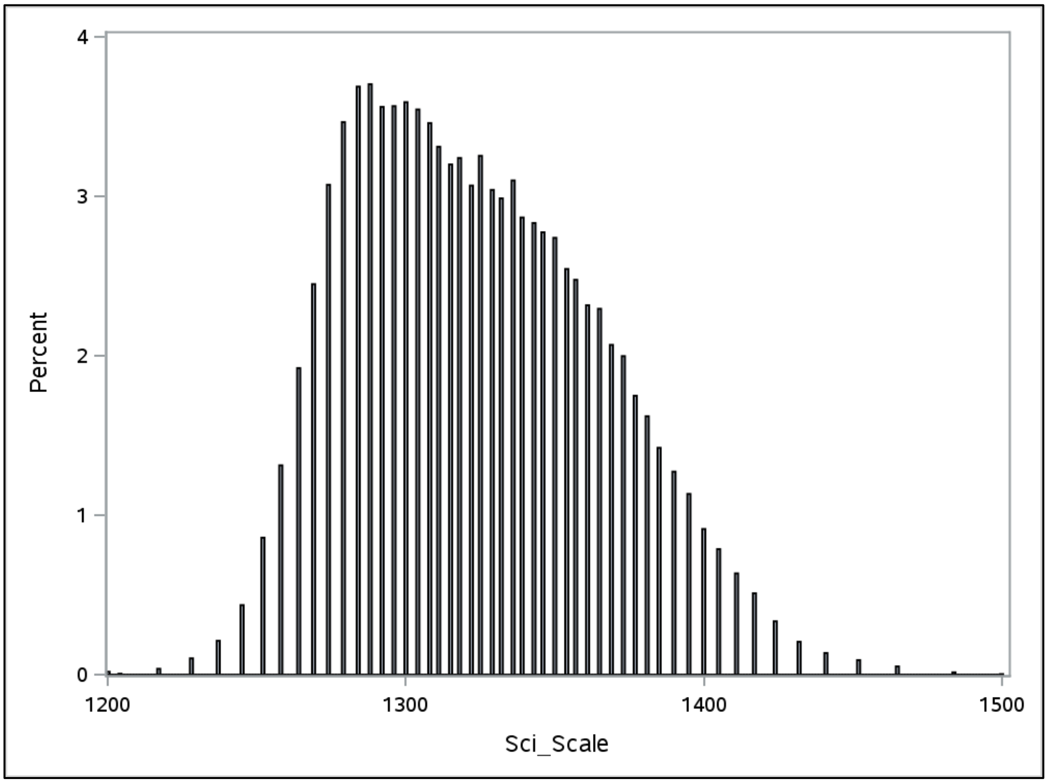


Figure C.2. Total Scale Score Distribution, Grade 8

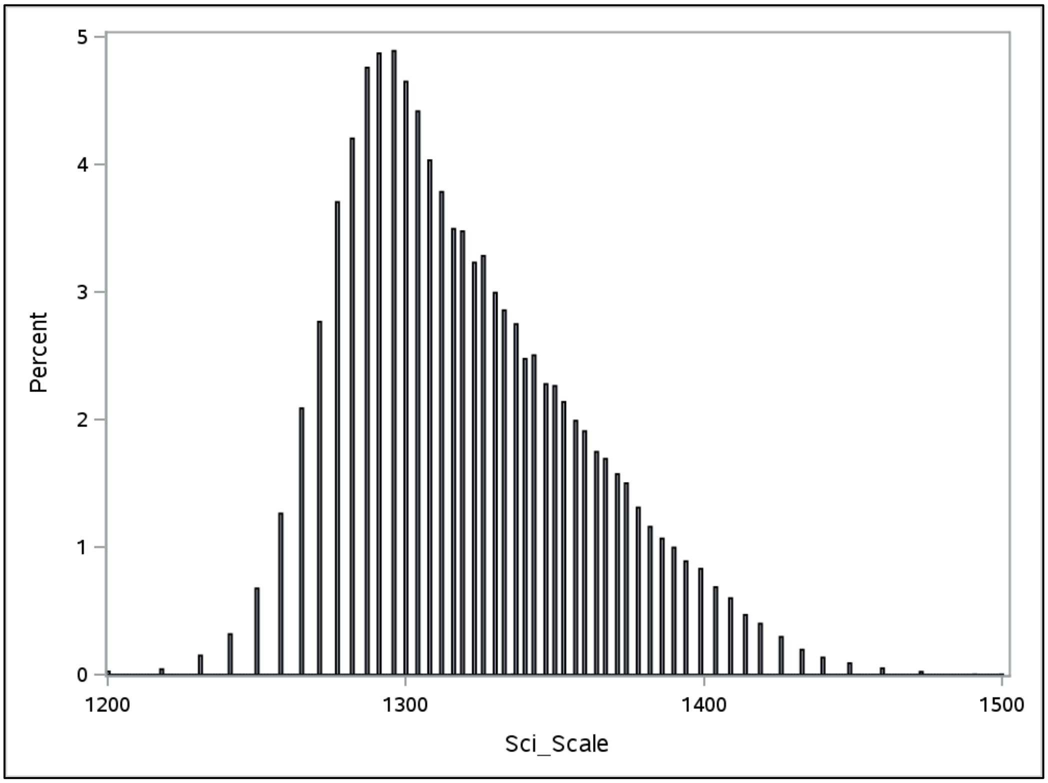
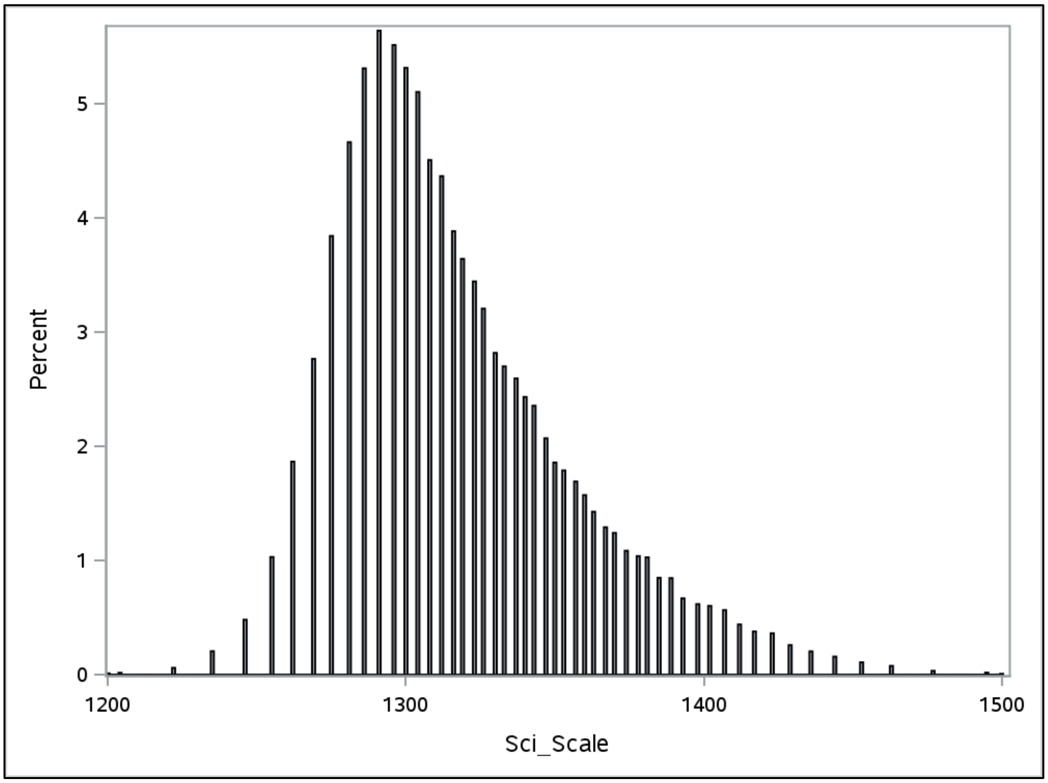


Figure C.3. Total Scale Score Distribution, Grade 11


Percent
Sci_Scale