AASA
Arizona ELA & Mathematics Assessments

2022
Technical Report

Submitted to the
Arizona Department of Education
November 2022

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 Academic Standards Assessment (AASA) 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 Academic Standards Assessment (AASA) is the statewide achievement test for Arizona students in English language arts (ELA) and mathematics in Grades 3–8. It is a criterion-referenced, computer-based assessment designed to measure student progress toward achievement of the Arizona Academic Standards adopted by the State Board of Education in December 2016. All Arizona public school students in Grades 3–8 take the grade-level AASA assessments.

Beginning in 2021–2022, AzM2 was renamed to AASA. The assessment is still aligned to the same 2016 standards and has the same cut scores. A Writing standalone field test (SAFT) was administered in Spring 2022 to all students in Grades 3–8 to build Arizona's item bank for extended writing items. Oral Reading Fluency (ORF) field test items were also embedded on the Grade 3 operational AASA test in Spring 2022 to enhance coverage of the Grade 3 ELA standards, although they will be field tested again in 2023 to further explore their functioning and performance.

1.2. 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 AASA assessments.

Arizona educators participated in meetings and provided feedback on assets developed for field testing in Spring 2022. These meetings were held virtually and included educators from across the state. The committee meetings included an ELA passage review that enabled educators to review ELA passages for content, grade-level appropriateness, and bias and sensitivity; a content and bias item review that enabled educators to review items for content, standard alignment, grade-level appropriateness, and bias and sensitivity; and a bias and sensitivity community review that enabled community members, including past and present Arizona educators, to evaluate items for bias and sensitivity concerns.

Table 1.1. Schedule of Major Events

Event	Date(s)
ELA Passage Review	April 24, 2021
Content and Bias Item Review	July 6–9, 2021
Technical Advisory Committee (TAC)	July 28, 2021
Bias and Sensitivity Community Review	August 10–11, 2021
Administration Training	November 30, 2021 – April 15, 2022
Writing SAFT Additional Order Window for Test Materials	January 14 – February 11, 2022
Writing SAFT Test Administration Window	January 24 – February 18, 2022
Technical Advisory Committee (TAC)	March 9, 2022
AASA Additional Order Window for Test Materials	March 24 – April 5, 2022

Event	Date(s)
Spring 2022 AASA Test Administration Window	April 4–29, 2022
Release of Grade 3 Electronic Score Reports	May 14, 2022
Release of Grades 4–8 Electronic Score Reports	May 25, 2022
Release of Grades 3–8 Paper Reports to Districts	June 15, 2022
Data Review	July 18–22, 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 Academic Standards

In 2016, the State Board of Education adopted new academic content standards in ELA and mathematics that reflect high expectations of all Arizona students and strive to ensure that high school graduates are college- and career-ready. The Arizona Academic Standards define the knowledge, understanding, and skills that need to be taught and learned so all students are ready to succeed in credit-bearing, college-entry courses and/or in the workplace. The ELA standards describe the reading, writing, language, speaking, and listening skills that students should acquire from Grades K–12, and the mathematics standards describe expectations for learning in Grades K–8 and the first three high school courses (Algebra I, Geometry, Algebra II; Mathematics 1, 2, 3), plus specific standards that could be included in a fourth high school credit mathematics course. The standards are located on the Arizona Department of Education (ADE) website at https://www.azed.gov/standards-practices.

The standards work together in a clear progression from Grades K–12. Each standard builds on the standard that came before and toward the standard that comes in the next grade level. They are the foundation to guide the construction and evaluation of programs in Arizona K–12 schools and the broader Arizona community. The Arizona Academic Standards are:

- Focused in coherent progressions across grades K–12
- Aligned with college and workforce expectations
- Inclusive of rigorous content and applications of knowledge through higher-order thinking
- Research and evidence based
- Broad in nature, allowing for the widest possible range of student learning, and
- Designed as an integrated approach to literacy (ELA)

2.2. Item Specifications

Item specifications are available for each grade and content area for the AASA assessments on the ADE website at https://www.azed.gov/assessment/aasa. These item specifications, refined by content experts at Pearson and ADE, strategically guide the item development process. They define the content limit, model tasks, and response types for a specific standard and are used by test development experts to guide the item development process. Item writers use the specifications while developing items to make the best use of the available item types. This document can also assist educators in understanding how the items are developed in alignment with the standards.

The descriptions of blueprints and Depth of Knowledge (DOK) in the item specifications are meant to provide an overview of the test. Item specifications are meant for the purposes of assessment, not instruction. They are not intended to be tools for instruction or the basis for curricula. AASA has a test blueprint that was developed by Arizona and is different from any other state or consortium test blueprint.

The item specifications were developed using a vertical alignment for each standard, wherein the suggested task demands and cognitive complexity of items build upon those of the previous grade level, just as the standards themselves do. The item specifications also provide models for item writers that include item samples that target different DOK and difficulty levels. These item models annotate the information to communicate the intent of the standard and DOK and clarify how to manipulate the item difficulty while keeping the cognitive demands the same for the writer. Detailed item specifications include the following:

- **Content Limits.** This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. For example, in Grade 3, fraction denominators are limited to 2, 3, 4, 6, and 8.
- **Acceptable Response Mechanisms.** This section identifies the various ways in which students may respond to a prompt (e.g., multiple choice, graphic response, equation response, matching, multi-select).
- **Task Demands.** In this section, the standards are broken down into specific task demands aligned to the standard. In addition, each task demand is assigned a common item format relevant to that particular task demand.

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

2.3. Test Blueprint

The test blueprint, in concert with the item specifications, defines the content and structure of the test. Table 2.1 – Table 2.4 present the blueprints based on the 2016 standards for Grades 3–8 in ELA and mathematics. The 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. Inherent in the number of points per test is the relative weighting associated with the standards and, in the case of AASA, the reporting categories being assessed.

Table 2.1. AASA ELA Blueprint, Grades 3–8

	Grades 3–5		Grades 6–8	
Reporting Category	Min.	Max.	Min.	Max.
Reading Standards for Literature	26%	35%	24%	31%
Reading Standards for Informational Text	26%	35%	30%	38%
Reading for Informational Text	26%	22%	30%	25%
Listening Comprehension	0%	13%	0%	13%
Writing and Language	26%	38%	30%	38%
Writing	13%	19%	17%	19%
Language	13%	19%	13%	19%

Table 2.2. AASA Mathematics Blueprint, Grades 3–5

Reporting Category	Grade 3		Grade 4		Grade 5	
	Min.	Max.	Min.	Max.	Min.	Max.
Operations & Algebraic Thinking and Numbers & Operations in Base Ten	49%	53%	46%	54%	38%	42%
Operations & Algebraic Thinking	38%	42%	22%	26%	31%	35%
Numbers in Base Ten	9%	13%	24%	28%	4%	8%
Numbers & Operations – Fractions	18%	22%	29%	33%	31%	35%
Measurement & Data and Geometry	26%	30%	15%	19%	24%	28%
Measurement & Data	26%	28%	9%	13%	18%	20%
Geometry	1%	4%	4%	7%	7%	11%

Table 2.3. AASA Mathematics Blueprint, Grades 6–7

Reporting Category	Grade 6		Grade 7	
	Min.	Max.	Min.	Max.
Ratios & Proportions	19%	23%	19%	23%
The Number System	28%	32%	19%	23%
Expressions & Equations	29%	33%	23%	27%
Geometry and Statistics & Probability	15%	19%	27%	35%
Geometry	6%	15%	15%	19%
Statistics & Probability	6%	11%	12%	16%

Table 2.4. AASA Mathematics Blueprint, Grade 8

Reporting Category	Grade 8	
	Min.	Max.
Functions	21%	25%
Expressions & Equations	29%	33%
Geometry	17%	21%
Statistics & Probability and The Number System	19%	27%
Statistics & Probability	4%	8%
The Number System	15%	19%

2.4. Depth of Knowledge

All items are aligned according to DOK, the cognitive complexity of the item and the cognitive demands on the student. DOK refers to the level of rigor or sophistication of the task in a given item designed to reflect the complexity of the Arizona Academic Standards. Table 2.5 presents a description of the DOK levels as provided in the item specifications documents, and Table 2.6 presents the percentage of points by DOK level as provided in the blueprint documents.

Table 2.5. DOK Levels

DOK Level	ELA	Mathematics
Level 1: Recall	Focuses on basic tasks such as correcting grammatical and spelling errors, defining terms, and locating details or facts in texts.	Focuses on the recall of information, such as definitions, terms, and simple procedures.
Level 2: Skill/Concept	Requires a greater degree of engagement and cognitive processing than DOK 1 items. DOK 2 items may require students to show relationships or identify examples, use context to identify meaning, identify structures or features of texts, or distinguish between facts and opinions.	Requires students to make decisions, solve problems, or recognize patterns. In general, DOK 2 items require a greater degree of engagement and cognitive processing than DOK 1 items.
Level 3: Strategic Thinking	Features higher-order cognitive tasks that assess students' capacities to read complex texts and think abstractly and focuses on critical thinking, developing, and assessing logical arguments, making inferences, and citing evidence to support claims or conclusions.	Features higher-order cognitive tasks that assess students' capacities to approach abstract or complex problems.
Level 4: Extended Thinking (Writing only)	Requires creativity, extensive planning, and/or sophisticated reasoning in the composition and organization of written essays.	N/A

Table 2.6. Percentage of Points by DOK Level

DOK Level	ELA	Mathematics
Level 1	10–20%	10–20%
Level 2	50–60%	60–70%
Level 3	15–25%	12–30%
Level 4	13–19% (Writing)	N/A

2.5. Item Types

The AASA assessments include traditional multiple-choice items and technology-enhanced items (TEI), as shown in Table 2.7. Examples of each item type may be found in the AASA sample tests accessed through TestNav (see Section 0 for more information).

TEIs are computer-delivered items that require students to interact with test content to select, construct, and/or support their responses and are better able to assess a deeper level of understanding. For paper-based assessments (including those for students with an IEP or 504 plan that specifies a paper-based accommodation), TEIs are modified or replaced with another item type that assesses the same standards so they can be scanned and scored electronically or hand scored. For example, gap match/gap match table, match – table grid, and short-constructed response items may be replaced with another item type that assesses the same standard and can be scanned and scored electronically. Inline choice items are modified so the student fills in a circle to indicate the correct word or phrase, and hot text items are modified so the student fills in a circle to indicate a selection.

Table 2.7. Item Types

Item Type	Description
Multiple-Choice (MC)	The student selects only one correct answer from among a number of options.
Multiple-Select (MS)	The student selects all of the correct answers from among a number of options.
Evidence-Based Selected Response (EBSR) (ELA only)	• MC/MS Format: The student answers a Part A multiple-choice item based on a passage and then provides evidence in support of that answer by completing another Part B multiple-choice item or a Part B multi-select item. • MC/TEI Format: The student answers a Part A multiple-choice item based on a passage and then provides evidence in support of that answer by completing a Part B technology-enhanced item.
Bar Graph (mathematics only)	The student drags bars vertically or horizontally along numerical values. Individual bars, histograms, and clusters are supported.
Equation Editor (mathematics only)	The student uses a palette of buttons to enter a numerical response or to create mathematical expressions.
Fraction Model (mathematics only)	The student divides a shape (circle or rectangle) into varying numbers of segments by clicking a 'Fewer' or 'More' button and selects those segments to shade those segments with a solid color.
Point Graph (mathematics only)	The student plots points, line segments, continuous lines, and/or polygons. Point graph items can use one or multiple graph interactions (composite graphs).
Shape Transformation (mathematics only)	The student chooses one of four variants of a single shape, drags it onto a four-quadrant grid, and positions it on the grid.
Inline Choice (IC)	The student selects a single text option from a drop-down menu within a table or inline text, similar to a fill-in-the-blank item. The item may contain multiple blanks.
Gap Match (GM)	Certain numbers, words, phrases, or sentences may be designated “draggable” in this item type. The student can click on the option, hold down the mouse button, and drag it to a graphic or other format.
Gap Match Table (GMT)	Same as the gap match item except the drop zone is in a table format.
Match – Table Grid (MTG)	The student selects radio buttons or checks boxes in cells to indicate if information from a column header matches information from a row.
Hot Text (HT) (ELA only)	The student selects one or more areas called hot spots on an image. For ELA, excerpted sentences from the text are presented in this item type. Certain words, phrases, or sentences are highlighted to indicate that the text is selectable (“hot”). The student can then click on an option to select it.
Hot Spot (mathematics only)	The student selects one or more areas called hot spots on an image. An example for mathematics is selecting a point on a number line. The student can click on an option to select it.
Short Constructed Response (SCR) (ELA only)	The student uses the keyboard to enter a response into a text field. These items can usually be answered in a sentence or two.
Writing Prompt (ELA only)	These items may require the student to use features of an online word processor. The student can perform various tasks within the online word processor such as bold text, use bullet points, underline, etc.

2.6. Test Designs

Table 2.8 and Table 2.9 present the test designs for the ELA and mathematics assessments. As shown in the tables, the AASA test consists of six test units for Grade 3 and five test units for Grades 4–8:

- ELA Oral Reading Fluency (ORF) test unit (Grade 3 only)
- ELA Writing test unit
- ELA Reading/Language Test Unit 1 and Test Unit 2
- Math Test Unit 1 and Test Unit 2

In Spring 2022, writing prompts were also field tested during a standalone field test administration to increase the number of eligible writing prompts in the item bank to be used operationally in future administrations of the ELA Writing test. For each grade level, a total of 6–7 writing prompts were field tested. The prompts field tested were comprised of a balance of informative and opinion/argumentative writing.

Table 2.8. AASA Test Design—ELA

		Overall				#Items by Test Unit											
			#Items			Writing			Reading/Language Test Unit 1			Reading/Language Test Unit 2			Oral Reading Fluency (ORF)		
Grade	#Forms	#Passages	OP	FT	Total	OP	FT	Total	OP	FT	Total	OP	FT	Total	OP	FT	Total
3	21	7	38	10	57	1	–	1	13	7	20	24	–	24	0	3	12
4	21	7	38	7	45	1	–	1	11	7	18	26	–	26	N/A	N/A	N/A
5	21	6	38	7	45	1	–	1	16	7	23	21	–	21	N/A	N/A	N/A
6	21	5	38	7	45	1	–	1	17	7	24	20	–	20	N/A	N/A	N/A
7	21	5	38	7	45	1	–	1	13	7	20	24	–	24	N/A	N/A	N/A
8	21	5	38	13	51	1	–	1	21	13	34	16	–	16	N/A	N/A	N/A

Note. Each writing prompt is worth 10 points. The test design for ELA is based on the number of items, and the total points per operational form vary from 52–56 points. The #Passages are specific to the two Reading test units.

Table 2.9. AASA Test Design—Mathematics

		#Items								
		Overall			Test Unit 1			Test Unit 2		
Grade	#Forms	Total	OP	FT	Total	OP	FT	Total	OP	FT
3	11	50	45	5	26	23	3	24	22	2
4	11	50	45	5	26	23	3	24	22	2
5	11	50	45	5	26	23	3	24	22	2
6	11	52	47	5	25	23	2	27	24	3
7	11	52	47	5	25	23	2	27	24	3
8	11	59	47	12	29	23	6	30	24	6

Note. Each operational item is worth 1 point for 50 points possible for each grade.

2.6.1. ELA

The ELA ORF test unit consists of three short passages that students read aloud to measure oral reading fluency. The ELA test has a Writing part and a Reading Part 1 and Part 2 for all grade levels. Writing consists of one writing prompt, which is an extended text/essay response. The Reading/Language is a long test, so it is split into two units. Each unit includes both reading and language items.

The ELA passages represent a variety of genres and topics. Pearson's content experts develop informational texts from multiple content areas, such as history, science, and technical subjects. Literary texts represent authentic pieces from multiple genres, including stories, poetry, and drama. The ratio of informational to literary texts increases at each grade band, with a greater percentage of informational texts in the upper grades. The AASA uses both single passages and passage sets in which students are asked to synthesize information across texts. The number of items associated with each varies depending on the actual set and what standards are assessed.

The AASA ELA assessment is designed to reflect the importance of using evidence and reading complex texts outlined in the Arizona Academic Standards. It includes extended writing tasks that provide students with meaningful contexts in which to construct their responses. Each writing prompt presents students with various stimuli (at least 2–3 per task) that serve as a springboard for an informed piece of writing. Students are given research articles, charts and graphs, and narratives to serve as the basis for their written responses. Students can then use this information, along with their own reasoning, to formulate an essay that is not only a clear and coherent expression of their own thinking but also grounded in research and evidence.

Each student is administered a single informative/explanatory or opinion/argumentative writing essay. Informative/explanatory writing is focused on conveying information accurately. Informative writing seeks to enlighten the reader about processes or procedures, phenomena, states of affairs, and terminology. To produce this kind of writing, students draw from what they already know and from primary and secondary sources. Students develop a main idea and a primary focus as they relate facts, details, and examples.

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

The reading level of the stimulus does not exceed the easy Lexile range for the grade level to enable the students to attend to the content of the passages and not struggle over unfamiliar language and non-content-related vocabulary. Moreover, this helps ensure that students are assessed on their writing skills and not their reading abilities.

2.6.2. Mathematics

Calculators are not allowed for the mathematics assessments in Grades 3–6. For the Grades 7 and 8 assessments, where calculator usage is allowable for some item types, the items are grouped into two units administered separately to students: calculator and no calculator. The construct of the items dictates in which section they are to be assessed.

Chapter 3: TEST DEVELOPMENT

This chapter addresses Standards 1.11, 3.2, 3.6, 4.0, 4.1, 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.

Items used to develop the Spring 2022 ELA operational test forms were drawn mainly from the item pool of Arizona-owned items that had been custom developed to align to the Arizona Academic Standards, and writing prompts were leased from the previous vendor. Pearson developed 931 ELA items (735 for the embedded field test + 196 practice items) across all grades for the Spring 2022 administration in partnership with experienced item/passage writers. For mathematics, Pearson developed 227 items (105 for the embedded field test + 122 practice items) across all grades for in partnership with an experienced vendor.

The items field tested in Spring 2022 for both ELA and mathematics were custom developed to align to the Arizona Academic Standards. A secondary source was a pool of items developed by ACT to meet blueprint requirements for operational testing, as shown in Table 3.1. For Grade 8, the items developed by ACT in both the operational and field test slots were used to establish a linking relationship between AASA and ACT.[1]

Table 3.1. Number of ACT Items per Form

Content Area	Grade	Total #OP Items	#OP ACT Items	%OP ACT Items	Total #FT Items	#FT ACT Items	%FT ACT Items
ELA	3	38	0	0.0	7	0	0.0
	4	38	5	13.2	7	0	0.0
	5	38	5	13.2	7	0	0.0
	6	38	0	0.0	7	0	0.0
	7	38	0	0.0	7	0	0.0
	8	38	6	15.8	13	6	46.2
Math	3	45	10	22.2	5	0	0.0
	4	45	5	11.1	5	0	0.0
	5	45	8	17.8	5	0	0.0
	6	47	5	10.6	5	0	0.0
	7	47	4	8.5	5	0	0.0
	8	47	16	34.0	12	7	58.3

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.

[1] Information about the linking study will be included in the 2022–2023 AASA technical report.

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 AASA tests based on the following steps:

1. Asset development plan
2. Item development
3. Item review
4. Field test candidate finalization
5. Preparation for item selection
6. Item selection and positioning
7. Sampling plan

3.2.1. Asset Development Plan

Pearson conducted a bank analysis at the start of the Spring 2022 test development cycle to identify gaps that were then used to inform creation of an asset development plan. An initial step was determining which items were ADE-owned, followed by a gap analysis process to determine the priorities for new item development.

For ELA, the gap analysis examined the Arizona-owned items in the bank eligible for operational use. A comparison to the blueprint requirements revealed the standards underrepresented in the bank as the focus for new development. Oral reading fluency was new for field testing in Spring 2022, so those items also required new development. In addition to development for field testing, sample items were developed to allow students to practice all interaction and item types that students could see during the spring operational testing.

For mathematics, Pearson and ADE worked together to identify the need for a new sample test. As a result, most new item development for the Spring 2022 administration focused on sample items designed to allow students to practice all interaction types and item types that would potentially be included on the spring test form. The remaining items developed as field test candidates in Spring 2022 were aimed at filling gaps in the bank.

The number of newly developed items varied by grade and content area depending on the needs of the bank, as shown in Table 3.2. Standards that were underrepresented in the item bank, or were represented by items with poorly performing statistics, were identified as candidates for item development. Blueprint requirements were also used to determine which standards most needed new item development.

Table 3.2. Number of Newly Developed Items

Content Area	Grade	#Items for FT	#Items for Sample Test	Total #Items
ELA	3	140	32	172
	4	119	36	155
	5	105	31	136
	6	124	32	156
	7	144	33	177
	8	103	32	135
	Total	735	196	931
Math	3	10	12	22
	4	11	10	21
	5	15	25	40
	6	15	25	40
	7	18	25	43
	8	36	25	61
	Total	105	122	227

3.2.2. Item Development

Item development for ELA began with the development of reading passages. To ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading, test developers adhere to detailed passage specifications. The passage specifications call for a close examination of both quantitative measures, such as word counts and Lexile readabilities as shown in Table 3.3, and qualitative measures such as passage structure and levels of meaning, all of which are defined as important measures of text complexity. For example, content experts use passage text complexity worksheets based on the passage specifications to analyze each passage in depth, as illustrated in Figure 3.1. Table 3.3 also presents the Lexile measures and word count for passages used in the Grade 3 Oral Reading Fluency (ORF) test.

Table 3.3. Passage Lexile Measures and Word Count

Grade	Lexile Range	Word Count Range	ORF Lexile Range	ORF Word Count
3	420–820	100–700	600–750	100–700
4	740–1010	100–900	–	–
5	740–1010	200–1,000	–	–
6	925–1185	200–1,100	–	–
7	925–1185	300–1,100	–	–
8	925–1185	350–1,200	–	–

Note. ORF = Oral Reading Fluency

Figure 3.1. Text Complexity Worksheet Example

UIN:	**Word Count:**	
Title:		
Genre:	**Sub-Genre:**	
Quantitative Measures	**Flesch-Kincaid:**	**Lexile:**
Qualitative Considerations		
Identify the theme and/or central message and describe how it is adequately developed. *(Theme and central message should be similar or the same across paired texts.)*		
Briefly describe how the characters are adequately developed, including how they respond to an event or how they change throughout the story.		
Describe the overall structure of a text and how it contributes to the development of the theme, setting, or plot.		
Briefly describe additional plot elements *(setting, climax, rising and falling action)* **that demonstrate clear plot development and how they are similar and/or different across multiple texts. (***Paired text only***.)**		
Explain how you, the author, develops the points of view from which each text is narrated.		
Compare/contrast the differences between the texts when considering genre, theme, and topic.		
Identify one higher level words used in the passage(s) and identify its text support for understanding meaning.		
List grade-level appropriate examples of literary devices used throughout the passage *(e.g., metaphor, onomatopoeia, flashback, foreshadowing, voice, irony, symbolism).*		
Identify a phrase from the text that has a figurative or connotative meaning and describe the text support.		
Holistically, this text should be considered: **ACCESSIBLE** **MODERATELY COMPLEX** **HIGHLY COMPLEX** for grade____.		

The next step of item development for ELA and the first step for mathematics was training item writers and introducing them to project requirements. Writers relied on existing item specifications and the Arizona Academic Standards to guide item development. The vendor submitted items in batches and revised as needed based on Pearson feedback. Throughout the writing process, there was open communication between Pearson and the vendor. Queries were addressed in a timely manner to facilitate a deeper understanding of the Arizona standards and ADE expectations.

Throughout all steps, Pearson responded to ADE feedback, revised, and resubmitted for approval as needed. An integral part of this process was a review by Pearson research librarians who verified accuracy of information 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.

3.2.3. Item Review

ADE pre-review was the first of several external reviews of the newly developed passages and items. Educators and community members also had opportunities to participate in review committees. Content and bias review allowed educators to apply their familiarity with Arizona students and the Arizona Academic Standards to provide feedback on the accuracy and appropriateness of the item and stimulus content. A bias and sensitivity community 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 style expectations and that no errors had unintentionally been introduced.

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 of review 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 selected statistics from the Spring 2021 administration, 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 item response theory difficulty (Rasch), item discrimination (point-biserial correlation by total score and by reporting category score), the Rasch model fit indices (infit/outfit), differential item functioning (DIF) flags as a measure of possible bias, coefficient alpha, kappa, 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. 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 a form that met psychometric approval was in place. 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. ADE reviewed the field test selections, and Pearson revised as needed.

3.2.7. Sampling Plan

All grades for ELA had 21 forms, and all grades for mathematics had 11 forms. The operational items were the same on all forms within a grade. 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 3.4. 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 3.4 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 3.5 presents the data review results based on the Spring 2022 data. Accepted items were added to the operational item pool for future use.

Table 3.4. 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 (multi-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 (multi-point items only)

Table 3.5. Data Review Results: Number of Field Tested Items

Content Area	Grade	#Accepted	#Accepted w/Edits	#Rejected
ELA	3	55	0	16
	4	43	0	17
	5	54	0	11
	6	68	0	11
	7	69	0	12
	8	70	0	12
Math	3	31	0	1
	4	24	2	1
	5	27	2	0
	6	32	0	3
	7	37	0	3
	8	35	0	1

3.4. Alignment

The AASA ELA and Mathematics assessments are rigorously examined in accordance with the guidelines provided in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence that is necessary to validate assessment scores for their intended purposes based on these standards. Although the validity of AASA test score interpretations is evaluated along several dimensions as a criterion-referenced system of tests, the meaning of test scores is critically evaluated by the degree to which test content is aligned with the standards.

Alignment of content standards is achieved through a rigorous, iterative test development process that proceeds from the content standards and includes ADE test developers, and educator and stakeholder committees. In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards is covered in each test administration. Thus, the test specification blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably).

Because the test blueprints determined how student achievement of the Arizona Academic Standards was evaluated, alignment of test blueprints with the content standards was critical. Alignment of test forms to the test blueprints is a thoughtful, careful task that involves collaboration among assessment specialists, psychometricians, and ADE.

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

3.5. Special Paper Version

Each grade and content area 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 AASA 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 3–8 participated in the Spring 2022 administration of the AASA test. Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them as eligible for an alternate assessment, Multi-State Alternate Assessment (MSAA), are excluded from the AASA test.

Test administrators were instructed to use the *Test Administration Directions* manual for the online administration of AASA, 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 order materials if needed. ADE reviewed all test forms in TestNav and approved them prior to the administration.

Table 4.1 presents the estimated time to complete each test unit. A test unit must be completed prior to starting the next one. All ELA Writing and Reading test units must be administered to receive an ELA score, and both mathematics test units must be administered to receive a mathematics score. The ELA Writing test must be administered on a separate day than the ELA Reading and mathematics units. ELA Reading and mathematics test units could be administered in any order, with no more than two test units plus the Grade 3 ORF unit in a single day. If two test units were administered on the same day, there must be a significant break between them. 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.

In Winter 2022 (Jan. 24 – Feb. 18, 2022), all students in Grades 3–8 took the Writing SAFT prior to the operational test administration. As part of the operational test administration, Grade 3 students also participated in the new ORF test unit that was field tested in Spring 2022; thus, the items were not included in scoring. Each student read three separate passages, with a time limit of one minute per passage. The ORF test unit was to be administered in small groups, with no more than six students testing simultaneously in a classroom or a computer lab environment.

Table 4.1. Estimated Testing Time by Test Unit

Unit	Testing Time
ELA Writing	60–90 minutes
ELA Writing SAFT	60–90 minutes
ELA Reading Test Unit 1	45–75 minutes
ELA Reading Test Unit 2	45–75 minutes
Grade 3 Oral Reading Fluency (ORF)	15 minutes
Math Test Unit 1	60–85 minutes
Math Test Unit 2	60–85 minutes

Note. The testing time is the same for the computer-based and paper-based administrations.

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 sent automatically to all participating schools, and a limited number were available for additional order 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 4.2 that were required for District Test Coordinators, School Test Coordinators, Test Administrators, and other school staff involved in testing or test results.

Table 4.2. Administration Trainings

Training	Description
AASA Test Administration	This training covered the Spring 2022 AASA test administration for Grades 3–8, 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 AASA 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 Sample Tests reflect the AASA test specifications and blueprints and had 1–25 items on each test, as shown in Table 4.3. The Sample Tests do not include an item for each of the aligned Arizona Academic 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 and content area. Every eligible item type was represented. An accompanying scoring guide identified standard and DOK alignment. The portal and scoring guides are both available on ADE website at https://www.azed.gov/assessment/aasa.

Table 4.3. Number of Items on the AASA Sample Tests

Grade	ELA	Writing	ORF	Mathematics
3	24	1	3	25
4	24	1	–	25
5	24	1	–	25
6	24	1	–	25
7	24	1	–	25
8	24	1	–	25

4.4. Accommodations

Accommodations are specific practices and procedures that provide students with equitable access during instruction and assessment. Accommodations are made to provide a student equal access to learning and equal opportunity to demonstrate what is known. They 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 4.4 presents the accommodations available to students while testing on AASA.

Table 4.4. AASA Available Accommodations

Available Accommodation	Description
Abacus	Students may use an abacus without restrictions for any mathematics test or a talking calculator for students taking Part 1 of the Grades 7 or 8 mathematics test.
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.

Available Accommodation	Description
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 4.5 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 4.5. Frequency of Accommodations Used

Content Area	Accommodation	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
ELA	Adult Transcription	24	21	17	18	17	14
	Assistive Technology	7	7	11	19	25	14
	Sign Test Content	19	22	18	38	22	27
	Simplified Directions	67	74	47	43	38	25
Math	Adult Transcription	15	21	9	20	4	6
	Assistive Technology	4	4	1	10	4	3
	Sign Test Content	19	22	17	36	20	26
	Simplified Directions	72	83	55	38	35	26

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 4.6 are available to all students taking the AASA assessment and cannot be disabled.

Table 4.6. 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.
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, test 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 test administration
- Allow students access to test content prior to testing
- Provide any reference sheets to students during the mathematics test administration or graphic organizers during the Writing test administration
- Allow students to share information during test administration
- Read any parts of the test to students, except as indicated in TAD or as part of an approved accommodation
- Influence students' responses by making any kind of gestures (e.g., pointing to items, holding up fingers to signify item numbers or answer options) while students are taking the test
- Instruct students to go back and reread/redo responses after they have finished their test since this instruction may only be given before the students take the test
- Review students' responses
- Change students' answer choices
- Read or review students' scratch paper
- Participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures

Figure 4.1. Test Security Agreement

Achievement Tests (AASA, AzSCI, ACT Aspire, and ACT)
School Year 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 Security Agreement.

Signed By: _____ Date: _____

Printed Name: _____

Title: _____ School: _____

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

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

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

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

Chapter 5: SCORING AND REPORTING

This chapter describes the procedures used by the Pearson Performance Scoring Center (PSC) to score the AASA writing, reading, and mathematics open-ended items. It also describes procedures used by Pearson's automated scoring team for scoring of the writing prompts. This section addresses Standards 2.7, 4.18, 4.19, 4.20, 6.8, and 6.9 (AERA et al., 2014) regarding the scoring of the assessments.

The AASA assessments were scored with maximum likelihood estimation (MLE) scoring, with an attemptedness rule that a student needed to answer one item in each operational unit. Both ELA and mathematics have their own scale score ranges. Students received a scale score in each content area, 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 is reported as one of three levels of mastery: *Below Mastery*, *At/Near Mastery*, or *Above Mastery*. Students who score *Below Mastery* demonstrate performance in the reporting category that 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.

5.1. Human Scoring of Open-Ended Items

The AASA assessments contain open-ended items that prompt students to write a short answer or extended response (i.e., a paragraph) that require scoring by professionally trained scorers. These items were the writing prompts on the ELA Writing test (both the operational and SAFT prompts in Spring 2022) and the paper-equivalent of the technology-enhanced (TE) items on the ELA Reading and mathematics assessments. Writing was scored via a distributed scoring model (i.e., scorers were trained in a self-paced model), whereas Reading and mathematics were scored using a synchronous model (i.e., scorers were trained by instructors). Human scoring was conducted in Pearson's scoring platform known as the Electronic Performance Evaluation Network (ePEN2).

5.1.1. Scorer Recruitment

Scorers are recruited by the Pearson Human Resources department, with scorers who have extensive experience scoring this type of rubric on previous projects being given first priority. Scorers receive performance ratings based on internal quality metrics of inter-rater reliability and validity. Those who have achieved a high performance rating on previous writing, reading, and mathematics responses are recruited for the AASA assessment. Upon being hired, scorers sign a confidentiality agreement in which they pledge to keep all information and student responses confidential.

Scoring supervisors are chosen based on demonstrated expertise in all facets of the scoring process, including strong organizational abilities and training, practical skills, leadership abilities, and sensitivity to interpersonal communication requirements. Supervisors also possess the essential capability of helping scorers understand the scoring requirements of the AASA. Supervisors perform a key role in that they provide continuous feedback to the scorers through the validity and calibration process, and they monitor the quality of their assigned scorers. All scoring, including the scorers and supervisors, for each content area is supervised by a content specialist who is responsible for training and leading the entirety of the project.

5.1.2. Training

Scorers and scoring supervisors were trained to learn the rubric and score responses according to the AASA scoring guidelines. At the beginning of the scoring project, all scoring supervisors and scorers completed project-specific training consisting of a review of the rubric and prompts for the items being scored and a review of the anchor responses selected and approved by ADE for each prompt.

Prior to scorer training, the scoring directors conducted supervisor training to ensure that the supervisors clearly understood the scoring rubrics and anchors. Scoring supervisors were then required to take one set of practice papers and two sets of qualification papers once they completed the item-specific modules. Supervisors must have passed one of the two qualification sets for the items they were assigned before they could score on the project based on the criteria in Table 5.1. Their scores were compared to the "true score" approved by ADE for each training response. These qualification standards were for the ELA writing prompts only. Because Reading and mathematics were all new and only 0,1 score point items, qualification was not created or required to score those open-ended items.

Table 5.1. Scoring Qualification Standards

Reporting Category	Score Points	Qualification %:Perfect/Adjacent Agreement	#Sets
Writing	1–4	70/90	1 of 2

Training for the ELA Writing prompts differed than the training for the Reading and mathematics open-ended items. Writing established training materials that could be inserted into modules for self-paced training, whereas training materials for Reading and mathematics were all new in Spring 2022 and needed to be created as the students completed testing. This could be accomplished because the Reading and mathematics open-ended items were only 0,1 score point items.

5.1.2.1. Writing

The training for ELA Writing was conducted in a live session using online modules designed to take scorers through the background of the assessment and the rubric and anchor sets for each item. A module is an online set of training materials that can be delivered to scorers individually at their own pace. These modules are embedded into the ePEN2 system and are set up so as not to allow scorers to advance in their training until all proceeding modules are complete and correct.

After the live training, scorers were required to take two sets of practice papers and two sets of qualification papers once they completed the item-specific modules. The scorer must have passed one of the two qualification sets based on the qualification standards in Table 5.1 to score the item or items associated with that module. Once the scorer completed the item-specific training and had qualified, they were allowed to score live responses for that item or set of items. Different scoring rubrics are used for the different item types on the AASA and are posted on the ADE website at https://www.azed.gov/assessment/aasa.

5.1.2.2. Reading and Mathematics

Prior to scorer training for ELA Reading and mathematics, scoring directors created anchor, bridge, and practice sets by selecting exemplars for training from actual student responses. The sets were shared with ADE and adjusted as needed for final approval. Anchor and practice sets were used for "prototype" items (as initial training to an item type), and shorter bridge sets were used for subsequent related items. During training development, the scoring directors and ADE established a process to introduce additional exemplars, as needed in scoring, to illustrate scoring decisions and to calibrate scorers when scoring teams encountered response approaches not covered in the anchor and bridge sets.

Training was conducted in the train-score-train-score model where scoring directors trained both supervisors and scorers on the content for a single item and worked with the team to score that item before moving to train the second item. There were two separate ELA teams and two separate mathematics team, each led by a scoring director and supported by a supervisor. Scoring directors conducted training live via online conferencing. After introducing the project, a scoring director began content training on a prototype item, covering the prompt, rubric, and the anchor set for the item. The team then took and discussed a practice set to test their knowledge of rubric application before moving into live scoring. Subsequent similar items were trained with bridge sets. For such items, the scoring director would prepare the team by covering the prompt, rubric, and bridge set.

5.1.3. Quality Control

A variety of reports are produced throughout the scoring process to allow scoring supervisory staff to monitor the progress of the project, the reliability of scores assigned, and individual scorers' work. Those reports include:

- Daily and Cumulative Interrater Reliability Reports by item and scorer. These reports provide information about how many times scorers were in exact agreement or assigned adjacent scores. The reliability is computed and is monitored daily and cumulatively for the project.
- Daily and Cumulative Validity Reports by item and scorer. These reports provide information about how many times scorers were in exact agreement or assigned adjacent scores to responses that are deemed True Scores. The validity is computed and is monitored daily and cumulatively for the project.
- Daily and Cumulative Frequency Distributions. These reports show how many times each score point has been assigned to the item being scored by readers. The frequency distributions are produced both on a daily basis and cumulatively for the entire scoring project. This report allows scoring supervisors and scoring directors to see whether scorers have a tendency to score consistently high or low.

The most immediate method of monitoring a scorer's performance is through backreading by scoring supervisors. If a scoring supervisor discovers that a scorer is consistently assigning scores other than those the scoring supervisor would assign, they can send a message to that scorer using the backreading function and through the instant messaging system in the ePEN2 scoring system.

With the help of the individual scorer reliability metrics and through backreading, the scoring staff can closely monitor each scorer's performance. Scorers are also monitored using the scorer exception process for validity and scoring rate. A scorer must meet and maintain the quality metrics established for AASA in the designated area to continue scoring the project. If a scorer fails to maintain the established validity perfect agreement and perfect plus adjacent agreement percentage, they will receive a targeted calibration set consisting of 10 anchor-type responses similar to a qualification set. If the scorer fails to pass the calibration set, they will be locked out of scoring and dismissed from the project. Scorer exception can also be set for scoring rate; they may receive up to three warnings before being locked out of the ePEN2 system. The scoring staff will then determine if the scorer will be unlocked and allowed to continue scoring based on how they are performing according to inter-rater reliability and validity statistics.

Scorers who have low inter-rater reliability or a lower- or higher-than-desired scoring rate are closely monitored in backreading and through reports. If, in the opinion of the scoring director and content specialist, these scorers are still performing below acceptable standards after receiving sufficient feedback and being given every reasonable opportunity to improve, they are manually locked out of the system and dismissed from the project.

5.1.4. Security

To ensure that test security is never compromised, the following safeguards are employed:

- All scorers must reside outside of Arizona.
- Scorers and scoring staff personnel must sign a non-disclosure and confidentiality form in which they agree not to use or divulge any information concerning the tests.
- Any and all contact with the press is handled through ADE.
- ePEN2 is accessed via a secure website with login credentials required for each user. Only Pearson supervisory staff can issue user IDs to scorers to access.

5.2. Automated Scoring for ELA Writing Prompts

Pearson's automated scoring engine, the Intelligent Essay Assessor (IEA), is the default option for scoring the AASA ELA writing prompts. Human scoring was applied to responses that were scored while IEA was being trained, c IEA needs to be trained anytime a new prompt is introduced. All the ELA prompts were scored at least in part by IEA in the spring.

For 10% of responses, a second reliability score was assigned to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. All reliability scoring was done by human scorers. This section describes the following concepts related to AASA automated writing scoring:

- Continuous flow

- Calibration of IEA using operational data
- Smart routing
- Confidence level
- Quality criteria for evaluating automated scoring
- Hierarchy of assigned scores for reporting
- Sampling responses used for training IEA
- Criteria for evaluating IEA performance

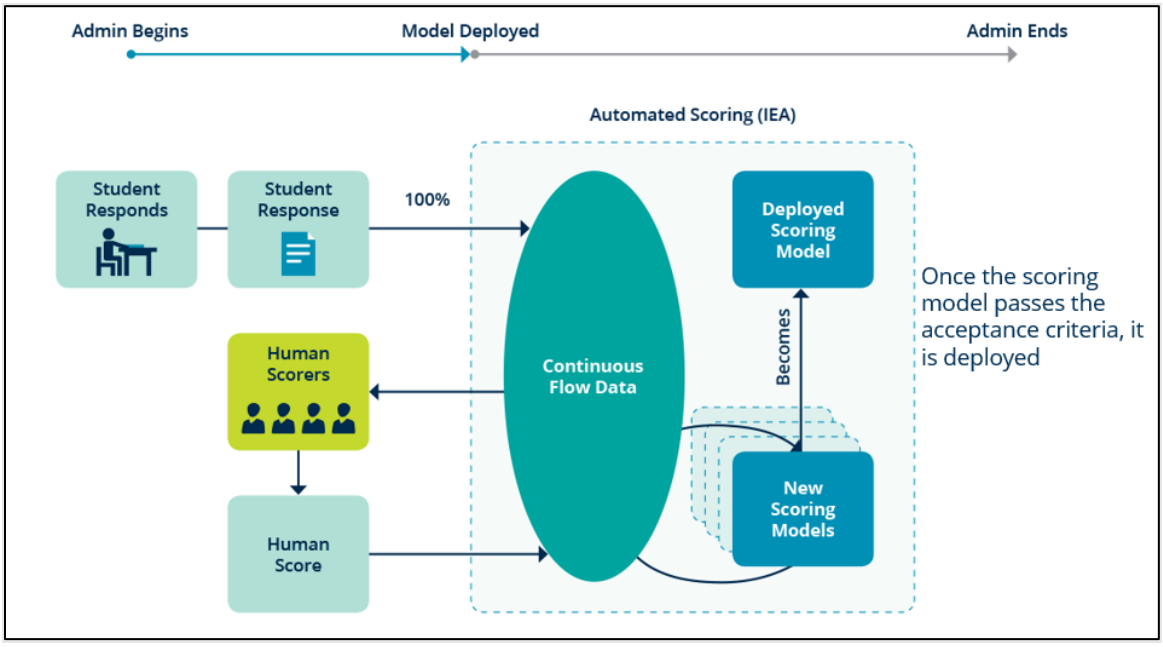
5.2.1. Continuous Flow

The Continuous Flow scoring solution is designed to optimize the quality and efficiency of scoring by flowing responses between human and automated scoring in real time. Responses can be scored by humans or IEA as appropriate based on whether the engine has been trained and is ready to score, the desired double scored percentage, and the confidence of the engine in scoring a particular response.

5.2.2. Calibration of IEA using Operational Data

With Continuous Flow, human scorers begin the scoring process and IEA learns from them. During scoring, student responses and corresponding human scores route directly to IEA. As the human-scored responses flow to IEA, the engine automatically builds potential scoring models, evaluating them against the criteria described in Section 5.2.8. IEA continuously analyzes and incorporates additional human scores until it creates an acceptable scoring model for a prompt. Once IEA obtains an acceptable scoring model, it can be "turned on" and becomes the primary source of scoring (although human scoring continues for the 10% reliability sample and other responses that may be routed accordingly). Figure 5.1 presents scoring model development and deployment in the continuous flow scoring approach.

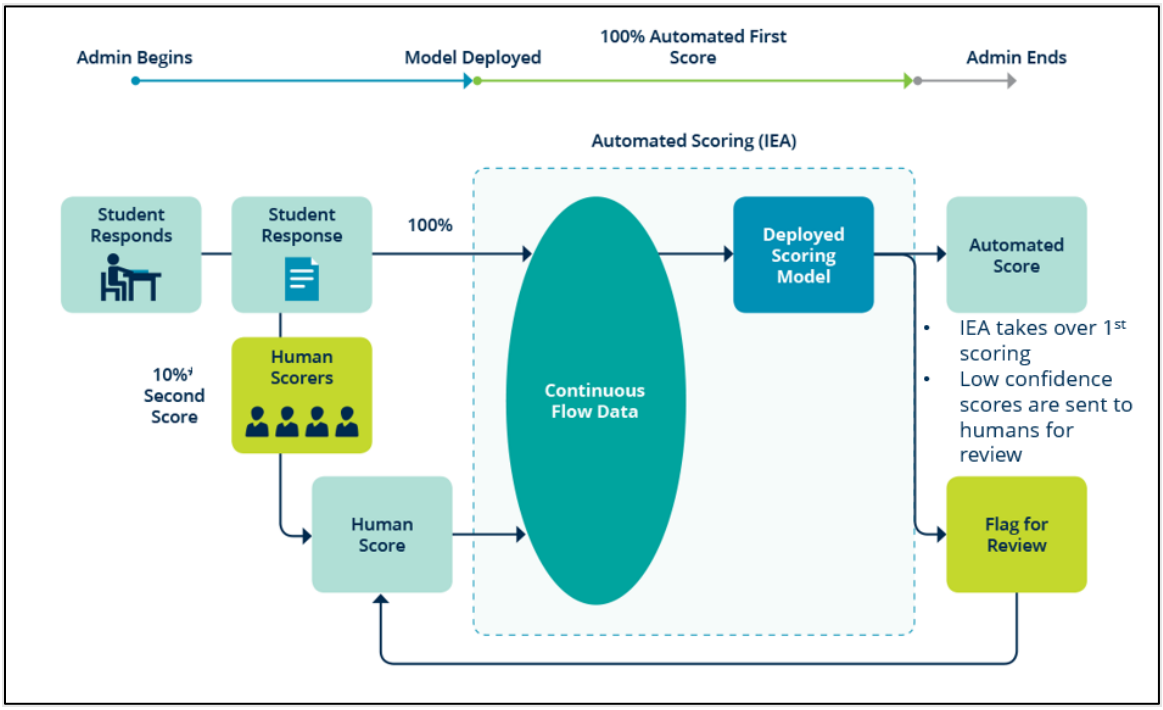
Figure 5.1. Dynamic Model Development and Deployment



5.2.3. Smart Routing

As illustrated in Figure 5.2, once IEA is trained, it takes over first scoring with human scorers providing the 10% second score for reliability. Smart routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores on those responses. Smart routing can be applied prompt by prompt to the extent needed to meet scoring quality criteria for automated scoring.

Figure 5.2. Smart Routing



5.2.4. Confidence Level

When the engine is less confident in scoring a response, the response is marked with a low confidence flag which automatically routes it for human scorers.

5.2.5. Quality Criteria for Evaluating Automated Scoring

The following industry-standard measures are computed between pairs of human scores, as well as between IEA and humans, to evaluate scoring performance: Pearson correlation, quadratic-weighted kappa, exact agreement, and standardized mean difference. Criteria for evaluating the training of IEA given these measures include the following:

- Pearson correlation between IEA-human should be at least 0.70 and within 0.1 of human-human.
- Quadratic-weighted kappa between IEA-human should be at least 0.70 and within 0.1 of human-human.
- Exact agreement between IEA-human should meet inter-rater reliability requirements (65%) and be within 5.25% of human-human.
- Standardized mean difference between IEA-human should be less than |0.15|.

The primary criterion for evaluating IEA was as follows: With smart routing applied as needed, IEA-human exact agreement is at least 65% and within 5.25% of human-human exact agreement.

5.2.6. Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the IEA score is reported operationally if it is a high confidence score. If the IEA score is low confidence, the human score is assigned.

5.2.7. Sampling Responses Used for Training IEA

The early performance of human scoring was closely monitored to verify that an appropriate set of data was available for training IEA. Several characteristics of the human scoring data were monitored, including:

- Exact agreement between human scorers (the goal was for this to be at least 65%)
- Exact agreement between human scores conditioned on score point (the goal was for this to be at least 50%)
- The number of responses at each score point
- The number of responses with two human scores assigned (IEA via Continuous Flow "ordered" additional scoring of responses during the sampling period as needed)

Although the desired characteristics of the training data were easily achieved for some prompts, they were more challenging to achieve for others. For some prompts, a subset of scores were reset and clarifying directions were provided to scorers to improve human-human agreement. A healthy percentage of responses were also backread during the sampling period. These scores in addition to the double human scores were all part of the data used to train IEA. See Section 9.1.2 for information on inter-rater reliability and the agreement rates.

5.2.8. Criteria for Evaluating IEA Performance

IEA performance on the writing prompts was evaluated based on IEA-human exact agreement and compared to agreement based on responses that were double-scored by humans. A portion of the data was held out for evaluating IEA-human exact agreement according to the following steps:

1. Determine exact agreement of the two human scores with each other.
2. Calculate agreement of the IEA scores with the human scores.
3. Compare the IEA-human agreement with the human-human agreement.
4. If the IEA-human agreement is within 5.25% of the human-human agreement, IEA can be deployed operationally.

In addition to the overall comparison, the following performance thresholds were targeted in the test data set: (1) at least 65% overall IEA-human agreement and (2) 50% IEA-human agreement by score point (i.e., conditioned on the human score).

5.3. Reporting

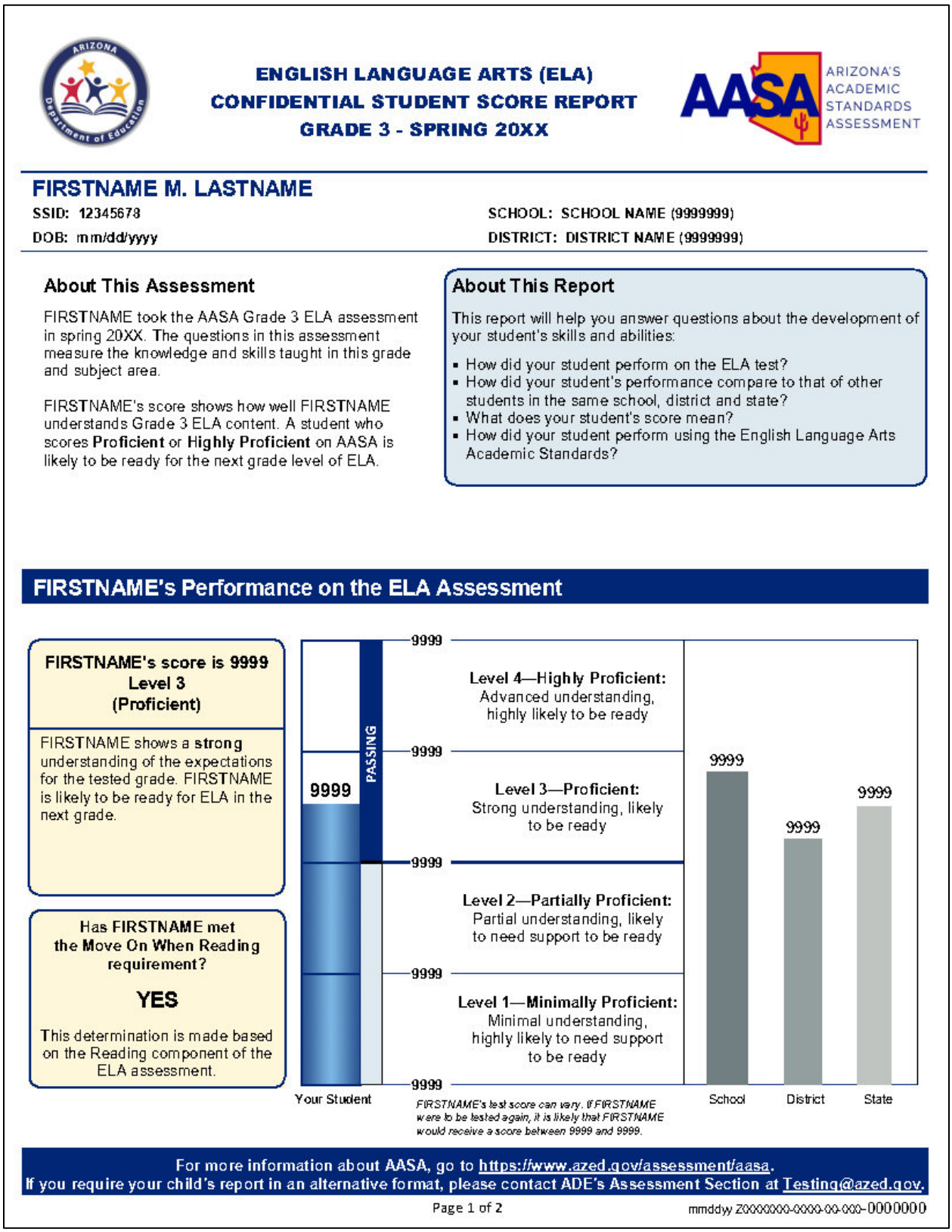
The following AASA reports were available online in PAN at https://az.pearsonaccessnext.com. PDF versions of the reports and district-wide electronic student data files were also available for downloading. District-level user roles provided access to all school-level reports and district-level reports, 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 5.3 and Figure 5.4 present sample reports.

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

AASA 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 time frame. 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 5.3. Sample Reports—Confidential Student Score Report



ENGLISH LANGUAGE ARTS (ELA)
CONFIDENTIAL STUDENT SCORE REPORT
GRADE 3 - SPRING 20XX

ARIZONA'S ACADEMIC STANDARDS ASSESSMENT

FIRSTNAME M. LASTNAME
SSID: 12345678
DOB: mm/dd/yyyy
SCHOOL: SCHOOL NAME (9999999)
DISTRICT: DISTRICT NAME (9999999)

About This Assessment

FIRSTNAME took the AASA Grade 3 ELA assessment in spring 20XX. The questions in this assessment measure the knowledge and skills taught in this grade and subject area.

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

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 on the ELA test?
- How did your student's performance compare to that of other students in the same school, district and state?
- What does your student's score mean?
- How did your student perform using the English Language Arts Academic Standards?

FIRSTNAME's Performance on the ELA Assessment

FIRSTNAME's score is 9999
Level 3
(Proficient)

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

Has FIRSTNAME met the Move On When Reading requirement?

YES

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

9999
PASSING
9999
Level 4—Highly Proficient: Advanced understanding, highly likely to be ready
9999
Level 3—Proficient: Strong understanding, likely to be ready
9999
Level 2—Partially Proficient: Partial understanding, likely to need support to be ready
9999
Level 1—Minimally Proficient: Minimal understanding, highly likely to need support to be ready
9999
Your Student
FIRSTNAME's test score can vary. If FIRSTNAME were to be tested again, it is likely that FIRSTNAME would receive a score between 9999 and 9999.
9999
School
9999
District
9999
State

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

Page 1 of 2
mmddyy Z0000000-0000-00-000-0000000

GRADE 3 **FIRSTNAME's ELA CONFIDENTIAL SCORE REPORT - page 2**

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

ELA Reporting Categories

Reading for Information

FIRSTNAME performed **above mastery** in Reading for Information.

What was assessed?
Students find the main idea and the supporting details of a text. They connect events, ideas, steps, sentences, paragraphs, and illustrations to one another. They find similarities and differences between two texts on the same topic.

What do these results mean?
Your student almost always finds connections between concepts, ideas, or events; uses the text and pictures to make conclusions to ask and answer questions; and finds the similarities and differences between important ideas and key details in two texts on the same topic.

Reading for Literature

FIRSTNAME performed **at or near mastery** in Reading for Literature.

What was assessed?
Students ask and answer questions about a text. They tell how characters and their actions affect a story. They explain how pictures help tell a story. They read two texts by one author and tell the similarities and differences. They find the central message of a story.

What do these results mean?
Your student can often find similarities and differences between the settings or plots of stories written by the same author; tell how one part of a story affects another part; use key details to retell a story and find the main idea; and tell the point of view in a story.

Writing and Language

FIRSTNAME performed **below mastery** in Writing and Language.

What was assessed?
Students write to give information or state opinions. They write on a topic giving supporting details or facts. They use correct capitalization, punctuation, and spelling. They use sentences, a glossary, or a dictionary to figure out the meaning of new words.

What do these results mean?
Your student may have trouble organizing writing for a purpose (like to give information or give opinions); using clues in a text to understand the meaning of new words; spelling commonly used words correctly; and writing simple sentences with correct capitalization and punctuation.

The Writing and Language portion of the ELA assessment requires that each student complete an essay. The essay is evaluated on three criteria.

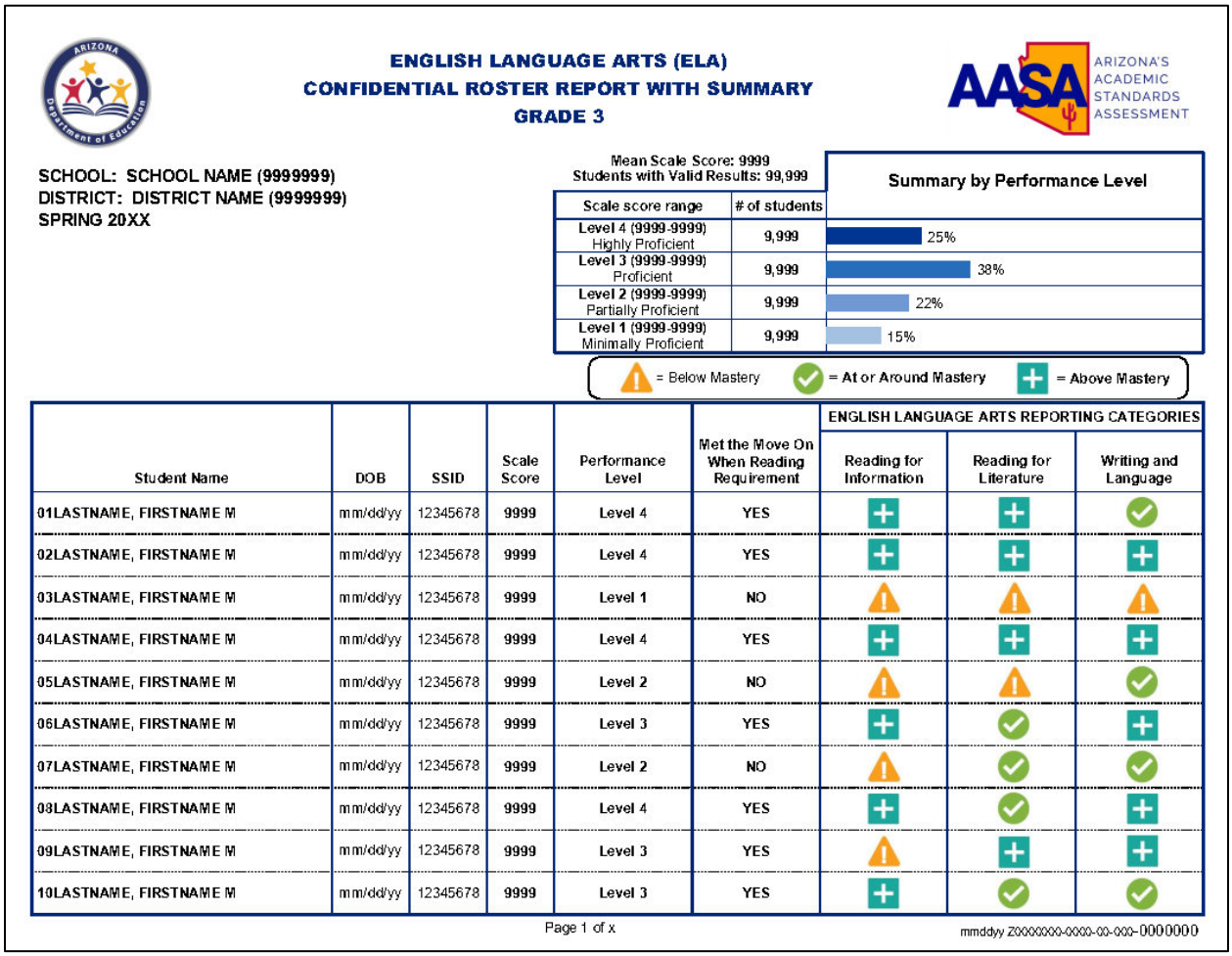
Writing Essay Performance

Statement of Purpose, Focus & Organization	Evidence & Elaboration	Conventions & Editing
Your student earned 3 out of 4 possible points. Your student's essay mainly stays on topic. The opinion is clearly stated and mostly focused. Context supporting the opinion fits the purpose. The response is organized and has few mistakes. There is some variety of transitions used. There is a clear progression of ideas within the essay. There is a clear beginning and end.	Your student earned 2 out of 4 possible points. Your student's essay includes details, facts, and sources that somewhat support its opinion. This evidence is unevenly integrated into the response. The words used are sometimes inappropriate for audience and purpose.	Your student earned 2 out of 2 possible points. Your student's essay shows a strong understanding of sentence structure and language conventions. There are few mistakes in punctuation, capitalization, and spelling present in the response.

For more information about AASA, go to https://www.azed.gov/assessment/aasa.
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 5.4. Sample Reports—Confidential Roster Report with Summary



ENGLISH LANGUAGE ARTS (ELA)
CONFIDENTIAL ROSTER REPORT WITH SUMMARY
GRADE 3

ARIZONA'S ACADEMIC STANDARDS ASSESSMENT

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

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

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

Summary by Performance Level

= Below Mastery = At or Around Mastery = Above Mastery

Student Name	DOB	SSID	Scale Score	Performance Level	Met the Move On When Reading Requirement	ENGLISH LANGUAGE ARTS REPORTING CATEGORIES		
						Reading for Information	Reading for Literature	Writing and Language
01LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 4	YES	+	+	✓
02LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 4	YES	+	+	+
03LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 1	NO	!	!	!
04LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 4	YES	+	+	+
05LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 2	NO	!	!	✓
06LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 3	YES	+	✓	+
07LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 2	NO	!	✓	✓
08LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 4	YES	+	✓	+
09LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 3	YES	!	+	+
10LASTNAME, FIRSTNAME M	mm/dd/yy	12345678	9999	Level 3	YES	+	✓	✓

Page 1 of x

mmddyy ZXXXXXXX-XXXX-XX-XXX-0000000

Chapter 6: CLASSICAL ITEM ANALYSIS

This chapter presents classical statistics for the data used for calibration, equating, and scaling for the Spring 2022 AASA assessments 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 and Table 6.2 present demographic information of the students included in the calibration sample for the Spring 2022 AASA 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—ELA

Subgroup		#Students by Grade					
		3	4	5	6	7	8
	All	74,555	74,639	75,546	74,588	77,690	81,428
Gender	Male	37,737	37,757	38,175	37,996	39,930	41,532
	Female	36,818	36,882	37,371	36,592	37,760	39,896
Ethnicity	Hispanic	35,637	35,356	35,933	35,986	37,587	39,912
	Non-Hispanic	38,918	39,283	39,613	38,602	40,103	41,516
Race	American Indian	4,053	3,998	4,155	4,117	4,522	4,792
	Asian	2,154	2,118	2,020	1,835	1,924	2,082
	Black or African American	5,347	5,268	5,265	5,254	5,302	5,524
	Multi-racial	4,487	4,470	4,348	4,180	4,079	4,169
	Native Hawaiian or Other Pacific Islander	411	388	514	390	388	429
	White	57,550	57,849	58,676	58,183	60,869	63,704
	Missing	553	548	568	629	606	728
Other	Special Ed.	9,693	10,226	10,366	9,936	9,539	9,585
	EL	7,927	7,570	6,806	6,619	6,923	7,364
	Low SES	30,708	30,488	30,751	30,399	30,931	31,743

Table 6.2. Frequency of Students by Subgroup—Mathematics

Subgroup		#Students by Grade					
		3	4	5	6	7	8
All		75,507	75,269	76,097	75,288	78,722	82,328
Gender	Male	38,362	38,163	38,482	38,376	40,506	42,050
	Female	37,145	37,106	37,615	36,912	38,216	40,278
Ethnicity	Hispanic	36,116	35,688	36,248	36,339	38,089	40,317
	Non-Hispanic	39,391	39,581	39,849	38,949	40,633	42,011
Race	American Indian	4,139	4,064	4,212	4,188	4,639	4,889
	Asian	2,173	2,135	2,030	1,844	1,942	2,098
	Black or African American	5,449	5,333	5,310	5,326	5,376	5,602
	Multi-racial	4,550	4,502	4,374	4,201	4,151	4,225
	Native Hawaiian or Other Pacific Islander	420	396	517	396	391	435
	White	58,202	58,257	59,054	58,679	61,580	64,319
	Missing	574	582	600	654	643	760
Other	Special Ed.	9,970	10,388	10,506	10,057	9,752	9,747
	EL	8,064	7,644	6,872	6,679	7,041	7,432
	Low SES	31,137	30,706	30,972	30,691	31,317	32,092

6.2. Descriptive Statistics

Table 6.3 presents descriptive statistics on total raw scores for the spring AASA assessment by content area and 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 minimum/maximum obtained raw score. (See Table 8.1 for the mean scale scores.)

Table 6.3. Classical Test Analysis Statistics

Content Area	Grade	#Students	#Items	Max. Possible Raw Score	Mean Raw Score	SD Raw Score	Min. Raw Score	Max. Raw Score
ELA	3	74,555	44	54	25.60	10.62	2	54
	4	74,639	44	56	28.64	10.93	2	56
	5	75,546	44	53	27.38	10.59	3	53
	6	74,588	44	56	27.64	10.24	2	55
	7	77,690	43	52	27.04	9.73	3	52
	8	81,428	44	55	27.75	10.25	3	55
Math	3	75,507	45	45	23.39	10.86	0	45
	4	75,269	45	45	21.95	11.71	0	45
	5	76,097	45	45	18.55	11.10	0	45
	6	75,288	47	47	19.46	10.82	0	47
	7	78,722	47	47	19.78	10.56	0	47
	8	82,328	47	47	16.85	10.07	0	47

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 AASA assessment targets *p*-values in the range of 0.20 to 0.90.

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

Table 6.4 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 3.4, the items will be reviewed during test construction of the next testing cycle for possible replacement in future administrations.

Table 6.4. Classical Item Analysis Summary

Content Area	Grade	#Items	Mean *P*-Value	Mean Point-Biserial
ELA	3	44	0.46	0.46
	4	44	0.51	0.47
	5	44	0.51	0.47
	6	44	0.48	0.43
	7	43	0.52	0.44
	8	44	0.48	0.44
Math	3	45	0.52	0.52
	4	45	0.49	0.56
	5	45	0.41	0.53
	6	47	0.41	0.50
	7	47	0.42	0.49
	8	47	0.36	0.47

Note. One item for ELA Grade 7 was omitted from scoring due to an error in the stimulus.

6.4. Distractor Analysis

Table 6.5 and Table 6.6 present the point-biserial correlations associated with a correct option and the incorrect options at various percentiles. As expected, the point-biserial correlation for a correct option was around 0.20 or higher for most items, whereas the point-biserial correlation for incorrect options was generally 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 associated with each response option was calculated as well. Typically, a negative point-biserial correlation is sought for distractors because less-proficient students should be more likely to choose an incorrect option.

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

Content Area	Grade	#MC Items	Min.	P25	P50	P75	Max.
ELA	3	24	0.26	0.37	0.43	0.50	0.59
	4	26	0.22	0.38	0.42	0.46	0.59
	5	19	0.29	0.37	0.42	0.49	0.55
	6	23	0.26	0.34	0.35	0.44	0.52
	7	26	0.26	0.38	0.42	0.49	0.57
	8	26	0.19	0.32	0.40	0.46	0.56
Math	3	20	0.20	0.35	0.45	0.54	0.65
	4	12	0.30	0.39	0.48	0.57	0.65
	5	12	0.25	0.36	0.44	0.49	0.58
	6	19	0.21	0.28	0.45	0.47	0.51
	7	21	0.28	0.34	0.40	0.46	0.57
	8	30	0.25	0.37	0.42	0.47	0.59

Note. Min.= minimum, P25 = 25th percentile, P50 = 50th percentile (median), P75 = 75th percentile, Max. = maximum. This analysis is conducted for MC items only.

Table 6.6. Distractor Analysis Summary: Point-Biserial Correlations for Incorrect Options

Content Area	Grade	#MC Items	Min.	P25	P50	P75	Max.
ELA	3	24	-0.35	-0.26	-0.20	-0.16	-0.02
	4	26	-0.32	-0.26	-0.20	-0.17	0.01
	5	19	-0.36	-0.27	-0.21	-0.14	0.03
	6	23	-0.33	-0.24	-0.20	-0.12	0.05
	7	26	-0.34	-0.25	-0.22	-0.16	0.00
	8	26	-0.38	-0.24	-0.20	-0.13	0.22
Math	3	20	-0.47	-0.25	-0.18	-0.11	-0.01
	4	12	-0.51	-0.27	-0.19	-0.15	-0.03
	5	12	-0.39	-0.24	-0.16	-0.12	0.03
	6	19	-0.39	-0.25	-0.18	-0.10	0.06
	7	21	-0.34	-0.24	-0.18	-0.14	0.04
	8	30	-0.30	-0.20	-0.16	-0.11	0.20

Note. Min.= minimum, P25 = 25th percentile, P50 = 50th percentile (median), P75 = 75th percentile, Max. = maximum. This analysis is conducted for MC items only.

Chapter 7: CALIBRATION, EQUATING, AND SCALING

This chapter describes the calibration, equating, and scaling procedures that took place for the Spring 2022 AASA assessments 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^{next} (PAN)
- Records that indicate the student took an accommodated form
- Records with non-valid attempts noted by less than one response
- Duplicate records (e.g., students indicated as taking the test more than once)
- Records in which a student was enrolled in an exclusionary school list from ADE

7.2. Calibration Methods

Item response theory (IRT) models were used in the item calibration. All tests were calibrated separately. If there was more than one operational form, all operational forms were calibrated concurrently. All calibration activities were replicated with 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 one-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 and 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 AASA 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 AASA assessment.

Table 7.1. IRT Statistics Summary

Content Area	Grade	#Items	Mean Rasch
ELA	3	44	0.23
	4	44	0.38
	5	44	0.15
	6	44	0.29
	7	43	0.25
	8	44	0.29
Math	3	45	0.38
	4	45	0.25
	5	45	0.31
	6	47	0.15
	7	47	0.16
	8	47	0.09

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.12 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.13 – Figure B.36. 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 $SE(\theta)$ is the CSEM, and $TI(\theta)$ 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. Equating

The Spring 2022 AASA tests were equated and placed on the operational AASA scale using a non-equivalent groups anchor item (NEAT) design. A set of anchor items was selected from the existing item bank. The anchor items were selected such that they contributed approximately 30% of the total score points and their content representation was as similar as possible to the blueprint. The location of all anchor items stayed within three positions from where they were in the previous year.

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

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

Table 7.2 presents the number of items for the initial anchor set of each grade and the number of items dropped from each initial anchor set for the Spring 2022 assessments.

Table 7.2. Summary of Anchor Items

Content Area	Grade	#Items in the Initial Anchor Set	#Items Dropped from Anchor
ELA	3	16	4
	4	15	0
	5	15	2
	6	15	2
	7	15	3
	8	15	3
Math	3	16	4
	4	16	2
	5	15	3
	6	17	2
	7	17	3
	8	16	2

7.5. Scaling Methods

The AASA reporting scale was established in 2015 when the first administration took place (known as the AzMERIT statewide achievement assessment at that time). These tests were placed on a vertical scale for the total score as a result of study previously completed (American Institutes for Research, 2015, Appendix J). Scaling constants for the total score were determined such that the vertically scaled theta score, based on the total test, was transformed by solving the following equations:

$$Scale\ Score = VS_A * \theta + VS_B$$

where VS_A and VS_B are scaling constants on the vertical scale that are used to transform θ, which are the performance level cuts on the theta (ability) scale, into scale scores. For reporting, θ is truncated at –3.5 and 3.5 for lower and upper ends, respectively.

The AASA reporting scale ranged from 2395 to 2658 across grades for ELA and from 3395 to 3776 across grades for mathematics. In addition to total score, a subscore was also calculated for each reporting category (which differ by grade) using the same formula. The scaling constants were applied to a theta score based on items associated with a reporting category to transform it to a scale score. Appendix B presents the raw-to-scale score conversion tables for each content area and grade.

7.6. IRT Assumptions

It is important to evaluate how the Rasch models applied for AASA 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.

7.6.1. Unidimensionality

An assumption under the Rasch models is unidimensionality, that there is exactly one latent variable (e.g., mathematics 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 content area and grade. Table 7.3 presents the first 10 eigenvalues from 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 content area and grade in Figure B.37 – Figure B.48 in Appendix B. The PCA results with the parallel analysis criterion show only one significant dimension for each grade, which supports unidimensionality.

Table 7.3. Eigenvalues from PCA

Content Area	Grade	1	2	3	4	5	6	7	8	9	10
ELA	3	15.09	1.91	1.27	0.97	0.91	0.87	0.86	0.85	0.83	0.80
	4	14.98	1.77	1.24	1.07	0.97	0.90	0.89	0.88	0.85	0.82
	5	15.27	1.57	1.31	1.07	0.97	0.92	0.88	0.86	0.83	0.82
	6	12.85	1.47	1.30	1.07	0.99	0.97	0.93	0.92	0.89	0.88
	7	13.78	1.60	1.22	1.02	0.99	0.92	0.89	0.87	0.86	0.85
	8	13.98	2.04	1.41	1.02	0.99	0.95	0.92	0.89	0.88	0.87
Math	3	20.50	2.06	1.11	1.01	0.96	0.87	0.82	0.81	0.80	0.76
	4	23.02	1.84	1.39	1.03	0.93	0.83	0.80	0.75	0.71	0.69
	5	20.46	1.77	1.42	1.16	0.96	0.92	0.85	0.82	0.79	0.78
	6	20.16	1.87	1.20	1.02	1.01	0.94	0.90	0.88	0.86	0.84
	7	19.76	1.54	1.34	1.00	0.97	0.94	0.88	0.86	0.83	0.82
	8	17.20	1.84	1.15	1.01	0.96	0.92	0.90	0.87	0.86	0.84

7.6.2. Local Item Independence

Local item independence is another assumption under the Rasch models that assumes any item pair is uncorrelated, conditioned on the latent trait (e.g., mathematics proficiency) an instrument is intended to measure. A violation of local item dependence assumption 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 the residual correlation greater than 0.70, only one item is needed on the test (Linacre, 2022a).

As shown in Table 7.4 that summarizes the distribution of the residual correlations, most residual correlations are slightly negative or slightly positive, and only two (out of more than 900 per grade) are greater than 0.70. The results of the residual correlations indicate that the local item independence assumption holds for the AASA tests.

Table 7.4. Q3 Statistics

Content Area	Grade	#Item Pairs	Mean	SD	Min.	P10	P25	P50	P75	P90	Max.	#Items Exceeding 0.70
ELA	3	946	-0.02	0.04	-0.11	-0.05	-0.04	-0.02	-0.01	0.00	0.76	1
	4	946	-0.02	0.04	-0.13	-0.06	-0.04	-0.03	-0.01	0.01	0.67	0
	5	946	-0.02	0.04	-0.11	-0.05	-0.04	-0.03	-0.01	0.01	0.60	0
	6	946	-0.02	0.04	-0.11	-0.05	-0.04	-0.02	-0.01	0.01	0.70	0
	7	903	-0.02	0.04	-0.10	-0.06	-0.04	-0.03	-0.01	0.00	0.58	0
	8	946	-0.02	0.04	-0.10	-0.06	-0.04	-0.03	-0.01	0.01	0.78	1
Math	3	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0
	4	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0
	5	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0
	6	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0
	7	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0
	8	990	-0.02	0.04	-0.12	-0.06	-0.04	-0.02	-0.01	0.02	0.15	0

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

7.6.3. Item Fit

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

Table 7.5 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 7.5. IRT Item Fit Summary Statistics

Content Area	Grade	#Items	#Flagged Items by Infit	%Flagged
ELA	3	44	0	0
	4	44	1	2
	5	44	0	0
	6	44	0	0
	7	43	0	0
	8	44	0	0
Math	3	45	2	4
	4	45	1	2
	5	45	0	0
	6	47	2	4
	7	47	0	0
	8	47	0	0

Chapter 8: TEST RESULTS

This chapter contains information about the results of the administration of the Spring 2022 AASA assessments, addressing Standards 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 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 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 content area and grade, including the mean and standard deviation of the 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 and Table 8.3 present the percentage of students in each level of mastery by reporting category, and Table 8.4 and Table 8.5 present the mean and standard deviation of 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 content area and grade in Appendix C.

Table 8.1. Overall Test Results

Content Area	Grade	N	Total Scale Score		% at Overall Performance Levels			
			Mean	SD	1	2	3	4
ELA	3	79,804	2500.64	35.45	48	12	26	15
	4	79,949	2519.10	34.04	42	14	31	13
	5	80,649	2529.81	35.65	40	21	29	10
	6	81,041	2542.68	30.12	36	25	35	4
	7	83,804	2554.47	34.18	38	19	32	11
	8	87,227	2558.97	33.04	42	22	26	10
Math	3	80,808	3515.80	44.57	33	27	28	12
	4	80,600	3545.07	51.24	38	23	25	14
	5	81,283	3577.68	44.21	39	24	26	11
	6	81,769	3607.68	41.96	48	21	20	11
	7	84,940	3626.67	41.14	56	17	14	13
	8	88,301	3653.19	36.48	55	18	17	10

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

Table 8.2. Performance Distributions by Reporting Category—ELA

Grade	Reporting Category	N	% at Levels of Mastery		
			1	2	3
3	Reading for Information	79,804	48	29	23
	Reading for Literature	79,804	49	24	27
	Writing and Language	79,804	39	34	27
4	Reading for Information	79,949	41	29	30
	Reading for Literature	79,949	45	25	30
	Writing and Language	79,949	33	38	29
5	Reading for Information	80,649	51	27	22
	Reading for Literature	80,649	48	26	26
	Writing and Language	80,649	38	42	20
6	Reading for Information	81,041	45	35	20
	Reading for Literature	81,041	39	36	25
	Writing and Language	81,041	45	30	25
7	Reading for Information	83,804	45	29	26
	Reading for Literature	83,804	39	30	31
	Writing and Language	83,804	39	31	30
8	Reading for Information	87,227	54	26	20
	Reading for Literature	87,227	45	34	21
	Writing and Language	87,227	47	30	23

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

Table 8.3. Performance Distributions by Reporting Category—Mathematics

Grade	Reporting Category	N	% at Levels of Mastery		
			1	2	3
3	Operations, Algebraic Thinking, & Numbers in Base Ten	80,808	48	23	29
	Numbers and Operations – Fractions	80,808	48	35	17
	Measurement, Data, & Geometry	80,808	41	36	23
4	Operations, Algebraic Thinking, & Numbers in Base Ten	80,600	48	20	32
	Numbers and Operations – Fractions	80,600	51	22	28
	Measurement, Data, & Geometry	80,600	43	37	20
5	Operations, Algebraic Thinking, & Numbers in Base Ten	81,283	53	21	27
	Numbers and Operations – Fractions	81,283	51	26	23
	Measurement, Data, & Geometry	81,283	45	33	22
6	Ratio and Proportional Relationships	81,769	48	31	21
	The Number System	81,769	53	27	20
	Expressions & Equations	81,769	58	22	20
	Geometry, Statistics & Probability	81,769	44	43	13
7	Ratio and Proportional Relationships	84,940	56	23	21
	The Number System	84,940	52	29	19
	Expressions & Equations	84,940	59	22	20
	Geometry, Statistics & Probability	84,940	59	29	12

Grade	Reporting Category	N	% at Levels of Mastery		
			1	2	3
8	Expressions and Equations	88,301	58	23	19
	Functions	88,301	58	28	14
	Geometry	88,301	56	32	12
	Statistics & Probability and The Number System	88,301	58	23	19

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

Table 8.4. Test Results by Accommodation—ELA

Grade	Accommodation	N	Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
3	Adult Transcription	24	2464.54	22.87	83	17	0	0
	Assistive Technology	7	*	*	*	*	*	*
	Sign Test Content	19	2468.32	39.84	84	–	5	11
	Simplified Directions	67	2473.43	25.02	84	9	6	1
4	Adult Transcription	21	2502.33	39.13	62	24	0	14
	Assistive Technology	7	*	*	*	*	*	*
	Sign Test Content	22	2473.27	13.22	100	0	0	0
	Simplified Directions	74	2492.43	25.33	78	12	8	1
5	Adult Transcription	17	2506.24	40.27	53	18	24	6
	Assistive Technology	11	2525.00	23.24	45	18	36	0
	Sign Test Content	18	2480.17	16.12	100	0	0	0
	Simplified Directions	47	2499.40	27.42	83	6	11	0
6	Adult Transcription	18	2512.94	38.73	78	0	22	0
	Assistive Technology	19	2513.68	30.01	74	21	5	0
	Sign Test Content	38	2504.55	14.85	92	8	0	0
	Simplified Directions	43	2518.79	27.96	67	19	14	0
7	Adult Transcription	17	2543.12	39.70	59	6	24	12
	Assistive Technology	25	2548.76	40.40	48	16	24	12
	Sign Test Content	22	2505.59	19.11	95	5	0	0
	Simplified Directions	38	2526.26	21.27	84	8	8	0
8	Adult Transcription	14	2534.50	26.33	71	14	14	0
	Assistive Technology	14	2531.57	26.17	79	7	14	0
	Sign Test Content	27	2515.70	15.57	96	4	0	0
	Simplified Directions	25	2534.40	29.17	64	28	8	0

Note. SD = standard deviation, 1 = *Minimally Proficient*, 2 = *Partially Proficient*, 3 = *Proficient*, 4 = *Highly Proficient*. Statistics for subgroups with less than 11 students are omitted in compliance with FERPA regulations and replaced with an asterisk (*).

Table 8.5. Test Results by Accommodation—Mathematics

Grade	Accommodation	N	Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
3	Adult Transcription	15	3477.20	34.73	73	20	7	0
	Assistive Technology	4	*	*	*	*	*	*
	Sign Test Content	19	3464.63	51.54	79	11	5	5
	Simplified Directions	72	3480.01	33.52	64	28	8	0
4	Adult Transcription	21	3498.86	45.78	81	10	10	0
	Assistive Technology	4	*	*	*	*	*	*
	Sign Test Content	22	3484.95	29.81	91	9	0	0
	Simplified Directions	83	3508.11	41.59	67	23	8	1
5	Adult Transcription	9	*	*	*	*	*	*
	Assistive Technology	1	*	*	*	*	*	*
	Sign Test Content	17	3533.65	39.05	76	18	6	0
	Simplified Directions	55	3556.42	35.58	58	20	22	0
6	Adult Transcription	20	3575.40	38.61	85	10	0	5
	Assistive Technology	10	*	*	*	*	*	*
	Sign Test Content	36	3565.42	25.77	94	0	6	0
	Simplified Directions	38	3574.37	28.66	89	5	5	0
7	Adult Transcription	4	*	*	*	*	*	*
	Assistive Technology	4	*	*	*	*	*	*
	Sign Test Content	20	3579.55	16.97	100	0	0	0
	Simplified Directions	35	3599.77	30.42	77	20	3	0
8	Adult Transcription	6	*	*	*	*	*	*
	Assistive Technology	3	*	*	*	*	*	*
	Sign Test Content	26	3620.46	18.12	96	4	0	0
	Simplified Directions	26	3635.31	22.03	73	23	4	0

Note. SD = standard deviation, 1 = *Minimally Proficient*, 2 = *Partially Proficient*, 3 = *Proficient*, 4 = *Highly Proficient*. Statistics for subgroups with less than 11 students are omitted in compliance with FERPA regulations and replaced with an asterisk (*).

Table 8.6 and Table 8.7 present 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 and content areas.

Table 8.6. Scale Score Distribution by Performance Level—ELA

Grade	Performance Level	N	Average Scale Score	%	Cumulative %
3	1	37,951	2470.11	47.56	47.56
	2	9,278	2502.45	11.63	59.18
	3	20,851	2522.92	26.13	85.31
	4	11,724	2558.36	14.69	100.00
4	1	33,548	2487.10	41.96	41.96
	2	11,021	2515.76	13.79	55.75
	3	24,610	2538.98	30.78	86.53
	4	10,770	2576.76	13.47	100.00
5	1	32,011	2494.40	39.69	39.69
	2	17,141	2530.14	21.25	60.95
	3	23,133	2556.12	28.68	89.63
	4	8,364	2591.91	10.37	100.00
6	1	29,069	2510.57	35.87	35.87
	2	20,588	2542.20	25.40	61.27
	3	28,507	2569.37	35.18	96.45
	4	2,877	2606.22	3.55	100.00
7	1	31,894	2519.92	38.06	38.06
	2	16,098	2550.53	19.21	57.27
	3	26,778	2577.46	31.95	89.22
	4	9,034	2615.37	10.78	100.00
8	1	36,731	2527.89	42.11	42.11
	2	19,496	2560.81	22.35	64.46
	3	22,522	2585.44	25.82	90.28
	4	8,478	2619.08	9.72	100.00

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

Table 8.7. Scale Score Distribution by Performance Level—Mathematics

Grade	Performance Level	N	Average Scale Score	%	Cumulative %
3	1	26,814	3466.11	33.18	33.18
	2	22,072	3511.18	27.31	60.50
	3	22,342	3547.79	27.65	88.14
	4	9,580	3590.88	11.86	100.00
4	1	30,424	3491.67	37.75	37.75
	2	18,608	3545.64	23.09	60.83
	3	20,458	3580.73	25.38	86.22
	4	11,110	3624.67	13.78	100.00
5	1	31,997	3534.08	39.36	39.36
	2	19,299	3577.00	23.74	63.11
	3	20,753	3611.11	25.53	88.64
	4	9,234	3655.02	11.36	100.00

Grade	Performance Level	N	Average Scale Score	%	Cumulative %
6	1	39,457	3572.50	48.25	48.25
	2	17,331	3614.59	21.20	69.45
	3	16,159	3643.63	19.76	89.21
	4	8,822	3685.60	10.79	100.00
7	1	47,262	3596.48	55.64	55.64
	2	14,728	3639.78	17.34	72.98
	3	12,241	3663.12	14.41	87.39
	4	10,709	3700.26	12.61	100.00
8	1	48,710	3627.18	55.16	55.16
	2	15,922	3659.84	18.03	73.20
	3	14,674	3685.60	16.62	89.81
	4	8,995	3729.44	10.19	100.00

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

Chapter 9: RELIABILITY AND VALIDITY

This chapter provides evidence supporting the reliability and validity of scores on the Spring 2022 AASA 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; 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 other words, 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. In addition, the range of certainty around the score should be small enough to support educational decisions.

9.1.1. Internal Consistency

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 and Table 9.2 present coefficient alphas and SEMs (computed based on the calibration sample) for the total and reporting category scores. The test-level and reporting category-level reliability coefficient alpha results suggest that the AASA assessments produce reliable scores.

Table 9.1. Coefficient Alpha and SEM by Total and Reporting Category Score—ELA

Grade	Reporting Category	N	#Items	Coefficient Alpha	SEM
3	Total	74,190	44	0.92	3.05
	Reading for Information	74,387	19	0.80	1.81
	Reading for Literature	74,388	16	0.78	1.73
	Writing and Language	74,190	9	0.84	1.60
4	Total	74,386	44	0.91	3.19
	Reading for Information	72,518	20	0.78	2.11
	Reading for Literature	74,523	15	0.82	1.76
	Writing and Language	74,386	9	0.80	1.51
5	Total	75,473	44	0.92	3.02
	Reading for Information	75,489	20	0.81	1.94
	Reading for Literature	75,498	15	0.80	1.66
	Writing and Language	75,473	9	0.81	1.52
6	Total	74,461	44	0.90	3.26
	Reading for Information	74,533	21	0.78	2.10
	Reading for Literature	74,519	14	0.72	1.80
	Writing and Language	74,461	9	0.80	1.63
7	Total	77,550	43	0.91	2.99
	Reading for Information	77,615	20	0.78	1.97
	Reading for Literature	77,639	13	0.75	1.55
	Writing and Language	77,550	10	0.80	1.56
8	Total	81,332	44	0.91	3.13
	Reading for Information	81,369	20	0.78	1.93
	Reading for Literature	81,364	15	0.77	1.68
	Writing and Language	81,332	9	0.83	1.63

Table 9.2. Coefficient Alpha and SEM by Total and Reporting Category Score—Mathematics

Grade	Reporting Category	N	#Items	Coefficient Alpha	SEM
3	Total	75,218	45	0.94	2.71
	Operations, Algebraic Thinking, & Numbers in Base Ten	75,143	23	0.92	1.89
	Numbers & Operations – Fractions	75,334	9	0.69	1.27
	Measurement, Data, & Geometry	75,218	13	0.76	1.41
4	Total	75,153	45	0.95	2.63
	Operations, Algebraic Thinking, & Numbers in Base Ten	75,140	23	0.91	1.87
	Numbers & Operations – Fractions	75,153	14	0.88	1.41
	Measurement, Data, & Geometry	75,075	8	0.74	1.12
5	Total	75,685	45	0.94	2.71
	Operations, Algebraic Thinking, & Numbers in Base Ten	75,685	18	0.89	1.69
	Numbers & Operations – Fractions	75,921	15	0.85	1.52
	Measurement, Data, & Geometry	75,983	12	0.77	1.42

Grade	Reporting Category	N	#Items	Coefficient Alpha	SEM
6	Total	75,119	47	0.93	2.78
	Ratio & Proportional Relationships	75,034	10	0.77	1.26
	The Number System	75,187	14	0.83	1.53
	Expressions & Equations	75,180	15	0.83	1.52
	Geometry, Statistics & Probability	75,119	8	0.61	1.18
7	Total	78,613	47	0.93	2.78
	Ratio & Proportional Relationships	78,484	10	0.78	1.31
	The Number System	78,622	10	0.79	1.15
	Expressions & Equations	78,623	12	0.81	1.36
	Geometry, Statistics & Probability	78,613	15	0.75	1.66
8	Total	82,215	47	0.92	2.85
	Expressions & Equations	81,842	15	0.84	1.54
	Functions	82,242	11	0.68	1.44
	Geometry	82,215	9	0.65	1.22
	Statistics & Probability and The Number System	82,253	12	0.77	1.44

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 upper and lower ends of the ability continuum because there are usually fewer items that measure those difficulty levels. Figure B.13 – Figure B.36 in Appendix B present the TCC and CSEM curves of the assessments. As expected, the CSEMs around the performance level cuts were the lowest.

9.1.2. Inter-rater Reliability

For the handscored ELA writing prompts, the consistency with which two raters assign scores to student responses is determined by inter-rater agreement, also referred to as rater agreement, which indicates the level of agreement between two scores assigned to student responses. It is the measure of how often scorers agree with each other. Rater agreement for the AASA ELA assessment is calculated between the human-scored and IEA-scored prompts. Rater agreement statistics include the percentage of exact and adjacent scores for each item that received two scores. For 10% of responses, a second "reliability" score was assigned by a second scorer.

The expectation is an inter-rater agreement of 65% or higher between the first and second scores. When IEA provided a high confidence score, the second reliability score was from a human rater. For the subset of responses where IEA provided a low confidence score, the first and second score were both from human raters. Pearson scoring staff used inter-rater agreement indices as one factor in determining the needs for continuing training and intervention on both individual and group levels.

Two other statistical indices are also used to measure reliability in the handscoring process: Cohen's kappa and intraclass correlation. The quadratic weighted kappa (Cohen, 1968) allows rater disagreements to be weighted differentially (e.g., magnitude of a one-point difference in ratings versus a two-point difference) and is calculated with the weighted differences included, which are defined by the following formulas:

$$w_{ij} = \frac{(|i-j|)^2}{(k-1)^2}$$

$$\kappa_w = 1 - \frac{\sum w_{ij}\, O_{ij}}{\sum w_{ij}\, E_{ij}}$$

where $|i-j|$is the number of categories by which raters disagree, and k is the total number of score categories, and w_{ij} is the weighted level of disagreement. E_{ij} is the expected matrix and O_{ij} is the observed matrix. The quadratic weighed kappa ranges from -1.0 to 1.0, with higher, more positive values indicative of greater rater agreement.

The intraclass correlation is defined by Shrout and Fleiss (1979) as "the correlation between one measurement (either a single rating or a mean of ratings) on a target and another measurement obtained on that target" (p. 422). In the context of the AASA assessments, the "target" was the student response and each measurement was obtained by a rater randomly assigned to that response. Therefore, *ICC*(1,1) was used to estimate the intraclass correlation. *ICC*(1,1) is estimated as follows (Shrout & Fleiss, 1979):

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

where *BMS* is the between-targets mean square, *WMS* is the within-targets mean square, and *k* is the number of raters rating each target.

Table 9.3 presents the quadratic weighted kappa and intraclass correlation by reporting category. Items with a kappa statistic lower than 0.20, considered as slight rater agreement (Landis & Koch, 1977) and of which there were none, were flagged for potential replacement in future administrations.

Table 9.3. Inter-rater Reliability Statistics

Grade	Trait	Score Range	N	Quadratic Kappa	ICC	%Exact Agreement	%Adjacent Agreement
3	Statement of Purpose, Focus & Organization	1–4	8,011	0.73	0.73	0.68	0.32
	Evidence & Elaboration	1–4	8,011	0.73	0.73	0.68	0.32
	Conventions & Editing	0–2	8,011	0.86	0.86	0.84	0.16
4	Statement of Purpose, Focus & Organization	1–4	8,059	0.69	0.69	0.68	0.32
	Evidence & Elaboration	1–4	8,059	0.71	0.71	0.73	0.27
	Conventions & Editing	0–2	8,059	0.89	0.89	0.89	0.10
5	Statement of Purpose, Focus & Organization	1–4	8,140	0.75	0.74	0.72	0.28
	Evidence & Elaboration	1–4	8,140	0.72	0.72	0.72	0.28
	Conventions & Editing	0–2	8,140	0.84	0.84	0.85	0.15
6	Statement of Purpose, Focus & Organization	1–4	8,171	0.74	0.74	0.65	0.34
	Evidence & Elaboration	1–4	8,171	0.70	0.70	0.66	0.34
	Conventions & Editing	0–2	8,171	0.82	0.82	0.82	0.17
	Human-scored Reading Item	0–1	7,423	0.94	0.94	0.97	0.03
7	Statement of Purpose, Focus & Organization	1–4	8,472	0.76	0.76	0.65	0.35
	Evidence & Elaboration	1–4	8,472	0.78	0.78	0.67	0.33
	Conventions & Editing	0–2	8,472	0.81	0.81	0.85	0.15
8	Statement of Purpose, Focus & Organization	1–4	8,725	0.73	0.73	0.68	0.32
	Evidence & Elaboration	1–4	8,725	0.73	0.73	0.68	0.31
	Conventions & Editing	0–2	8,725	0.82	0.82	0.84	0.16

Note. ICC = intraclass correlation

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., mathematics 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 AASA assessments, 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:

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

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

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

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

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

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

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

Table 9.4 presents the summary of DIF classification criteria for both the MH method and SMD. An alpha level of 0.05 was used for all SMD statistics.

Table 9.4. DIF Flag Categories

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

The 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 9.5 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 9.5. Number of Items Exhibiting Strong DIF

Content Area	Grade	#Items	#Items with Strong DIF
ELA	3	44	1
	4	44	0
	5	44	2
	6	44	3
	7	43	1
	8	44	2
Math	3	45	0
	4	45	2
	5	45	0
	6	47	2
	7	47	1
	8	47	3

9.3. Correlations Among Reporting Categories

Correlations were examined between on total raw score and the reporting category raw scores. The data used to calculate the correlations were based on the calibration sample described in Chapter 7. Table 9.6 – Table 9.9 present the test correlations and disattenuated correlations between the total raw score and the reporting category raw score for each content area and grade. 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 and Table 9.2, were used to calculate the corrected correlation coefficients for attenuation. The disattenuated correlations could be greater than 1.00.

Table 9.6. Correlations and Disattenuated Correlations between Total and Reporting Category Raw Score—ELA

Grade	Score	Total	Reading for Information	Reading for Literature	Writing and Language
3	Total	1.00	0.93	0.91	0.89
	Reading for Information	1.08	1.00	0.79	0.73
	Reading for Literature	1.07	1.00	1.00	0.70
	Writing and Language	1.01	0.89	0.86	1.00
4	Total	1.00	0.92	0.91	0.87
	Reading for Information	1.09	1.00	0.75	0.71
	Reading for Literature	1.05	0.94	1.00	0.69
	Writing and Language	1.02	0.90	0.85	1.00
5	Total	1.00	0.93	0.91	0.87
	Reading for Information	1.08	1.00	0.77	0.70
	Reading for Literature	1.06	0.96	1.00	0.70
	Writing and Language	1.01	0.86	0.87	1.00
6	Total	1.00	0.92	0.89	0.86
	Reading for Information	1.10	1.00	0.74	0.67
	Reading for Literature	1.11	0.99	1.00	0.66
	Writing and Language	1.01	0.85	0.87	1.00
7	Total	1.00	0.92	0.90	0.88
	Reading for Information	1.09	1.00	0.76	0.70
	Reading for Literature	1.09	0.99	1.00	0.70
	Writing and Language	1.03	0.89	0.90	1.00
8	Total	1.00	0.90	0.88	0.88
	Reading for Information	1.07	1.00	0.70	0.67
	Reading for Literature	1.05	0.90	1.00	0.65
	Writing and Language	1.01	0.83	0.81	1.00

Table 9.7. Correlations and Disattenuated Correlations between Total and Reporting Category Raw Score—Mathematics Grades 3–5

Grade	Score	Total	Operations, Algebraic Thinking, & Numbers in Base Ten	Numbers & Operations – Fractions	Measurement, Data, & Geometry
3	Total	1.00	0.97	0.82	0.90
	Operations, Algebraic Thinking, & Numbers in Base Ten	1.04	1.00	0.71	0.82
	Numbers & Operations – Fractions	1.02	0.89	1.00	0.67
	Measurement, Data, & Geometry	1.06	0.98	0.93	1.00
4	Total	1.00	0.96	0.93	0.86
	Operations, Algebraic Thinking, & Numbers in Base Ten	1.03	1.00	0.81	0.75
	Numbers & Operations – Fractions	1.02	0.91	1.00	0.76
	Measurement, Data, & Geometry	1.03	0.91	0.94	1.00

Grade	Score	Total	Operations, Algebraic Thinking, & Numbers in Base Ten	Numbers & Operations – Fractions	Measurement, Data, & Geometry
5	Total	1.00	0.95	0.92	0.88
	Operations, Algebraic Thinking, & Numbers in Base Ten	1.04	1.00	0.80	0.77
	Numbers & Operations – Fractions	1.03	0.92	1.00	0.74
	Measurement, Data, & Geometry	1.03	0.93	0.91	1.00

Table 9.8. Correlations and Disattenuated Correlations between Total and Reporting Category Raw Score—Mathematics Grades 6 and 7

Grade	Score	Total	Ratio & Proportional Relationships	The Number System	Expressions & Equations	Geometry, Statistics & Probability
6	Total	1.00	0.90	0.94	0.93	0.78
	Ratio & Proportional Relationships	1.06	1.00	0.80	0.78	0.62
	The Number System	1.07	1.00	1.00	0.81	0.66
	Expressions & Equations	1.06	0.98	0.98	1.00	0.65
	Geometry, Statistics & Probability	1.04	0.90	0.93	0.91	1.00
7	Total	1.00	0.88	0.90	0.92	0.91
	Ratio & Proportional Relationships	1.03	1.00	0.73	0.75	0.73
	The Number System	1.05	0.93	1.00	0.80	0.76
	Expressions & Equations	1.06	0.94	1.00	1.00	0.77
	Geometry, Statistics & Probability	1.09	0.95	0.99	0.99	1.00

Table 9.9. Correlations and Disattenuated Correlations between Total and Reporting Category Raw Score—Mathematics Grade 8

Grade	Score	Total	Expressions and Equations	Functions	Geometry	Statistics & Probability and The Number System
8	Total	1.00	0.94	0.87	0.79	0.91
	Expressions and Equations	1.07	1.00	0.75	0.66	0.80
	Functions	1.10	0.99	1.00	0.59	0.71
	Geometry	1.02	0.89	0.89	1.00	0.63
	Statistics & Probability and The Number System	1.08	0.99	0.98	0.89	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 AASA tests is to assess the ELA and mathematics proficiency of students based on the Arizona Academic 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, p. 14). AASA measures a student's level of ELA and mathematics proficiency based on the skills specified in the Arizona Academic Standards. Thus, an alignment of the AASA test to the standards is critical.

Item specifications and test blueprints are the core documents that ensure the assessments are aligned to the Arizona Academic Standards. The item specifications define the content limit, model tasks, and response types for a specific standard, and 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.

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 items for 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 AASA assessments meet the test blueprints 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 Academic 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). As presented in Chapter 3, all newly developed items for the Spring 2022 AASA assessment went through a rigorous item review process, including content, bias, and sensitivity committees. During the review process, a group of educators were trained to evaluate whether the items were aligned to the Arizona Academic Standards and to assess important knowledge or skills identified by the standards and item specifications. The items deemed to be acceptable by the committees were eligible to be used on the AASA test.

A standalone field test was administered in Spring 2022 for the ELA Writing test to increase the number of eligible writing prompts in the item bank for operational use in future administrations. New items were field tested in Grade 3 for to assess students Oral Reading Fluency (ORF), which is one of reading foundation standards, in Spring 2022. The ORF items were designed to align with low, medium, and high levels of difficulty (based on Lexiles) and gauge students' ability to read aloud words. Although the ELA Grade 3 ORF items are intended to be used operationally in future administrations to enhance coverage of Grade 3 ELA standards, they will be field tested again in 2023 to further explore their functioning and performance.

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). AASA is designed to measure students' overall ELA and mathematics proficiency based on the Arizona Academic Standards, which are composed of various reporting categories for each content area. AASA items across all reporting categories 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 PCA analysis with the parallel analysis (Horn, 1965) criterion indicated that there is one dominant dimension for both ELA and mathematics 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 7.4, showed that most statistics are either slightly negative or slightly positive, which indicates that the item independence assumption generally holds for the AASA tests.

In addition to the total scale score, the scale score for each reporting category is reported individually. The scale scores for the reporting categories are generated by including the items associated with each reporting category and using the item parameter estimates from the concurrent calibration across all reporting categories. Details about scaling methods are described in Section 7.5. Correlations between the total score and reporting category score presented in Section 9.3 show 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 6.5 and Table 6.6 show that all the multiple-choice items have negative point-biserial correlations, except for 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 that students possess at different performance levels (Egan et al., 2012). The PLDs are the foundation of standard setting meetings. The PLDs for AASA, provided on the ADE website at https://www.azed.gov/assessment/aasa, were drafted prior to the 2015 standard setting workshop and included educator input. ADE considered any need for clarification or revision that arose throughout the standard setting process prior to publishing the final versions (American Institutes for Research, 2015). See Section 10.1 for more details on standard setting.

9.4.5. Evidence Based on Relation 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). Because both the ELA and mathematics AASA assessments are administered to all eligible Arizona students, scores on the tests are expected to be positively correlated. Table 9.10 presents the correlation between AASA ELA and mathematics scale scores from the Spring 2022 administration. The correlations range from 0.73 to 0.78.

Table 9.10. Correlation between AASA ELA and Mathematics Scale Scores

Grade	N	Correlation
3	79,809	0.78
4	79,936	0.78
5	80,688	0.76
6	81,053	0.76
7	83,791	0.78
8	87,175	0.73

9.4.6. Summary

Different aspects of validity evidence have been collected to evaluate the use of AASA scores for their intended purposes. Overall, the evidence provided above supports the use of AASA scores. The PCA revealed unidimensionality of AASA, which supports the use of unidimensional Rasch models. The AASA ELA and mathematics scores were also positively correlated. 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 AASA 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 AASA 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 AASA 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 July 2015. Analyses were conducted to examine the consistency and accuracy with which students who took the Spring 2022 AASA assessment were assigned to the performance levels.

10.1. Standard Setting

Standard setting for the AASA tests was conducted from July 13–16, 2015, following the first operational administration of the AASA in Spring 2015 (known as the AzMERIT assessments at that time) using the bookmark standard setting procedure. The State Board of Education adopted the panelist-recommended performance standards on August 14, 2015. See the standard setting report for a detailed account of the workshop process and outcomes (American Institutes for Research, 2015).

Table 10.1 presents the final scale score ranges for the AASA performance levels, and Table 10.2 presents the scale score and associated CSEM at the performance level cuts. The CSEM is very similar across all grades and content areas within each cut (i.e., 9 or 10 for *Partially Proficient* and *Proficient* and between 10 and 4 for *Highly Proficient*).

Table 10.1. Performance Level Cut Scores

Content Area	Grade	*Minimally Proficient*	*Partially Proficient*	*Proficient*	*Highly Proficient*
ELA	3	2395–2496	2497–2508	2509–2540	2541–2605
	4	2400–2509	2510–2522	2523–2558	2559–2610
	5	2419–2519	2520–2542	2543–2577	2578–2629
	6	2431–2531	2532–2552	2553–2596	2597–2641
	7	2438–2542	2543–2560	2561–2599	2600–2648
	8	2448–2550	2551–2571	2572–2603	2604–2658
Math	3	3395–3494	3495–3530	3531–3572	3573–3605
	4	3435–3529	3530–3561	3562–3605	3606–3645
	5	3478–3562	3563–3594	3595–3634	3635–3688
	6	3512–3601	3602–3628	3629–3662	3663–3722
	7	3529–3628	3629–3651	3652–3679	3680–3739
	8	3566–3649	3650–3672	3673–3704	3705–3776

Note. The scale score cut for Move on When Reading (MOWR) in Grade 3 is 2446.

Table 10.2. CSEM at Performance Level Cuts

Content Area	Grade	*Partially Proficient* Cut		*Proficient* Cut		*Highly Proficient* Cut	
		Scale Score	CSEM	Scale Score	CSEM	Scale Score	CSEM
ELA	3	2497	9	2509	9	2541	10
	4	2510	9	2523	9	2559	11
	5	2520	9	2543	10	2578	12
	6	2532	9	2553	9	2597	11
	7	2543	10	2561	10	2600	11
	8	2551	9	2572	9	2604	11
Math	3	3495	10	3531	10	3573	13
	4	3530	11	3562	10	3606	13
	5	3563	10	3595	10	3635	11
	6	3602	10	3629	10	3663	11
	7	3629	10	3652	10	3680	11
	8	3650	10	3673	9	3705	10

Performance classifications for reporting categories are determined by student performance on the reporting categories compared to the respective *Proficient* performance standard. For each reporting category, a mid-range band is established by extending one CSEM below and above the *Proficient* performance standard scale score cut. If a student's scale score for a reporting category is fallen into the mid-range band, the student performance is classified as *At/Near Mastery* for the reporting category. On the other hand, if a student's scale score is above or below the mid-range band, the student performance is classified as *Above Mastery* or *Below Mastery*, respectively.

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 AASA tests. As a form of reliability, classification consistency represents how reliably students can be classified into performance levels.

For tests such as the AASA 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 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	Not 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 AASA 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 AASA 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 content area and 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

Content Area	Grade	N	Consistency	Inconsistency	Chance	κ	Accuracy	False Positive	False Negative
ELA	3	74,555	0.74	0.26	0.34	0.61	0.81	0.11	0.09
	4	74,639	0.72	0.28	0.31	0.60	0.79	0.11	0.10
	5	75,546	0.72	0.28	0.30	0.60	0.79	0.11	0.09
	6	74,588	0.72	0.28	0.32	0.59	0.80	0.11	0.09
	7	77,690	0.72	0.28	0.30	0.59	0.79	0.11	0.10
	8	81,428	0.72	0.28	0.31	0.59	0.79	0.11	0.09
Math	3	75,507	0.75	0.25	0.28	0.66	0.83	0.09	0.08
	4	75,269	0.76	0.24	0.28	0.67	0.83	0.09	0.08
	5	76,097	0.77	0.23	0.29	0.67	0.83	0.09	0.08
	6	75,288	0.77	0.23	0.34	0.65	0.83	0.09	0.08
	7	78,722	0.78	0.22	0.39	0.64	0.84	0.09	0.07
	8	82,328	0.78	0.22	0.39	0.65	0.84	0.09	0.07

10.3. MOWR Policy

Arizona's Move On When Reading (MOWR) policy is designed to provide students with evidence-based, effective reading instruction in Grades K–3 to position them for success as they progress through school, college, and career. The heart of the legislation emphasizes early identification and immediate intervention for struggling readers. Grade 3 students must meet the MOWR cut score of 2446 on the AASA ELA Reading portion, as established by the State Board of Education, to be promoted to Grade 4, with some exemptions. Students who are retained receive an extra year of specialized support so they are ready to enter Grade 4 as strong readers. For more information, refer to the ADE website at https://www.azed.gov/mowr/.

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.

American Institutes for Research (AIR). (2015). *Recommending AzMERIT performance standards: English language arts grades 3–11, math grades 3–8, Algebra I, Geometry, and Algebra II*. https://www.azed.gov/sites/default/files/2016/12/spring-2015-azmerit-standard-setting_091415.pdf?id=5846d5b4aadebe0cf0337f5e

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.

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.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213.

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.

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.

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.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). *Handbook of test development*. Routledge.

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

Miller, E, G., Ourania, R., & Twing, J, S. (2004). Evaluation of the 0.3 logits screening criterion in common item equating. *Journal of Applied Measurement, 5*, 172–177.

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.

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.12 present the item-level CTT statistics for each content area and grade, including 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.13 – Table A.24 present the item-level distractor analysis for multiple-choice items, including the percentage of students who selected the 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, ELA Grade 3

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	74,577	0.25	0.73
2	OE	4	74,577	0.26	0.72
3	OE	2	74,577	0.62	0.69
4	MC	1	74,441	0.70	0.42
5	MC	1	74,417	0.54	0.41
6	MX	1	74,451	0.33	0.47
7	MC	1	74,428	0.59	0.48
8	MC	1	74,356	0.62	0.57
9	MC	1	74,352	0.41	0.37
10	MC	1	74,387	0.66	0.58
11	MC	1	74,370	0.61	0.51
12	MC	1	69,401	0.33	0.42
13	MX	1	74,401	0.23	0.37
14	MC	1	74,370	0.33	0.27
15	XI	1	74,390	0.56	0.48
16	XI	1	74,036	0.84	0.38
17	XI	2	74,185	0.63	0.62
18	XI	1	73,963	0.72	0.40
19	MC	1	74,481	0.72	0.43
20	MX	1	74,489	0.18	0.32
21	MC	1	74,481	0.76	0.45
22	MC	1	70,967	0.25	0.42
23	MC	1	74,467	0.45	0.37
24	XI	1	74,477	0.32	0.50
25	MC	1	74,438	0.55	0.48
26	MX	1	74,449	0.22	0.55
27	MC	1	74,408	0.42	0.51
28	MC	1	74,419	0.48	0.37
29	MC	1	74,418	0.50	0.53
30	MC	1	74,401	0.40	0.37
31	MX	1	74,413	0.20	0.36
32	MC	1	74,402	0.35	0.43

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
33	MC	1	74,403	0.38	0.26
34	MC	1	74,414	0.59	0.47
35	MC	1	74,404	0.51	0.41
36	MX	1	74,420	0.29	0.53
37	MC	1	74,399	0.62	0.59
38	MC	1	74,384	0.36	0.43
39	MC	1	74,410	0.28	0.36
40	MX	1	74,415	0.14	0.26
41	MC	1	74,413	0.27	0.38
42	XI	1	74,273	0.72	0.43
43	XI	2	74,310	0.48	0.55
44	XI	2	74,216	0.63	0.64

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, ELA Grade 4

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	74,639	0.20	0.65
2	OE	4	74,639	0.17	0.63
3	OE	2	74,639	0.54	0.69
4	MC	1	74,561	0.67	0.43
5	MC	1	74,556	0.52	0.39
6	MX	1	74,561	0.37	0.37
7	MC	1	74,547	0.50	0.42
8	MC	1	74,523	0.51	0.38
9	MC	1	74,509	0.63	0.41
10	MC	1	74,489	0.42	0.30
11	MX	2	74,525	0.36	0.43
12	MC	1	74,515	0.33	0.22
13	MC	1	74,522	0.49	0.40
14	XI	1	74,309	0.41	0.33
15	XI	2	74,390	0.68	0.59
16	XI	2	74,358	0.63	0.47
17	MC	1	74,553	0.61	0.35
18	MC	1	71,937	0.44	0.49
19	MC	1	74,558	0.47	0.34
20	MC	1	74,591	0.80	0.43
21	MC	1	74,537	0.58	0.44
22	MC	1	74,532	0.55	0.44
23	MC	1	74,543	0.70	0.52
24	MC	1	74,524	0.50	0.54
25	MC	1	74,515	0.45	0.42
26	MC	1	72,518	0.62	0.58
27	MC	1	74,525	0.58	0.45
28	MX	2	74,562	0.47	0.58
29	MC	1	74,510	0.45	0.46

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
30	MC	1	74,525	0.46	0.34
31	MC	1	74,535	0.49	0.51
32	MC	1	74,509	0.62	0.59
33	MC	1	74,534	0.50	0.48
34	XI	1	74,438	0.46	0.68
35	MC	1	74,542	0.83	0.49
36	MC	1	74,547	0.58	0.38
37	MC	1	74,523	0.39	0.40
38	MC	1	72,197	0.33	0.47
39	MX	1	74,500	0.24	0.50
40	MC	1	71,797	0.36	0.58
41	MC	1	74,523	0.61	0.40
42	XI	1	74,432	0.72	0.46
43	XI	2	74,448	0.71	0.60
44	XI	1	74,386	0.70	0.43

Note. MC = multiple-choice, MX = multiple-select, and 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, ELA Grade 5

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	75,543	0.30	0.72
2	OE	4	75,543	0.28	0.70
3	OE	2	75,543	0.65	0.69
4	MC	1	74,684	0.74	0.51
5	MC	1	75,494	0.55	0.37
6	MC	1	74,602	0.42	0.50
7	MX	1	75,498	0.25	0.36
8	MC	1	75,486	0.40	0.41
9	XI	1	75,345	0.43	0.45
10	MX	1	75,482	0.28	0.42
11	MC	1	74,253	0.62	0.51
12	XI	1	75,258	0.46	0.51
13	MC	1	75,473	0.57	0.41
14	MC	1	75,477	0.53	0.40
15	XI	1	75,232	0.31	0.48
16	MC	1	75,483	0.43	0.35
17	MC	1	75,478	0.48	0.43
18	MC	1	75,486	0.43	0.42
19	XI	1	75,441	0.80	0.49
20	XI	2	75,460	0.80	0.56
21	XI	1	75,411	0.76	0.48
22	MC	1	74,877	0.47	0.48
23	MC	1	75,511	0.51	0.33
24	MC	1	75,502	0.42	0.37
25	XI	1	75,473	0.24	0.37

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
26	MC	1	75,521	0.60	0.38
27	MC	1	75,512	0.79	0.44
28	XI	1	75,469	0.52	0.59
29	MC	1	75,483	0.39	0.29
30	MC	1	75,493	0.68	0.48
31	MC	1	75,486	0.67	0.54
32	XI	1	75,469	0.62	0.51
33	MC	1	75,494	0.66	0.49
34	MC	1	75,494	0.49	0.39
35	MC	1	75,498	0.57	0.54
36	MC	1	74,705	0.50	0.61
37	MX	1	75,472	0.24	0.48
38	MC	1	75,484	0.64	0.53
39	XI	1	75,498	0.18	0.26
40	MC	1	74,886	0.50	0.59
41	MC	1	75,489	0.54	0.42
42	XI	1	75,470	0.60	0.40
43	XI	1	75,460	0.40	0.39
44	XI	2	75,473	0.59	0.49

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. Item-Level CTT Statistics, ELA Grade 6

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	74,581	0.31	0.72
2	OE	4	74,581	0.26	0.70
3	OE	2	74,581	0.65	0.70
4	MC	1	74,527	0.78	0.45
5	MC	1	74,550	0.64	0.44
6	MC	1	74,536	0.69	0.38
7	MC	1	74,534	0.43	0.44
8	MX	1	74,526	0.25	0.33
9	MX	1	74,511	0.50	0.50
10	MC	1	74,515	0.70	0.43
11	MC	1	74,502	0.47	0.35
12	MC	1	74,487	0.38	0.41
13	MC	1	74,489	0.63	0.47
14	MC	1	74,490	0.39	0.34
15	MC	1	74,487	0.73	0.52
16	MX	1	74,508	0.24	0.49
17	MC	1	74,520	0.39	0.56
18	XI	1	74,525	0.21	0.26
19	MC	1	74,502	0.58	0.34
20	XI	1	74,407	0.78	0.27
21	XI	2	74,457	0.57	0.35
22	XI	2	74,417	0.53	0.48

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
23	MC	1	74,539	0.56	0.35
24	MX	1	74,553	0.23	0.33
25	MX	1	74,533	0.20	0.49
26	MC	1	74,526	0.44	0.34
27	OE	1	74,053	0.54	0.56
28	MC	1	74,519	0.52	0.39
29	MC	1	74,538	0.54	0.32
30	MC	1	74,503	0.35	0.26
31	MX	2	74,524	0.44	0.62
32	MC	1	74,515	0.51	0.35
33	MC	1	74,523	0.58	0.44
34	MC	1	74,519	0.44	0.32
35	MC	1	74,517	0.48	0.41
36	MC	1	74,530	0.43	0.35
37	MC	1	74,538	0.46	0.31
38	MX	2	74,528	0.35	0.44
39	MC	1	74,529	0.41	0.31
40	MC	1	74,524	0.32	0.33
41	MC	1	74,533	0.43	0.45
42	XI	1	74,490	0.53	0.46
43	XI	2	74,509	0.71	0.52
44	XI	1	74,461	0.78	0.49

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.5. Item-Level CTT Statistics, ELA Grade 7

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	77,688	0.28	0.73
2	OE	4	77,688	0.23	0.69
3	OE	2	77,688	0.74	0.67
4	MC	1	77,616	0.70	0.38
5	MC	1	77,620	0.60	0.43
6	MC	1	77,634	0.62	0.31
7	MC	1	77,619	0.71	0.53
8	MC	1	77,612	0.51	0.49
9	MC	1	76,808	0.19	0.37
10	MC	1	77,578	0.53	0.41
11	MC	1	76,665	0.44	0.53
12	MC	1	77,608	0.63	0.42
13	MC	1	77,588	0.45	0.35
14	MC	1	77,591	0.47	0.42
15	MC	1	77,597	0.57	0.39
16	XI	1	77,544	0.74	0.39
17	XI	1	77,569	0.70	0.41
18	XI	1	77,534	0.80	0.49
19	MC	1	77,614	0.29	0.26

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
20	MC	1	77,636	0.66	0.44
21	MX	1	77,646	0.21	0.32
22	MC	1	77,628	0.43	0.31
23	MC	1	77,622	0.53	0.38
24	MC	1	77,631	0.78	0.49
25	MC	1	77,645	0.76	0.45
26	MX	1	77,639	0.41	0.58
27	MC	1	77,625	0.21	0.36
28	MX	1	77,619	0.31	0.57
29	MC	1	77,639	0.47	0.47
30	MC	1	77,610	0.49	0.49
31	MC	1	77,603	0.62	0.57
32	MC	1	77,606	0.39	0.30
33	MC	1	77,614	0.54	0.42
34	MC	1	77,608	0.56	0.40
35	MC	1	77,615	0.62	0.54
36	MC	1	77,621	0.71	0.54
37	MX	2	77,573	0.44	0.44
38	MC	1	77,617	0.55	0.43
39	XI	1	77,435	0.24	0.30
40	MC	1	77,615	0.40	0.33
41	XI	1	77,583	0.36	0.38
42	XI	1	77,539	0.69	0.45
43	XI	2	77,550	0.67	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. One item for ELA Grade 7 was omitted from scoring due to an error in the stimulus.

Table A.6. Item-Level CTT Statistics, ELA Grade 8

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	OE	4	81,424	0.31	0.69
2	OE	4	81,424	0.27	0.70
3	OE	2	81,424	0.64	0.66
4	MC	1	81,362	0.47	0.44
5	MC	1	81,384	0.64	0.39
6	MC	1	81,384	0.52	0.46
7	MC	1	81,372	0.61	0.41
8	MC	1	81,383	0.52	0.31
9	MC	1	81,355	0.45	0.42
10	MC	1	81,352	0.36	0.30
11	MC	1	81,360	0.55	0.47
12	MC	1	81,358	0.44	0.38
13	MC	1	81,355	0.45	0.38
14	MC	1	81,372	0.41	0.37
15	MC	1	81,360	0.43	0.40
16	MC	1	81,367	0.36	0.47

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
17	XI	1	81,370	0.31	0.45
18	MC	1	81,353	0.59	0.54
19	XI	1	81,326	0.49	0.55
20	MC	1	81,363	0.65	0.32
21	MC	1	81,369	0.72	0.56
22	MC	1	81,370	0.19	0.19
23	MC	1	81,364	0.67	0.55
24	XI	1	81,314	0.86	0.43
25	XI	2	81,321	0.72	0.55
26	XI	2	81,309	0.63	0.44
27	MC	1	81,384	0.78	0.50
28	MC	1	81,367	0.48	0.30
29	MC	1	81,376	0.43	0.48
30	MC	1	81,380	0.60	0.45
31	MC	1	81,368	0.42	0.33
32	MC	1	81,364	0.35	0.27
33	MC	1	81,370	0.47	0.41
34	MC	1	81,364	0.39	0.25
35	MX	1	81,366	0.20	0.27
36	MC	1	81,362	0.42	0.39
37	MC	1	81,375	0.37	0.41
38	MC	1	81,372	0.20	0.54
39	MX	1	81,370	0.23	0.42
40	MC	1	81,350	0.23	0.51
41	MX	1	81,369	0.23	0.43
42	XI	1	81,344	0.53	0.48
43	XI	2	81,352	0.76	0.60
44	XI	2	81,332	0.78	0.56

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.7. Item-Level CTT Statistics, Mathematics Grade 3

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	75,452	0.87	0.31
2	XI	1	75,439	0.44	0.58
3	MC	1	75,413	0.63	0.52
4	XI	1	75,386	0.78	0.35
5	XI	1	75,286	0.63	0.62
6	MC	1	75,387	0.63	0.55
7	XI	1	74,822	0.91	0.35
8	MC	1	75,366	0.58	0.34
9	XI	1	75,357	0.64	0.56
10	MC	1	75,368	0.33	0.63
11	XI	1	75,348	0.62	0.53
12	MC	1	75,303	0.32	0.29
13	MC	1	75,350	0.38	0.42

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
14	XI	1	75,237	0.41	0.60
15	MC	1	75,306	0.39	0.37
16	MC	1	75,278	0.42	0.52
17	XI	1	75,246	0.50	0.69
18	MC	1	75,172	0.50	0.44
19	MC	1	75,287	0.58	0.56
20	MC	1	75,271	0.37	0.46
21	MC	1	75,285	0.29	0.55
22	XI	1	75,242	0.69	0.63
23	XI	1	75,165	0.67	0.63
24	MC	1	75,443	0.74	0.43
25	MC	1	75,439	0.60	0.50
26	XI	1	75,431	0.68	0.65
27	XI	1	75,255	0.35	0.66
28	XI	1	75,363	0.65	0.66
29	MC	1	75,396	0.55	0.57
30	MC	1	75,383	0.75	0.53
31	XI	1	75,287	0.11	0.44
32	XI	1	75,319	0.63	0.59
33	MC	1	75,391	0.41	0.62
34	MX	1	75,295	0.44	0.62
35	MC	1	75,353	0.31	0.20
36	XI	1	75,333	0.77	0.60
37	MC	1	75,377	0.59	0.37
38	XI	1	75,010	0.28	0.51
39	MC	1	75,312	0.57	0.28
40	XI	1	75,255	0.42	0.61
41	MC	1	75,334	0.36	0.65
42	XI	1	75,226	0.60	0.63
43	MC	1	75,334	0.68	0.59
44	XI	1	75,143	0.34	0.64
45	XI	1	75,218	0.10	0.44

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.8. Item-Level CTT Statistics, Mathematics Grade 4

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	XI	1	74,884	0.40	0.67
2	XI	1	75,141	0.52	0.63
3	XI	1	75,106	0.72	0.61
4	XI	1	75,120	0.18	0.38
5	XI	1	75,061	0.60	0.58
6	XI	1	75,127	0.62	0.54
7	MC	1	75,178	0.40	0.37
8	MC	1	75,191	0.38	0.36
9	XI	1	75,013	0.48	0.62

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
10	XI	1	74,929	0.30	0.57
11	MC	1	75,156	0.80	0.51
12	XI	1	74,911	0.40	0.59
13	XI	1	74,816	0.39	0.68
14	MC	1	75,125	0.29	0.64
15	XI	1	75,063	0.50	0.62
16	MC	1	75,081	0.43	0.44
17	MC	1	75,150	0.36	0.58
18	XI	1	74,961	0.68	0.58
19	XI	1	74,775	0.19	0.53
20	MC	1	75,092	0.69	0.60
21	XI	1	74,903	0.47	0.49
22	XI	1	74,879	0.15	0.48
23	MC	1	75,053	0.55	0.60
24	MC	1	75,219	0.80	0.47
25	XI	1	75,094	0.61	0.67
26	XI	1	75,170	0.64	0.63
27	MC	1	75,192	0.70	0.50
28	XI	1	75,093	0.48	0.63
29	XI	1	74,866	0.17	0.43
30	XI	1	75,041	0.54	0.61
31	XI	1	74,939	0.43	0.68
32	XI	1	75,096	0.40	0.67
33	MC	1	75,128	0.25	0.58
34	XI	1	75,162	0.80	0.59
35	MC	1	75,149	0.55	0.54
36	XI	1	75,092	0.72	0.48
37	XI	1	75,083	0.75	0.59
38	MC	1	75,135	0.29	0.61
39	XI	1	74,996	0.38	0.53
40	XI	1	75,121	0.61	0.66
41	MC	1	75,155	0.80	0.42
42	MC	1	75,176	0.42	0.64
43	XI	1	75,075	0.48	0.52
44	MC	1	75,140	0.37	0.30
45	XI	1	75,153	0.34	0.65

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.9. Item-Level CTT Statistics, Mathematics Grade 5

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	XI	1	76,010	0.58	0.44
2	XI	1	75,943	0.44	0.56
3	MC	1	76,026	0.50	0.43
4	MC	1	76,011	0.41	0.44
5	XI	1	75,813	0.46	0.66

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
6	XI	1	75,840	0.19	0.47
7	XI	1	76,024	0.68	0.57
8	XI	1	75,986	0.46	0.45
9	MC	1	76,019	0.37	0.30
10	MC	1	76,039	0.46	0.51
11	XI	1	75,942	0.50	0.59
12	XI	1	75,972	0.35	0.63
13	XI	1	75,913	0.29	0.69
14	XI	1	75,834	0.40	0.58
15	MC	1	75,998	0.36	0.49
16	XI	1	75,848	0.47	0.68
17	XI	1	75,957	0.22	0.47
18	XI	1	75,900	0.28	0.60
19	XI	1	75,868	0.26	0.52
20	XI	1	75,773	0.48	0.67
21	XI	1	75,921	0.19	0.54
22	MC	1	75,970	0.25	0.25
23	MC	1	76,011	0.57	0.43
24	MC	1	76,078	0.69	0.37
25	XI	1	76,045	0.76	0.29
26	XI	1	75,998	0.50	0.53
27	XI	1	76,016	0.36	0.63
28	XI	1	76,003	0.41	0.67
29	MC	1	76,069	0.58	0.51
30	XI	1	75,991	0.57	0.42
31	MC	1	76,035	0.32	0.48
32	MC	1	76,011	0.33	0.46
33	XI	1	75,967	0.32	0.62
34	XI	1	75,999	0.38	0.66
35	XI	1	75,898	0.19	0.56
36	MC	1	76,029	0.46	0.44
37	MC	1	76,016	0.32	0.58
38	XI	1	75,944	0.36	0.61
39	MC	1	76,033	0.42	0.36
40	MC	1	76,033	0.57	0.50
41	XI	1	75,785	0.37	0.65
42	XI	1	75,921	0.16	0.61
43	XI	1	75,983	0.65	0.64
44	XI	1	75,968	0.58	0.60
45	XI	1	75,685	0.15	0.51

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

Table A.10. Item-Level CTT Statistics, Mathematics Grade 6

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	XI	1	75,155	0.50	0.59
2	MC	1	75,163	0.63	0.49
3	MC	1	75,251	0.64	0.27
4	MC	1	75,205	0.23	0.36
5	XI	1	75,182	0.59	0.56
6	MC	1	75,212	0.61	0.47
7	MC	1	75,219	0.80	0.45
8	XI	1	74,953	0.32	0.59
9	XI	1	74,997	0.30	0.55
10	XI	1	75,068	0.34	0.61
11	MC	1	75,196	0.23	0.28
12	XI	1	75,018	0.20	0.28
13	XI	1	75,010	0.09	0.46
14	MC	1	75,170	0.32	0.35
15	MC	1	75,175	0.35	0.25
16	XI	1	75,020	0.28	0.66
17	MC	1	75,178	0.71	0.46
18	XI	1	75,165	0.47	0.61
19	MC	1	75,166	0.49	0.27
20	XI	1	75,115	0.53	0.54
21	MC	1	75,185	0.69	0.47
22	XI	1	74,848	0.33	0.58
23	MC	1	75,169	0.70	0.45
24	XI	1	75,050	0.59	0.65
25	XI	1	75,146	0.51	0.63
26	MC	1	75,243	0.46	0.32
27	XI	1	75,161	0.51	0.69
28	MC	1	75,236	0.39	0.50
29	XI	1	75,055	0.20	0.58
30	XI	1	75,187	0.51	0.53
31	XI	1	75,075	0.34	0.51
32	MC	1	75,223	0.34	0.51
33	XI	1	75,114	0.45	0.58
34	MC	1	75,200	0.28	0.63
35	XI	1	75,047	0.12	0.48
36	MC	1	75,214	0.35	0.46
37	MC	1	75,202	0.35	0.20
38	XI	1	74,936	0.27	0.61
39	XI	1	75,034	0.34	0.65
40	MC	1	75,175	0.28	0.47
41	XI	1	75,083	0.28	0.66
42	MC	1	75,180	0.42	0.48
43	XI	1	75,055	0.28	0.64
44	XI	1	75,102	0.43	0.58
45	MC	1	75,206	0.67	0.51

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
46	MC	1	75,187	0.60	0.41
47	XI	1	75,119	0.18	0.53

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.11. Item-Level CTT Statistics, Mathematics Grade 7

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	78,683	0.54	0.36
2	XI	1	78,263	0.20	0.57
3	XI	1	78,461	0.45	0.56
4	XI	1	78,502	0.48	0.64
5	MC	1	78,641	0.65	0.48
6	MC	1	78,622	0.37	0.36
7	XI	1	78,426	0.42	0.58
8	XI	1	78,473	0.40	0.50
9	XI	1	78,399	0.39	0.56
10	MC	1	78,587	0.54	0.45
11	MC	1	78,588	0.52	0.39
12	XI	1	78,422	0.34	0.62
13	MC	1	78,588	0.41	0.46
14	XI	1	78,430	0.19	0.45
15	XI	1	78,460	0.48	0.66
16	XI	1	78,325	0.26	0.66
17	XI	1	78,344	0.14	0.52
18	XI	1	78,302	0.48	0.54
19	MC	1	78,596	0.73	0.40
20	MC	1	78,575	0.38	0.30
21	XI	1	78,363	0.35	0.65
22	MC	1	78,559	0.66	0.48
23	XI	1	78,508	0.58	0.62
24	MC	1	78,697	0.81	0.42
25	XI	1	78,609	0.29	0.59
26	MC	1	78,690	0.78	0.34
27	XI	1	78,449	0.35	0.66
28	XI	1	78,415	0.34	0.49
29	MC	1	78,659	0.43	0.32
30	MC	1	78,664	0.45	0.47
31	MC	1	78,644	0.42	0.28
32	XI	1	77,908	0.05	0.36
33	XI	1	78,269	0.27	0.54
34	XI	1	78,456	0.24	0.60
35	XI	1	78,291	0.17	0.51
36	XI	1	78,285	0.14	0.46
37	MC	1	78,646	0.38	0.46
38	XI	1	78,479	0.29	0.57
39	XI	1	78,279	0.16	0.55

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
40	XI	1	78,484	0.56	0.62
41	XI	1	78,516	0.28	0.62
42	MC	1	78,637	0.71	0.28
43	MC	1	78,622	0.62	0.39
44	MC	1	78,613	0.53	0.40
45	MC	1	78,623	0.52	0.41
46	MC	1	78,622	0.70	0.57
47	MC	1	78,613	0.39	0.33

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.12. Item-Level CTT Statistics, Mathematics Grade 8

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
1	MC	1	82,288	0.36	0.41
2	MC	1	82,273	0.45	0.47
3	MC	1	82,299	0.64	0.28
4	XI	1	82,104	0.50	0.61
5	MC	1	82,294	0.43	0.40
6	MC	1	82,270	0.31	0.59
7	XI	1	82,037	0.67	0.52
8	MC	1	82,273	0.31	0.40
9	XI	1	81,979	0.18	0.46
10	MC	1	82,247	0.27	0.46
11	MC	1	82,248	0.26	0.29
12	MC	1	82,234	0.27	0.44
13	MC	1	82,230	0.29	0.25
14	XI	1	81,899	0.28	0.62
15	MC	1	82,240	0.40	0.37
16	MC	1	82,230	0.40	0.47
17	XI	1	81,993	0.15	0.52
18	MC	1	82,236	0.48	0.43
19	MC	1	82,232	0.55	0.47
20	XI	1	81,700	0.21	0.63
21	MC	1	82,221	0.24	0.46
22	MC	1	82,246	0.60	0.43
23	XI	1	82,145	0.33	0.60
24	XI	1	81,273	0.47	0.54
25	MC	1	82,295	0.36	0.39
26	MC	1	82,283	0.49	0.48
27	MC	1	82,276	0.42	0.36
28	MC	1	82,264	0.39	0.32
29	MC	1	82,277	0.45	0.38
30	MC	1	82,265	0.33	0.45
31	MC	1	82,270	0.30	0.33
32	MC	1	82,279	0.28	0.47
33	XI	1	82,027	0.20	0.47

Item Number	Item Type	Max. Points	N	*P*-Value	Point-Biserial
34	MC	1	82,263	0.34	0.45
35	MC	1	82,258	0.34	0.40
36	XI	1	82,219	0.38	0.49
37	MC	1	82,220	0.18	0.63
38	MC	1	82,253	0.41	0.49
39	XI	1	81,943	0.30	0.61
40	MC	1	82,242	0.25	0.52
41	MC	1	82,246	0.45	0.35
42	MC	1	82,245	0.46	0.38
43	XI	1	82,074	0.36	0.66
44	XI	1	81,645	0.31	0.59
45	XI	1	81,543	0.20	0.53
46	XI	1	81,842	0.16	0.58
47	MC	1	82,215	0.47	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.13. Distractor Analysis of Multiple-Choice Items, ELA Grade 3

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
4	70.0	0.42	15.4	-0.20	5.9	-0.26	8.7	-0.21
5	54.0	0.41	16.5	-0.29	19.3	-0.12	10.3	-0.16
7	59.3	0.48	13.7	-0.20	13.1	-0.26	14.0	-0.23
8	62.2	0.57	15.0	-0.31	10.5	-0.30	12.2	-0.22
9	41.4	0.37	19.1	-0.20	17.9	-0.18	21.6	-0.09
10	66.4	0.58	12.7	-0.29	13.0	-0.31	7.9	-0.28
11	60.6	0.51	13.0	-0.20	17.2	-0.28	9.2	-0.27
14	33.3	0.27	27.9	-0.06	17.1	-0.15	21.7	-0.10
19	71.8	0.43	13.4	-0.28	4.9	-0.22	9.8	-0.17
21	76.1	0.45	8.6	-0.19	8.6	-0.28	6.7	-0.24
23	45.1	0.37	22.1	-0.15	15.0	-0.24	17.8	-0.09
25	55.1	0.48	13.7	-0.29	22.6	-0.17	8.6	-0.23
27	42.4	0.51	22.0	-0.17	19.4	-0.20	16.1	-0.28
28	47.5	0.37	12.4	-0.30	26.7	-0.04	13.3	-0.19
29	49.8	0.53	22.3	-0.32	13.8	-0.21	14.1	-0.18
30	40.3	0.37	22.4	-0.11	25.8	-0.19	11.5	-0.16
32	34.7	0.43	28.1	-0.06	17.6	-0.21	19.6	-0.25
33	38.4	0.26	30.1	-0.07	19.1	-0.10	12.4	-0.17
34	58.7	0.48	14.3	-0.27	16.3	-0.24	10.7	-0.17
35	50.8	0.41	13.5	-0.23	26.3	-0.14	9.3	-0.21
37	62.0	0.59	19.4	-0.35	11.3	-0.29	7.3	-0.22
38	36.1	0.43	21.0	-0.10	22.5	-0.15	20.3	-0.27
39	28.3	0.36	20.2	-0.11	28.4	-0.11	23.1	-0.17
41	27.2	0.38	30.5	-0.02	23.5	-0.17	18.8	-0.22

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.14. Distractor Analysis of Multiple-Choice Items, ELA Grade 4

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
4	66.6	0.43	17.3	-0.30	10.7	-0.11	5.3	-0.25
5	52.4	0.40	26.1	-0.18	5.4	-0.20	16.1	-0.20
7	50.4	0.42	22.7	-0.13	17.5	-0.29	9.4	-0.17
8	50.6	0.38	11.2	-0.20	21.8	-0.16	16.4	-0.17
9	63.4	0.41	14.6	-0.22	12.9	-0.21	9.1	-0.17
10	41.7	0.30	16.7	-0.19	17.6	-0.22	24.1	0.01
12	32.8	0.22	18.2	-0.04	35.0	-0.09	13.9	-0.12
13	48.9	0.40	21.4	-0.24	13.9	-0.28	15.9	-0.02
17	60.6	0.35	5.3	-0.24	29.6	-0.18	4.5	-0.17
19	46.8	0.34	12.3	-0.19	30.1	-0.11	10.8	-0.18
20	79.5	0.43	2.6	-0.17	4.6	-0.25	13.3	-0.27
21	57.5	0.44	9.3	-0.22	19.6	-0.20	13.7	-0.21
22	55.3	0.44	11.5	-0.20	25.0	-0.17	8.2	-0.30
23	70.2	0.52	13.4	-0.29	8.9	-0.25	7.6	-0.25
24	50.3	0.54	16.0	-0.17	17.0	-0.27	16.7	-0.29
25	45.0	0.42	25.3	-0.09	14.1	-0.28	15.6	-0.20
27	58.1	0.45	11.9	-0.31	18.0	-0.17	12.1	-0.17
29	45.3	0.46	14.5	-0.16	16.1	-0.28	24.1	-0.17
30	46.4	0.34	13.4	-0.26	23.8	-0.10	16.3	-0.11
31	48.9	0.51	20.1	-0.19	18.1	-0.27	12.9	-0.22
32	61.5	0.59	12.5	-0.32	13.3	-0.28	12.7	-0.27
33	49.8	0.48	18.7	-0.22	6.5	-0.29	25.0	-0.19
35	83.2	0.49	7.5	-0.27	5.0	-0.29	4.2	-0.25
36	57.6	0.38	8.3	-0.20	18.4	-0.18	15.8	-0.17
37	39.1	0.40	27.6	-0.26	21.2	-0.11	12.1	-0.12
41	61.1	0.40	20.8	-0.13	9.3	-0.31	8.9	-0.19

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.15. Distractor Analysis of Multiple-Choice Items, ELA Grade 5

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
5	54.7	0.37	9.6	-0.25	15.9	-0.24	19.9	-0.05
8	40.4	0.42	21.3	-0.18	27.4	-0.17	10.9	-0.17
13	56.5	0.41	14.6	-0.17	13.1	-0.29	15.8	-0.13
14	52.5	0.40	21.8	-0.05	15.3	-0.29	10.5	-0.23
16	43.2	0.35	20.9	-0.14	14.2	-0.28	21.6	-0.05
17	48.2	0.43	19.8	-0.29	22.6	-0.15	9.4	-0.13
18	43.4	0.42	20.3	-0.16	15.0	-0.22	21.2	-0.16
23	51.4	0.33	18.0	-0.22	18.5	-0.14	12.0	-0.08
24	42.4	0.37	23.9	-0.11	18.1	-0.21	15.5	-0.14
26	59.5	0.38	11.7	-0.12	8.2	-0.32	20.6	-0.15
27	78.7	0.44	7.5	-0.24	8.3	-0.30	5.5	-0.15
29	39.3	0.29	10.7	-0.19	12.0	-0.30	38.0	0.03
30	68.1	0.49	5.6	-0.25	11.4	-0.27	14.9	-0.23
31	66.7	0.54	14.3	-0.33	7.7	-0.33	11.3	-0.17
33	65.8	0.49	6.3	-0.11	17.9	-0.33	10.0	-0.27
34	48.5	0.39	22.1	-0.07	13.9	-0.35	15.5	-0.12
35	57.2	0.55	20.6	-0.25	13.0	-0.26	9.1	-0.28
38	63.6	0.53	10.3	-0.17	16.2	-0.26	10.0	-0.36
41	54.2	0.42	12.6	-0.34	16.8	-0.21	16.4	-0.04

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.16. Distractor Analysis of Multiple-Choice Items, ELA Grade 6

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
4	78.4	0.45	11.2	-0.28	5.8	-0.28	4.5	-0.16
5	64.4	0.44	13.5	-0.28	11.1	-0.11	11.0	-0.26
6	69.1	0.38	5.3	-0.21	8.1	-0.30	17.5	-0.13
10	69.7	0.43	9.5	-0.17	10.4	-0.25	10.4	-0.23
11	46.9	0.35	19.1	-0.18	23.3	-0.12	10.7	-0.18
13	63.3	0.47	6.6	-0.28	14.1	-0.27	15.9	-0.17
14	38.5	0.34	12.0	-0.12	18.2	-0.21	31.4	-0.10
15	73.1	0.52	9.0	-0.22	8.7	-0.30	9.3	-0.30
19	57.5	0.34	12.7	-0.28	8.9	-0.33	20.9	0.04
23	55.8	0.35	15.7	-0.11	23.2	-0.22	5.3	-0.20
26	44.2	0.35	13.9	-0.20	36.0	-0.10	6.0	-0.23
28	52.2	0.39	13.7	-0.17	14.3	-0.27	19.8	-0.11
29	53.6	0.32	7.8	-0.24	26.0	-0.10	12.6	-0.16
30	34.7	0.26	26.7	0.05	16.7	-0.21	21.9	-0.16
32	50.8	0.35	13.7	-0.15	13.7	-0.26	21.7	-0.08
33	58.2	0.44	20.2	-0.12	15.5	-0.30	6.1	-0.23
34	43.5	0.32	16.5	-0.17	26.6	-0.04	13.5	-0.24
35	47.5	0.41	13.2	-0.08	20.9	-0.19	18.4	-0.25
36	42.6	0.35	10.5	-0.25	23.0	-0.20	23.9	-0.03
37	46.3	0.31	9.8	-0.22	24.8	-0.13	19.2	-0.08
39	41.2	0.32	23.4	-0.08	25.9	-0.14	9.5	-0.22
40	31.5	0.34	17.5	-0.22	41.6	-0.06	9.4	-0.16
41	42.7	0.45	13.0	-0.22	28.5	-0.13	15.8	-0.24

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.17. Distractor Analysis of Multiple-Choice Items, ELA Grade 7

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
4	69.7	0.38	19.4	-0.19	5.7	-0.23	5.2	-0.21
5	60.3	0.43	10.2	-0.22	10.7	-0.25	18.9	-0.17
6	61.8	0.31	15.5	-0.05	7.5	-0.28	15.3	-0.16
7	71.0	0.54	4.6	-0.22	11.4	-0.31	13.0	-0.29
8	50.7	0.49	12.9	-0.23	19.6	-0.29	16.8	-0.15
10	53.0	0.41	14.1	-0.18	21.4	-0.19	11.5	-0.21
12	62.9	0.42	11.3	-0.24	13.3	-0.28	12.5	-0.09
13	45.2	0.35	30.7	-0.06	12.3	-0.28	11.8	-0.17
14	47.2	0.42	12.2	-0.22	12.6	-0.27	27.9	-0.10
15	57.1	0.39	15.7	-0.16	16.7	-0.23	10.5	-0.16
19	29.1	0.26	12.0	-0.22	35.3	-0.08	23.6	-0.02
20	65.9	0.44	14.6	-0.21	12.1	-0.26	7.4	-0.20
22	43.1	0.31	14.4	-0.19	17.8	-0.23	24.6	0.00
23	52.5	0.38	15.1	-0.20	20.2	-0.12	12.2	-0.22
24	78.1	0.49	5.4	-0.25	8.4	-0.31	8.1	-0.22
25	76.2	0.45	14.3	-0.26	5.6	-0.25	3.9	-0.23
29	47.3	0.47	11.1	-0.15	15.0	-0.25	26.6	-0.22
30	49.0	0.49	14.9	-0.15	16.7	-0.21	19.4	-0.29
31	61.6	0.57	16.3	-0.30	17.5	-0.34	4.6	-0.18
32	38.5	0.30	14.3	-0.30	27.0	-0.04	20.2	-0.06
33	53.6	0.42	16.0	-0.14	19.5	-0.27	10.9	-0.17
34	55.9	0.40	10.1	-0.17	18.7	-0.19	15.3	-0.20
35	61.6	0.54	11.5	-0.29	14.7	-0.26	12.2	-0.25
36	70.5	0.54	12.1	-0.31	9.9	-0.26	7.5	-0.25
38	54.9	0.43	18.4	-0.16	11.2	-0.25	15.5	-0.20
40	39.5	0.34	23.9	-0.14	21.8	-0.24	14.9	-0.02

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.18. Distractor Analysis of Multiple-Choice Items, ELA Grade 8

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
4	47.3	0.44	14.9	-0.19	17.6	-0.30	20.2	-0.10
5	63.6	0.39	12.9	-0.15	9.2	-0.23	14.3	-0.20
6	52.1	0.46	20.4	-0.22	17.6	-0.22	9.9	-0.19
7	60.9	0.41	9.3	-0.26	22.6	-0.15	7.3	-0.24
8	52.3	0.31	26.6	-0.04	12.7	-0.30	8.4	-0.14
9	45.0	0.42	29.4	-0.18	13.0	-0.16	12.6	-0.22
10	36.4	0.30	18.1	-0.12	25.3	-0.13	20.2	-0.12
11	54.7	0.47	12.3	-0.22	14.8	-0.24	18.2	-0.20
12	44.0	0.38	19.0	-0.18	23.9	-0.09	13.1	-0.23
13	44.7	0.38	19.7	-0.08	20.5	-0.20	15.1	-0.21
14	41.1	0.38	10.0	-0.26	31.6	-0.14	17.3	-0.11
15	42.6	0.40	14.6	-0.28	21.0	-0.16	21.7	-0.09
18	58.9	0.54	12.4	-0.23	10.3	-0.32	18.5	-0.24
20	64.6	0.32	21.5	-0.01	10.8	-0.38	3.1	-0.20
21	72.4	0.56	5.9	-0.24	9.2	-0.25	12.4	-0.38
22	18.8	0.19	15.1	-0.30	12.2	-0.24	53.9	0.22
23	66.7	0.55	11.6	-0.29	12.2	-0.29	9.5	-0.25
27	77.8	0.50	9.3	-0.29	5.9	-0.27	7.0	-0.22
28	48.2	0.30	9.7	-0.18	16.9	-0.20	25.2	-0.05
29	43.3	0.48	17.1	-0.25	22.6	-0.18	17.1	-0.19
30	59.5	0.45	9.9	-0.19	15.0	-0.21	15.6	-0.25
31	41.7	0.33	13.4	-0.23	22.0	-0.13	22.9	-0.08
32	34.5	0.27	11.0	-0.19	28.3	-0.09	26.2	-0.07
33	46.9	0.41	14.9	-0.15	22.7	-0.09	15.5	-0.31
34	38.7	0.25	22.0	-0.07	27.6	-0.09	11.7	-0.16
36	42.4	0.40	18.7	-0.12	25.8	-0.18	13.1	-0.22

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.19. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 3

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
1	86.5	0.31	6.0	-0.22	2.7	-0.15	4.8	-0.13	–	–
3	63.3	0.52	6.0	-0.18	14.2	-0.29	6.6	-0.16	9.9	-0.22
6	62.5	0.55	15.7	-0.30	17.2	-0.32	4.6	-0.17	–	–
8	58.3	0.34	21.9	-0.19	10.4	-0.22	9.4	-0.08	–	–
12	31.5	0.29	32.9	-0.10	15.8	-0.14	19.8	-0.09	–	–
13	37.6	0.42	34.1	-0.24	5.5	-0.21	22.8	-0.10	–	–
15	38.8	0.37	24.8	-0.21	16.5	-0.13	9.3	-0.07	10.6	-0.06
16	42.0	0.52	20.3	-0.29	22.3	-0.18	15.5	-0.18	–	–
18	49.5	0.44	17.6	-0.32	9.1	-0.11	18.1	-0.08	5.7	-0.15
19	57.8	0.56	15.6	-0.31	11.1	-0.29	10.9	-0.10	4.6	-0.20
20	36.5	0.46	19.8	-0.22	9.7	-0.22	27.9	-0.07	6.1	-0.14
24	74.3	0.43	14.3	-0.31	7.1	-0.18	4.3	-0.16	–	–
25	59.5	0.50	15.3	-0.16	17.6	-0.30	4.3	-0.19	3.4	-0.20
29	54.8	0.57	24.4	-0.36	15.8	-0.28	5.0	-0.11	–	–
30	75.1	0.53	9.2	-0.33	10.6	-0.27	5.0	-0.24	–	–
35	31.3	0.20	19.4	-0.07	15.8	-0.16	16.0	-0.02	17.5	-0.01
37	59.3	0.37	23.4	-0.22	6.5	-0.15	10.9	-0.16	–	–
39	56.8	0.28	18.2	-0.23	20.6	-0.07	4.4	-0.09	–	–
41	36.4	0.65	13.5	-0.27	9.2	-0.15	34.8	-0.35	6.1	-0.05
43	68.3	0.59	19.6	-0.47	5.6	-0.28	6.5	-0.09	–	–

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.20. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 4

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
7	40.2	0.37	29.3	-0.19	15.8	-0.16	14.7	-0.09	–	–
8	38.0	0.36	28.8	-0.15	7.5	-0.14	6.7	-0.18	19.0	-0.07
11	79.5	0.51	4.3	-0.27	7.3	-0.25	2.1	-0.18	6.9	-0.24
16	43.2	0.44	23.5	-0.13	16.3	-0.27	17.0	-0.17	–	–
20	68.9	0.60	8.9	-0.36	10.5	-0.34	11.7	-0.23	–	–
23	55.3	0.60	12.9	-0.28	19.4	-0.32	12.3	-0.24	–	–
24	79.7	0.47	16.4	-0.40	2.7	-0.19	1.1	-0.11	–	–
27	69.5	0.50	8.0	-0.24	16.1	-0.32	6.4	-0.19	–	–
35	55.2	0.54	10.4	-0.26	17.3	-0.17	9.1	-0.27	7.9	-0.19
41	80.1	0.42	7.4	-0.28	7.0	-0.25	5.6	-0.14	–	–
42	42.3	0.65	39.6	-0.51	8.5	-0.16	3.3	-0.07	6.4	-0.05
44	37.2	0.30	25.3	-0.03	14.3	-0.19	9.6	-0.17	13.7	-0.05

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.21. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 5

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
3	50.3	0.43	18.4	-0.17	22.5	-0.24	8.8	-0.16	–	–
4	41.1	0.44	9.9	-0.17	29.3	-0.12	11.5	-0.19	8.2	-0.20
9	36.7	0.30	14.4	-0.08	17.2	-0.16	15.0	-0.16	16.7	0.01
10	45.9	0.51	18.4	-0.30	27.0	-0.25	8.7	-0.10	–	–
15	36.1	0.49	17.4	-0.28	16.7	-0.26	29.9	-0.07	–	–
22	25.4	0.25	13.2	0.00	39.9	-0.14	14.8	-0.13	6.7	0.03
24	68.5	0.37	26.6	-0.29	3.5	-0.17	1.4	-0.12	–	–
31	31.6	0.48	13.4	-0.16	24.4	-0.33	14.5	-0.11	16.1	0.03
36	45.6	0.45	31.0	-0.24	15.9	-0.17	7.5	-0.19	–	–
37	31.6	0.58	26.9	-0.05	29.9	-0.39	7.1	-0.17	4.5	-0.13
39	41.5	0.36	21.2	-0.09	8.4	-0.16	19.1	-0.14	9.8	-0.13
40	56.8	0.50	9.7	-0.20	21.9	-0.27	11.6	-0.24	–	–

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.22. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 6

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
2	62.8	0.49	9.0	-0.18	11.2	-0.22	9.8	-0.22	7.2	-0.19
3	63.8	0.27	20.5	-0.17	5.6	-0.16	10.1	-0.08	–	–
4	22.7	0.36	20.3	0.01	13.1	-0.12	4.3	-0.10	39.5	-0.19
6	61.2	0.47	23.1	-0.32	6.1	-0.23	9.6	-0.14	–	–
7	80.4	0.45	7.4	-0.23	5.9	-0.23	6.3	-0.27	–	–
11	23.1	0.28	32.7	-0.18	19.8	-0.09	24.4	0.00	–	–
14	31.7	0.35	14.0	-0.07	19.3	-0.12	27.9	-0.16	7.1	-0.09
15	34.5	0.25	14.1	-0.08	23.4	-0.21	28.0	0.00	–	–
17	70.8	0.46	10.4	-0.30	7.8	-0.28	11.1	-0.14	–	–
19	49.2	0.27	19.0	-0.06	13.9	-0.30	17.9	-0.01	–	–
21	68.9	0.47	8.3	-0.21	10.7	-0.25	12.1	-0.26	–	–
23	70.2	0.45	14.9	-0.19	8.4	-0.28	6.5	-0.25	–	–
26	45.8	0.32	24.6	-0.04	16.5	-0.27	13.1	-0.13	–	–
28	39.4	0.50	9.7	-0.19	5.8	-0.18	4.1	-0.12	41.0	-0.25
36	34.9	0.46	25.9	-0.39	28.0	-0.03	11.3	-0.10	–	–
37	34.9	0.21	13.2	-0.11	25.4	-0.12	11.5	-0.10	14.9	0.06
42	42.1	0.48	16.0	-0.11	29.8	-0.31	12.1	-0.17	–	–
45	66.9	0.51	7.0	-0.22	11.0	-0.30	15.1	-0.25	–	–
46	60.4	0.41	17.9	-0.25	17.2	-0.20	4.6	-0.15	–	–

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.23. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 7

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
1	53.7	0.36	14.4	-0.17	20.1	-0.17	11.8	-0.16	–	–
5	65.3	0.48	17.3	-0.26	14.1	-0.30	3.3	-0.16	–	–
6	36.9	0.36	13.6	0.03	16.3	-0.16	26.1	-0.18	7.1	-0.18
10	53.6	0.45	11.0	-0.23	20.8	-0.15	9.8	-0.21	4.8	-0.13
11	51.7	0.39	11.5	-0.19	30.1	-0.28	6.7	-0.03	–	–
13	40.7	0.46	20.2	0.04	16.7	-0.25	18.6	-0.30	3.8	-0.15
19	72.5	0.40	12.3	-0.27	7.7	-0.24	7.5	-0.10	–	–
20	37.7	0.30	26.8	-0.13	26.1	-0.18	9.4	-0.03	–	–
22	65.6	0.48	3.3	-0.15	25.3	-0.34	5.7	-0.22	–	–
24	80.9	0.42	12.8	-0.31	3.1	-0.19	3.1	-0.16	–	–
26	77.9	0.34	6.0	-0.17	8.3	-0.20	7.8	-0.18	–	–
29	43.4	0.32	18.8	-0.08	34.6	-0.25	3.3	-0.06	–	–
30	44.8	0.47	22.8	-0.17	22.7	-0.26	9.8	-0.18	–	–
31	41.6	0.28	18.1	-0.10	23.5	-0.21	16.8	-0.03	–	–
37	38.4	0.46	25.6	-0.23	15.4	-0.18	20.6	-0.14	–	–
42	71.1	0.28	1.6	-0.08	5.2	-0.14	22.1	-0.21	–	–
43	62.4	0.39	10.0	-0.21	19.6	-0.20	8.0	-0.18	–	–
44	52.8	0.40	11.3	-0.22	25.0	-0.17	11.0	-0.17	–	–
45	52.4	0.41	11.8	-0.26	16.3	-0.27	19.5	-0.05	–	–
46	69.7	0.57	13.9	-0.31	7.2	-0.27	9.2	-0.30	–	–
47	39.3	0.33	27.2	-0.24	26.1	-0.08	7.4	-0.09	–	–

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.24. Distractor Analysis of Multiple-Choice Items, Mathematics Grade 8

Item Number	Correct Option		Distractor 1		Distractor 2		Distractor 3		Distractor 4	
	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.	%	Pt. Bis.
1	36.4	0.41	29.0	-0.13	16.6	-0.23	18.1	-0.14	–	–
2	45.2	0.47	6.6	-0.11	19.0	-0.21	19.1	-0.19	10.1	-0.18
3	64.1	0.28	26.1	-0.17	5.1	-0.17	4.7	-0.11	–	–
5	43.1	0.40	13.5	-0.16	14.7	-0.21	28.7	-0.16	–	–
6	30.6	0.59	16.2	-0.19	33.2	-0.26	7.5	-0.15	12.5	-0.12
8	31.2	0.40	28.2	-0.14	21.7	-0.22	19.0	-0.08	–	–
10	27.3	0.46	16.5	-0.10	12.7	-0.20	25.9	-0.11	17.7	-0.14
11	26.2	0.29	41.3	-0.14	28.1	-0.12	4.5	-0.03	–	–
12	26.8	0.45	17.3	-0.23	36.7	-0.12	7.6	-0.12	11.5	-0.06
13	28.9	0.25	21.7	-0.24	15.5	-0.18	9.7	-0.11	24.2	0.20
15	39.7	0.37	29.3	-0.18	19.6	-0.19	11.4	-0.08	–	–
16	40.0	0.47	10.6	-0.19	28.9	-0.17	12.6	-0.17	7.9	-0.17
18	48.1	0.43	11.7	-0.05	21.0	-0.26	10.5	-0.22	8.7	-0.09
19	54.7	0.47	13.8	-0.25	17.4	-0.21	9.8	-0.16	4.2	-0.10
21	23.6	0.46	7.0	-0.15	38.0	-0.30	26.2	-0.03	5.2	0.00
22	59.5	0.43	5.2	-0.12	10.9	-0.22	15.8	-0.15	8.5	-0.21
25	36.2	0.39	15.5	-0.15	22.9	-0.10	13.4	-0.12	12.1	-0.16
26	48.5	0.48	21.0	-0.11	10.0	-0.24	15.1	-0.27	5.4	-0.12
27	42.1	0.36	12.3	-0.19	28.8	-0.19	16.8	-0.08	–	–
28	39.3	0.32	9.6	-0.12	30.9	-0.24	20.1	-0.03	–	–
29	45.0	0.38	15.8	-0.17	22.5	-0.20	16.7	-0.11	–	–
30	33.4	0.45	18.9	-0.21	24.7	-0.13	23.0	-0.18	–	–
31	29.9	0.33	15.7	-0.07	19.7	-0.05	18.2	-0.19	16.5	-0.09
34	34.4	0.45	19.3	-0.18	22.4	-0.17	23.9	-0.17	–	–
35	33.8	0.40	26.3	-0.23	30.2	-0.19	9.7	0.00	–	–
38	41.1	0.49	13.3	-0.14	22.3	-0.20	23.2	-0.26	–	–
40	24.9	0.52	12.6	-0.14	21.0	-0.24	12.3	-0.24	29.3	-0.01
41	45.2	0.35	12.4	-0.07	24.8	-0.27	17.6	-0.09	–	–
42	46.1	0.38	20.7	-0.20	21.1	-0.13	12.1	-0.17	–	–
47	47.4	0.47	18.5	-0.20	25.7	-0.30	8.4	-0.10	–	–

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 statistics:

- Table B.1 – Table B.12 present the IRT statistics, including item type, Rasch difficulty, standard error (SE) of Rasch, and infit values.
- Table B.13 – Table B.24 present the raw-to-scale score conversion tables.
- Figure B.1 – Figure B.12 present the item-person map for each post-equated operational form.
- Figure B.13 – Figure B.36 present the test characteristic curve (TCC) and conditional standard error of measurement (CSEM) curve for each post-equated operational form.
- Figure B.37 – Figure B.48 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, ELA Grade 3

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	1.7464	0.0060	0.77
2	OE	1.7032	0.0061	0.79
3	OE	-0.5928	0.0057	0.80
4	MC	-1.0914	0.0089	0.99
5	MC	-0.2251	0.0083	1.07
6	MX	0.9109	0.0088	0.98
7	MC	-0.5026	0.0084	0.96
8	MC	-0.6569	0.0085	0.85
9	MC	0.2279	0.0084	1.12
10	MC	-0.8899	0.0087	0.82
11	MC	-0.5722	0.0085	0.92
12	MC	1.0918	0.0093	1.08
13	MX	1.5143	0.0097	1.07
14	MC	0.7924	0.0087	1.20
15	XI	-0.3018	0.0084	0.98
16	XI	-2.0213	0.0107	0.94
17	XI	-0.6709	0.0058	0.78
18	XI	-1.1756	0.0090	1.00
19	MC	-1.2038	0.0090	0.99
20	MX	1.8785	0.0105	1.12
21	MC	-1.2578	0.0091	0.86
22	MC	1.6354	0.0101	1.08
23	MC	0.2325	0.0084	1.13
24	XI	0.9551	0.0089	0.94
25	MC	-0.2860	0.0083	0.99
26	MX	1.5892	0.0098	0.84
27	MC	0.3728	0.0084	0.94
28	MC	-0.1070	0.0083	1.13
29	MC	0.2382	0.0084	0.95
30	MC	0.4854	0.0085	1.11

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
31	MX	1.7682	0.0102	1.05
32	MC	0.7947	0.0087	1.03
33	MC	0.3615	0.0084	1.24
34	MC	-0.6059	0.0085	0.99
35	MC	-0.0629	0.0083	1.08
36	MX	1.1496	0.0091	0.89
37	MC	-0.6467	0.0085	0.83
38	MC	0.7163	0.0087	1.04
39	MC	0.9817	0.0089	1.03
40	MX	2.3215	0.0117	1.10
41	MC	1.2425	0.0093	1.04
42	XI	-1.1878	0.0090	0.97
43	XI	0.2835	0.0059	1.01
44	XI	-0.6979	0.0061	0.88

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, ELA Grade 4

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	2.8463	0.0072	0.80
2	OE	3.1511	0.0074	0.84
3	OE	0.1960	0.0061	0.79
4	MC	-0.4466	0.0086	0.97
5	MC	0.2970	0.0082	1.07
6	MX	1.2824	0.0087	1.16
7	MC	0.4019	0.0082	1.04
8	MC	0.3901	0.0082	1.09
9	MC	-0.2710	0.0085	1.03
10	MC	0.8749	0.0084	1.19
11	MX	0.9695	0.0056	1.46
12	MC	1.3307	0.0088	1.26
13	MC	0.2998	0.0083	1.07
14	XI	0.9022	0.0084	1.14
15	XI	-0.4609	0.0063	0.99
16	XI	-0.4217	0.0068	1.06
17	MC	-0.1222	0.0084	1.11
18	MC	0.8056	0.0085	0.96
19	MC	0.5815	0.0083	1.15
20	MC	-1.2374	0.0098	0.93
21	MC	0.2231	0.0083	1.01
22	MC	0.1503	0.0083	1.01
23	MC	-0.6438	0.0088	0.88
24	MC	0.4060	0.0083	0.90
25	MC	0.6750	0.0083	1.05
26	MC	-0.1881	0.0085	0.84
27	MC	-0.1009	0.0084	1.01

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
28	MX	0.5217	0.0053	1.13
29	MC	0.6853	0.0083	1.00
30	MC	0.6029	0.0083	1.14
31	MC	0.4876	0.0083	0.94
32	MC	-0.1719	0.0084	0.83
33	MC	0.4293	0.0083	0.97
34	XI	0.6517	0.0083	0.75
35	MC	-1.5209	0.0105	0.83
36	MC	0.0345	0.0083	1.08
37	MC	0.7343	0.0083	1.04
38	MC	1.3287	0.0089	0.97
39	MX	1.9051	0.0096	0.90
40	MC	0.9234	0.0086	0.81
41	MC	-0.1457	0.0084	1.05
42	XI	-0.6218	0.0088	0.90
43	XI	-0.6247	0.0061	0.89
44	XI	-0.3391	0.0085	0.93

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, ELA Grade 5

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	1.6162	0.0059	0.77
2	OE	1.8724	0.0062	0.80
3	OE	-0.6838	0.0063	0.78
4	MC	-1.1909	0.0093	0.88
5	MC	-0.0881	0.0083	1.14
6	MC	0.5748	0.0084	0.96
7	MX	1.5753	0.0094	1.08
8	MC	0.8288	0.0085	1.07
9	XI	0.5023	0.0084	1.03
10	MX	1.3922	0.0091	1.02
11	MC	-0.4448	0.0085	0.94
12	XI	0.4586	0.0084	0.97
13	MC	-0.1823	0.0083	1.08
14	MC	0.0258	0.0083	1.10
15	XI	1.2033	0.0089	0.95
16	MC	0.5083	0.0084	1.14
17	MC	0.2479	0.0083	1.05
18	MC	0.4969	0.0084	1.07
19	XI	-1.5766	0.0099	0.88
20	XI	-1.4163	0.0067	0.91
21	XI	-1.0147	0.0090	0.84
22	MC	0.3303	0.0083	0.99
23	MC	0.0809	0.0083	1.19
24	MC	0.3012	0.0083	1.13

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
25	XI	1.6204	0.0095	1.06
26	MC	-0.3456	0.0084	1.11
27	MC	-1.5020	0.0098	0.93
28	XI	0.0752	0.0083	0.85
29	MC	0.7196	0.0085	1.20
30	MC	-0.7333	0.0087	0.94
31	MC	-0.7411	0.0087	0.89
32	XI	-0.4691	0.0085	0.95
33	MC	-0.6872	0.0086	0.96
34	MC	0.2312	0.0083	1.11
35	MC	-0.2218	0.0084	0.91
36	MC	-0.0183	0.0083	0.84
37	MX	1.8757	0.0100	1.05
38	MC	-0.7338	0.0087	0.95
39	XI	1.8036	0.0098	1.04
40	MC	0.1683	0.0083	0.86
41	MC	-0.1059	0.0083	1.07
42	XI	-0.2395	0.0084	1.08
43	XI	0.6930	0.0084	1.08
44	XI	-0.3126	0.0062	1.16

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. Item-Level IRT Statistics, ELA Grade 6

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	1.2907	0.0054	0.73
2	OE	1.8881	0.0058	0.75
3	OE	-0.5294	0.0060	0.72
4	MC	-1.3381	0.0096	0.90
5	MC	-0.5108	0.0084	0.97
6	MC	-0.7668	0.0087	1.01
7	MC	0.5508	0.0082	0.98
8	MX	1.6526	0.0094	1.10
9	MX	0.2267	0.0081	0.92
10	MC	-0.7990	0.0087	0.97
11	MC	0.3566	0.0081	1.08
12	MC	0.8305	0.0083	1.00
13	MC	-0.4561	0.0084	0.93
14	MC	0.4843	0.0081	1.06
15	MC	-1.0592	0.0091	0.89
16	MX	1.6025	0.0093	0.88
17	MC	0.7713	0.0083	0.85
18	XI	1.8046	0.0097	1.10
19	MC	-0.2555	0.0082	1.08
20	XI	-1.2737	0.0095	1.09
21	XI	-0.0381	0.0064	1.02

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
22	XI	0.0630	0.0059	1.09
23	MC	-0.0765	0.0081	1.07
24	MX	1.6406	0.0094	1.03
25	MX	1.8380	0.0098	0.87
26	MC	0.4923	0.0081	1.08
27	OE	0.0161	0.0081	0.85
28	MC	0.0998	0.0081	1.03
29	MC	-0.1186	0.0081	1.12
30	MC	0.9722	0.0084	1.15
31	MX	0.4475	0.0049	0.94
32	MC	0.3866	0.0081	1.09
33	MC	-0.1934	0.0082	0.98
34	MC	0.5261	0.0081	1.10
35	MC	0.3298	0.0081	1.01
36	MC	0.4663	0.0081	1.06
37	MC	0.3644	0.0081	1.12
38	MX	0.8424	0.0055	1.14
39	MC	0.6423	0.0082	1.10
40	MC	1.1476	0.0086	1.04
41	MC	0.5672	0.0082	0.97
42	XI	0.1085	0.0081	0.96
43	XI	-0.8239	0.0063	1.03
44	XI	-1.2797	0.0095	0.87

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.5. Item-Level IRT Statistics, ELA Grade 7

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	1.6728	0.0057	0.75
2	OE	1.9493	0.0058	0.81
3	OE	-0.9836	0.0063	0.74
4	MC	-0.7242	0.0086	1.04
5	MC	-0.1683	0.0082	1.01
6	MC	-0.2914	0.0083	1.14
7	MC	-0.7991	0.0087	0.86
8	MC	0.2812	0.0081	0.96
9	MC	2.1435	0.0101	1.00
10	MC	-0.0824	0.0082	1.08
11	MC	0.6316	0.0082	0.91
12	MC	-0.2964	0.0083	1.01
13	MC	0.5614	0.0081	1.13
14	MC	0.4573	0.0081	1.05
15	MC	0.2098	0.0081	1.08
16	XI	-0.9896	0.0090	1.00
17	XI	-0.7386	0.0086	0.99
18	XI	-1.1784	0.0093	0.79

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
19	MC	1.4457	0.0089	1.18
20	MC	-0.3577	0.0083	0.96
21	MX	1.8722	0.0095	1.06
22	MC	0.6689	0.0082	1.18
~~23~~	–	–	–	–
23	MC	0.1890	0.0081	1.09
24	MC	-1.2487	0.0094	0.87
25	MC	-1.1217	0.0092	0.91
26	MX	0.7813	0.0082	0.86
27	MC	2.0191	0.0098	1.05
28	MX	1.3607	0.0087	0.84
29	MC	0.5796	0.0081	1.00
30	MC	0.1822	0.0081	0.98
31	MC	-0.2786	0.0082	0.85
32	MC	0.9098	0.0083	1.19
33	MC	0.1344	0.0081	1.04
34	MC	0.0143	0.0081	1.07
35	MC	-0.5070	0.0084	0.95
36	MC	-0.7703	0.0087	0.86
37	MX	0.7974	0.0072	1.07
38	MC	0.1479	0.0081	1.03
39	XI	1.7971	0.0094	1.13
40	MC	0.8596	0.0083	1.13
41	XI	1.0448	0.0084	1.06
42	XI	-0.6537	0.0086	0.97
43	XI	-0.6810	0.0064	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. One item for ELA Grade 7 was omitted from scoring due to an error in the stimulus.

Table B.6. Item-Level IRT Statistics, ELA Grade 8

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	OE	1.4944	0.0055	0.82
2	OE	1.6380	0.0056	0.79
3	OE	-0.5723	0.0057	0.81
4	MC	0.2891	0.0079	1.01
5	MC	-0.3595	0.0080	1.02
6	MC	-0.0717	0.0079	1.00
7	MC	-0.2582	0.0079	1.02
8	MC	0.0395	0.0079	1.15
9	MC	0.3584	0.0079	1.03
10	MC	0.8516	0.0081	1.14
11	MC	-0.0802	0.0079	0.97
12	MC	0.5562	0.0080	1.09
13	MC	0.2403	0.0079	1.07
14	MC	0.6039	0.0080	1.09

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
15	MC	0.5266	0.0079	1.05
16	MC	0.8965	0.0082	0.98
17	XI	0.8547	0.0081	0.93
18	MC	-0.2957	0.0079	0.89
19	XI	0.2037	0.0079	0.89
20	MC	-0.5945	0.0081	1.10
21	MC	-1.0375	0.0086	0.82
22	MC	1.9602	0.0098	1.15
23	MC	-0.7087	0.0082	0.85
24	XI	-1.9917	0.0106	0.87
25	XI	-1.0197	0.0061	0.93
26	XI	-0.7317	0.0061	1.24
27	MC	-1.3775	0.0091	0.85
28	MC	0.2446	0.0079	1.17
29	MC	0.4930	0.0079	0.96
30	MC	-0.3255	0.0080	0.98
31	MC	0.5739	0.0080	1.13
32	MC	0.9586	0.0082	1.19
33	MC	0.3112	0.0079	1.04
34	MC	0.6173	0.0080	1.19
35	MX	1.9038	0.0097	1.12
36	MC	0.5358	0.0079	1.05
37	MC	0.8469	0.0081	1.03
38	MC	1.8458	0.0095	0.81
39	MX	1.9162	0.0097	1.05
40	MC	1.7787	0.0094	0.90
41	MX	1.6357	0.0091	0.95
42	XI	0.0218	0.0079	0.97
43	XI	-0.9939	0.0059	0.84
44	XI	-1.2173	0.0061	0.89

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.7. Item-Level IRT Statistics, Mathematics Grade 3

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-2.0209	0.0118	1.08
2	XI	0.8701	0.0088	0.97
3	MC	-0.2876	0.0090	1.02
4	XI	-1.2342	0.0100	1.15
5	XI	-0.2355	0.0089	0.85
6	MC	-0.2389	0.0089	0.98
7	XI	-2.5135	0.0137	0.93
8	MC	0.1885	0.0088	1.32
9	XI	-0.0200	0.0088	0.93
10	MC	1.3296	0.0091	0.81
11	XI	-0.1982	0.0089	1.00

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
12	MC	1.6179	0.0094	1.33
13	MC	1.2265	0.0090	1.19
14	XI	1.0546	0.0089	0.94
15	MC	1.1536	0.0090	1.27
16	MC	1.0418	0.0089	1.05
17	XI	0.2538	0.0088	0.80
18	MC	0.5299	0.0088	1.18
19	MC	0.0425	0.0088	0.98
20	MC	1.2971	0.0091	1.12
21	MC	1.7598	0.0095	0.95
22	XI	-0.6431	0.0093	0.80
23	XI	-0.5069	0.0091	0.81
24	MC	-1.0498	0.0097	1.09
25	MC	-0.0591	0.0088	1.08
26	XI	-0.5535	0.0092	0.77
27	XI	1.3903	0.0092	0.80
28	XI	-0.3647	0.0090	0.77
29	MC	-0.0694	0.0089	1.02
30	MC	-1.0541	0.0097	0.92
31	XI	3.4836	0.0136	0.89
32	XI	-0.2805	0.0090	0.89
33	MC	1.0490	0.0089	0.89
34	MX	0.9142	0.0089	0.89
35	MC	1.6260	0.0094	1.43
36	XI	-1.2121	0.0100	0.75
37	MC	0.1016	0.0088	1.28
38	XI	1.8770	0.0097	0.94
39	MC	0.1790	0.0088	1.43
40	XI	1.0942	0.0089	0.94
41	MC	1.3037	0.0091	0.80
42	XI	-0.0545	0.0089	0.85
43	MC	-0.7637	0.0094	0.92
44	XI	1.4909	0.0092	0.82
45	XI	3.4992	0.0136	0.88

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.8. Item-Level IRT Statistics, Mathematics Grade 4

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	XI	0.8284	0.0092	0.84
2	XI	0.0581	0.0091	0.93
3	XI	-1.4982	0.0103	0.92
4	XI	2.3966	0.0111	1.17
5	XI	-0.5777	0.0093	1.03
6	XI	-0.5871	0.0093	1.05
7	MC	0.7896	0.0092	1.37

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
8	MC	0.9277	0.0093	1.36
9	XI	0.2926	0.0091	0.96
10	XI	1.7826	0.0101	1.06
11	MC	-1.8874	0.0109	0.97
12	XI	0.6264	0.0092	0.99
13	XI	0.8866	0.0093	0.81
14	MC	1.4015	0.0096	0.78
15	XI	0.4884	0.0091	0.97
16	MC	0.5977	0.0091	1.26
17	MC	0.9731	0.0093	0.98
18	XI	-0.9748	0.0097	0.95
19	XI	2.3918	0.0111	0.86
20	MC	-1.0490	0.0097	0.92
21	XI	0.3382	0.0091	1.21
22	XI	2.7834	0.0121	0.94
23	MC	-0.2944	0.0092	1.00
24	MC	-1.9117	0.0109	1.05
25	XI	-0.4933	0.0093	0.82
26	XI	-0.6800	0.0094	0.88
27	MC	-0.8873	0.0096	1.09
28	XI	0.3307	0.0091	0.94
29	XI	2.5611	0.0115	1.00
30	XI	-0.0977	0.0091	0.97
31	XI	0.8235	0.0092	0.86
32	XI	0.7836	0.0092	0.84
33	MC	1.8549	0.0102	0.88
34	XI	-1.9376	0.0110	0.77
35	MC	-0.1515	0.0091	1.10
36	XI	-1.2685	0.0100	1.10
37	XI	-1.5010	0.0103	0.85
38	MC	1.6863	0.0100	0.88
39	XI	0.9581	0.0093	1.08
40	XI	-0.5645	0.0093	0.84
41	MC	-1.9371	0.0110	1.16
42	MC	0.6589	0.0092	0.87
43	XI	0.1922	0.0091	1.15
44	MC	0.9851	0.0093	1.49
45	XI	1.0134	0.0093	0.79

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.9. Item-Level IRT Statistics, Mathematics Grade 5

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	XI	-0.7129	0.0087	1.12
2	XI	0.1010	0.0087	0.99
3	MC	-0.1505	0.0087	1.20

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
4	MC	0.2765	0.0088	1.17
5	XI	-0.0308	0.0087	0.80
6	XI	1.7796	0.0107	1.02
7	XI	-1.2771	0.0091	0.83
8	XI	-0.0003	0.0087	1.16
9	MC	0.5431	0.0090	1.39
10	MC	-0.0011	0.0087	1.06
11	XI	-0.2395	0.0087	0.91
12	XI	0.6516	0.0091	0.86
13	XI	1.0428	0.0095	0.75
14	XI	0.1633	0.0088	0.93
15	MC	0.5796	0.0090	1.08
16	XI	-0.0492	0.0087	0.78
17	XI	1.5900	0.0103	1.04
18	XI	1.0037	0.0094	0.84
19	XI	1.2333	0.0097	1.00
20	XI	-0.1130	0.0087	0.79
21	XI	1.7810	0.0107	0.93
22	MC	1.2912	0.0098	1.36
23	MC	-0.6564	0.0087	1.14
24	MC	-1.3704	0.0091	1.19
25	XI	-1.7938	0.0097	1.18
26	XI	-0.2199	0.0087	1.01
27	XI	0.5642	0.0090	0.87
28	XI	0.3424	0.0088	0.82
29	MC	-0.3931	0.0086	1.01
30	XI	-0.6364	0.0087	1.17
31	MC	0.8663	0.0093	1.08
32	MC	0.7877	0.0092	1.11
33	XI	0.7354	0.0091	0.84
34	XI	0.4783	0.0089	0.83
35	XI	1.6408	0.0104	0.81
36	MC	0.2311	0.0088	1.19
37	MC	0.8664	0.0093	0.91
38	XI	0.5884	0.0090	0.89
39	MC	0.2545	0.0088	1.30
40	MC	-0.5909	0.0087	1.06
41	XI	0.5561	0.0090	0.84
42	XI	2.0419	0.0113	0.76
43	XI	-1.0993	0.0089	0.74
44	XI	-0.6794	0.0087	0.85
45	XI	2.1496	0.0116	0.92

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.10. Item-Level IRT Statistics, Mathematics Grade 6

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	XI	-0.4127	0.0086	0.88
2	MC	-1.1064	0.0087	1.00
3	MC	-1.3975	0.0089	1.36
4	MC	1.3003	0.0102	1.17
5	XI	-0.9079	0.0086	0.90
6	MC	-1.0139	0.0087	1.04
7	MC	-2.2420	0.0102	0.87
8	XI	0.7581	0.0094	0.93
9	XI	0.9796	0.0096	1.04
10	XI	0.5220	0.0091	0.89
11	MC	1.2711	0.0101	1.28
12	XI	1.5049	0.0106	1.29
13	XI	2.7909	0.0146	0.87
14	MC	0.6730	0.0092	1.24
15	MC	0.2708	0.0089	1.37
16	XI	0.9002	0.0095	0.79
17	MC	-1.5789	0.0091	0.98
18	XI	-0.1050	0.0086	0.87
19	MC	-0.1202	0.0086	1.42
20	XI	-0.7099	0.0086	0.95
21	MC	-1.6040	0.0091	1.02
22	XI	0.6159	0.0092	0.93
23	MC	-1.5454	0.0091	0.98
24	XI	-1.0345	0.0087	0.75
25	XI	-0.4526	0.0086	0.81
26	MC	-0.1634	0.0086	1.31
27	XI	-0.4474	0.0086	0.73
28	MC	0.2004	0.0088	1.04
29	XI	1.5575	0.0107	0.86
30	XI	-0.4621	0.0086	0.98
31	XI	0.7045	0.0093	1.09
32	MC	0.5155	0.0091	1.04
33	XI	-0.0054	0.0087	0.93
34	MC	1.0454	0.0097	0.88
35	XI	2.3880	0.0130	0.97
36	MC	0.4734	0.0090	1.09
37	MC	0.4689	0.0090	1.48
38	XI	0.9921	0.0097	0.86
39	XI	0.5496	0.0091	0.82
40	MC	0.9524	0.0096	1.06
41	XI	0.9330	0.0096	0.78
42	MC	0.0472	0.0087	1.08
43	XI	0.9346	0.0096	0.81
44	XI	-0.2134	0.0086	0.91
45	MC	-1.3462	0.0089	0.92

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
46	MC	-0.9714	0.0086	1.13
47	XI	1.6617	0.0109	0.91

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.11. Item-Level IRT Statistics, Mathematics Grade 7

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	-0.5590	0.0083	1.21
2	XI	1.5471	0.0104	0.85
3	XI	-0.0865	0.0084	0.94
4	XI	-0.2400	0.0084	0.82
5	MC	-1.2031	0.0085	0.97
6	MC	0.3774	0.0087	1.23
7	XI	0.0956	0.0085	0.92
8	XI	0.1853	0.0085	1.03
9	XI	0.5154	0.0088	1.02
10	MC	-0.5558	0.0083	1.09
11	MC	-0.4679	0.0083	1.18
12	XI	0.5323	0.0088	0.86
13	MC	0.1555	0.0085	1.09
14	XI	1.6358	0.0105	1.06
15	XI	-0.2643	0.0084	0.78
16	XI	1.1259	0.0096	0.79
17	XI	2.0849	0.0117	0.88
18	XI	-0.2364	0.0084	0.96
19	MC	-1.6349	0.0090	1.01
20	MC	0.3282	0.0086	1.32
21	XI	0.4675	0.0088	0.81
22	MC	-1.2100	0.0086	0.99
23	XI	-0.7828	0.0084	0.79
24	MC	-2.2161	0.0099	0.91
25	XI	0.9471	0.0093	0.92
26	MC	-1.9922	0.0095	1.05
27	XI	0.4711	0.0088	0.80
28	XI	0.5576	0.0089	1.04
29	MC	0.0064	0.0084	1.29
30	MC	-0.0247	0.0084	1.08
31	MC	0.2320	0.0086	1.38
32	XI	3.5835	0.0184	0.89
33	XI	0.6839	0.0090	0.89
34	XI	1.2389	0.0098	0.86
35	XI	1.7949	0.0109	0.94
36	XI	2.2889	0.0123	1.09
37	MC	0.1505	0.0085	1.08
38	XI	0.8655	0.0092	0.95
39	XI	1.9468	0.0113	0.84

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
40	XI	-0.6649	0.0083	0.81
41	XI	0.9212	0.0093	0.85
42	MC	-1.5448	0.0089	1.19
43	MC	-0.9877	0.0084	1.11
44	MC	-0.5406	0.0083	1.17
45	MC	-0.7717	0.0083	1.15
46	MC	-1.4616	0.0088	0.80
47	MC	0.2368	0.0086	1.28

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.12. Item-Level IRT Statistics, Mathematics Grade 8

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
1	MC	0.0068	0.0082	1.08
2	MC	-0.4637	0.0079	0.99
3	MC	-1.4875	0.0080	1.13
4	XI	-0.6574	0.0078	0.79
5	MC	-0.1720	0.0081	1.12
6	MC	0.3467	0.0086	0.86
7	XI	-1.5961	0.0081	0.81
8	MC	0.3288	0.0086	1.10
9	XI	1.2933	0.0104	1.02
10	MC	0.5532	0.0089	1.02
11	MC	0.5570	0.0089	1.20
12	MC	0.5850	0.0090	1.04
13	MC	0.4494	0.0088	1.28
14	XI	0.4998	0.0089	0.82
15	MC	-0.1728	0.0081	1.12
16	MC	-0.1897	0.0081	1.00
17	XI	1.4931	0.0110	0.89
18	MC	-0.6109	0.0078	1.03
19	MC	-0.9451	0.0078	0.95
20	XI	1.1524	0.0101	0.89
21	MC	0.8078	0.0094	1.00
22	MC	-1.1863	0.0079	0.97
23	XI	0.2057	0.0085	0.85
24	XI	-0.5492	0.0079	0.89
25	MC	0.0185	0.0082	1.11
26	MC	-0.6298	0.0078	0.96
27	MC	-0.3013	0.0080	1.13
28	MC	-0.3429	0.0080	1.16
29	MC	-0.2105	0.0081	1.16
30	MC	0.1770	0.0084	1.04
31	MC	0.3878	0.0087	1.19
32	MC	0.4853	0.0088	1.02
33	XI	0.8341	0.0094	0.91

Item Number	Item Type	Rasch Difficulty	SE	MNSQ Infit
34	MC	0.1202	0.0084	1.03
35	MC	0.1515	0.0084	1.10
36	XI	-0.2529	0.0080	0.95
37	MC	1.3312	0.0105	0.82
38	MC	-0.2507	0.0080	0.97
39	XI	0.3671	0.0087	0.85
40	MC	0.7188	0.0092	0.94
41	MC	-0.4611	0.0079	1.14
42	MC	-0.4381	0.0079	1.11
43	XI	0.0565	0.0083	0.77
44	XI	0.4641	0.0088	0.90
45	XI	0.9230	0.0096	0.86
46	XI	1.3971	0.0107	0.85
47	MC	-0.5740	0.0079	0.98

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.13. Raw-to-Scale Score Conversion, ELA Grade 3

Raw Score	Scale Score	CSEM	Performance Level
2	2395	23	1
3	2395	23	1
4	2398	22	1
5	2412	18	1
6	2421	16	1
7	2429	15	1
8	2436	14	1
9	2441	13	1
10	2447	12	1
11	2451	12	1
12	2456	11	1
13	2460	11	1
14	2463	11	1
15	2467	10	1
16	2471	10	1
17	2474	10	1
18	2477	10	1
19	2480	10	1
20	2483	10	1
21	2486	9	1
22	2489	9	1
23	2492	9	1
24	2495	9	1
25	2498	9	2
26	2501	9	2
27	2504	9	2
28	2507	9	2

Raw Score	Scale Score	CSEM	Performance Level
29	2510	9	3
30	2513	9	3
31	2516	9	3
32	2519	10	3
33	2522	10	3
34	2525	10	3
35	2528	10	3
36	2531	10	3
37	2535	10	3
38	2538	10	3
39	2542	10	4
40	2545	11	4
41	2549	11	4
42	2553	11	4
43	2558	12	4
44	2562	12	4
45	2567	12	4
46	2572	13	4
47	2578	14	4
48	2585	15	4
49	2593	16	4
50	2601	17	4
51	2605	18	4
52	2605	18	4
53	2605	18	4
54	2605	18	4

Table B.14. Raw-to-Scale Score Conversion, ELA Grade 4

Raw Score	Scale Score	CSEM	Performance Level
2	2400	25	1
3	2400	25	1
4	2409	22	1
5	2423	18	1
6	2432	16	1
7	2440	15	1
8	2447	14	1
9	2453	13	1
10	2458	12	1
11	2463	12	1
12	2467	11	1
13	2471	11	1
14	2475	10	1
15	2478	10	1
16	2482	10	1
17	2485	10	1
18	2488	10	1

Raw Score	Scale Score	CSEM	Performance Level
19	2491	9	1
20	2494	9	1
21	2497	9	1
22	2500	9	1
23	2503	9	1
24	2505	9	1
25	2508	9	1
26	2511	9	2
27	2513	9	2
28	2516	9	2
29	2518	9	2
30	2521	9	2
31	2524	9	3
32	2526	9	3
33	2529	9	3
34	2532	9	3
35	2535	9	3
36	2537	9	3
37	2540	9	3
38	2543	10	3
39	2546	10	3
40	2550	10	3
41	2553	10	3
42	2557	11	3
43	2561	11	4
44	2565	11	4
45	2569	12	4
46	2574	13	4
47	2580	13	4
48	2586	14	4
49	2593	15	4
50	2601	16	4
51	2610	17	4
52	2610	17	4
53	2610	17	4
54	2610	17	4
55	2610	17	4
56	2610	17	4

Table B.15. Raw-to-Scale Score Conversion, ELA Grade 5

Raw Score	Scale Score	CSEM	Performance Level
2	2419	23	1
3	2419	23	1
4	2422	22	1
5	2435	18	1
6	2444	16	1
7	2452	15	1
8	2459	14	1
9	2464	13	1
10	2470	12	1
11	2474	12	1
12	2479	11	1
13	2483	11	1
14	2486	11	1
15	2490	10	1
16	2494	10	1
17	2497	10	1
18	2500	10	1
19	2503	10	1
20	2507	10	1
21	2510	10	1
22	2513	9	1
23	2516	9	1
24	2518	9	1
25	2521	9	2
26	2524	9	2
27	2527	9	2
28	2530	9	2
29	2533	9	2
30	2536	9	2
31	2539	10	2
32	2543	10	3
33	2545	10	3
34	2548	10	3
35	2552	10	3
36	2555	10	3
37	2558	10	3
38	2562	10	3
39	2565	11	3
40	2569	11	3
41	2573	11	3
42	2578	12	4
43	2582	12	4
44	2587	12	4
45	2593	13	4
46	2599	14	4

Raw Score	Scale Score	CSEM	Performance Level
47	2605	15	4
48	2613	16	4
49	2622	17	4
50	2629	19	4
51	2629	19	4
52	2629	19	4
53	2629	19	4

Table B.16. Raw-to-Scale Score Conversion, ELA Grade 6

Raw Score	Scale Score	CSEM	Performance Level
2	2431	24	1
3	2431	24	1
4	2437	22	1
5	2450	18	1
6	2460	16	1
7	2468	15	1
8	2475	14	1
9	2480	13	1
10	2486	12	1
11	2490	12	1
12	2495	11	1
13	2499	11	1
14	2502	11	1
15	2506	10	1
16	2509	10	1
17	2513	10	1
18	2516	10	1
19	2519	9	1
20	2522	9	1
21	2525	9	1
22	2528	9	1
23	2530	9	1
24	2533	9	2
25	2536	9	2
26	2538	9	2
27	2541	9	2
28	2543	9	2
29	2546	9	2
30	2549	9	2
31	2551	9	2
32	2554	9	3
33	2556	9	3
34	2559	9	3
35	2562	9	3
36	2564	9	3
37	2567	9	3

Raw Score	Scale Score	CSEM	Performance Level
38	2570	9	3
39	2573	9	3
40	2576	10	3
41	2579	10	3
42	2582	10	3
43	2586	10	3
44	2589	11	3
45	2593	11	3
46	2597	11	4
47	2602	12	4
48	2607	12	4
49	2612	13	4
50	2618	14	4
51	2625	15	4
52	2634	17	4
53	2641	18	4
54	2641	18	4
55	2641	18	4
56	2641	18	4

Table B.17. Raw-to-Scale Score Conversion, ELA Grade 7

Raw Score	Scale Score	CSEM	Performance Level
2	2438	24	1
3	2438	24	1
4	2443	22	1
5	2457	18	1
6	2466	16	1
7	2474	15	1
8	2481	14	1
9	2487	13	1
10	2492	12	1
11	2497	12	1
12	2501	11	1
13	2506	11	1
14	2509	11	1
15	2513	11	1
16	2517	10	1
17	2520	10	1
18	2524	10	1
19	2527	10	1
20	2530	10	1
21	2533	10	1
22	2537	10	1
23	2540	10	1
24	2543	10	2
25	2546	10	2

Raw Score	Scale Score	CSEM	Performance Level
26	2549	10	2
27	2552	10	2
28	2555	10	2
29	2558	10	2
30	2561	10	3
31	2565	10	3
32	2568	10	3
33	2571	10	3
34	2575	10	3
35	2578	10	3
36	2582	10	3
37	2585	11	3
38	2589	11	3
39	2593	11	3
40	2597	11	3
41	2602	12	4
42	2606	12	4
43	2611	12	4
44	2616	13	4
45	2622	13	4
46	2628	14	4
47	2635	15	4
48	2644	16	4
49	2648	17	4
50	2648	17	4
51	2648	17	4
52	2648	17	4

Table B.18. Raw-to-Scale Score Conversion, ELA Grade 8

Raw Score	Scale Score	CSEM	Performance Level
2	2448	23	1
3	2448	23	1
4	2450	22	1
5	2463	18	1
6	2473	16	1
7	2481	15	1
8	2487	14	1
9	2493	13	1
10	2498	12	1
11	2503	12	1
12	2507	11	1
13	2511	11	1
14	2515	11	1
15	2519	10	1
16	2522	10	1
17	2525	10	1

Raw Score	Scale Score	CSEM	Performance Level
18	2529	10	1
19	2532	10	1
20	2535	10	1
21	2538	9	1
22	2541	9	1
23	2544	9	1
24	2547	9	1
25	2550	9	1
26	2552	9	2
27	2555	9	2
28	2558	9	2
29	2561	9	2
30	2564	9	2
31	2567	9	2
32	2569	9	2
33	2572	9	3
34	2575	9	3
35	2578	10	3
36	2582	10	3
37	2585	10	3
38	2588	10	3
39	2591	10	3
40	2595	10	3
41	2598	11	3
42	2602	11	3
43	2606	11	4
44	2610	11	4
45	2615	12	4
46	2619	12	4
47	2625	13	4
48	2630	13	4
49	2637	14	4
50	2644	15	4
51	2652	17	4
52	2658	18	4
53	2658	18	4
54	2658	18	4
55	2658	18	4

Table B.19. Raw-to-Scale Score Conversion, Mathematics Grade 3

Raw Score	Scale Score	CSEM	Performance Level
0	3395	25	1
1	3395	25	1
2	3401	23	1
3	3415	19	1
4	3426	17	1
5	3434	15	1
6	3442	14	1
7	3448	13	1
8	3454	13	1
9	3459	12	1
10	3464	12	1
11	3468	12	1
12	3473	11	1
13	3477	11	1
14	3481	11	1
15	3485	11	1
16	3488	10	1
17	3492	10	1
18	3495	10	2
19	3499	10	2
20	3502	10	2
21	3506	10	2
22	3509	10	2
23	3513	10	2
24	3516	10	2
25	3519	10	2
26	3523	10	2
27	3526	10	2
28	3531	10	3
29	3534	11	3
30	3537	11	3
31	3541	11	3
32	3545	11	3
33	3549	11	3
34	3554	12	3
35	3558	12	3
36	3563	12	3
37	3569	13	3
38	3575	14	4
39	3581	15	4
40	3589	16	4
41	3598	17	4
42	3605	19	4
43	3605	19	4
44	3605	19	4
45	3605	19	4

Table B.20. Raw-to-Scale Score Conversion, Mathematics Grade 4

Raw Score	Scale Score	CSEM	Performance Level
0	3435	23	1
1	3435	23	1
2	3436	22	1
3	3450	19	1
4	3460	17	1
5	3468	15	1
6	3476	14	1
7	3482	14	1
8	3488	13	1
9	3493	12	1
10	3498	12	1
11	3503	12	1
12	3507	11	1
13	3512	11	1
14	3516	11	1
15	3520	11	1
16	3524	11	1
17	3527	11	1
18	3531	10	2
19	3535	10	2
20	3538	10	2
21	3542	10	2
22	3545	10	2
23	3549	10	2
24	3553	10	2
25	3556	10	2
26	3560	10	2
27	3563	10	3
28	3567	11	3
29	3571	11	3
30	3574	11	3
31	3578	11	3
32	3582	11	3
33	3587	11	3
34	3591	12	3
35	3596	12	3
36	3601	12	3
37	3606	13	4
38	3612	14	4
39	3618	14	4
40	3626	15	4
41	3634	17	4
42	3645	19	4
43	3645	19	4
44	3645	19	4
45	3645	19	4

Table B.21. Raw-to-Scale Score Conversion, Mathematics Grade 5

Raw Score	Scale Score	CSEM	Performance Level
0	3478	26	1
1	3478	26	1
2	3489	22	1
3	3503	18	1
4	3513	16	1
5	3521	15	1
6	3528	14	1
7	3534	13	1
8	3539	12	1
9	3544	12	1
10	3548	11	1
11	3553	11	1
12	3557	11	1
13	3560	11	1
14	3564	10	2
15	3568	10	2
16	3571	10	2
17	3574	10	2
18	3578	10	2
19	3581	10	2
20	3584	10	2
21	3587	10	2
22	3591	10	2
23	3595	10	3
24	3597	10	3
25	3600	10	3
26	3603	10	3
27	3606	10	3
28	3610	10	3
29	3613	10	3
30	3616	10	3
31	3620	10	3
32	3624	11	3
33	3628	11	3
34	3632	11	3
35	3636	11	4
36	3640	12	4
37	3645	12	4
38	3651	13	4
39	3657	14	4
40	3663	15	4
41	3671	16	4
42	3681	18	4
43	3688	20	4
44	3688	20	4
45	3688	20	4

Table B.22. Raw-to-Scale Score Conversion, Mathematics Grade 6

Raw Score	Scale Score	CSEM	Performance Level
0	3512	23	1
1	3512	23	1
2	3513	22	1
3	3526	19	1
4	3537	16	1
5	3545	15	1
6	3552	14	1
7	3558	13	1
8	3563	13	1
9	3568	12	1
10	3573	12	1
11	3578	11	1
12	3582	11	1
13	3586	11	1
14	3590	11	1
15	3593	10	1
16	3597	10	1
17	3600	10	1
18	3604	10	2
19	3607	10	2
20	3610	10	2
21	3614	10	2
22	3617	10	2
23	3620	10	2
24	3623	10	2
25	3626	10	2
26	3630	10	3
27	3633	10	3
28	3636	10	3
29	3639	10	3
30	3643	10	3
31	3646	10	3
32	3650	10	3
33	3653	11	3
34	3657	11	3
35	3661	11	3
36	3665	11	4
37	3670	12	4
38	3674	12	4
39	3679	13	4
40	3685	13	4
41	3691	14	4
42	3698	15	4
43	3706	16	4
44	3716	19	4

Raw Score	Scale Score	CSEM	Performance Level
45	3722	20	4
46	3722	20	4
47	3722	20	4

Table B.23. Raw-to-Scale Score Conversion, Mathematics Grade 7

Raw Score	Scale Score	CSEM	Performance Level
0	3529	22	1
1	3529	22	1
2	3530	22	1
3	3543	19	1
4	3553	16	1
5	3562	15	1
6	3569	14	1
7	3575	13	1
8	3580	13	1
9	3585	12	1
10	3590	12	1
11	3595	11	1
12	3599	11	1
13	3603	11	1
14	3606	11	1
15	3610	10	1
16	3614	10	1
17	3617	10	1
18	3621	10	1
19	3624	10	1
20	3627	10	1
21	3630	10	2
22	3634	10	2
23	3637	10	2
24	3640	10	2
25	3643	10	2
26	3647	10	2
27	3650	10	2
28	3653	10	3
29	3657	10	3
30	3660	10	3
31	3663	10	3
32	3667	11	3
33	3671	11	3
34	3675	11	3
35	3680	11	4
36	3683	12	4
37	3688	12	4
38	3693	12	4
39	3698	13	4

Raw Score	Scale Score	CSEM	Performance Level
40	3704	14	4
41	3710	14	4
42	3718	15	4
43	3726	17	4
44	3737	19	4
45	3739	20	4
46	3739	20	4
47	3739	20	4

Table B.24. Raw-to-Scale Score Conversion, Mathematics Grade 8

Raw Score	Scale Score	CSEM	Performance Level
0	3566	24	1
1	3566	24	1
2	3573	22	1
3	3586	18	1
4	3596	16	1
5	3604	15	1
6	3610	14	1
7	3616	13	1
8	3621	12	1
9	3626	12	1
10	3630	11	1
11	3634	11	1
12	3638	11	1
13	3642	10	1
14	3645	10	1
15	3648	10	1
16	3652	10	2
17	3655	10	2
18	3658	9	2
19	3661	9	2
20	3664	9	2
21	3666	9	2
22	3669	9	2
23	3673	9	3
24	3675	9	3
25	3678	9	3
26	3681	9	3
27	3684	9	3
28	3687	9	3
29	3690	9	3
30	3693	10	3
31	3696	10	3
32	3699	10	3
33	3702	10	3
34	3706	10	4

Raw Score	Scale Score	CSEM	Performance Level
35	3709	10	4
36	3713	11	4
37	3717	11	4
38	3721	12	4
39	3726	12	4
40	3731	13	4
41	3736	13	4
42	3743	15	4
43	3751	16	4
44	3760	18	4
45	3774	22	4
46	3776	23	4
47	3776	23	4

Figure B.1. Item-Person Map, ELA Grade 3

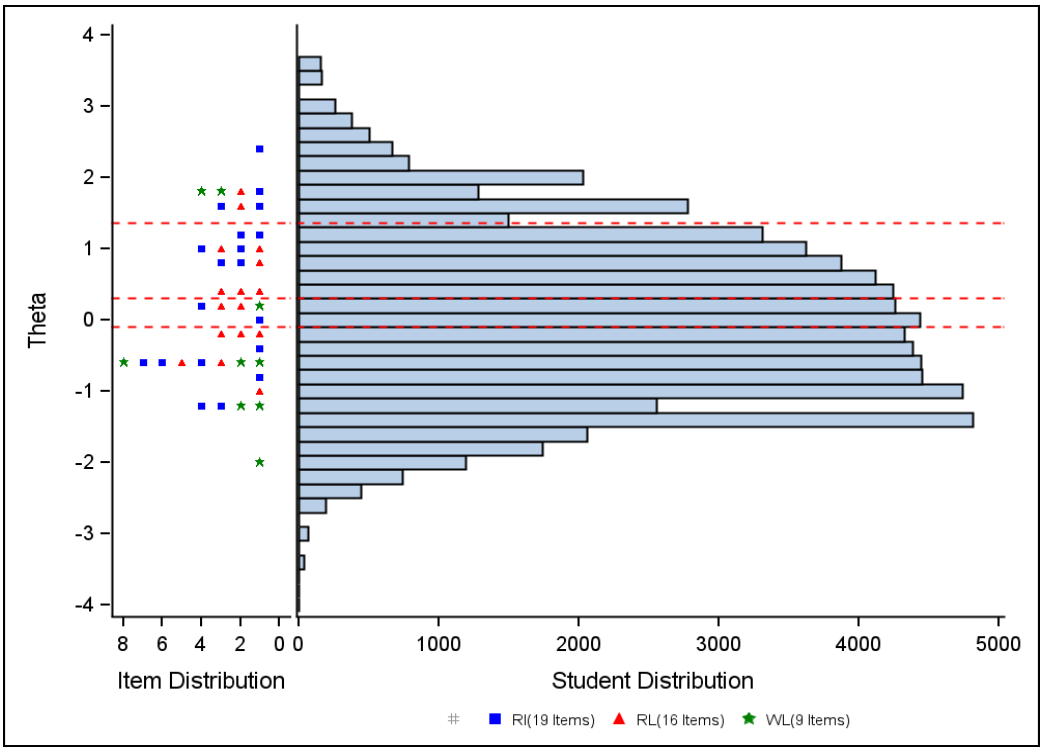


Figure B.2. Item-Person Map, ELA Grade 4

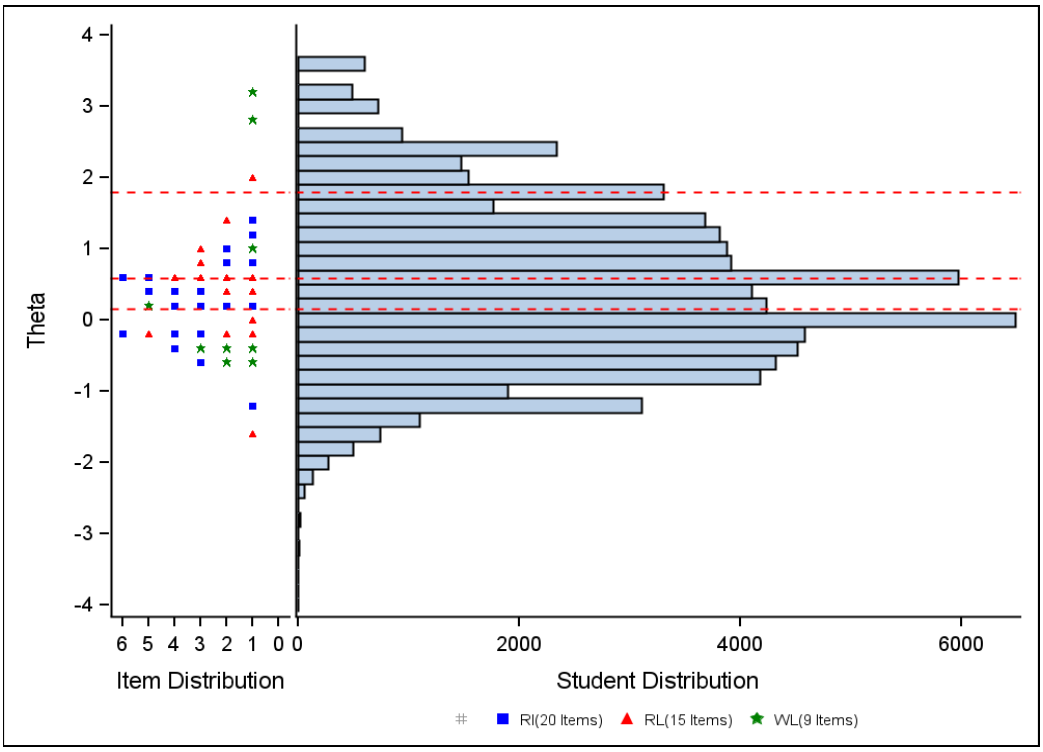


Figure B.3. Item-Person Map, ELA Grade 5

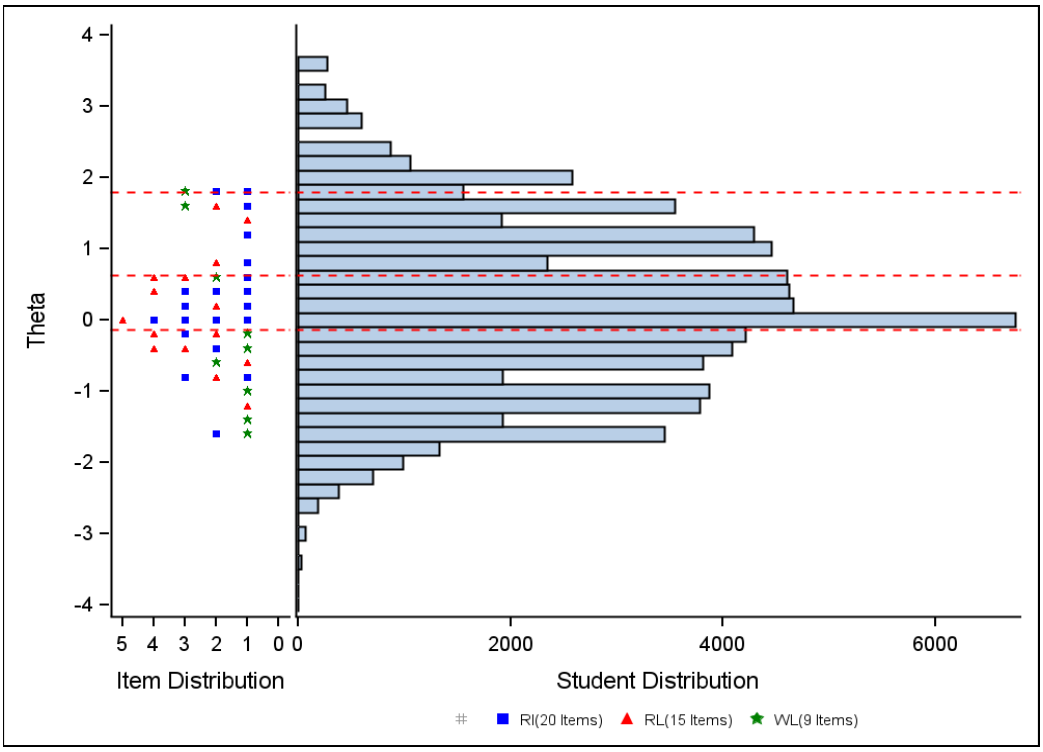


Figure B.4. Item-Person Map, ELA Grade 6

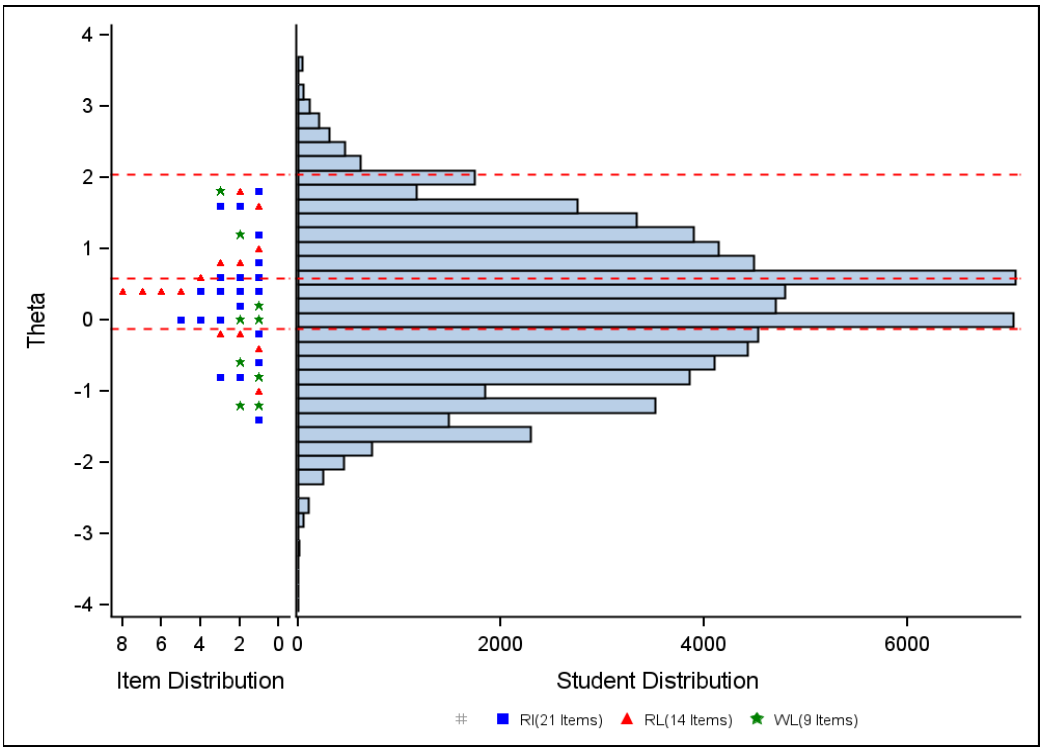


Figure B.5. Item-Person Map, ELA Grade 7

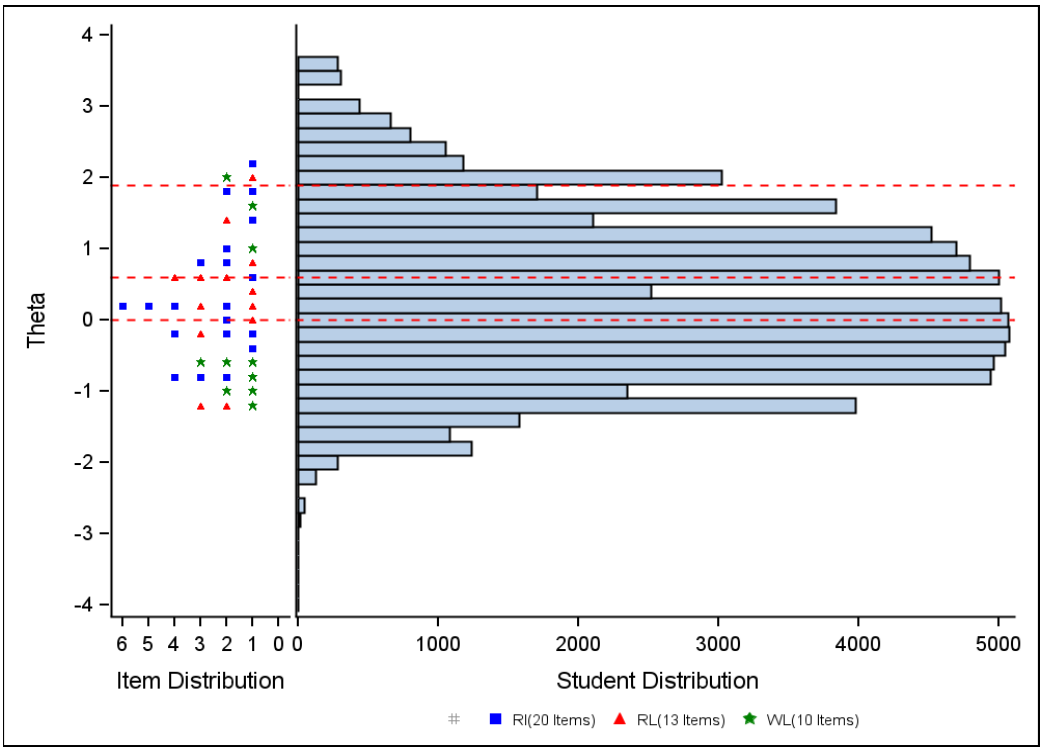


Figure B.6. Item-Person Map, ELA Grade 8

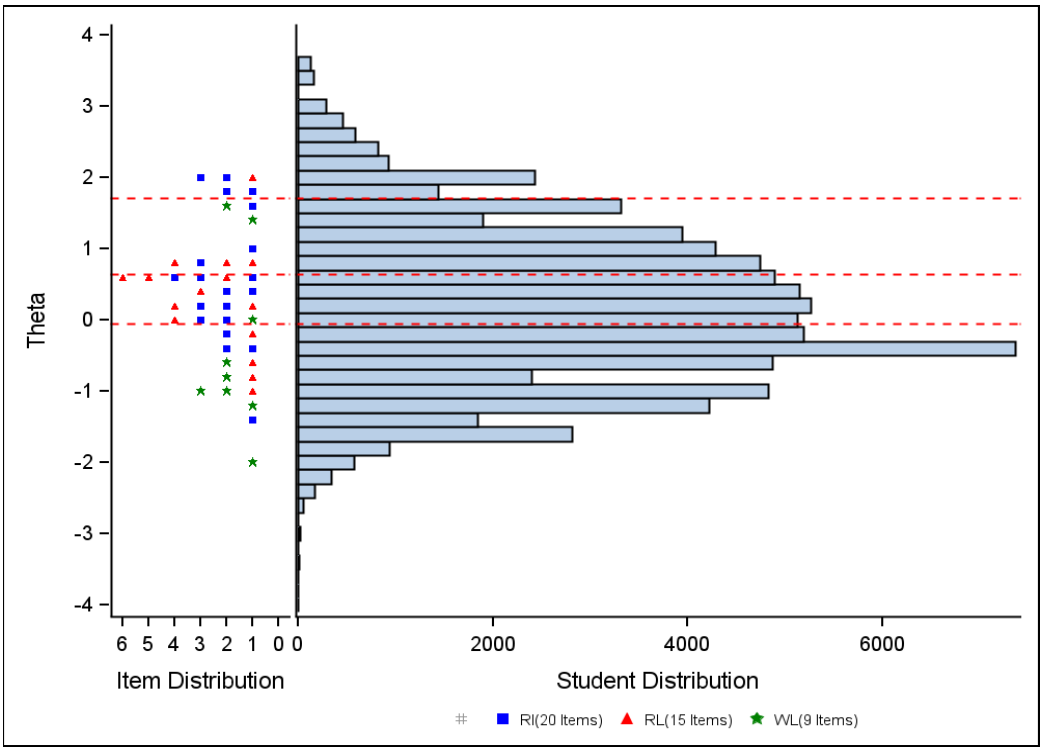


Figure B.7. Item-Person Map, Mathematics Grade 3

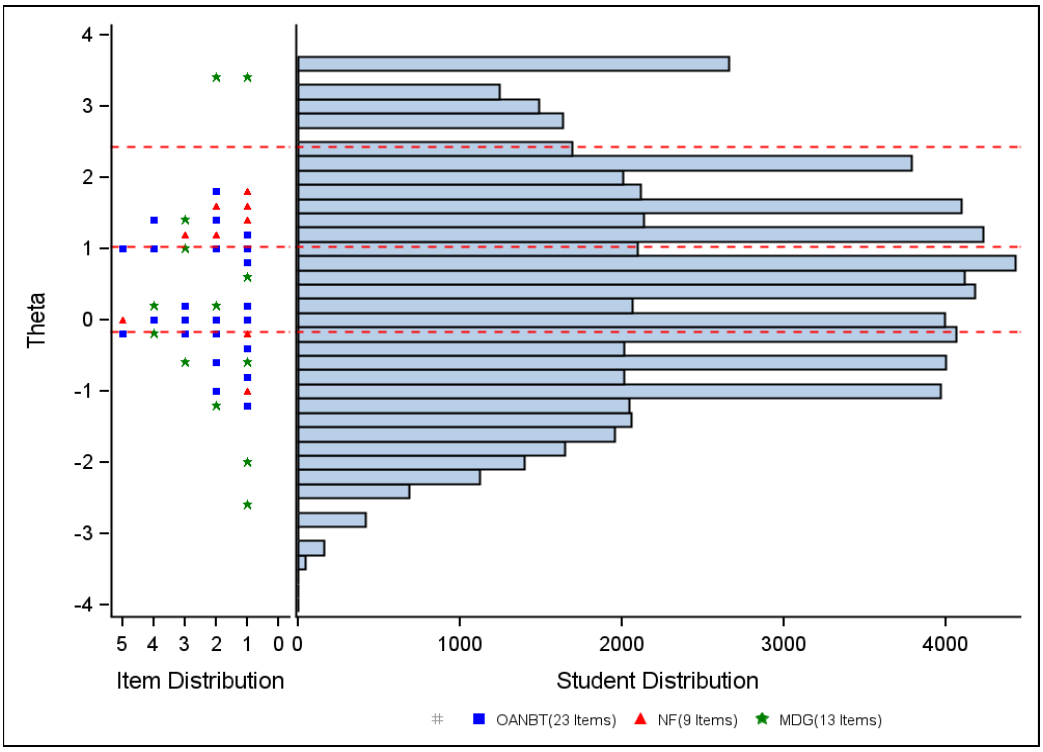


Figure B.8. Item-Person Map, Mathematics Grade 4

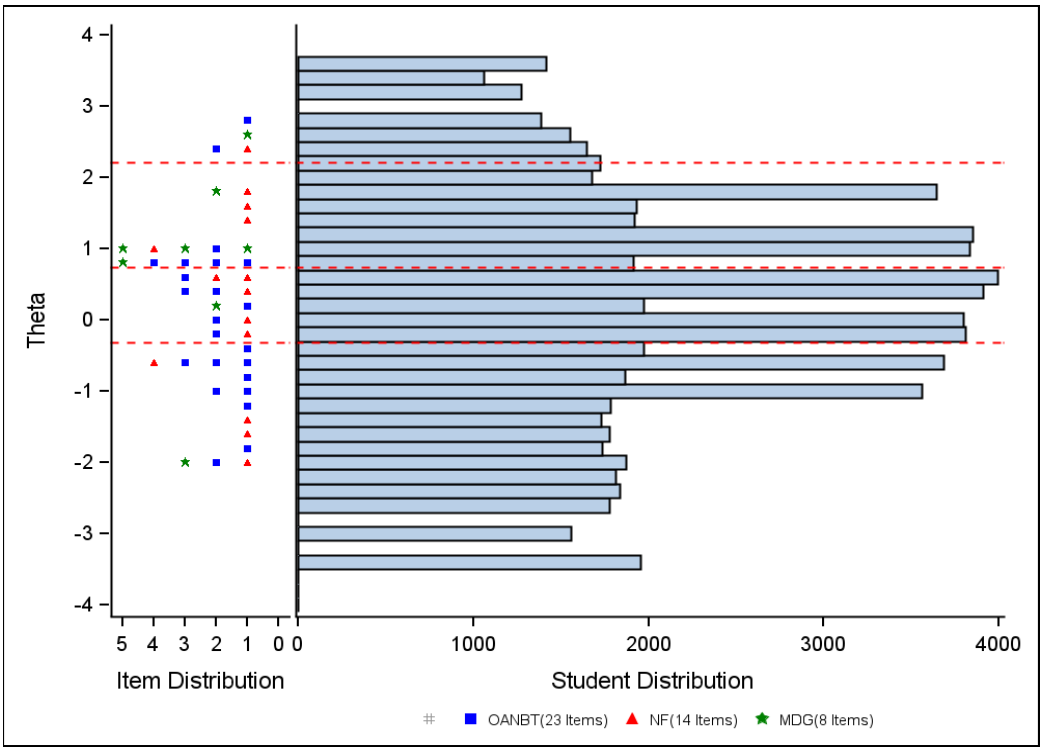


Figure B.9. Item-Person Map, Mathematics Grade 5

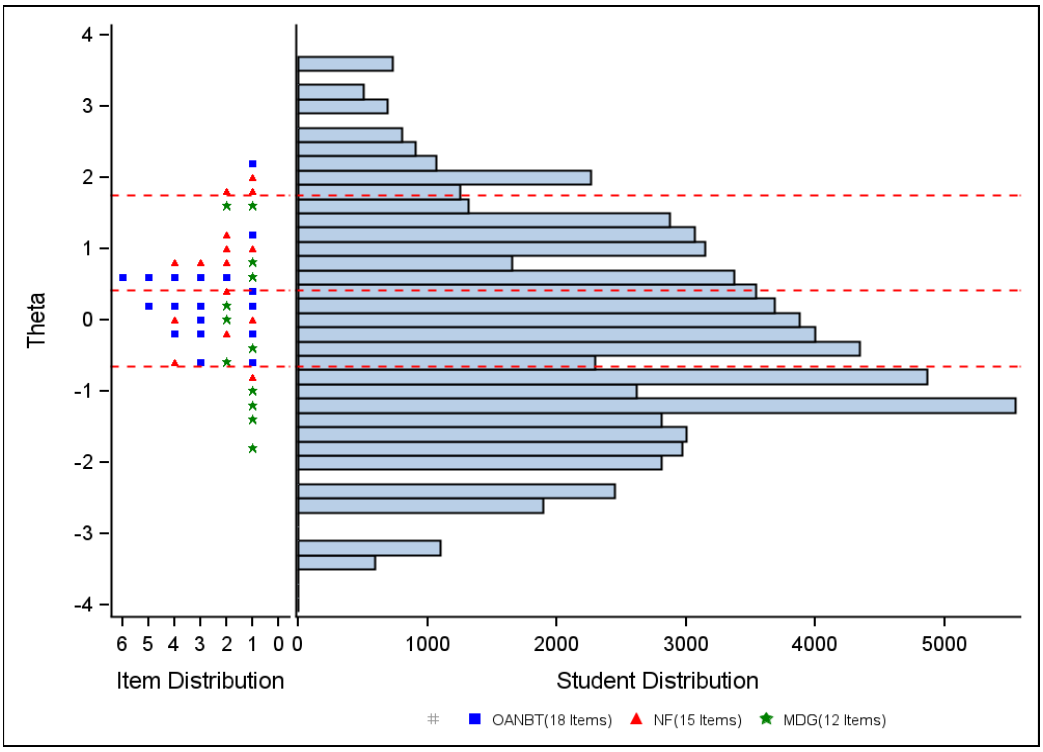


Figure B.10. Item-Person Map, Mathematics Grade 6

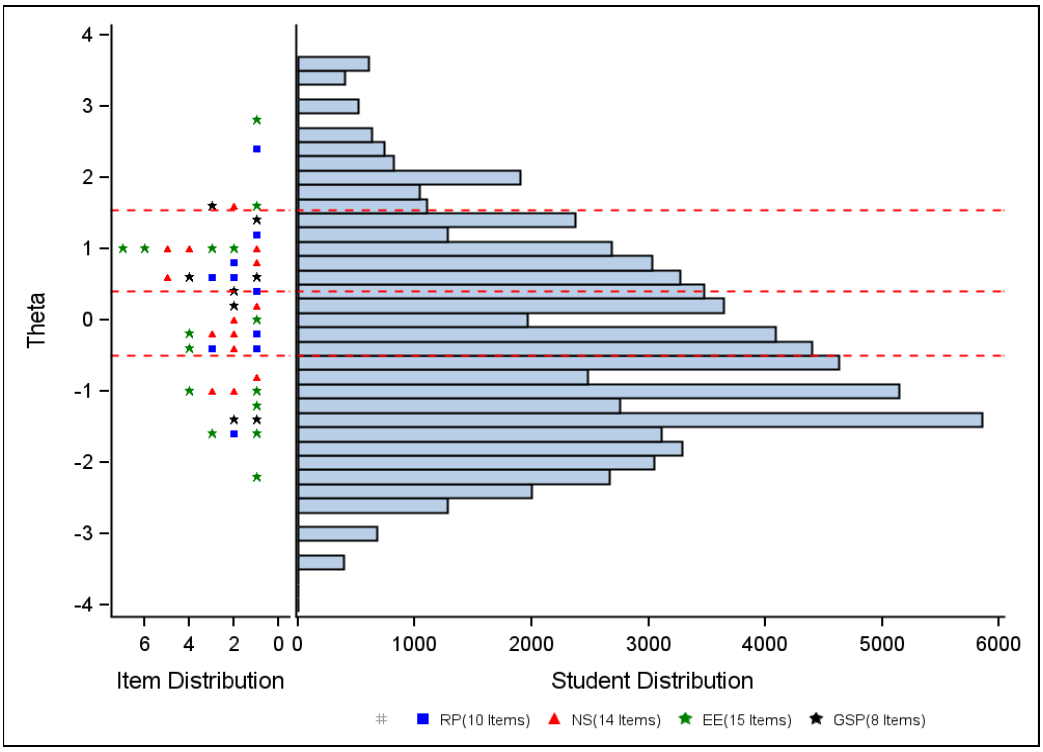


Figure B.11. Item-Person Map, Mathematics Grade 7

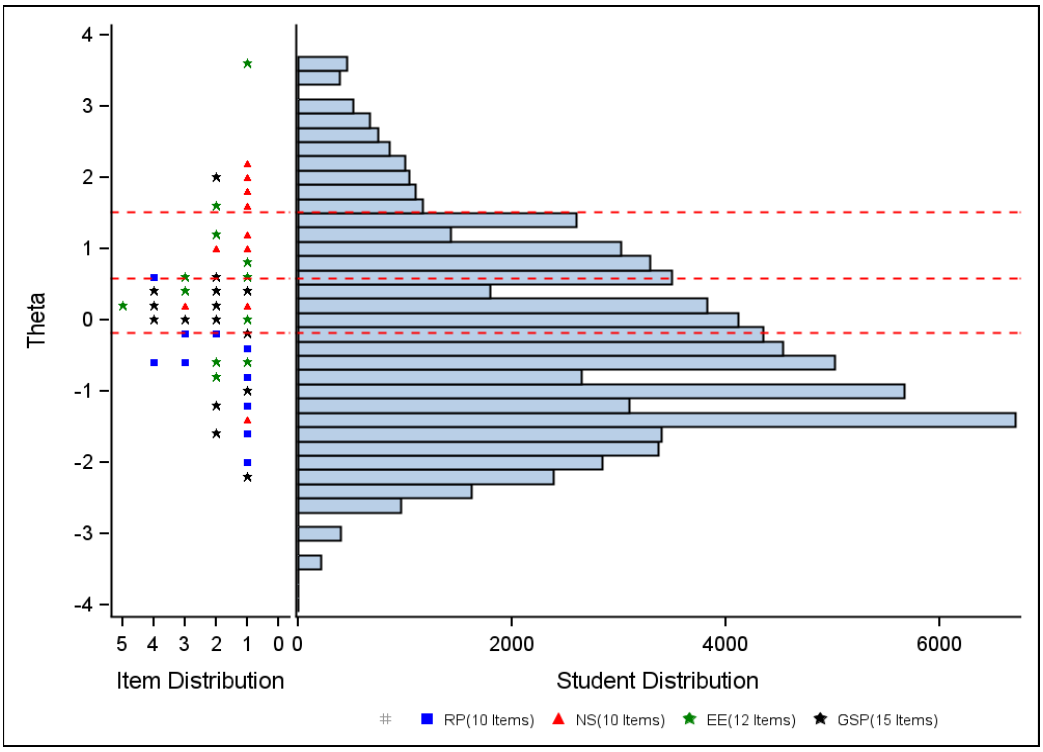


Figure B.12. Item-Person Map, Mathematics Grade 8

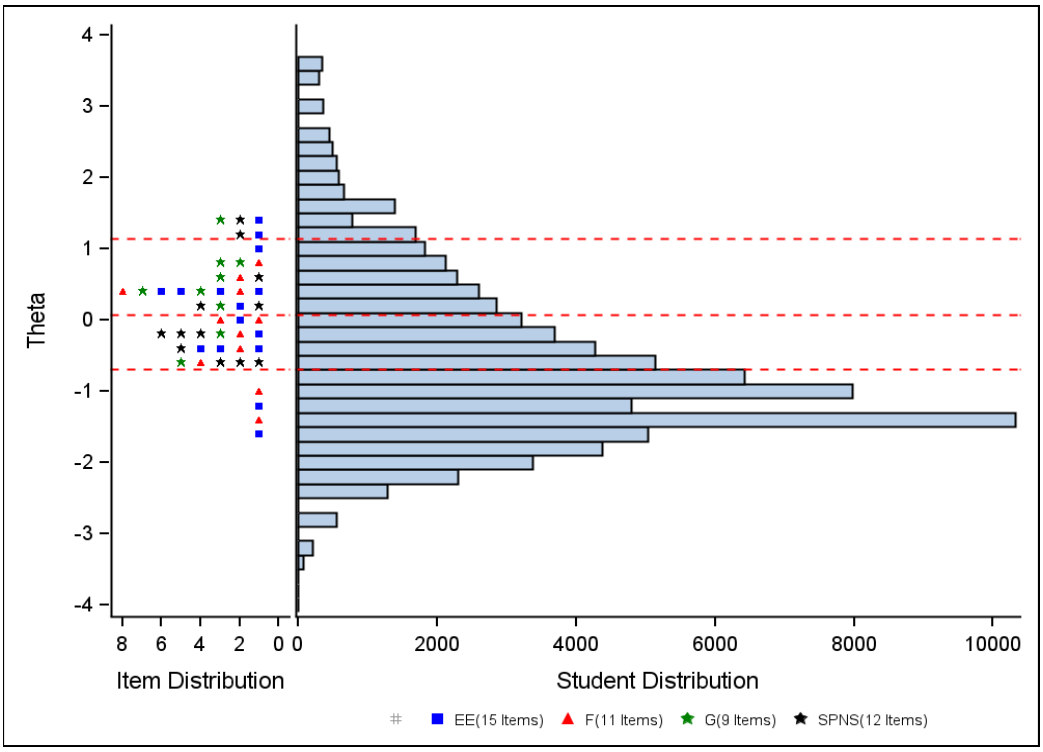


Figure B.13. TCC, ELA Grade 3

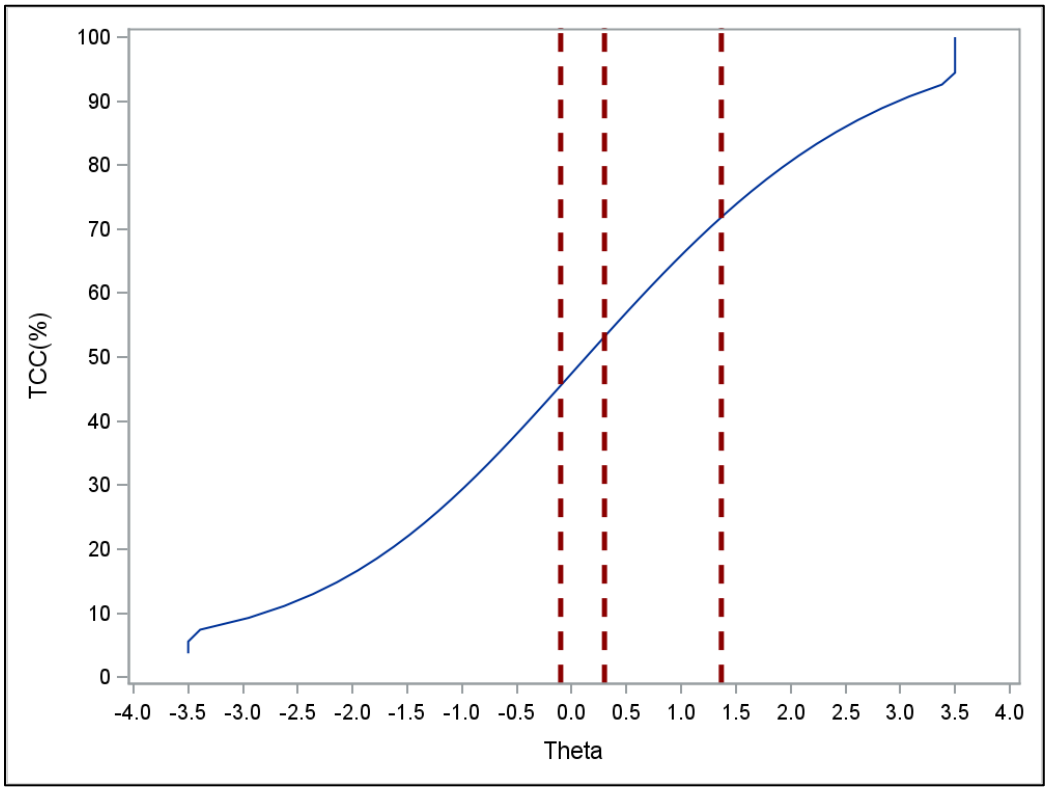


Figure B.14. CSEM, ELA Grade 3

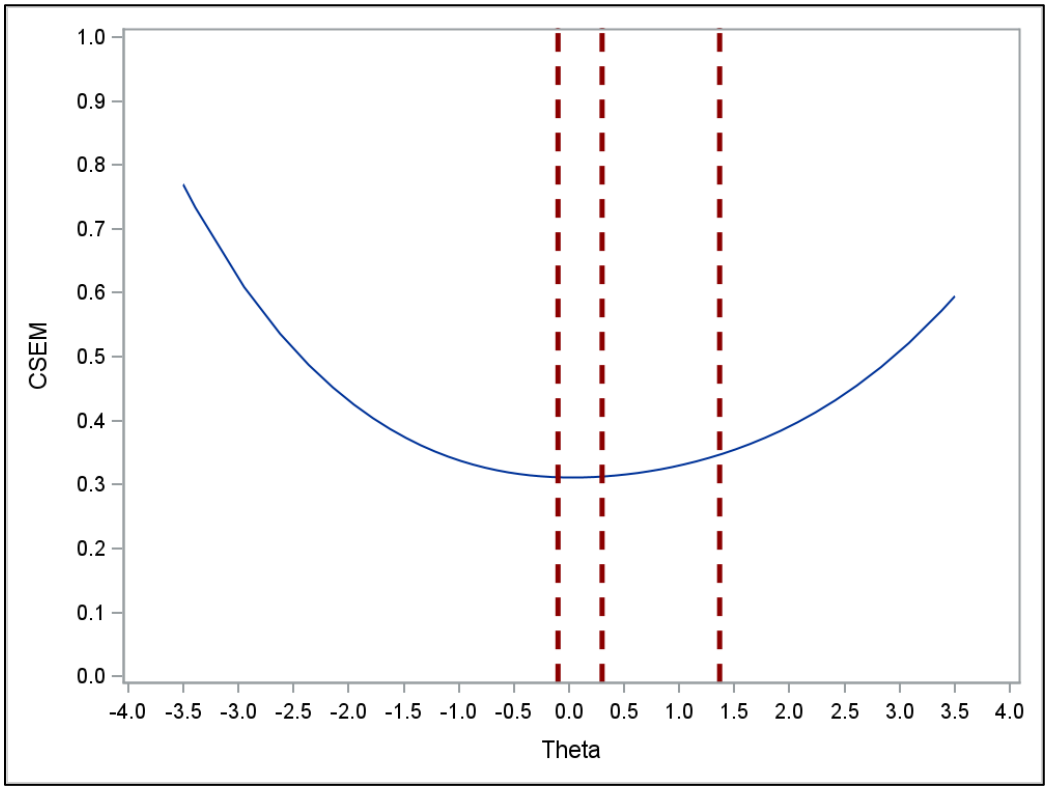


Figure B.15. TCC, ELA Grade 4

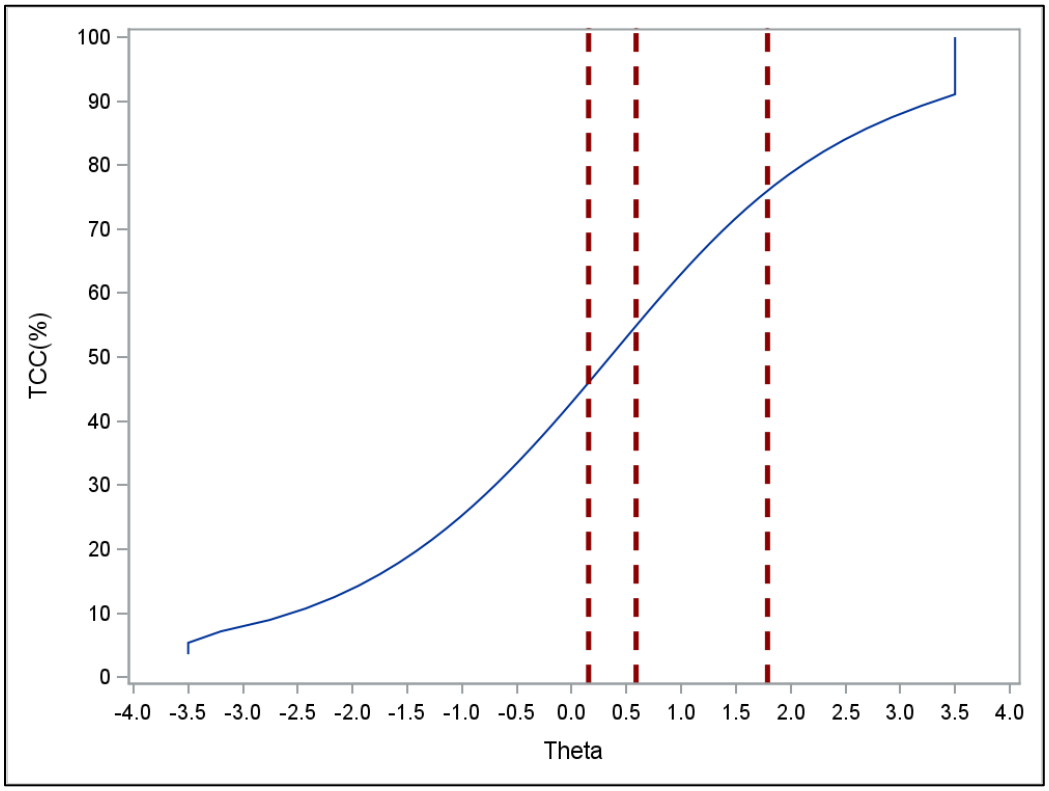


Figure B.16. CSEM, ELA Grade 4

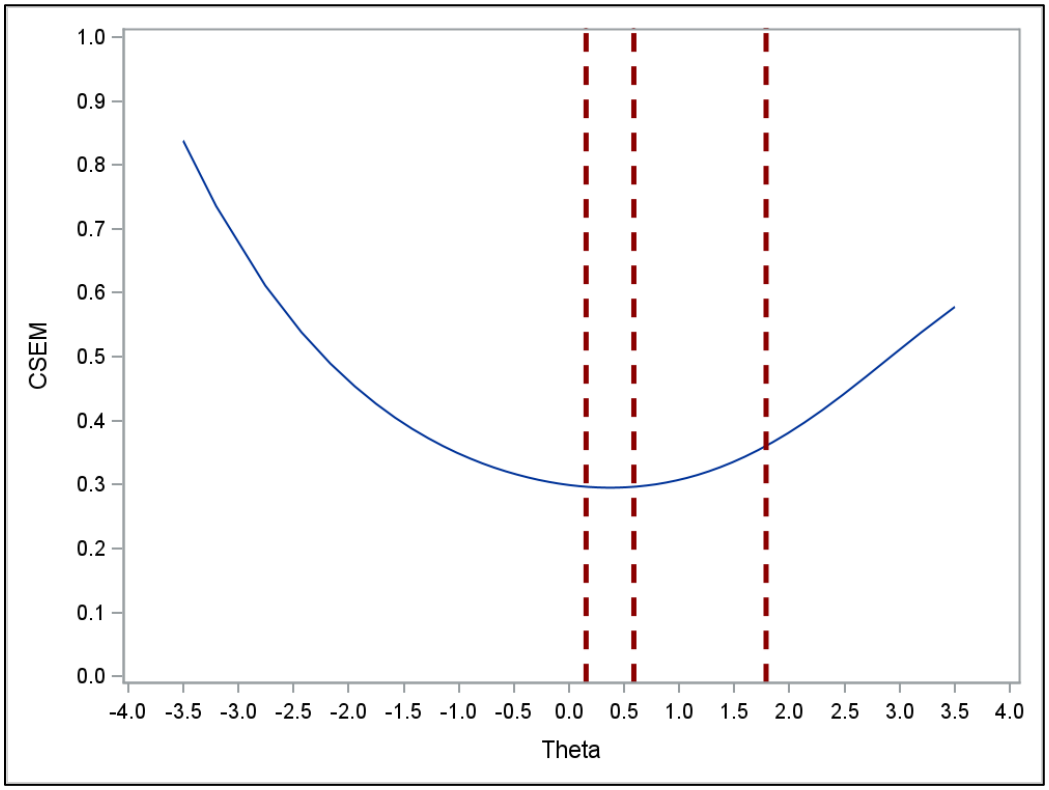


Figure B.17. TCC, ELA Grade 5

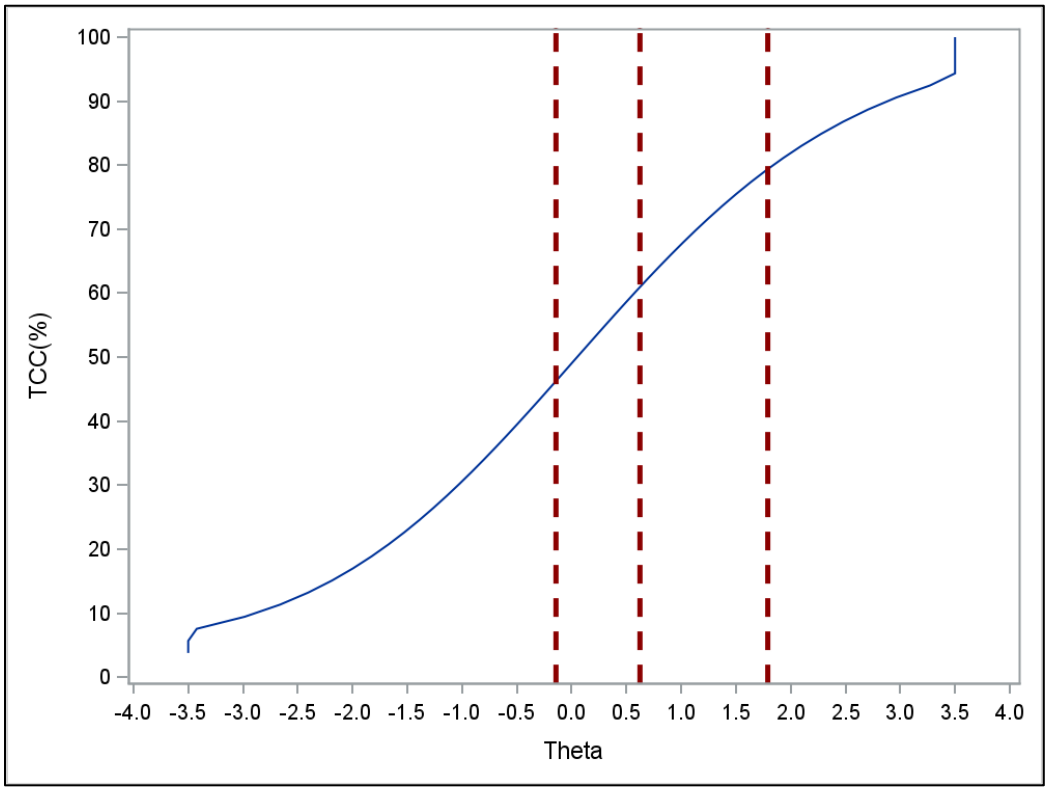


Figure B.18. CSEM, ELA Grade 5

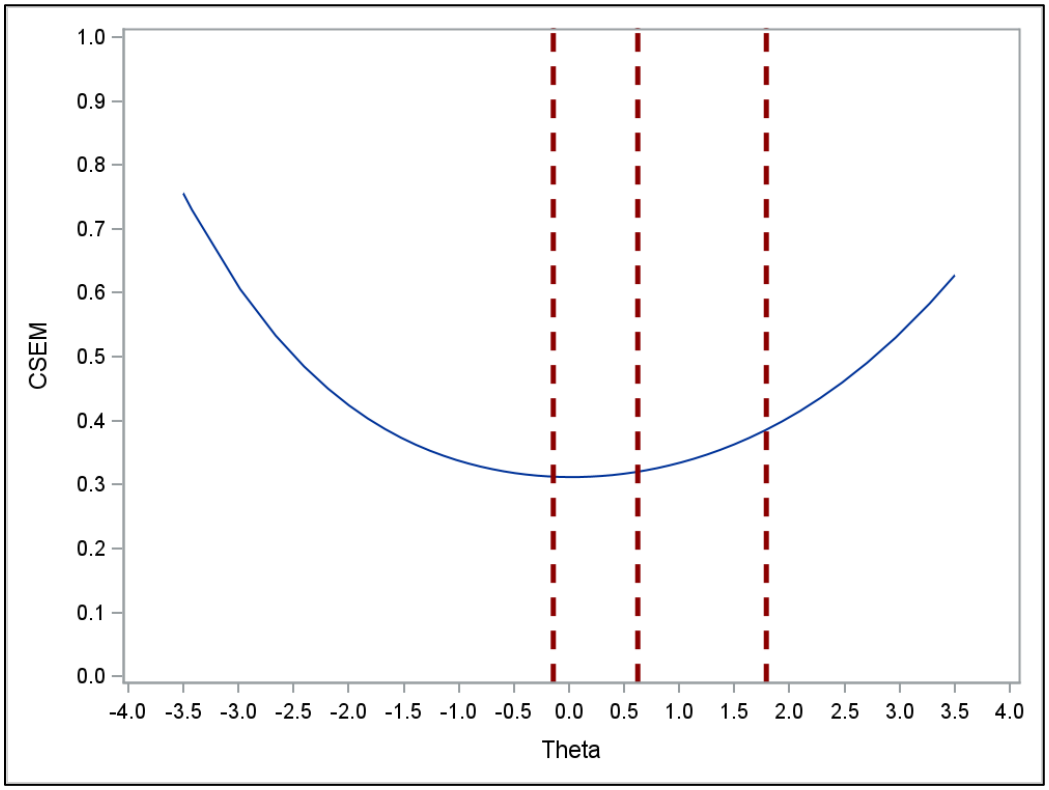


Figure B.19. TCC, ELA Grade 6

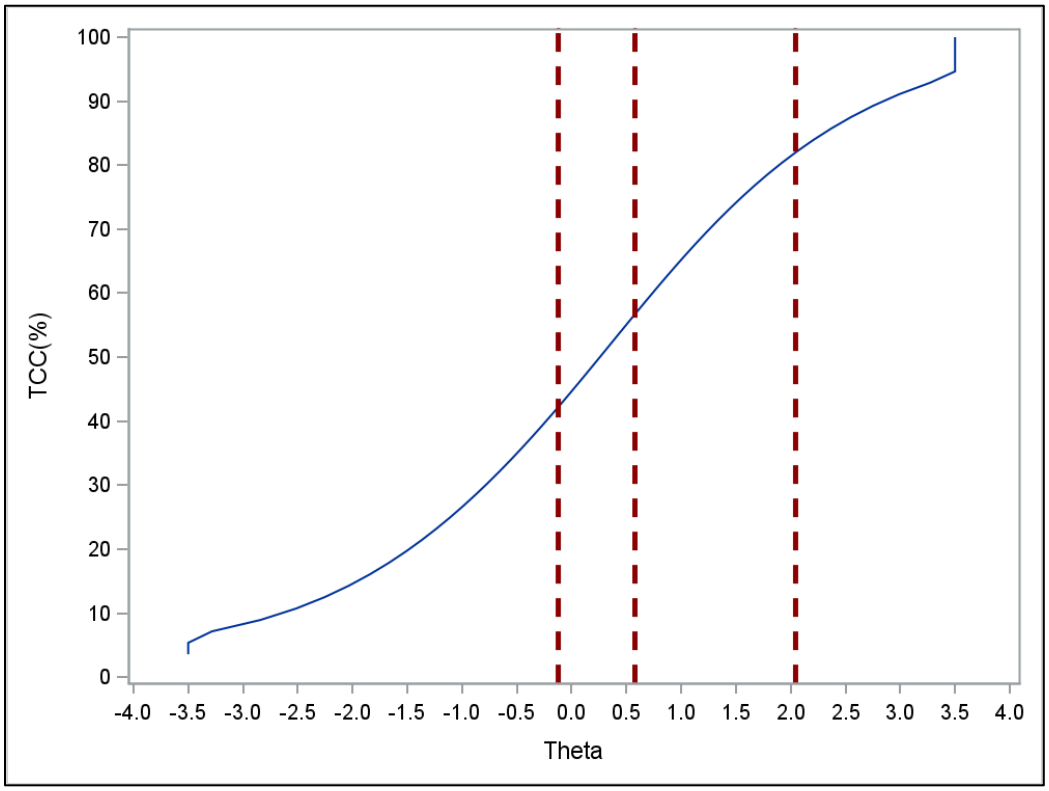


Figure B.20. CSEM, ELA Grade 6

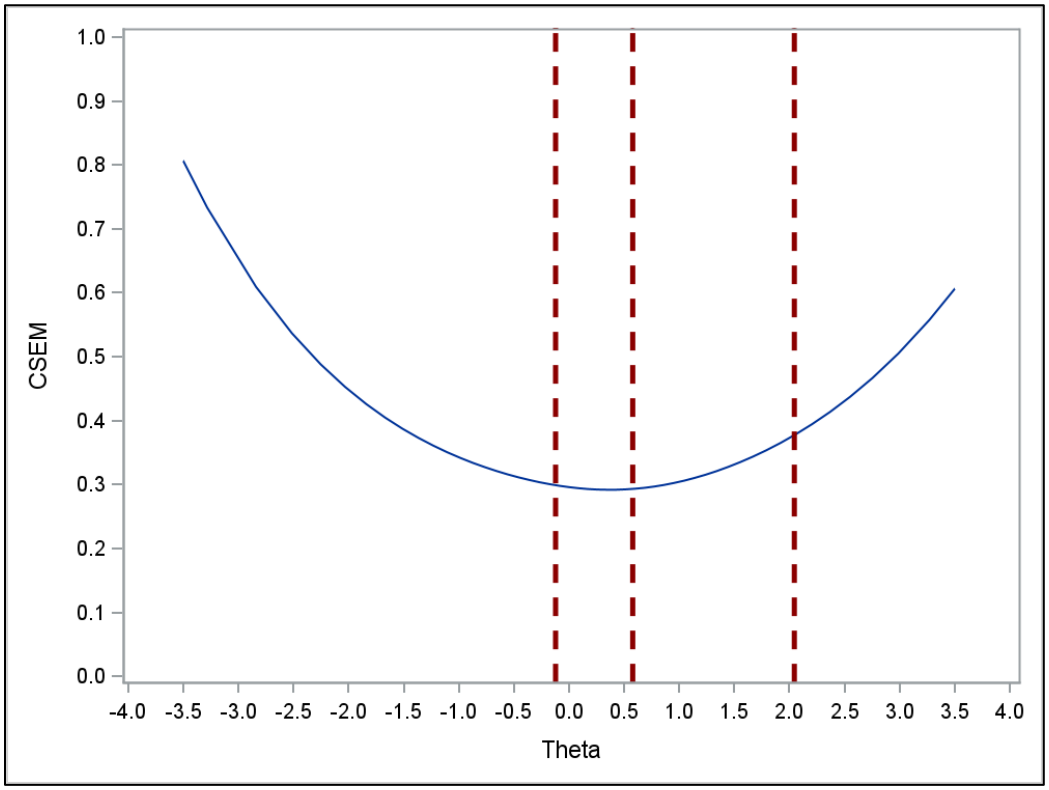


Figure B.21. TCC, ELA Grade 7

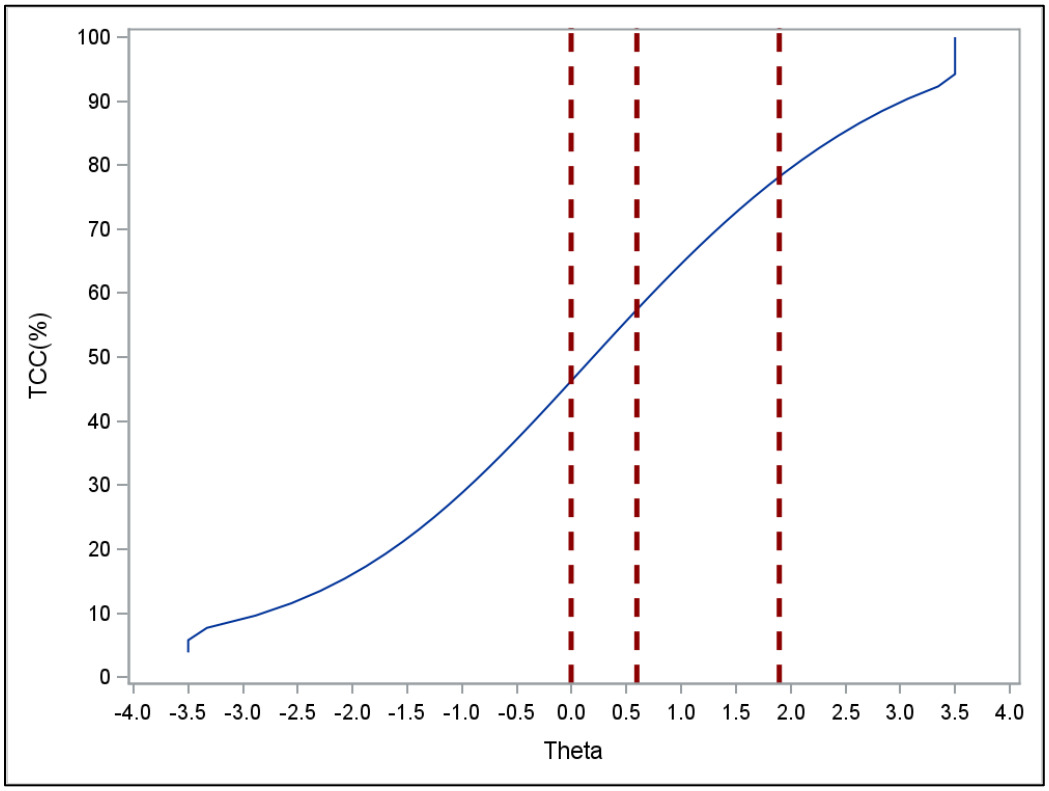


Figure B.22. CSEM, ELA Grade 7

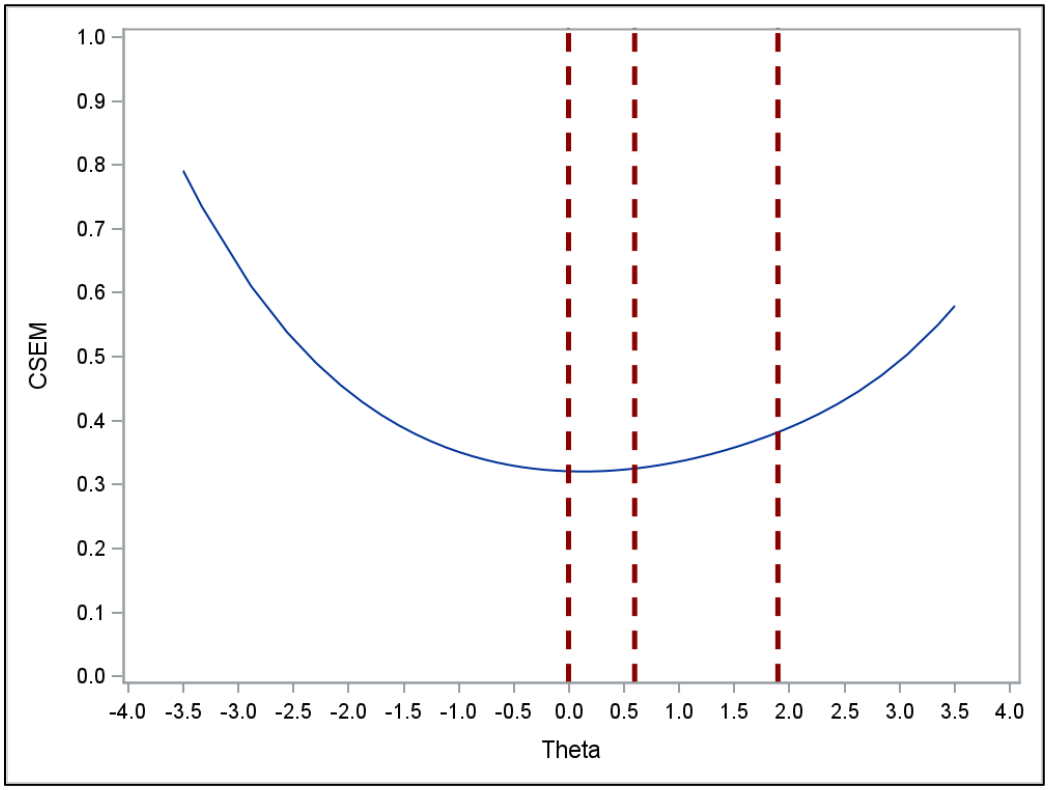


Figure B.23. TCC, ELA Grade 8

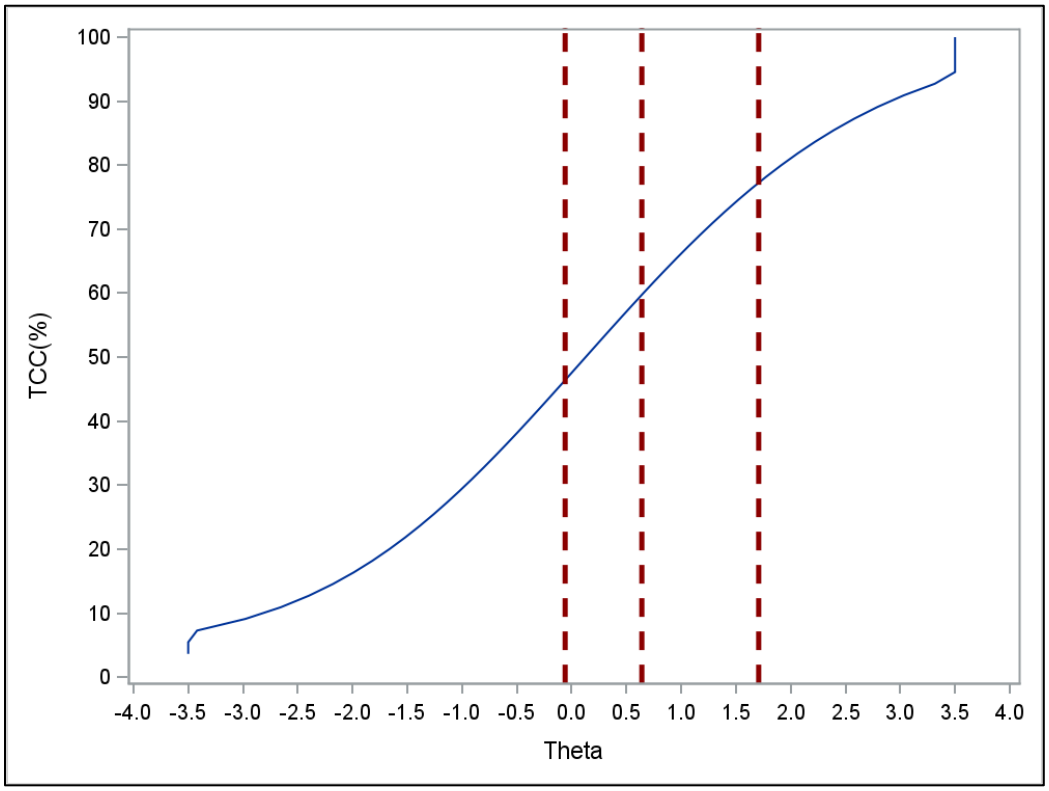


Figure B.24. CSEM, ELA Grade 8

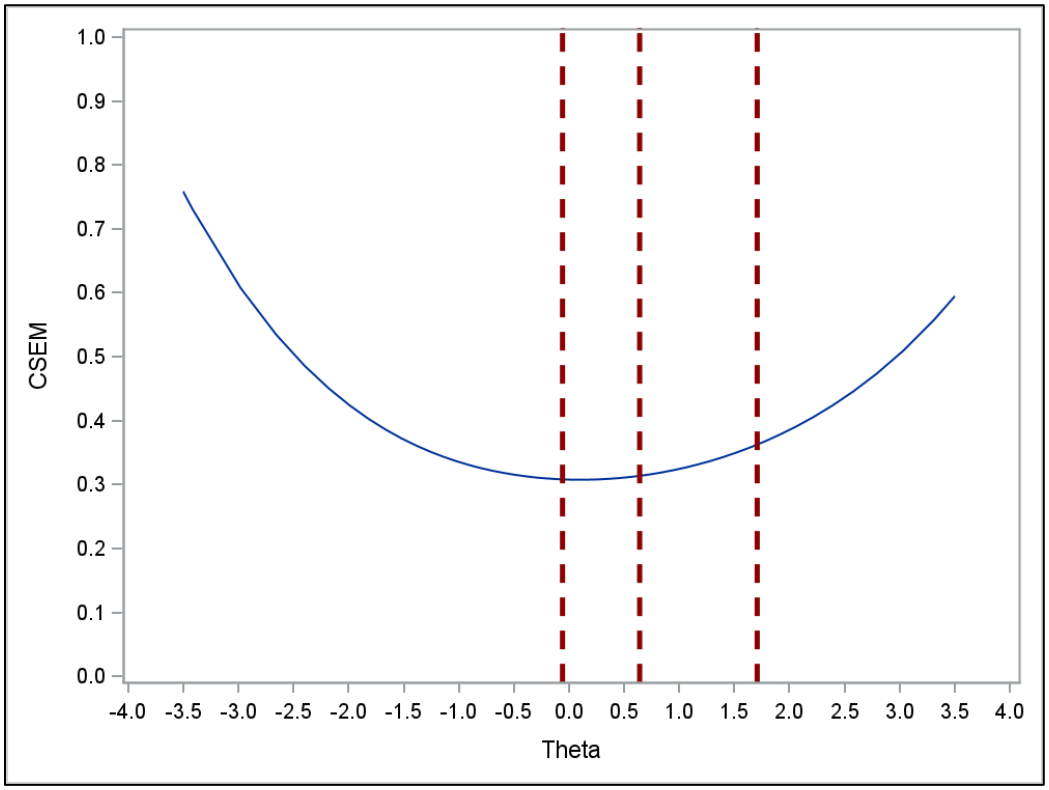


Figure B.25. TCC, Mathematics Grade 3

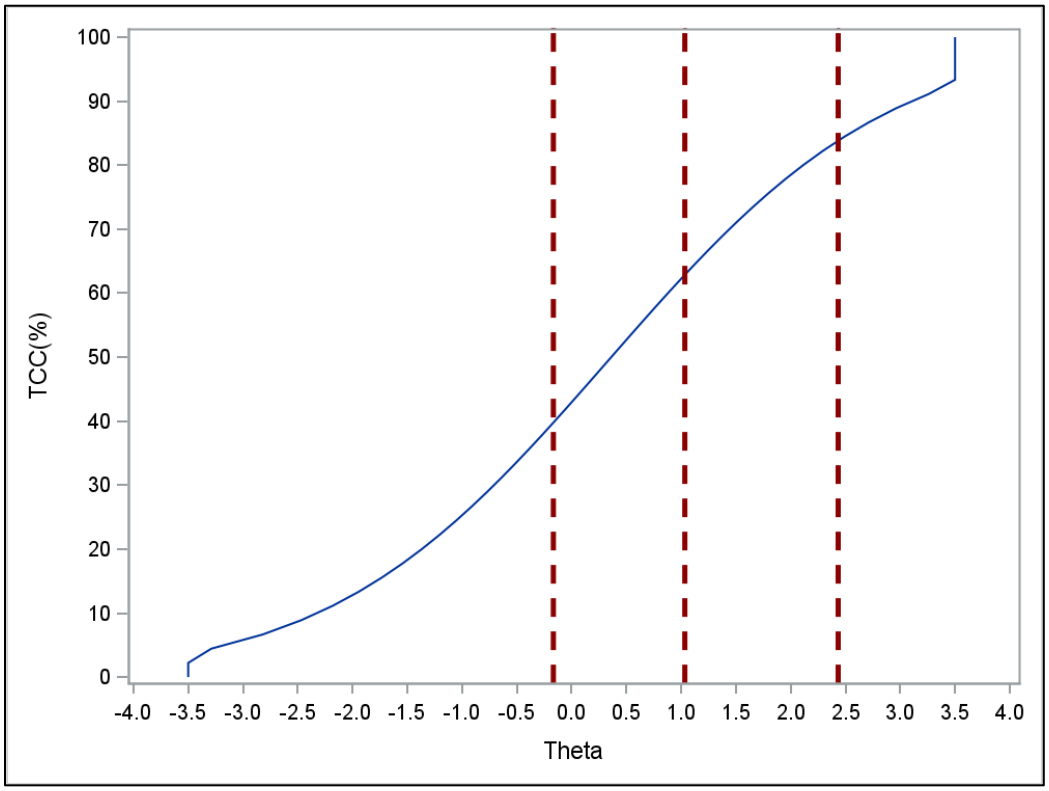


Figure B.26. CSEM, Mathematics Grade 3

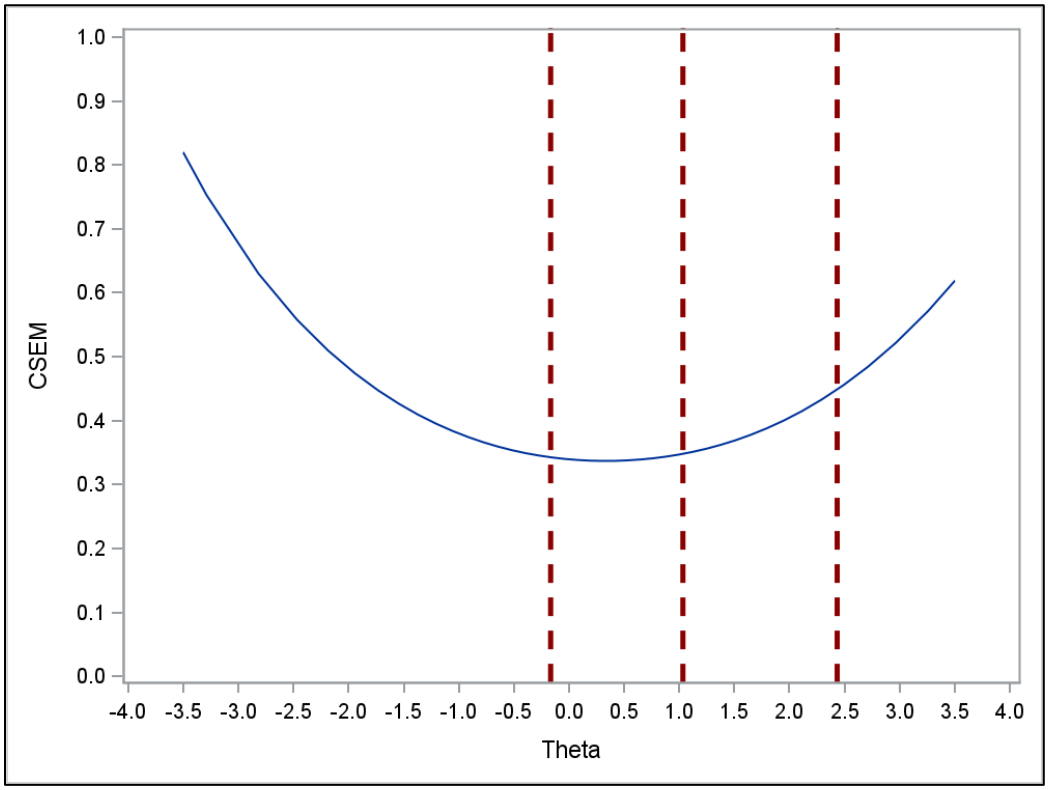


Figure B.27. TCC, Mathematics Grade 4

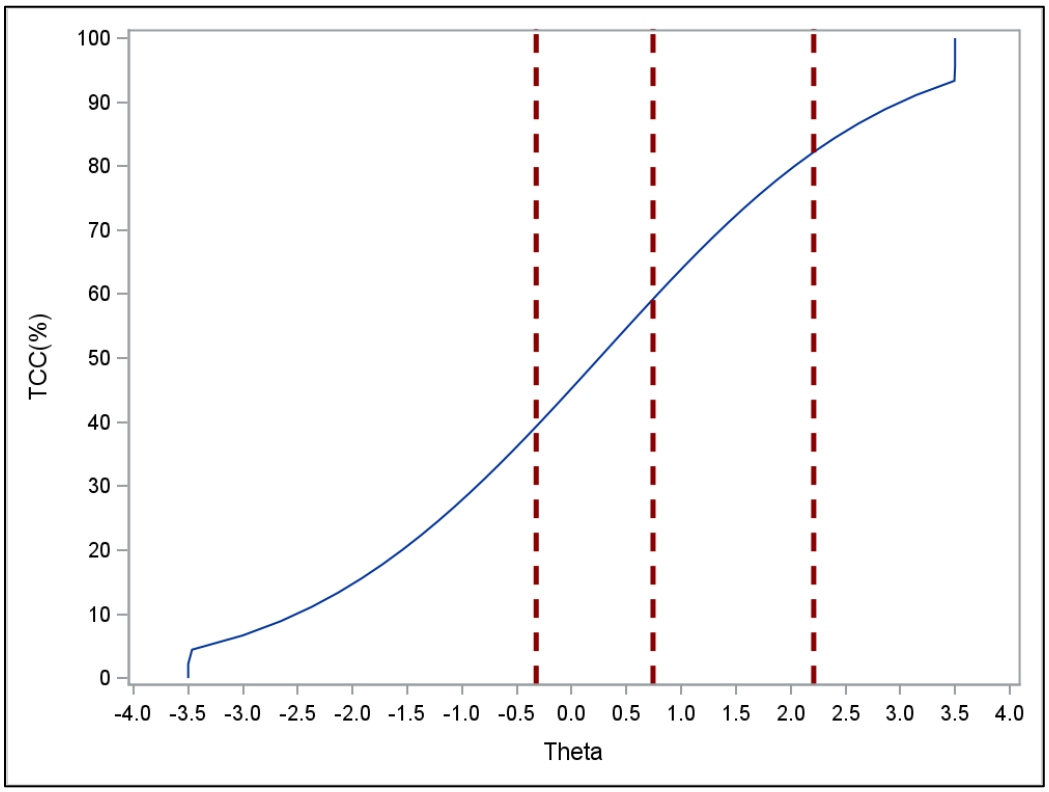


Figure B.28. CSEM, Mathematics Grade 4

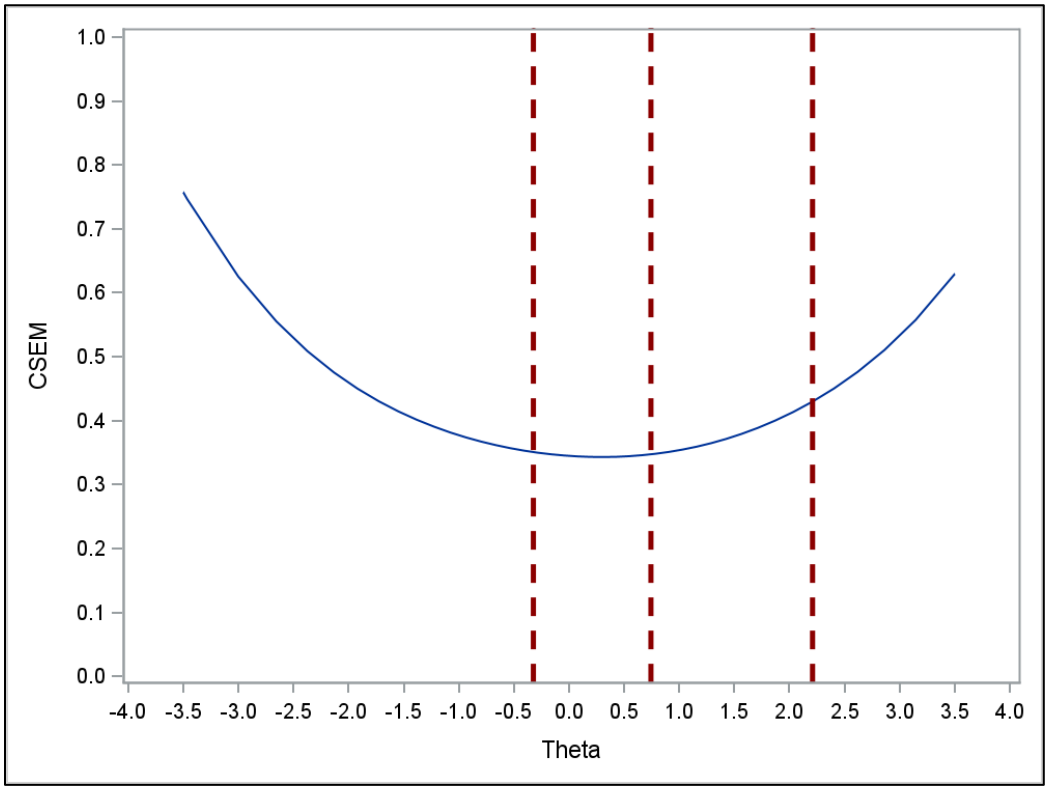


Figure B.29. TCC, Mathematics Grade 5

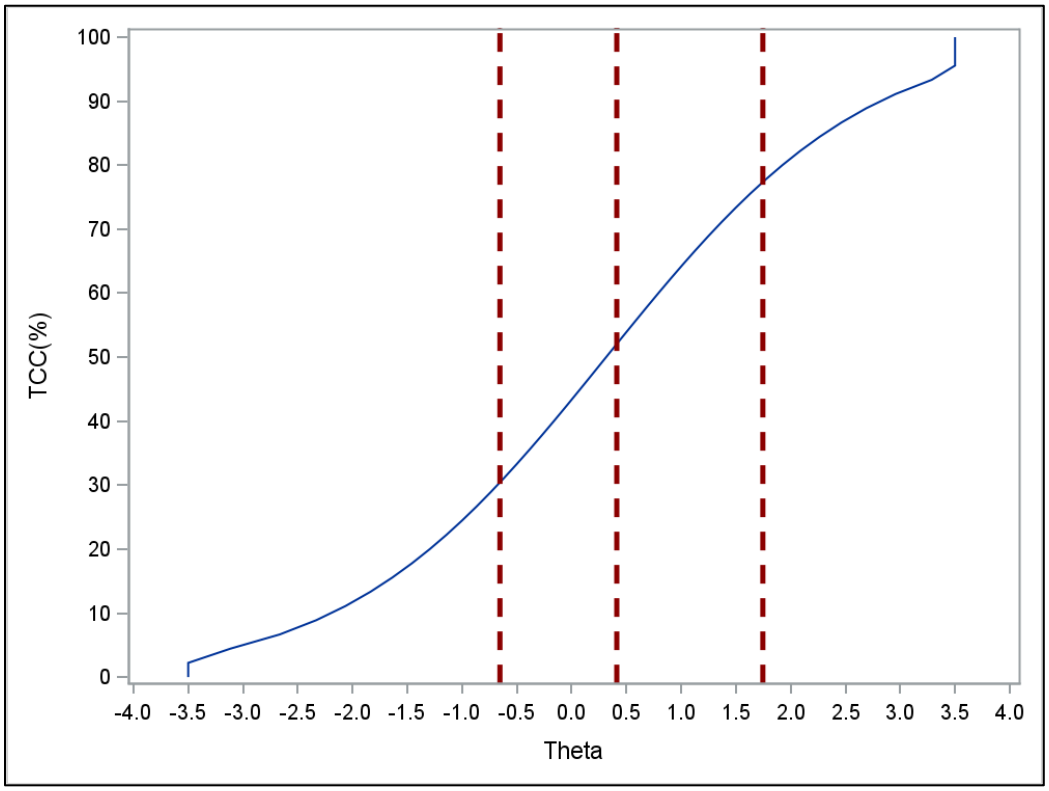


Figure B.30. CSEM, Mathematics Grade 5

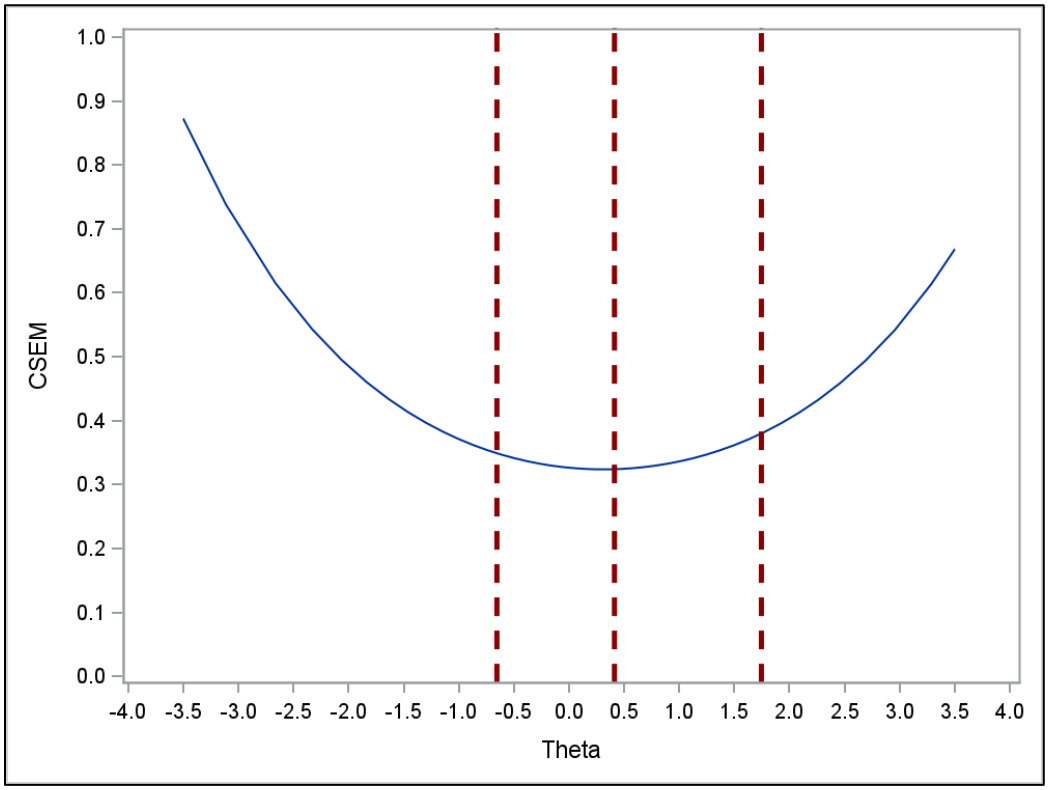


Figure B.31. TCC, Mathematics Grade 6

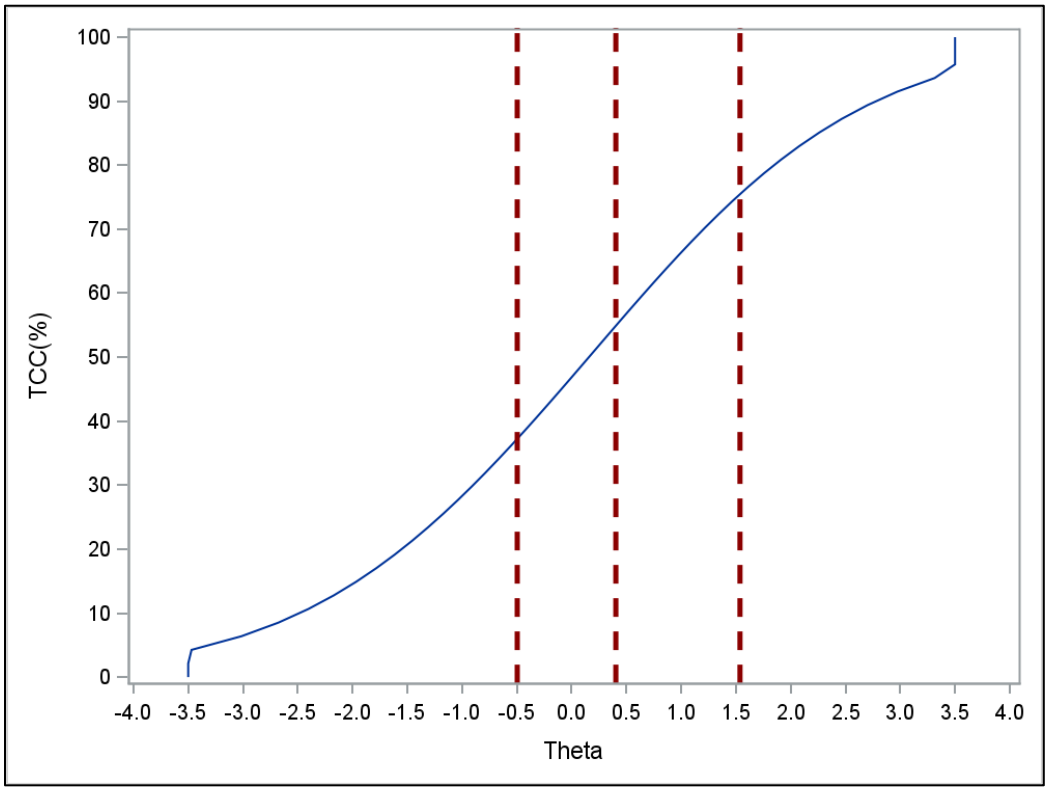


Figure B.32. CSEM, Mathematics Grade 6

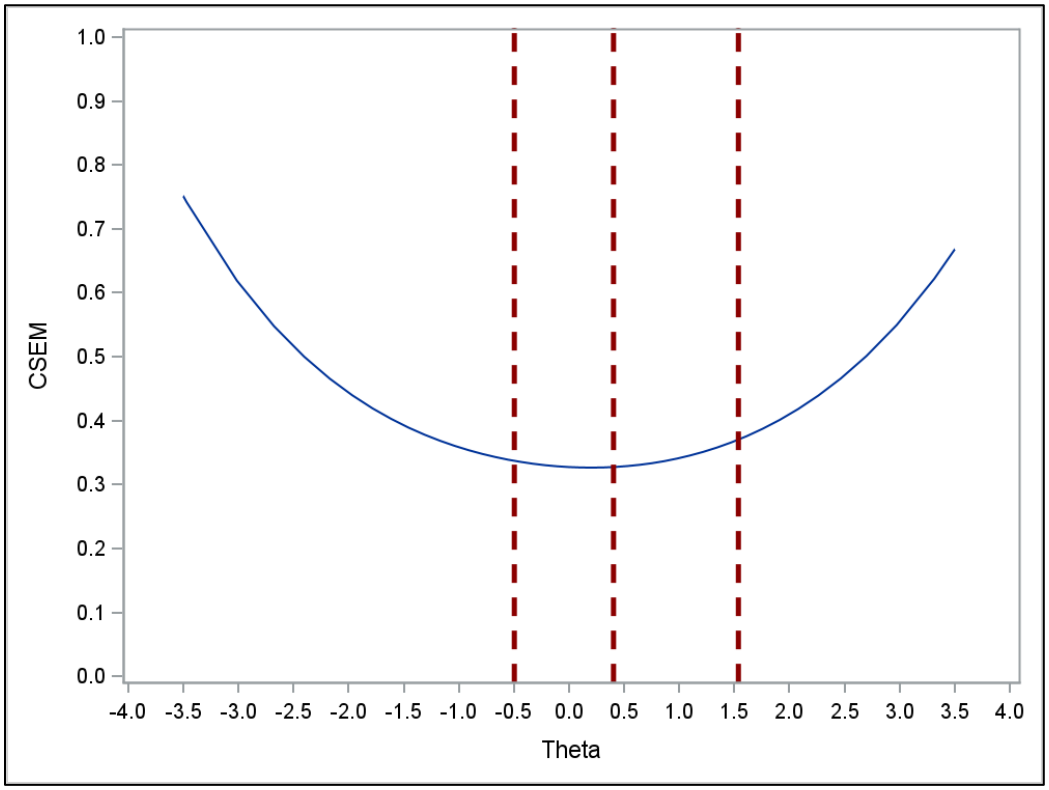


Figure B.33. TCC, Mathematics Grade 7

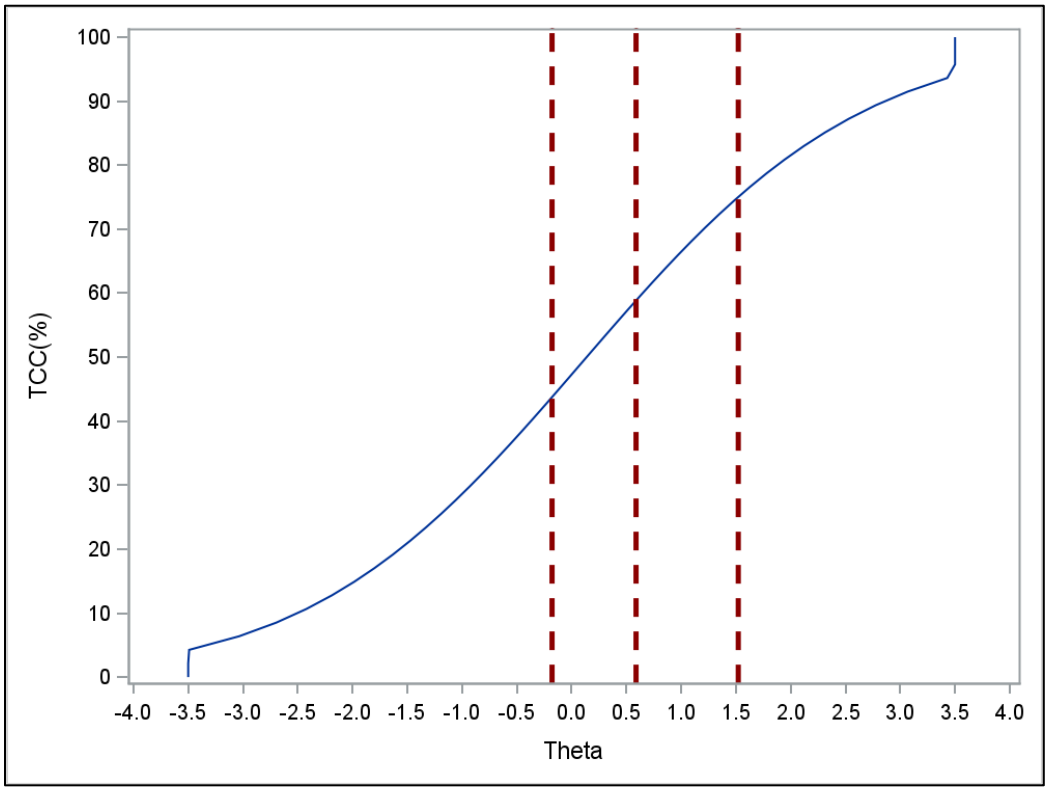


Figure B.34. CSEM, Mathematics Grade 7

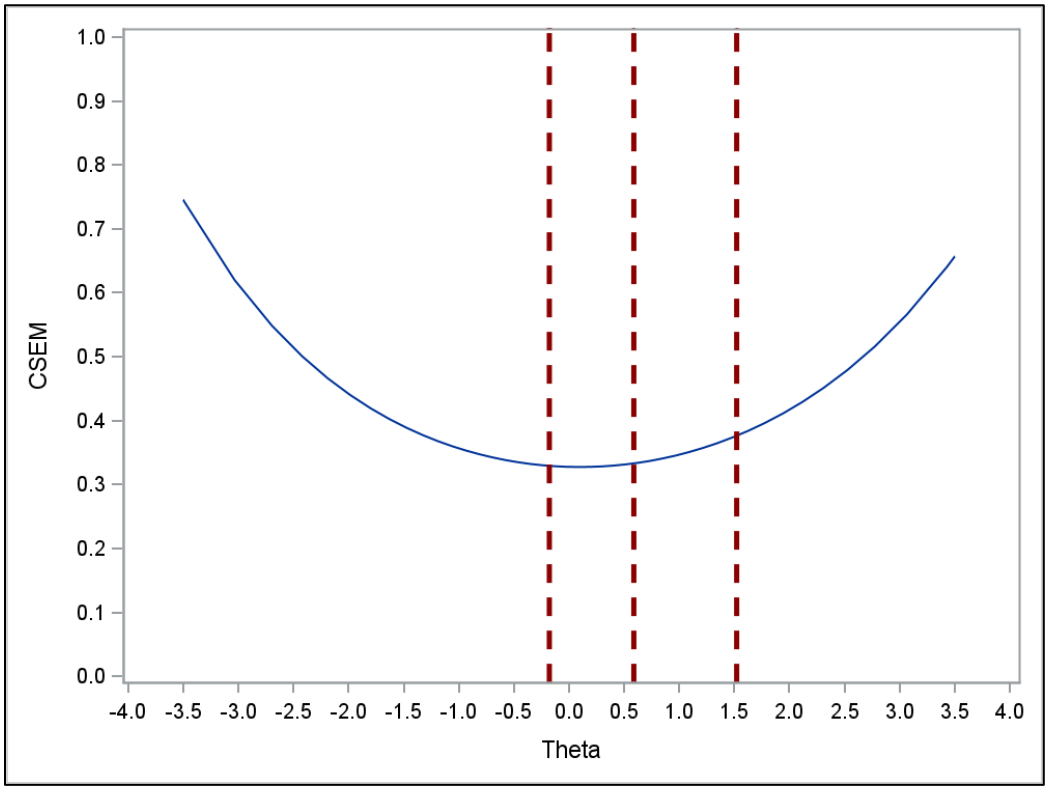


Figure B.35. TCC, Mathematics Grade 8

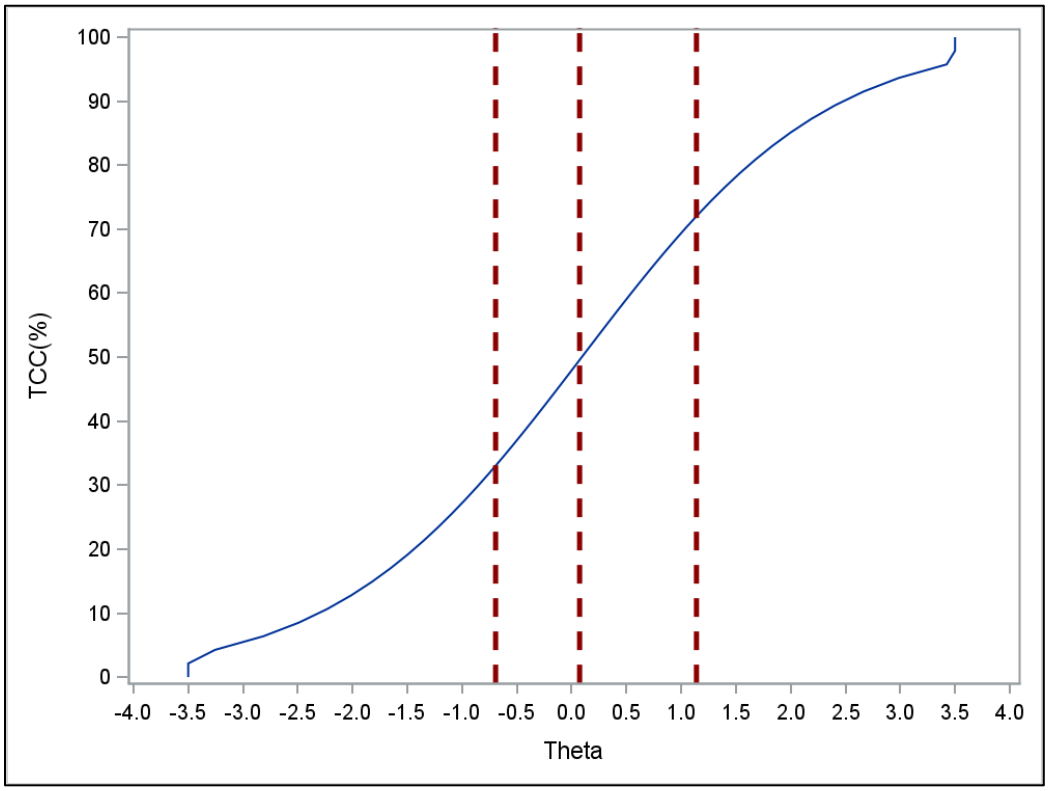


Figure B.36. CSEM, Mathematics Grade 8

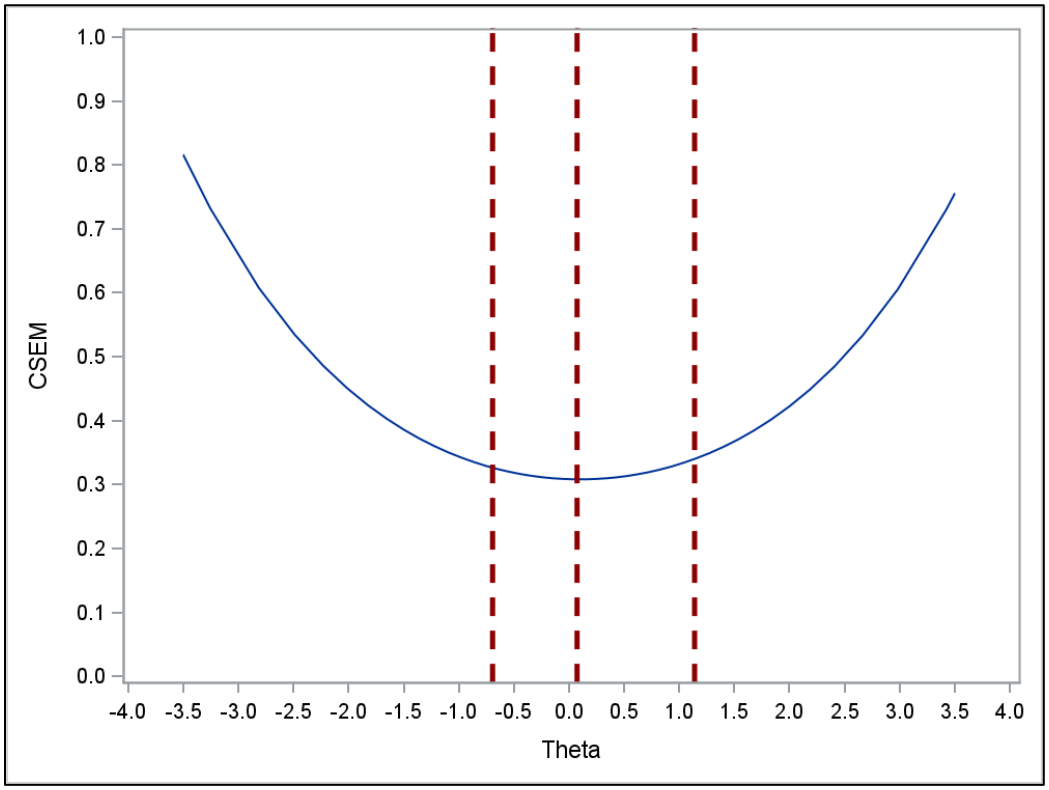


Figure B.37. Scree Plot, ELA Grade 3

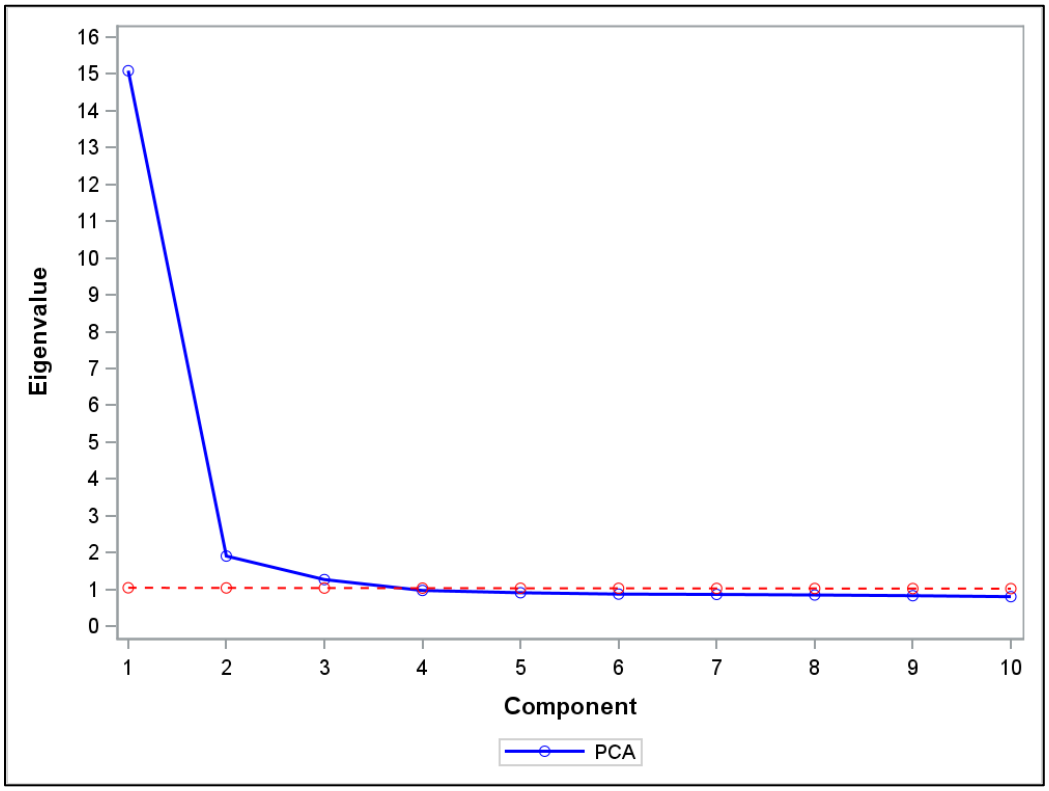


Figure B.38. Scree Plot, ELA Grade 4

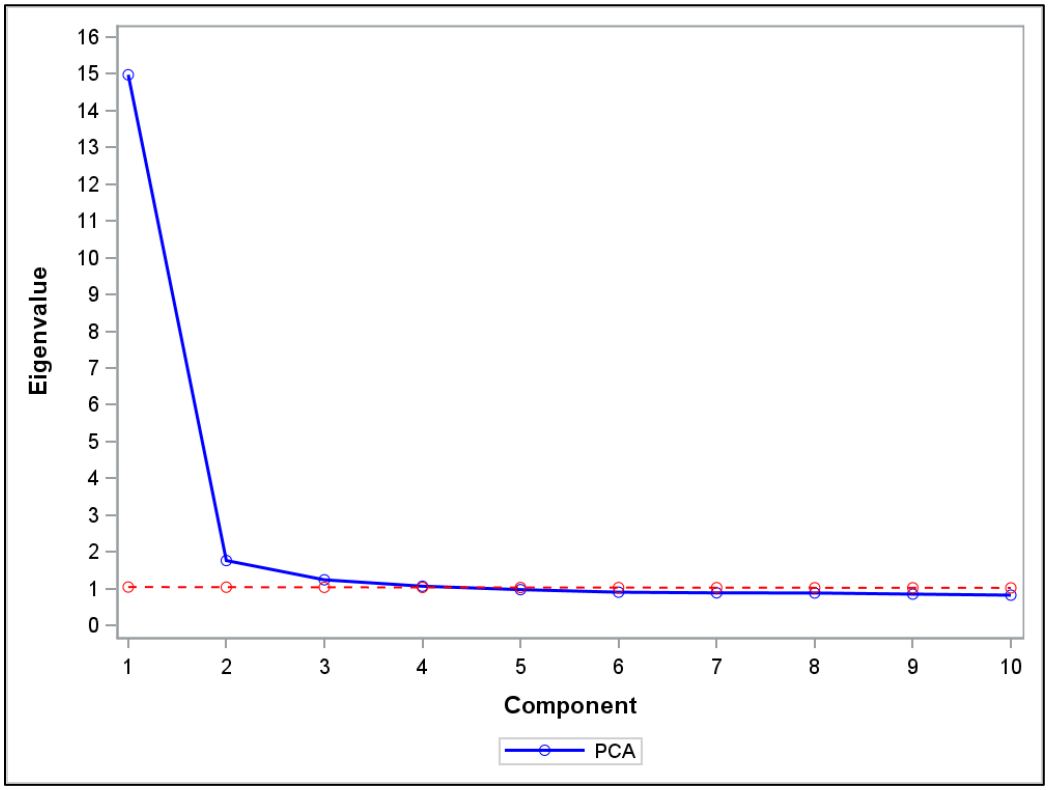


Figure B.39. Scree Plot, ELA Grade 5

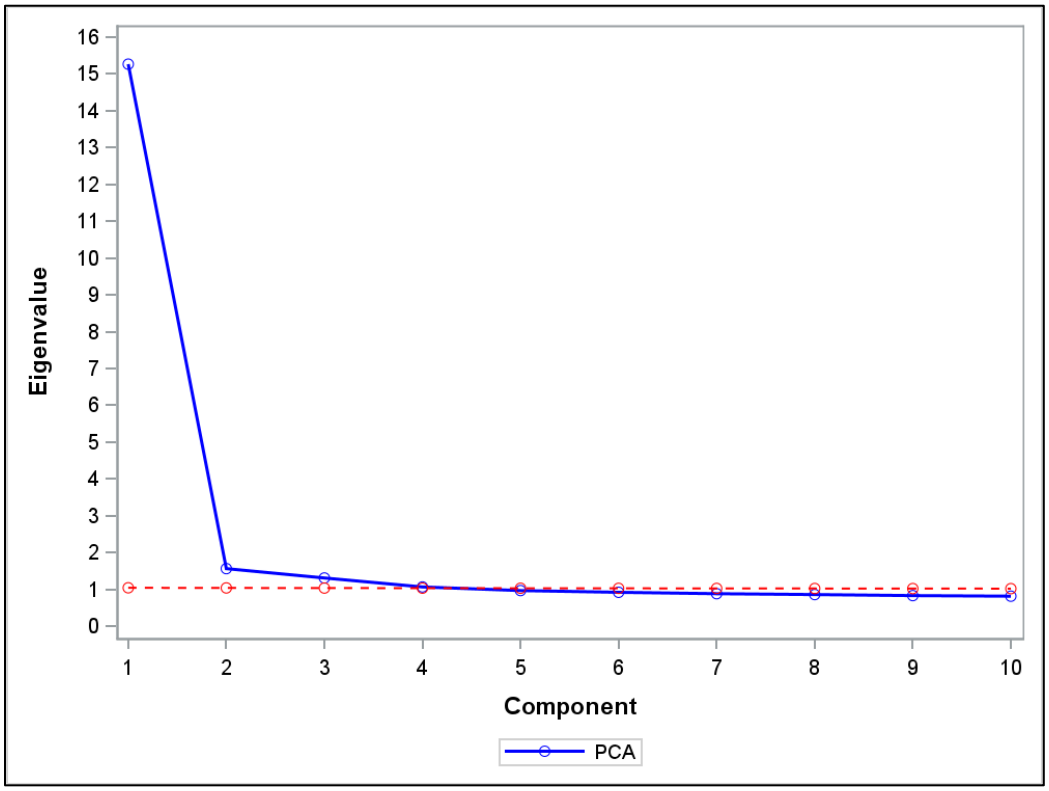


Figure B.40. Scree Plot, ELA Grade 6

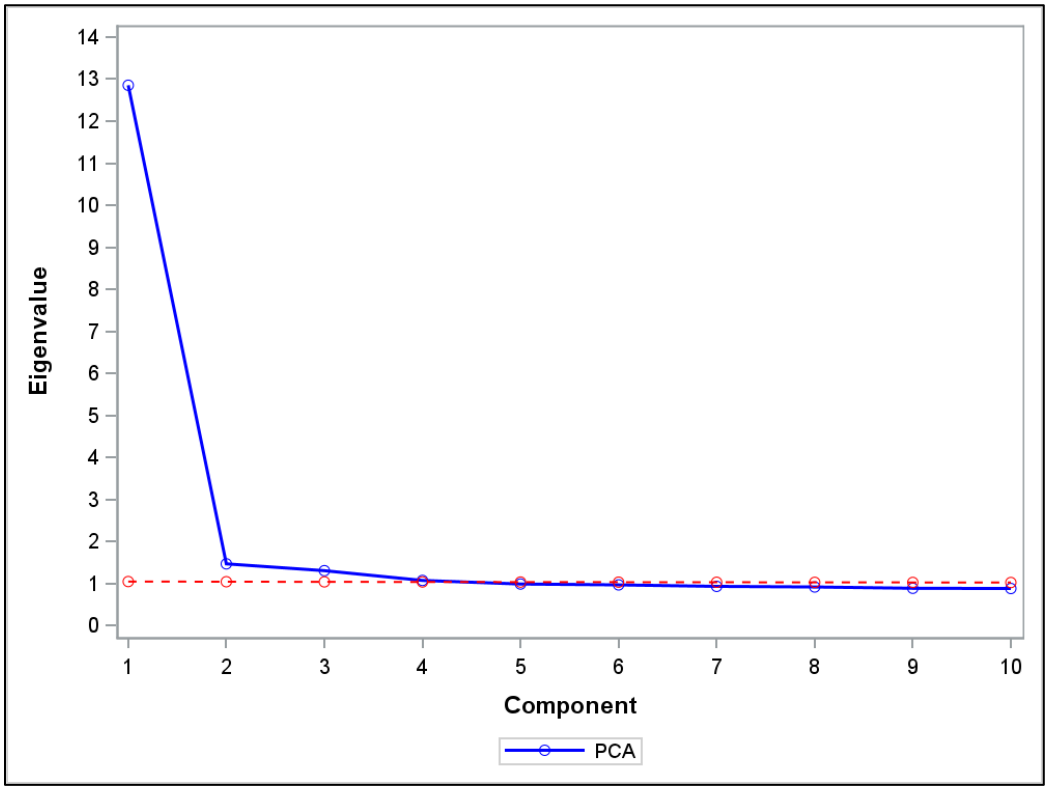


Figure B.41. Scree Plot, ELA Grade 7

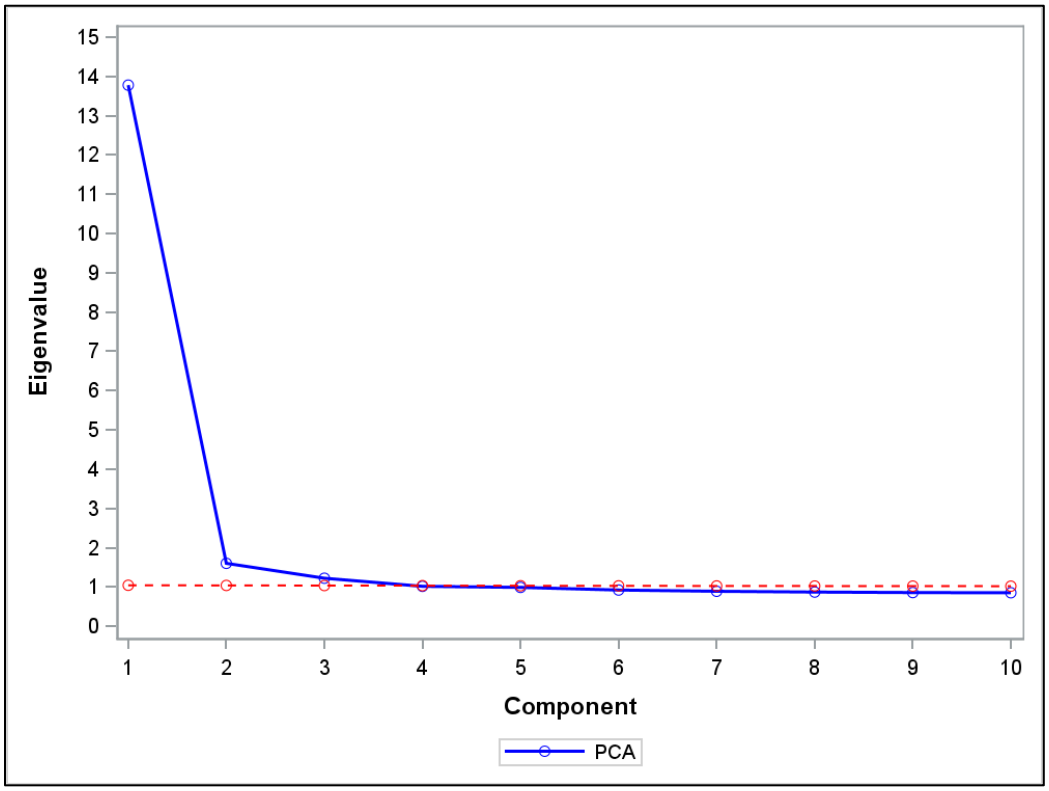


Figure B.42. Scree Plot, ELA Grade 8

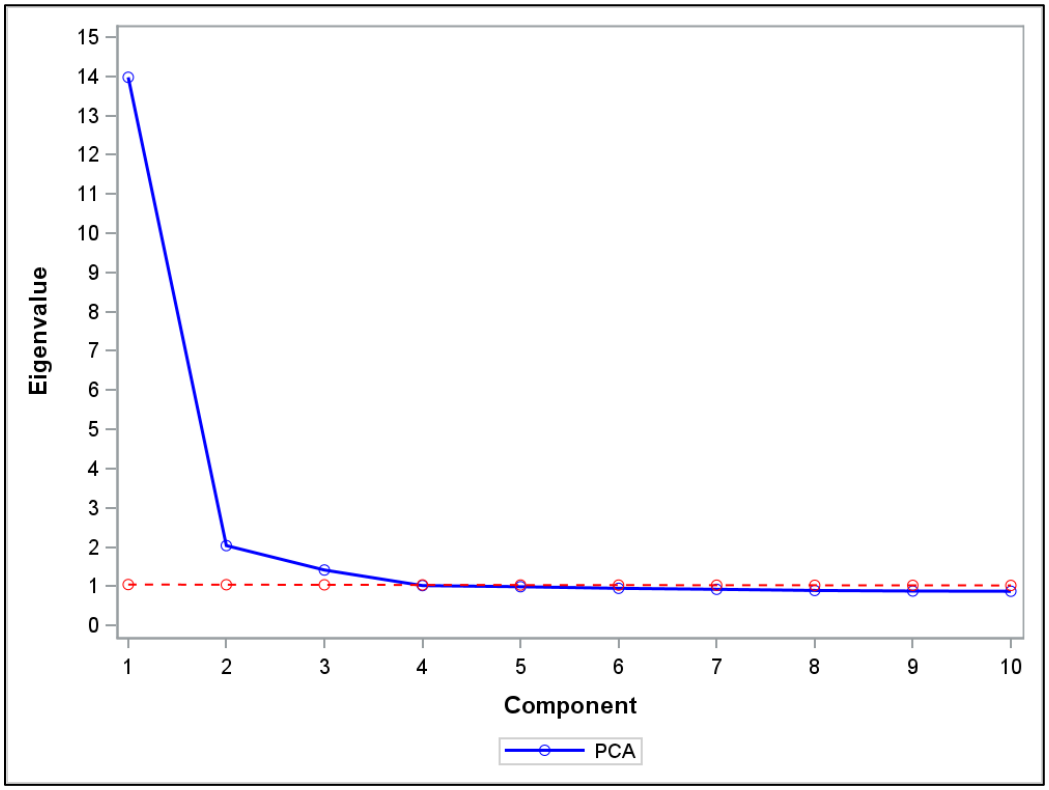


Figure B.43. Scree Plot, Mathematics Grade 3

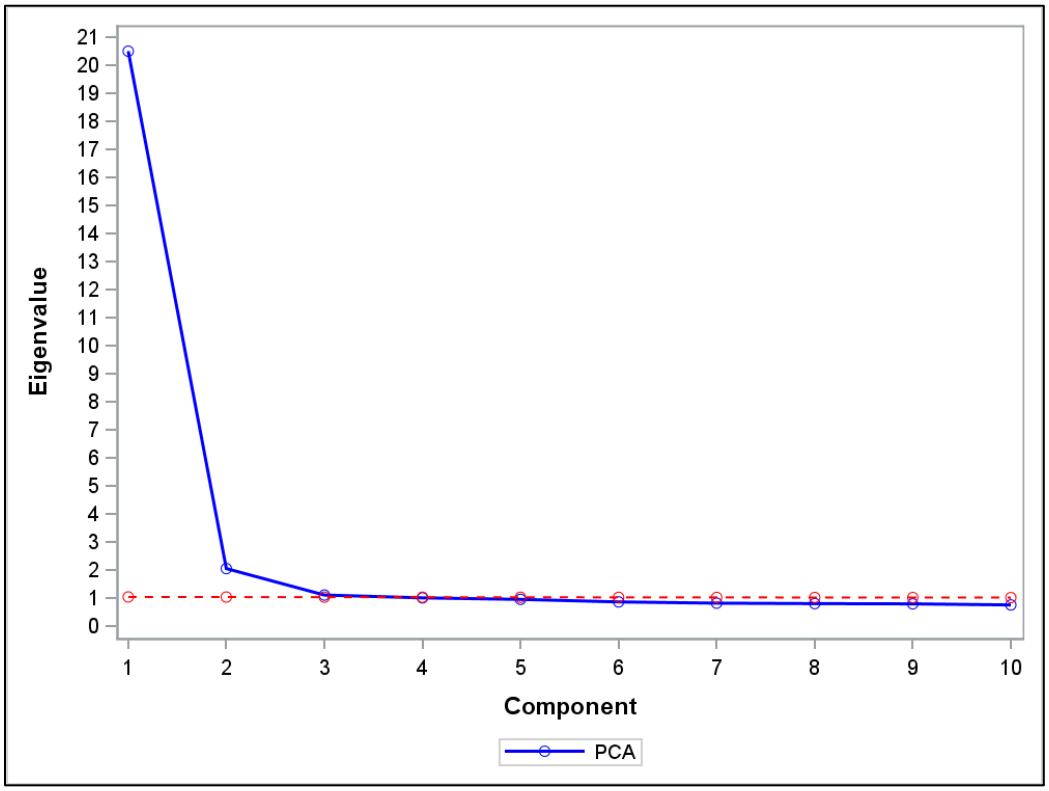


Figure B.44. Scree Plot, Mathematics Grade 4

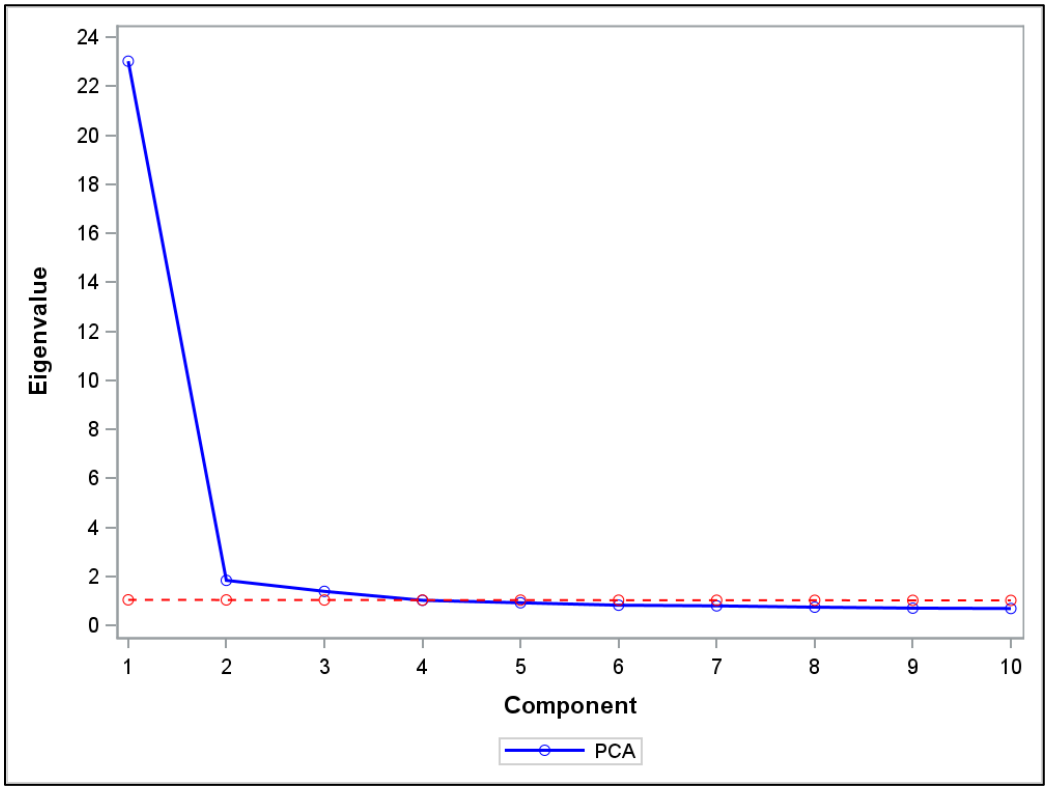


Figure B.45. Scree Plot, Mathematics Grade 5

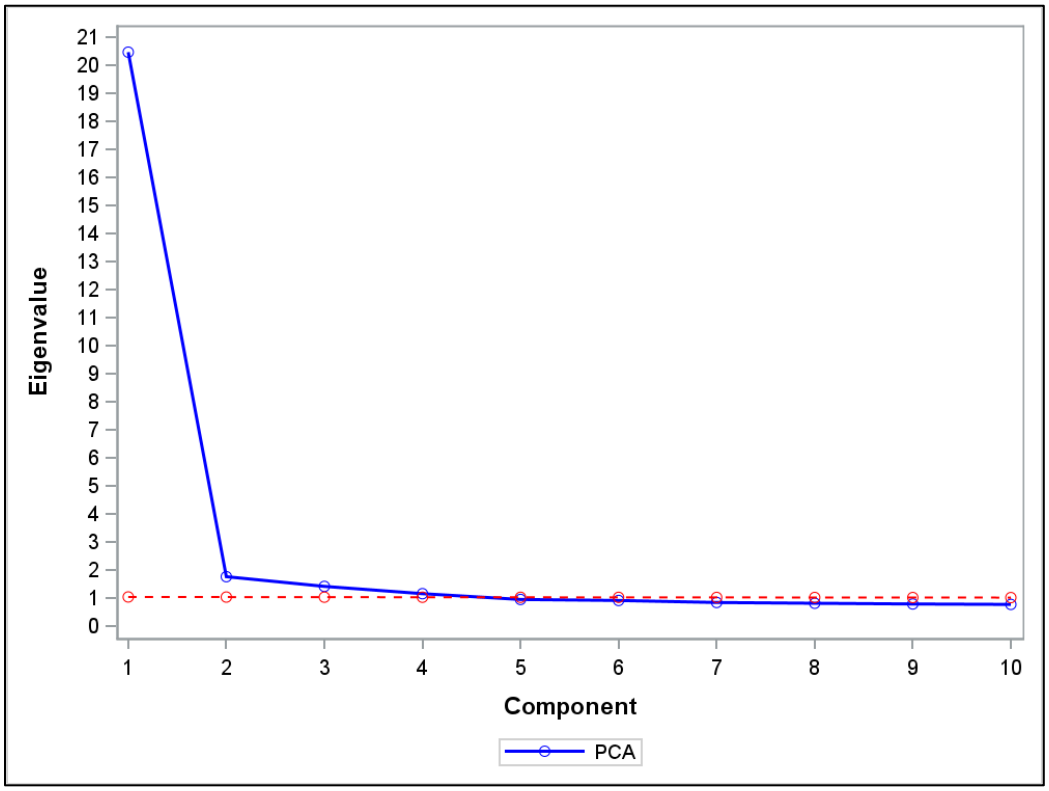


Figure B.46. Scree Plot, Mathematics Grade 6

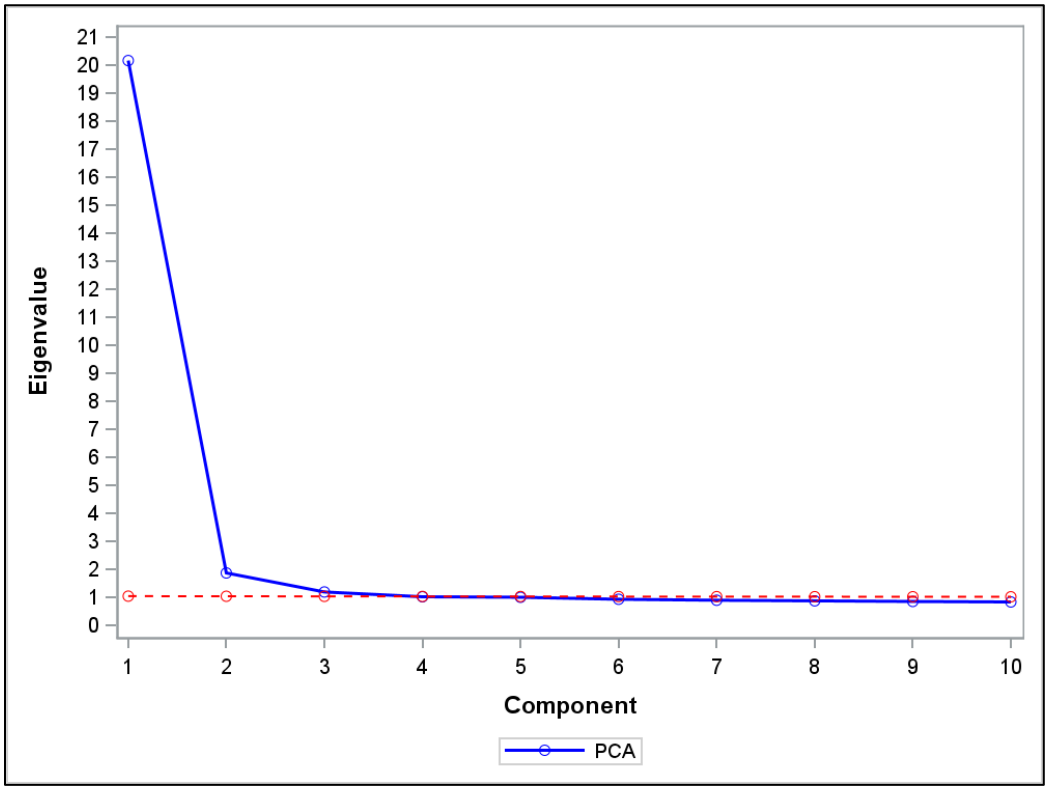


Figure B.47. Scree Plot, Mathematics Grade 7

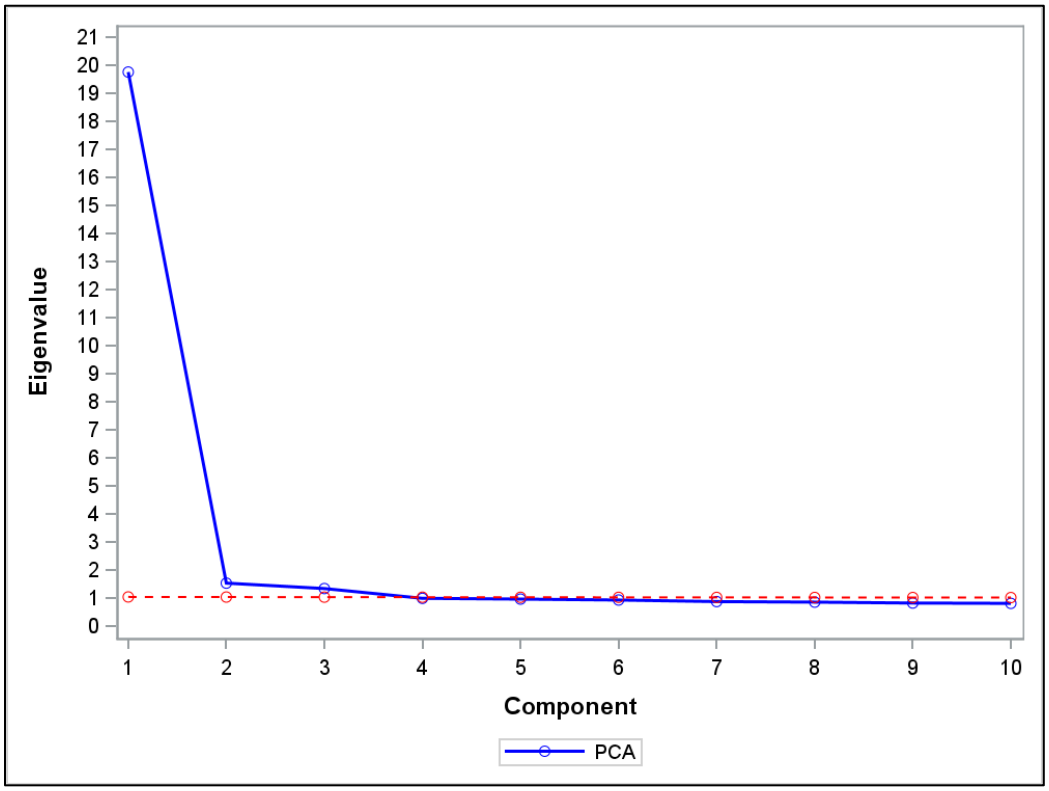
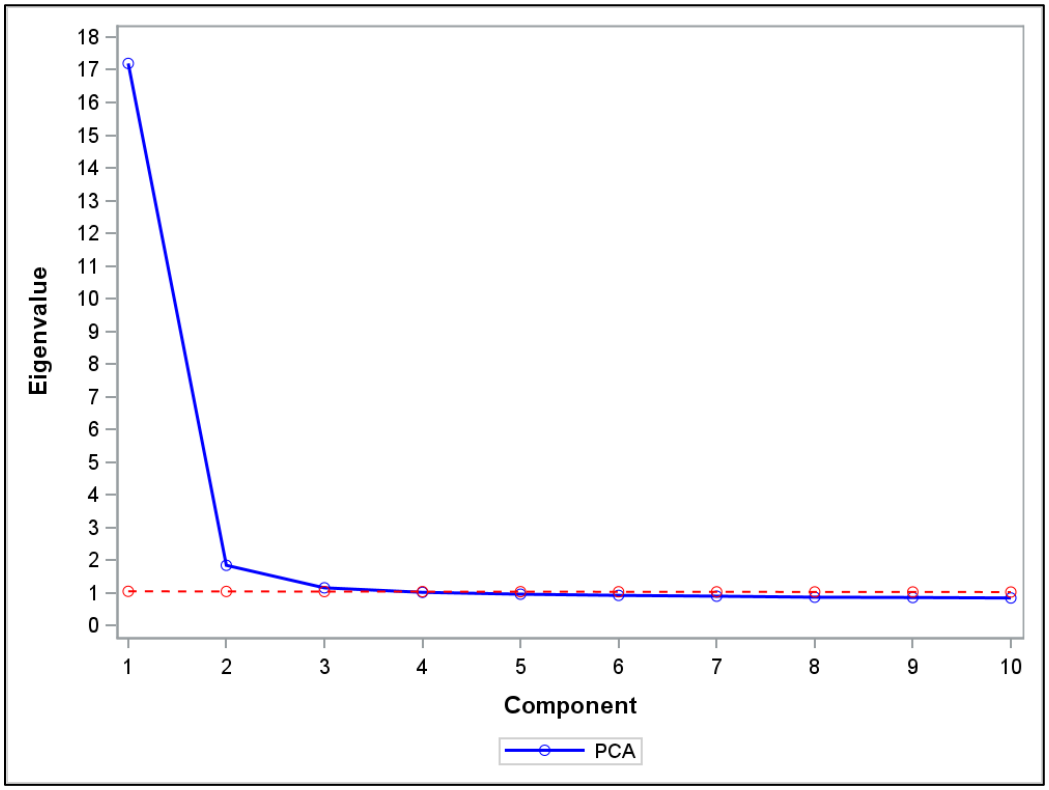


Figure B.48. Scree Plot, Mathematics Grade 8



Appendix C: ADMINISTRATION RESULTS

This appendix presents the Spring 2022 AASA results for all students and subgroups. Specifically:

- Table C.1 – Table C.12 present the overall results by subgroup, including the sample size, mean and standard deviation (SD) of the total scale score, and percentage of students at each performance level overall.
- Figure C.1 – Figure C.12 present histograms of the total scale score distribution.

Table C.1. Test Results by Subgroup, ELA Grade 3

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		79,804	2500.64	35.45	48	12	26	15
Gender	Male	40,297	2498.25	35.17	50	11	25	13
	Female	39,399	2503.11	35.57	45	12	27	16
	Missing	108	2488.58	33.03	62	9	19	10
Ethnicity	Hispanic	37,499	2491.84	32.52	58	12	22	8
	Non-Hispanic	42,196	2508.48	36.11	38	11	30	20
Race	American Indian	4,240	2482.63	29.34	71	10	15	4
	Asian	2,782	2522.20	36.15	24	10	34	32
	Black or African American	5,772	2490.88	32.06	59	12	21	8
	Multi-racial	4,913	2506.07	35.38	41	13	28	18
	Native Hawaiian or Other Pacific Islander	442	2496.48	31.90	52	14	24	10
	White	61,541	2501.44	35.36	46	12	27	15
	Missing	114	2488.19	32.77	62	10	18	10
Other	Special Ed.	11,530	2475.41	30.79	79	6	11	4
	EL	8,076	2468.12	21.67	90	5	4	1
	Low SES	31,407	2488.39	31.68	63	11	19	7

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

Table C.2. Test Results by Subgroup, ELA Grade 4

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		79,949	2519.10	34.04	42	14	31	13
Gender	Male	40,437	2517.21	34.14	44	13	30	13
	Female	39,408	2521.07	33.83	40	14	32	14
	Missing	104	2507.21	29.44	61	10	24	6
Ethnicity	Hispanic	37,338	2510.12	30.80	53	14	26	7
	Non-Hispanic	42,504	2527.02	34.78	33	13	35	19
Race	American Indian	4,210	2501.71	27.66	64	14	19	3
	Asian	2,834	2541.20	34.07	18	11	40	31
	Black or African American	5,632	2508.13	30.92	55	14	24	7
	Multi-racial	4,910	2523.82	33.81	36	14	34	16
	Native Hawaiian or Other Pacific Islander	417	2516.05	31.38	45	15	31	10
	White	61,836	2519.94	33.91	41	14	32	14
	Missing	110	2509.02	30.45	58	11	24	7
Other	Special Ed.	11,513	2493.34	29.48	76	8	12	4
	EL	7,519	2486.39	20.99	87	8	5	0
	Low SES	30,952	2507.06	30.20	57	14	23	6

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

Table C.3. Test Results by Subgroup, ELA Grade 5

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		80,649	2529.81	35.65	40	21	29	10
Gender	Male	40,640	2527.10	35.91	43	21	27	9
	Female	39,887	2532.60	35.17	37	22	30	11
	Missing	122	2522.11	31.68	51	21	24	4
Ethnicity	Hispanic	37,747	2521.08	33.36	50	22	23	6
	Non-Hispanic	42,780	2537.55	35.83	31	21	34	15
Race	American Indian	4,392	2510.74	30.27	63	20	15	2
	Asian	2,806	2552.20	33.84	16	18	42	24
	Black or African American	5,691	2517.80	32.94	53	21	21	4
	Multi-racial	4,733	2535.79	35.15	32	22	32	13
	Native Hawaiian or Other Pacific Islander	488	2526.89	32.43	43	22	28	7
	White	62,417	2530.83	35.47	38	21	29	11
	Missing	122	2522.11	31.68	51	21	24	4
Other	Special Ed.	11,303	2499.51	31.38	76	12	9	2
	EL	6,639	2493.36	23.60	87	10	3	0
	Low SES	31,252	2517.47	32.71	54	21	20	4

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

Table C.4. Test Results by Subgroup, ELA Grade 6

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		81,041	2542.68	30.12	36	25	35	4
Gender	Male	41,155	2539.37	30.40	40	25	32	3
	Female	39,776	2546.13	29.44	31	26	39	4
	Missing	110	2533.94	26.91	49	26	23	2
Ethnicity	Hispanic	38,173	2535.66	28.40	45	27	27	2
	Non-Hispanic	42,757	2548.97	30.23	28	24	42	5
Race	American Indian	4,391	2527.05	27.28	57	25	17	1
	Asian	2,616	2562.20	28.56	14	20	55	11
	Black or African American	5,742	2533.68	27.71	48	25	26	1
	Multi-racial	4,677	2546.92	29.73	30	27	39	4
	Native Hawaiian or Other Pacific Islander	431	2539.79	27.85	39	25	34	2
	White	63,072	2543.50	29.93	35	26	36	4
	Missing	112	2533.77	27.09	49	26	23	2
Other	Special Ed.	10,743	2515.14	26.49	76	14	9	1
	EL	6,431	2511.48	21.41	83	14	3	0
	Low SES	30,860	2532.75	28.07	49	26	24	1

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

Table C.5. Test Results by Subgroup, ELA Grade 7

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		83,804	2554.47	34.18	38	19	32	11
Gender	Male	42,954	2550.58	34.35	43	18	29	9
	Female	40,774	2558.61	33.51	33	20	35	12
	Missing	76	2539.42	30.53	50	26	21	3
Ethnicity	Hispanic	39,764	2546.05	31.21	47	21	26	6
	Non-Hispanic	43,964	2562.12	34.95	30	18	37	15
Race	American Indian	4,669	2536.47	28.03	60	19	18	2
	Asian	2,680	2580.78	34.31	14	12	43	31
	Black or African American	5,728	2544.77	30.92	49	20	26	5
	Multi-racial	4,538	2559.82	33.80	31	20	35	14
	Native Hawaiian or Other Pacific Islander	434	2551.89	33.75	39	21	29	10
	White	65,677	2555.19	33.97	37	19	33	11
	Missing	78	2539.79	30.41	50	26	22	3
Other	Special Ed.	10,139	2524.05	26.49	79	11	9	1
	EL	6,544	2519.30	20.09	87	9	4	0
	Low SES	31,303	2543.12	30.65	51	20	24	5

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

Table C.6. Test Results by Subgroup, ELA Grade 8

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		87,227	2558.97	33.04	42	22	26	10
Gender	Male	44,288	2554.10	33.11	49	21	23	8
	Female	42,849	2564.01	32.21	35	24	29	12
	Missing	90	2553.77	32.61	50	20	21	9
Ethnicity	Hispanic	39,562	2551.54	30.75	51	23	21	5
	Non-Hispanic	47,576	2565.16	33.60	35	22	30	13
Race	American Indian	4,964	2543.96	28.55	62	21	15	3
	Asian	2,735	2583.15	33.60	17	18	35	30
	Black or African American	5,893	2550.44	31.01	52	23	20	5
	Multi-racial	4,830	2562.64	32.58	38	22	28	11
	Native Hawaiian or Other Pacific Islander	508	2554.61	30.54	49	22	22	7
	White	68,176	2559.62	32.82	41	23	27	10
	Missing	121	2549.90	32.95	56	18	18	7
Other	Special Ed.	10,087	2528.19	25.29	83	11	5	1
	EL	6,897	2526.09	20.30	88	10	2	0
	Low SES	32,039	2548.93	30.35	55	22	19	4

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

Table C.7. Test Results by Subgroup, Mathematics Grade 3

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		80,808	3515.80	44.57	33	27	28	12
Gender	Male	40,953	3517.81	45.61	32	27	28	13
	Female	39,740	3513.77	43.38	35	28	27	10
	Missing	115	3498.90	41.66	50	27	18	5
Ethnicity	Hispanic	38,002	3504.58	41.03	42	30	22	6
	Non-Hispanic	42,690	3525.83	45.21	25	25	32	17
Race	American Indian	4,332	3492.13	38.90	55	27	15	3
	Asian	2,816	3548.74	42.21	11	17	39	32
	Black or African American	5,881	3499.43	41.68	48	28	19	5
	Multi-racial	4,975	3520.58	44.60	29	27	30	14
	Native Hawaiian or Other Pacific Islander	450	3510.17	42.77	37	32	22	9
	White	62,233	3517.19	44.00	32	28	29	12
	Missing	121	3498.04	41.67	51	26	17	5
Other	Special Ed.	11,850	3486.52	42.95	61	21	13	4
	EL	8,251	3480.32	33.52	68	23	8	1
	Low SES	31,862	3500.21	40.80	46	29	20	5

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

Table C.8. Test Results by Subgroup, Mathematics Grade 4

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		80,600	3545.07	51.24	38	23	25	14
Gender	Male	40,860	3547.43	53.46	37	21	26	16
	Female	39,635	3542.71	48.72	39	25	25	11
	Missing	105	3518.88	50.52	59	17	20	4
Ethnicity	Hispanic	37,676	3531.90	47.82	48	25	20	7
	Non-Hispanic	42,816	3556.73	51.32	29	22	30	20
Race	American Indian	4,268	3517.39	45.91	61	21	14	4
	Asian	2,855	3583.43	46.70	13	15	34	38
	Black or African American	5,699	3524.22	47.48	54	23	17	5
	Multi-racial	4,950	3550.32	49.96	33	24	29	15
	Native Hawaiian or Other Pacific Islander	426	3541.20	46.57	41	26	24	10
	White	62,291	3546.77	50.63	36	23	26	14
	Missing	111	3520.13	51.37	59	16	21	5
Other	Special Ed.	11,716	3508.33	50.06	68	16	11	5
	EL	7,611	3500.95	40.39	76	16	7	1
	Low SES	31,236	3527.26	47.52	52	24	18	6

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

Table C.9. Test Results by Subgroup, Mathematics Grade 5

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		81,283	3577.68	44.21	39	24	26	11
Gender	Male	40,989	3578.73	45.73	39	22	26	13
	Female	40,170	3576.65	42.59	40	25	25	10
	Missing	124	3561.82	36.95	54	24	19	3
Ethnicity	Hispanic	38,069	3566.36	40.08	49	25	20	6
	Non-Hispanic	43,090	3587.73	45.27	31	23	30	16
Race	American Indian	4,448	3554.04	36.49	62	23	13	2
	Asian	2,862	3613.83	42.34	13	16	38	33
	Black or African American	5,738	3559.61	39.25	56	23	17	4
	Multi-racial	4,770	3583.16	43.58	34	24	28	13
	Native Hawaiian or Other Pacific Islander	492	3573.75	42.39	41	28	23	8
	White	62,849	3579.00	43.79	38	24	26	12
	Missing	124	3561.82	36.95	54	24	19	3
Other	Special Ed.	11,461	3545.73	39.52	71	16	10	3
	EL	6,727	3541.74	32.84	76	16	6	1
	Low SES	31,499	3562.56	39.84	53	24	18	5

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

Table C.10. Test Results by Subgroup, Mathematics Grade 6

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		81,769	3607.68	41.96	48	21	20	11
Gender	Male	41,540	3609.18	43.29	47	21	20	12
	Female	40,111	3606.18	40.50	50	22	19	9
	Missing	118	3590.26	32.98	62	27	8	3
Ethnicity	Hispanic	38,532	3596.71	37.29	59	21	15	5
	Non-Hispanic	43,118	3617.54	43.46	38	21	24	16
Race	American Indian	4,461	3585.09	33.82	72	16	10	2
	Asian	2,633	3641.97	43.68	18	18	31	32
	Black or African American	5,817	3590.54	35.72	66	19	11	4
	Multi-racial	4,706	3612.26	41.43	43	23	23	12
	Native Hawaiian or Other Pacific Islander	440	3602.63	39.02	52	25	14	9
	White	63,592	3609.14	41.63	46	22	21	11
	Missing	120	3590.29	33.26	62	27	9	3
Other	Special Ed.	10,871	3575.27	33.93	82	10	6	2
	EL	6,519	3571.22	27.31	87	9	3	1
	Low SES	31,172	3592.87	36.48	63	20	13	4

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

Table C.11. Test Results by Subgroup, Mathematics Grade 7

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		84,940	3626.67	41.14	56	17	14	13
Gender	Male	43,594	3628.76	42.45	53	18	15	14
	Female	41,267	3624.50	39.61	58	17	14	11
	Missing	79	3610.63	34.00	67	20	11	1
Ethnicity	Hispanic	40,289	3615.75	36.47	67	16	11	6
	Non-Hispanic	44,572	3636.58	42.61	45	19	18	18
Race	American Indian	4,789	3604.49	32.41	79	12	6	3
	Asian	2,724	3663.93	44.23	23	16	21	40
	Black or African American	5,814	3609.90	34.90	74	14	8	5
	Multi-racial	4,619	3630.66	41.25	51	19	15	14
	Native Hawaiian or Other Pacific Islander	439	3622.06	39.96	57	20	14	9
	White	66,474	3627.99	40.63	54	18	15	13
	Missing	81	3610.77	33.75	67	21	11	1
Other	Special Ed.	10,375	3593.95	31.06	87	7	3	2
	EL	6,689	3589.89	24.86	93	5	1	1
	Low SES	31,743	3612.50	35.76	71	15	9	5

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

Table C.12. Test Results by Subgroup, Mathematics Grade 8

Subgroup		N	Total Scale Score		% at Performance Levels			
			Mean	SD	1	2	3	4
All		88,301	3653.19	36.48	55	18	17	10
Gender	Male	44,901	3654.23	38.05	55	17	16	12
	Female	43,311	3652.13	34.75	56	19	17	9
	Missing	89	3646.36	32.02	69	15	11	6
Ethnicity	Hispanic	39,972	3643.93	30.84	66	17	12	5
	Non-Hispanic	48,241	3660.88	38.93	46	19	20	15
Race	American Indian	5,065	3637.02	27.09	77	13	8	3
	Asian	2,850	3691.56	45.68	21	15	25	39
	Black or African American	5,968	3640.61	29.21	71	15	10	4
	Multi-racial	4,901	3655.62	36.71	52	20	17	11
	Native Hawaiian or Other Pacific Islander	515	3648.20	30.87	59	19	17	5
	White	68,883	3653.77	35.95	54	19	17	10
	Missing	119	3642.79	31.34	74	12	9	5
Other	Special Ed.	10,268	3627.92	24.60	87	7	4	2
	EL	6,984	3625.75	19.66	91	6	2	0
	Low SES	32,416	3641.64	29.90	70	16	11	4

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

Figure C.1. Total Scale Score Distribution, ELA Grade 3

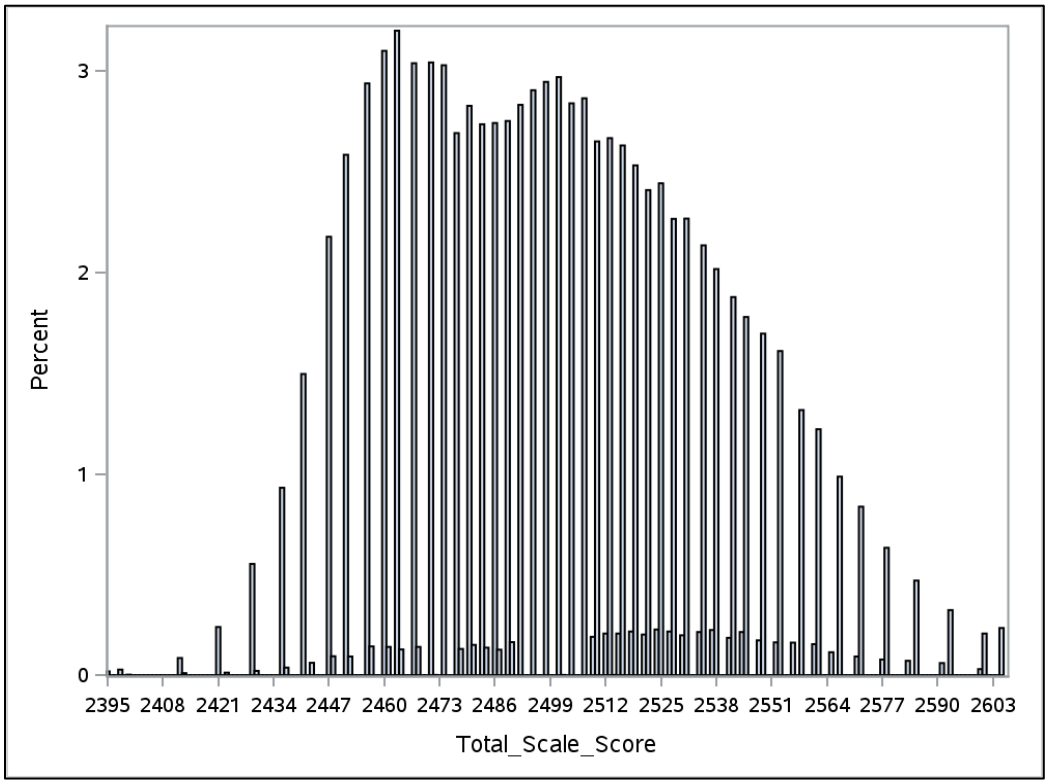


Figure C.2. Total Scale Score Distribution, ELA Grade 4

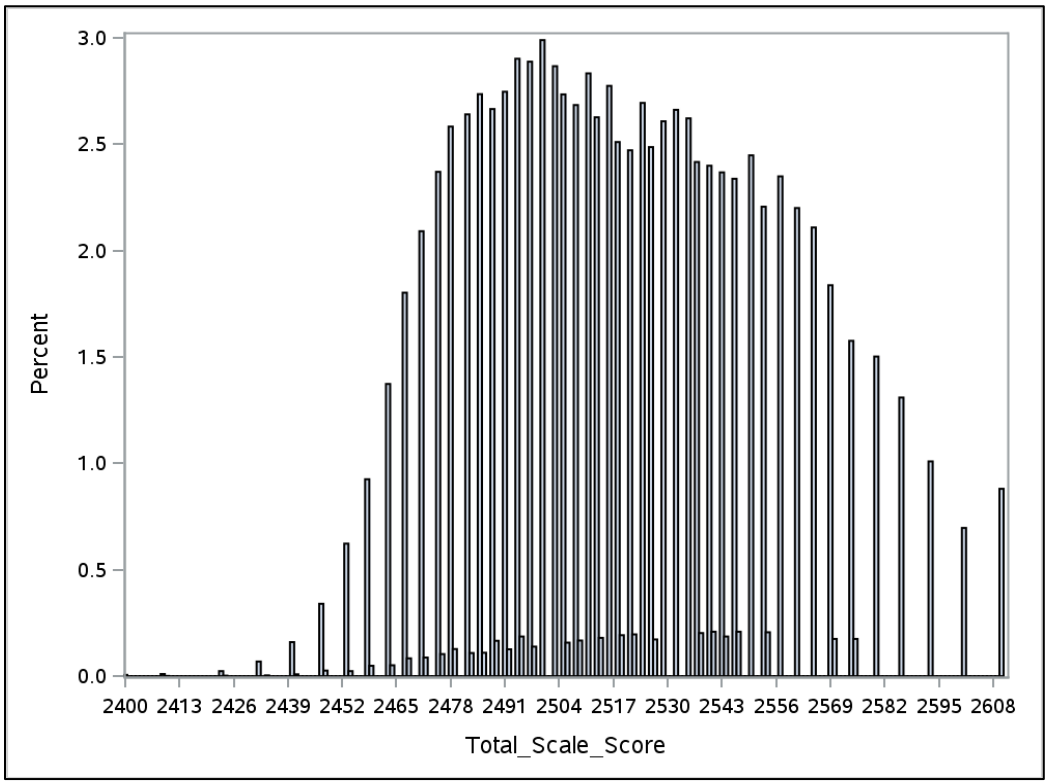


Figure C.3. Total Scale Score Distribution, ELA Grade 5

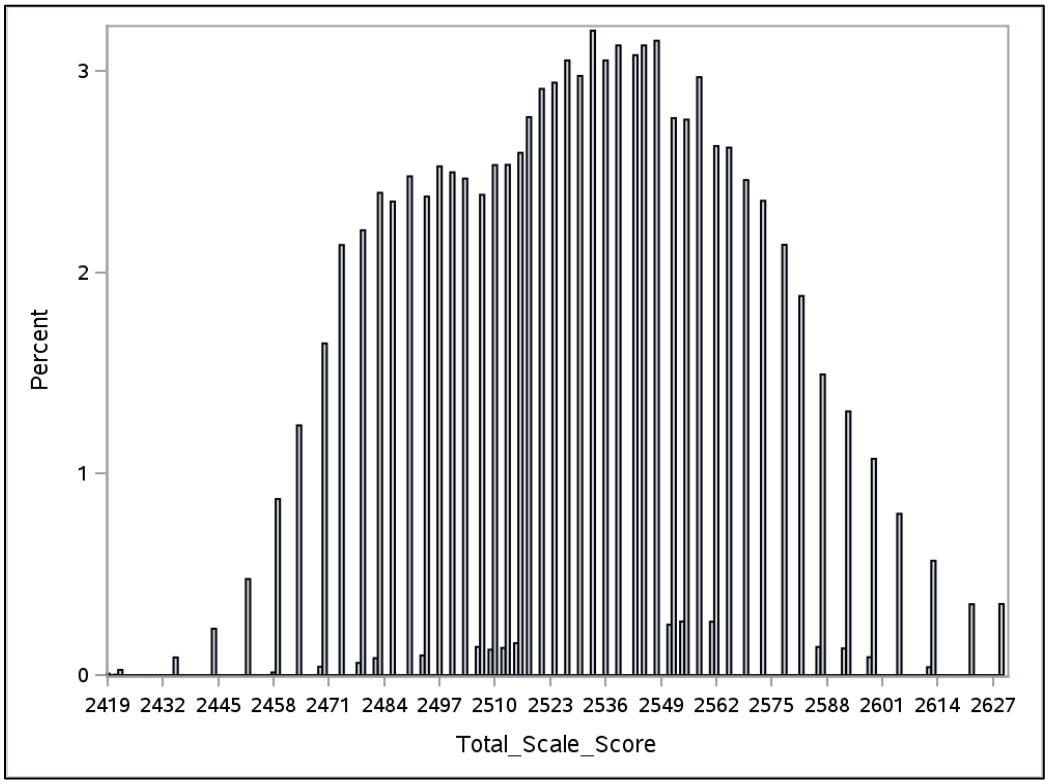


Figure C.4. Total Scale Score Distribution, ELA Grade 6

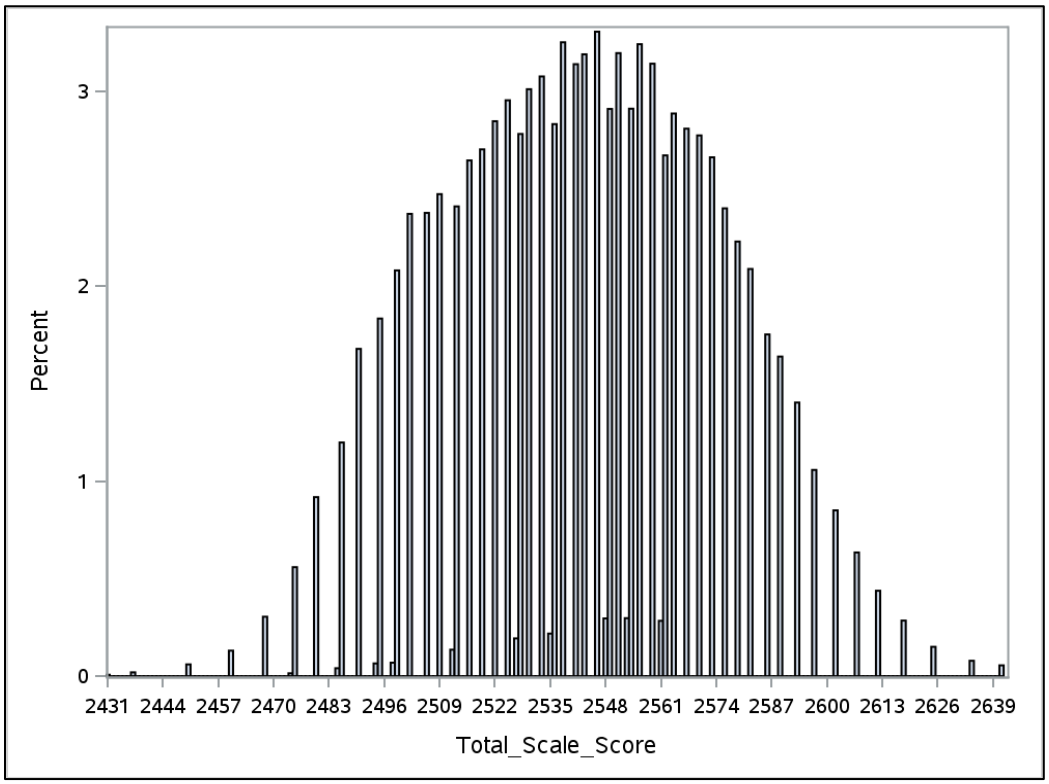


Figure C.5. Total Scale Score Distribution, ELA Grade 7

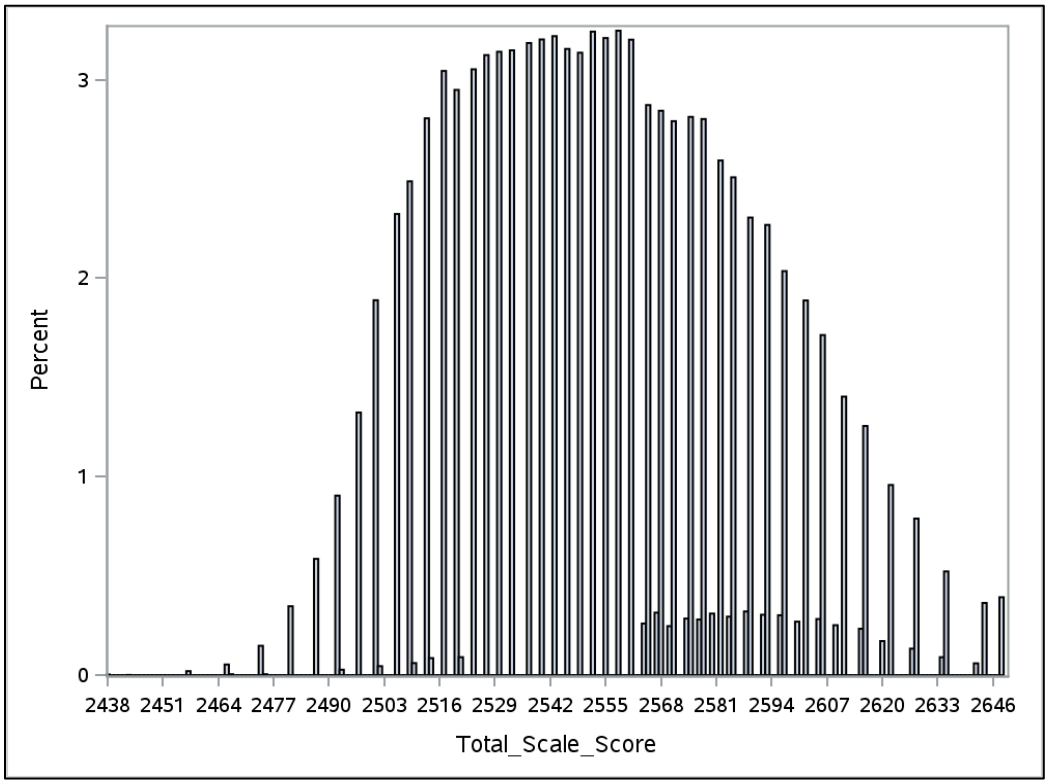


Figure C.6. Total Scale Score Distribution, ELA Grade 8

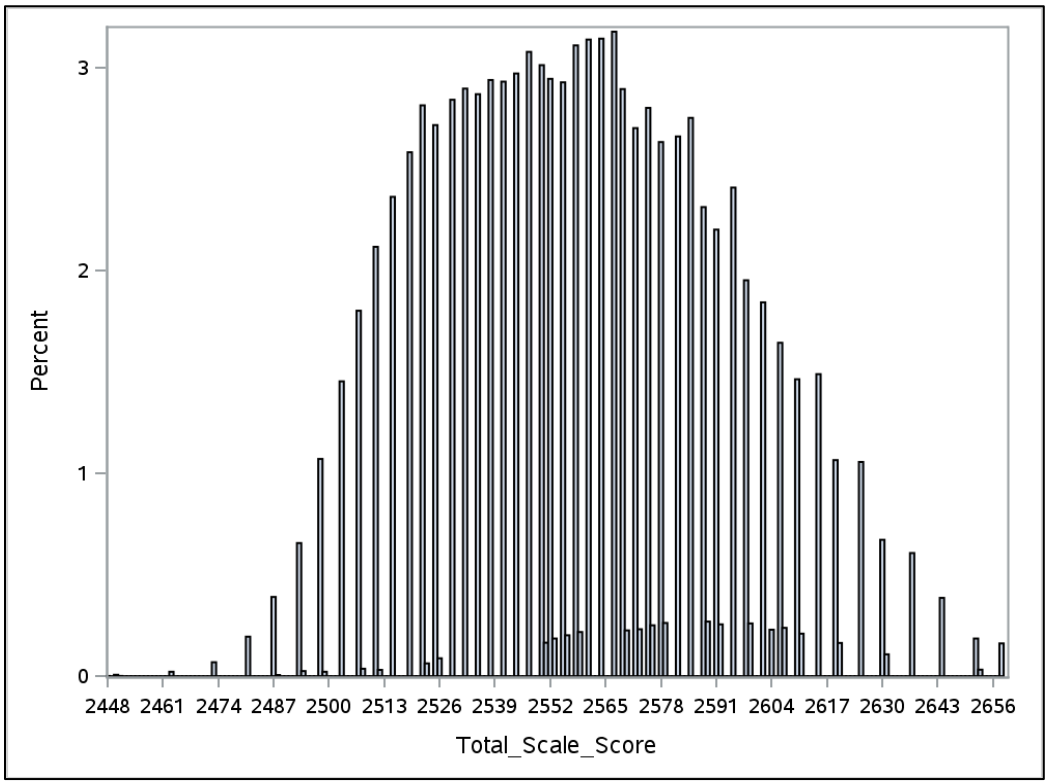


Figure C.7. Total Scale Score Distribution, Mathematics Grade 3

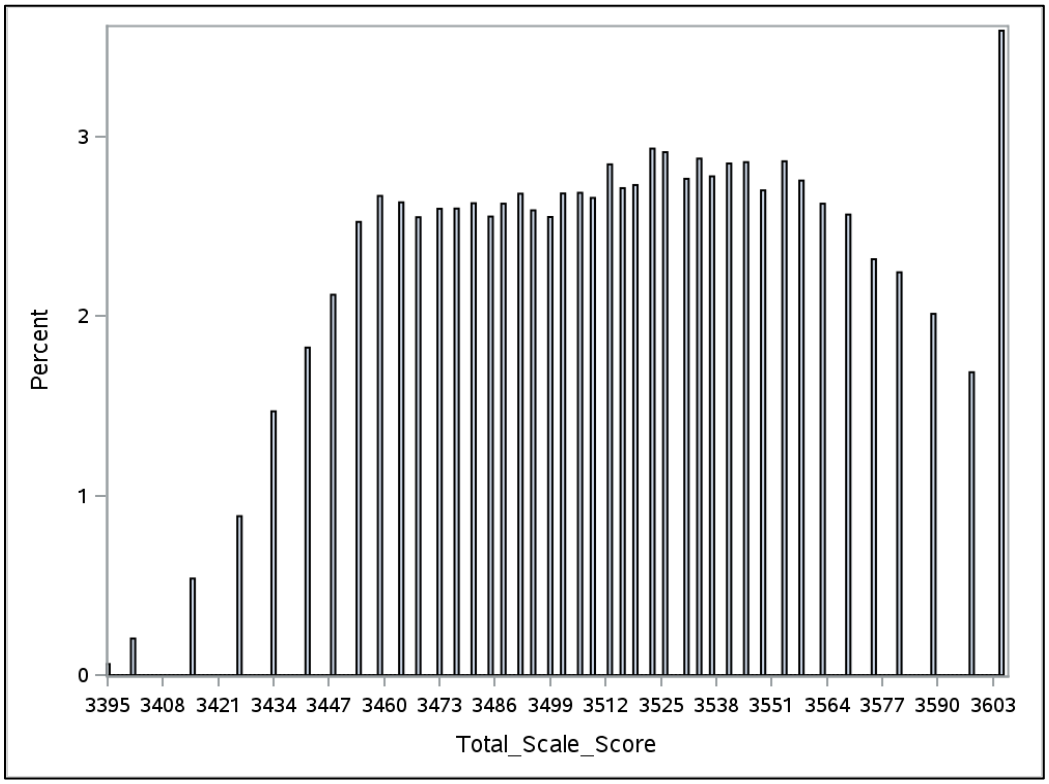


Figure C.8. Total Scale Score Distribution, Mathematics Grade 4

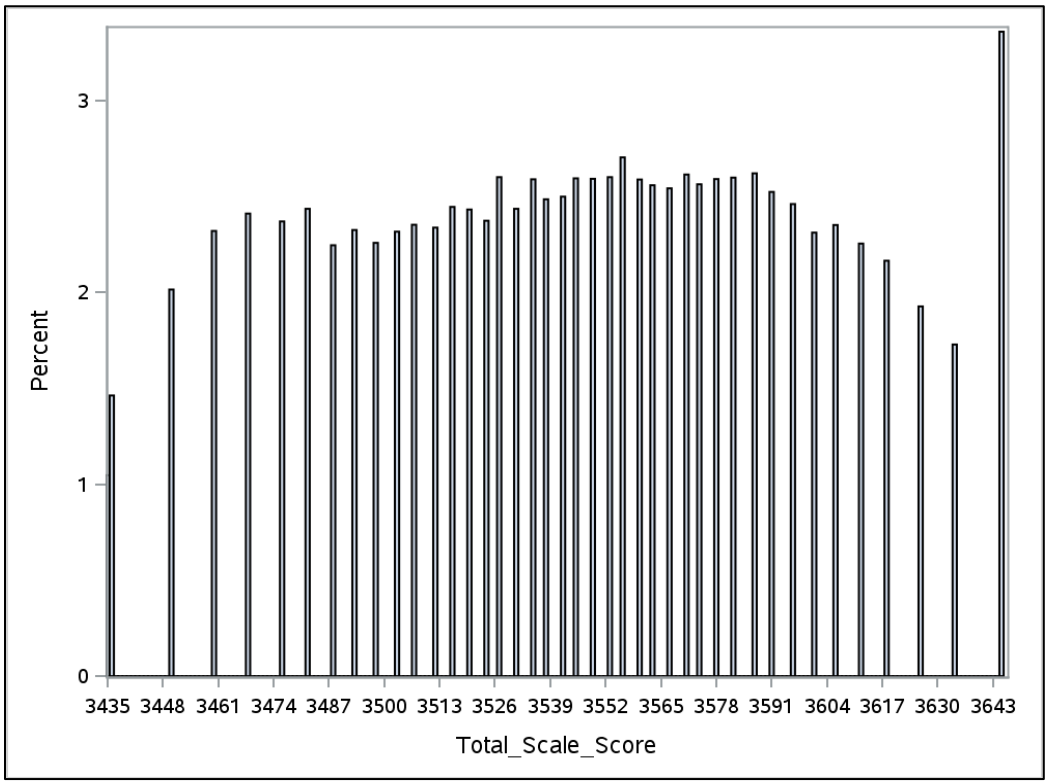


Figure C.9. Total Scale Score Distribution, Mathematics Grade 5

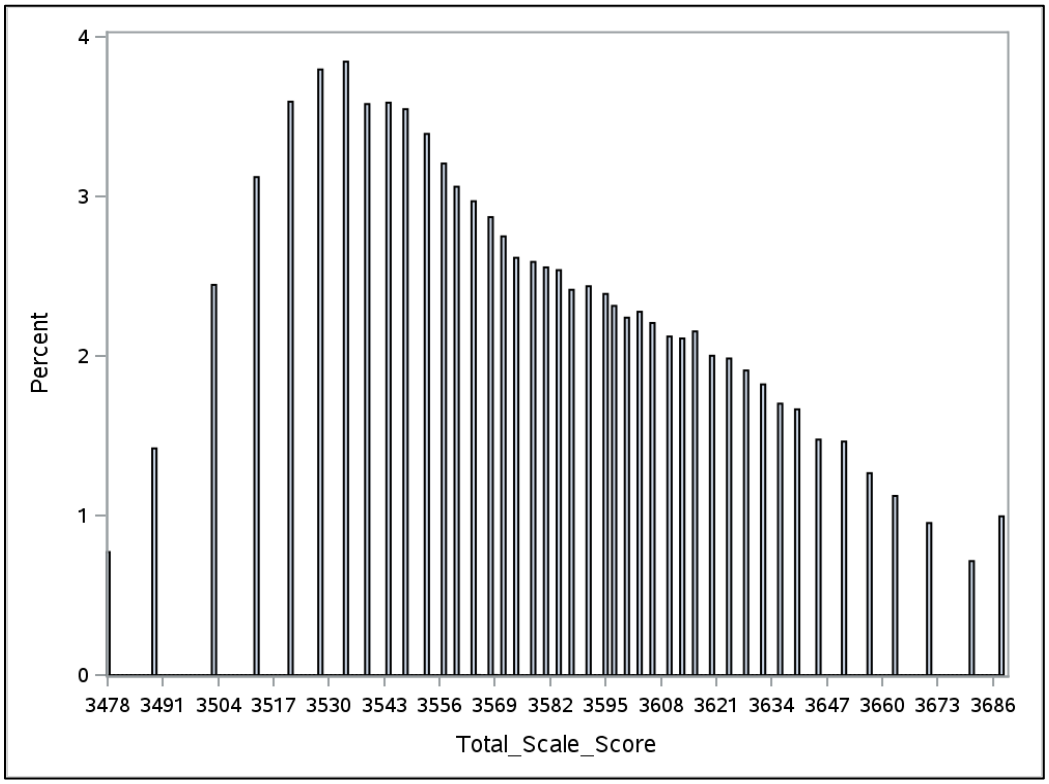


Figure C.10. Total Scale Score Distribution, Mathematics Grade 6

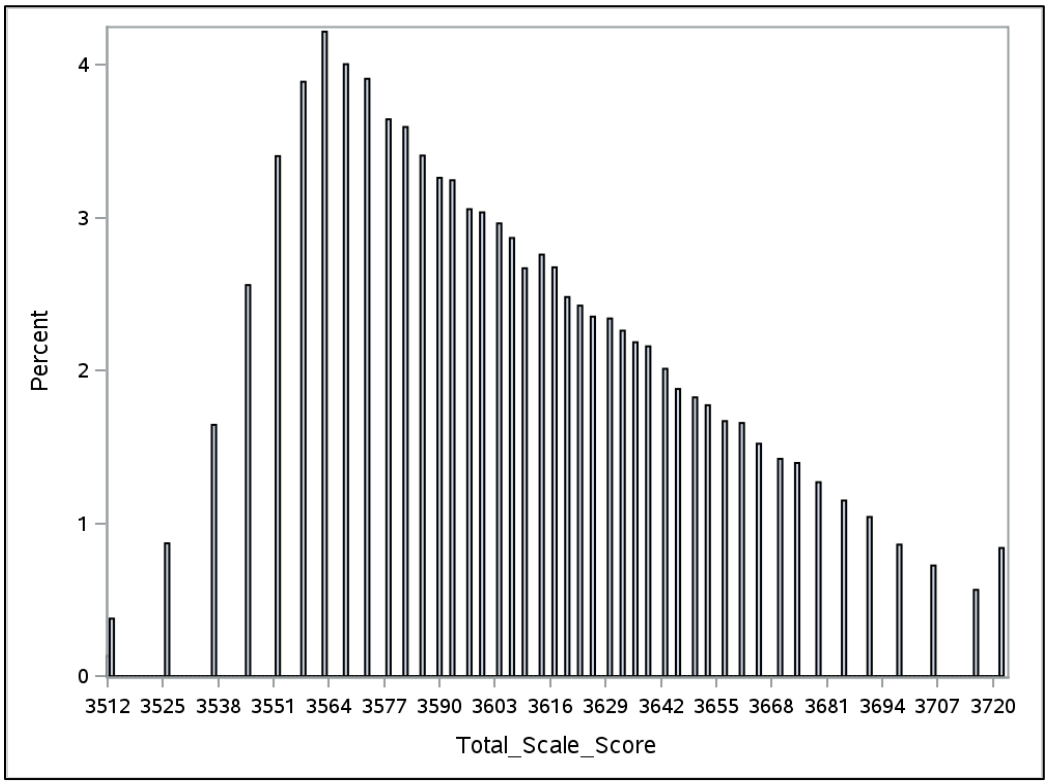


Figure C.11. Total Scale Score Distribution, Mathematics Grade 7

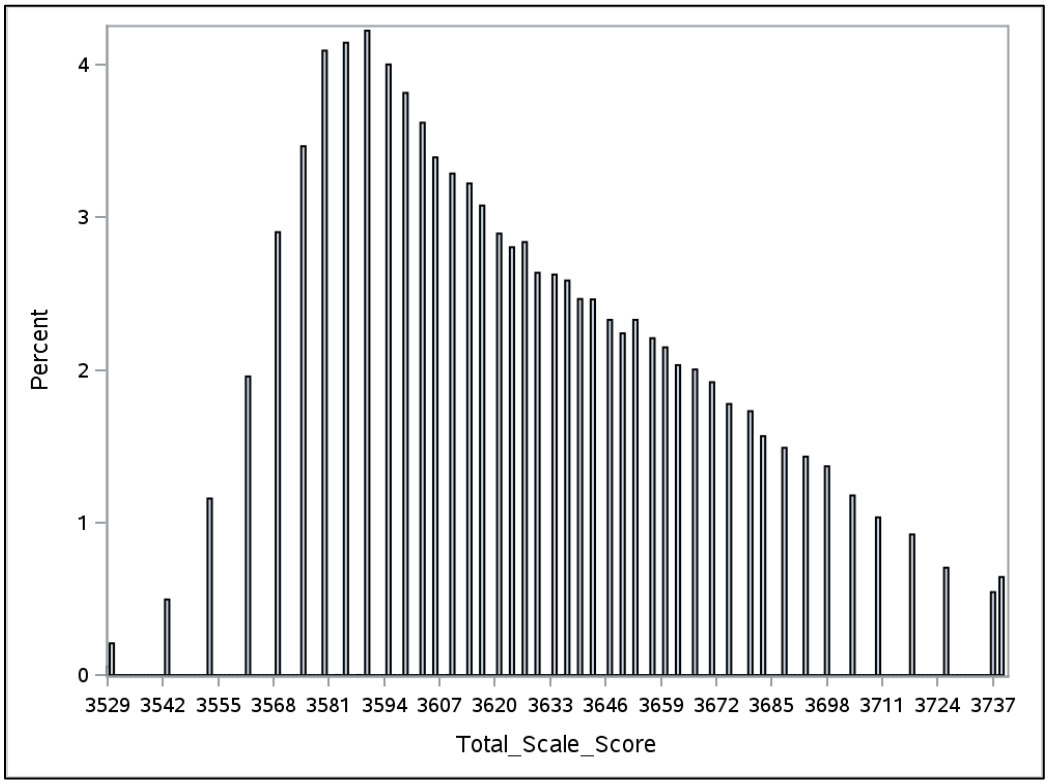


Figure C.12. Total Scale Score Distribution, Mathematics Grade 8