Illinois Assessment of Readiness (IAR) Technical Report 2023–2024

Prepared by Pearson for the Illinois State Board of Education (ISBE)
November 2024

Table of Contents

List of Tables

List of Figures

Section 1: Introduction

This technical report documents the evidence of reliability and validity to support test users in evaluating the intended purposes, uses, and interpretations of the test scores for the Spring 2024 administration of the Illinois Assessment of Readiness (IAR) assessments in English language arts/literacy (ELA/L) and mathematics. The evidence includes descriptions of the test design, development, and administration procedures; the student test results; and psychometric analyses including calibration, equating, and scaling to ensure that the test results can be compared across different test forms and administrations. The information 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).

1.1. Assessment Overview

The IAR assessments are Illinois' statewide summative assessments administered each spring to measure student performance on the Illinois Learning Standards in ELA/L and mathematics incorporating the Common Core State Standards (CCSS) in grades 3–8. The primary purpose of the IAR is to allow students to demonstrate what they know and can do in ELA/L and mathematics, assist educators in supporting student learning, make use of technology in assessments, advance accountability at all levels, and provide a measure of college and career readiness for students.

The assessments are administered online with paper accommodated forms available as needed, along with a wide range of accessibility features for all students and accommodations for students with disabilities, including screen readers, assistive technology, braille, large print, and text-to-speech (TTS). Student results are reported as an overall scale score and performance level with subclaim performance indicators. The five performance levels are Level 5: *Exceeded Expectations*, Level 4: *Met Expectations*, Level 3: *Approached Expectations*, Level 2: *Partially Met Expectations*, and Level 1: *Did Not Yet Meet Expectations*. Students performing at Levels 4 and 5 met or exceeded expectations, have demonstrated readiness for the next grade level/course, and are likely on track for college and careers.

1.2. Background

Illinois joined the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium in 2010 and administered its first PARCC summative assessments for ELA/L and mathematics in 2014–2015. Before this, Illinois administered the Illinois Standards Achievement Test (ISAT) for grades 3–8 and the Prairie State Achievement Examination (PSAE) for high school students in reading, mathematics, and science.

In 2013, the PARCC Governing Board established Parcc Inc. to support test delivery. After the contract with Parcc Inc. ended in June 2017, the Council of Chief State School Officers (CCSSO) took over the intellectual property and contracted with New Meridian to manage item development, forms construction, and governance. From 2017 to 2023, Illinois licensed New Meridian content for its assessments. In 2020, Illinois took steps toward greater independence in assessment development by starting to create custom content using the existing test blueprint and psychometric procedures but focusing exclusively on Illinois students. The Spring 2020 administration was canceled due to the COVID-19 pandemic, but testing resumed in Spring 2021 with items licensed from New Meridian while Illinois continued to develop its own IAR items.

Field testing of Illinois' custom-developed IAR items began in 2022, with some of those items included on the 2023 operational forms. By 2024, Illinois completed its transition to fully independent test content, with all items on the IAR sourced from the state's custom-developed bank. This marked Illinois' complete shift from shared consortia assessments to a state-specific assessment system tailored to its students, field tested and scaled exclusively within Illinois under the original PARCC and New Meridian frameworks.

1.3. Student Participation

As stated in the *Accessibility Features and Accommodations Manual* available online at https://www.isbe.net/iar, all students, including students with disabilities and English learners (ELs), are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for certain students with disabilities who have been identified by the Individualized Education Program (IEP) team to take their state's alternate assessment. All other students will participate in the ELA/L and mathematics assessments. Federal laws governing student participation in statewide assessments include the Every Student Succeeds Act of 2015 (ESSA), the Individuals with Disabilities Education Improvement Act of 2004 (IDEA), Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008), and the Elementary and Secondary Education Act (ESEA) of 1965, as amended.

1.4. Organizations and Groups Involved

As the assessment vendor for Illinois, Pearson is responsible for producing all testing materials, packaging and distribution, receiving and scanning of materials, and scoring, as well as program management and customer service. Pearson psychometrics is responsible for all psychometric analyses of the test data, including classical item analyses, differential item functioning (DIF) analyses, item calibrations based on item response theory (IRT), scaling, and development of all conversion tables. Educators participate in various item development activities to ensure that the assessments accurately reflect the content standards and student population, and Pearson uses services from several subcontractors. For example, the Human Resources Research Organization (HumRRO) provides third-party replication, MetaMetrics provides Lexile® and Quantile licensing and professional development services, the Center for Assessment calculates the student growth percentiles, and edCount conducts expert accommodation reviews, facilitators bias committees, and provides resource, training support, and professional development.

Section 2: Test Design

The IAR assessments are aligned to the Illinois Learning Standards, available online at https://www.isbe.net/Pages/Standards-Courses.aspx. They incorporate the CCSS and are designed to elicit evidence from students that supports valid and reliable claims about the extent to which they are college and career ready or on track toward that goal and are making expected academic gains based on the standards. The tests are timed, administered in two or three units, and contain selected-response items, brief and extended constructed-response items, technology-enhanced items, and performance tasks.

2.1. Claims and Subclaims

The assessments are designed to measure and report results in categories referred to as master claims and subclaims, as shown in Table 2.1 and Table 2.2. This claim structure, grounded in the CCSS, undergirds the design and development of the ELA/L and mathematics assessments. The master claim is the overall performance goal for the assessments reported as an overall scale score and performance level, whereas the subclaims further explicate what is measured on the assessments and include claims about student performance on the standards and evidence outlined in the evidence tables for both ELA/L (including Reading and Writing) and mathematics.

Table 2.1. Claims and Subclaims—ELA/L

Type	Description
Master Claim	Students must demonstrate that they are college and career ready or on track to readiness as demonstrated through reading and comprehending of grade-level texts of appropriate complexity and writing effectively when using and/or analyzing sources.
Major Claims	(1) Reading and comprehending a range of sufficiently complex texts independently (2) Writing effectively when using and/or analyzing sources
Subclaims	The claims and evidence are grouped into the following categories: • Reading: Literary Text • Reading: Informational Text • Reading: Vocabulary • Writing: Written Expression • Writing: Knowledge of Language and Conventions

Table 2.2. Claims and Subclaims—Mathematics

Type	Description
Master Claim	Students solve grade-level problems aligned to the Standards for Mathematical Content with connections to the Standards for Mathematical Practice to determine the degree to which a student is college or career ready or on track to being ready in mathematics.
Subclaims	The claims and evidence are grouped into the following categories: • Subclaim A: Major Content with Connections to Practices • Subclaim B: Additional and Supporting Content with Connections to Practices • Subclaim C: Highlighted Practices with Connections to Content: Expressing mathematical reasoning by constructing viable arguments, critiquing the reasoning of others, and/or attending to precision when making mathematical statements • Subclaim D: Highlighted Practice with Connections to Content: Modeling/Application by solving real-world problems by applying knowledge and skills articulated in the standards

2.2. Test Blueprints

Each content area and grade-level assessment was based on the test blueprints that determine the range and distribution of content and the distribution of points across the subclaims and item types. The blueprints guided how each test was built. Table 2.3 and Table 2.4 present a high-level overview of the IAR blueprints that show the percentage of points for each subclaim. Public-facing blueprints can be found online at https://www.isbe.net/iar. Content developers use additional documents with more detailed blueprint information and sequencing guides when building test forms to ensure consistency in content and psychometric properties.

For grade 3, the reading subclaim constitutes 55% of the total score points, whereas for grades 4–8, it accounts for 60%. Conversely, the writing subclaim makes up 44% of the total score points for grade 3 and 40% for grades 4–8.

Table 2.3. High-Level Blueprint—ELA/L

Grade	Reading: Literary Text	Reading: Informational Text	Reading: Vocabulary	Writing: Written Expression	Writing: Knowledge of Language and Conventions
3	20%	20%	15%	33%	11%
4	16–24%	22–30%	14%	32%	8%
5	22%	24%	14%	32%	8%
6	16–24%	22–30%	14%	32%	8%
7	16–24%	22–30%	14%	32%	8%
8	16–24%	22–30%	14%	32%	8%

Table 2.4. High-Level Blueprint—Mathematics

Grade	Major Content	Additional and Supporting Content	Reasoning	Modeling
3	39%	19%	19%	23%
4	40–44%	14–17%	19%	23%
5	39%	19%	19%	23%
6	39%	19%	19%	23%
7	39%	19%	19%	23%
8	39%	19%	19%	23%

2.3. Item Types

The assessments contain selected-response items, brief and extended constructed-response items, technology-enabled and technology-enhanced items, and task types in both ELA/L and mathematics (“tasks” for ELA/L refers to passage sets, whereas “tasks” for mathematics refers to specific items). Technology-enabled items are single-response or constructed-response items that involve a digital stimulus or open-ended response box with which the students engage in answering items, whereas technology-enhanced items involve specialized student interactions for collecting performance data (i.e., the act of performing the task is the way in which data are collected). Students may be asked, among other interactions, to categorize information, organize or classify data, order a series of events, plot data, generate equations, highlight text, or fill in a blank. One example of a technology-enhanced item is an interaction in which students drag response options onto a Venn diagram to show the relationship among ideas. Examples of the item types are provided in the practice items, available online at https://il.mypearsonsupport.com/practice-items/.

Each ELA/L test form has three units, two operational and one field test. Within each unit, students are presented with one or more of the following tasks:

- Literary Analysis Task (LAT): Students analyze two literary texts for similarities and differences. This task has one expository prose constructed-response (PCR) item.
- Research Simulation Task (RST): Students analyze and synthesize two or three informational texts. This task has one expository PCR.
- Narrative Writing Task (NWT): Students analyze one literary text for reading comprehension. This task includes one narrative PCR.
- Short, Long, or Paired Passage Set: Students respond to evidence-based selected-response (EBSR) and technology-enhanced constructed-response (TECR) items that assess reading. There is no writing prompt. EBSR and TECR items are worth 2 points each, whereas the PCR items are worth 12–19 points depending on the task type.

Mathematics tasks are identified by type, as shown below. Each task type can be assessed with multiple-choice, multiple-select, fill-in-the-blank, or technology-enhanced interactions. All tasks are standalone.

- Type 1 items assess concepts, skills, and procedures and are worth 1, 2, or 4 points.
- Type 2 items assess mathematical reasoning and are worth 3 or 4 points.
- Type 3 items assess modeling or application and are worth 3 or 6 points.

2.4. Test Units and Testing Times

Each assessment consists of multiple units, as shown in Table 2.5. The ELA/L assessments consist of two operational units and one field test unit, whereas the mathematics assessments consist of three operational units with embedded field testing. A field test sampling plan determines the total number of ELA/L students required to take the field test, with only those students who are selected participating in the third field test unit.

The IAR is a timed assessment, with the testing time limited to the unit testing times presented in Table 2.5 (except for an extended time accommodation). The unit testing time is the amount of time that must be provided to any student who needs it to complete the unit, and the total testing time reflects the operational testing time only. A new unit cannot be started until all students in the testing environment are finished or until the unit testing time has expired. If all students have completed testing before the end of the unit testing time, the unit may end. Once the unit testing time has elapsed, the unit must end (except for students with extended time accommodations).

Table 2.5. Test Units and Testing Times

Assessment(s)	Unit(s)	Testing Time per Unit	Total Testing Time
ELA/L 3	Units 1–2	75 minutes	150 minutes or 225 minutes (for schools assigned to be in the field test)
	Unit 3 (field test) – only given to students in field test sample. Schools are eligible to be in the field test sample once every three years.	75 minutes	

Assessment(s)	Unit(s)	Testing Time per Unit	Total Testing Time
ELA/L 4–8	Units 1–2	90 minutes	180 minutes or 270 minutes (for schools assigned to be in the field test)
	Unit 3 (field test) – only given to students in field test sample. Schools are eligible to be in the field test sample once every three years.	90 minutes	
Mathematics 3–5	Units 1–3 (non-calculator)	60 minutes	180 minutes
Mathematics 6–7	Unit 1 (calculator + non-calculator) Units 2–3 (calculator)	60 minutes 60 minutes	180 minutes
Mathematics 8	Units 1 (non-calculator) Units 2–3 (calculator)	60 minutes 60 minutes	180 minutes

2.5. Test Design

The ELA/L assessments focus on reading and comprehending a range of sufficiently complex literary and informational passages independently and writing effectively when analyzing text. Each passage set has 4–8 brief comprehension and vocabulary items, and the PCR items include three types of tasks: Literary Analysis, Research Simulation, and Narrative Writing. The PCR traits contribute to different claims, and the aggregate of the traits contributes to the summative scale score. For each performance-based task, students read one or more texts, answer several comprehension and vocabulary items, and then write an essay (extended response) based on the material they read.

All ELA/L assessments include a Research Simulation Task and either the Literary Analysis Task or the Narrative Writing Task. The Literary Analysis Task and the Research Simulation Task are scored for three traits: Reading Comprehension, Written Expression, and Knowledge of Conventions. The Narrative Writing Task is scored for two traits: Written Expression and Knowledge of Conventions. All traits are initially scored as either 0–3 or 0–4 points, with the Written Expression traits then multiplied by 3 (or weighted) to increase their contribution to the total score, making possible subclaim scores 0, 3, 6, and 9 or 0, 3, 6, 9, and 12. Table 2.6 presents the maximum possible points for the PCR items.

Table 2.6. Contribution of PCR Items in ELA/L: Number of Possible Points by Task

Grade(s)	Score	Literary Analysis	Research Simulation	Narrative Writing
3	Reading	3	3	0
	Written Expression	9	9	9
	Knowledge of Conventions	3	3	3
	Total	15	15	12
4–5	Reading	4	4	0
	Written Expression	12	12	9
	Knowledge of Conventions	3	3	3
	Total	19	19	12
6–8	Reading	4	4	0
	Written Expression	12	12	12
	Knowledge of Conventions	3	3	3
	Total	19	19	15

The mathematics assessments include tasks that measure a combination of conceptual understanding, applications, skills, and procedures. Each grade-level assessment includes both short- and extended-response items focused on applying skills and concepts to solve problems that require demonstration of the mathematical practices from the Illinois Learning Standards with a focus on modeling and reasoning with precision. Mathematics constructed-response items consist of tasks designed to assess a student's ability to use mathematics to solve real-life problems. Some of the tasks require students to describe how they solved a problem, while other tasks measure conceptual understanding and the ability to apply concepts by means of selected-response or technology-enhanced items. Students are also required to demonstrate their skills and knowledge by answering innovative selected-response and short-answer items that measure concepts and skills.

Section 3: Test Development

Pearson constructed the Spring 2024 test forms with custom-developed IAR items from the operationally ready item pool.

3.1. Asset Development Plan

The item bank houses passages and items at each assessed grade level and subject and supports the administration of the assessments, along with item release and practice tests. Prior to the annual item development cycle, the item development teams evaluated the strengths of the bank and considered the needs for future tests to establish an asset development plan.

3.2. Passage Selection

ELA/L tests are based on authentic texts, including multimedia stimuli, that are not developed for the purposes of the assessment or to achieve a particular readability metric but reflect the original language of the authors. Using the *Passage Selection Guidelines* that provided a text complexity framework and guidance on selecting a variety of text types and passages, ELA/L subject matter experts were trained to search for appropriate passages to support an annual pool of passages for consideration. Content experts then reviewed the passages for adherence to the *Passage Selection Guidelines* and the annual asset development plan in the number and distribution of genres and topics. Next, a Text Review Committee provided feedback about the grade-level appropriateness, content, and potential bias concerns and reached consensus about which passages would move forward for development. ELA/L asset development was not conducted until after the texts were approved by this committee.

3.3. Item Development

Item writers were recruited and trained to develop the number of items specified in the asset development plan. The items were then reviewed internally for content accuracy, alignment to the standards, range of difficulty, adherence to Universal Design principles that maximize the participation of the widest possible range of students, bias and sensitivity, and copy editing to enable the accurate measurement of the standards.

Next, external review committees reviewed every newly developed item to ensure that they aligned to the standards and were fair for all student populations. The committees included the Content Item Review Committee, Bias and Sensitivity Review Committee, and Editorial Review Committee that reviewed up to 10% of the items for grammar, punctuation, clarity, and adherence to the style guide. The meetings were conducted either in person or virtually and included large group training on the expectations and processes of each meeting, followed by breakout meetings by content and grade where additional training was provided.

The content review committees reviewed and edited test items for adherence to the foundational documents, Universal Design principles, accessibility guidelines, associated item metadata, and the style guide and verified that the appropriate scoring rule had been applied to each item. The bias and sensitivity review committees confirmed that the items did not have any bias or sensitivity issues that would interfere with a student’s ability to achieve their best performance, evaluating adherence to the *Fairness and Sensitivity Guidelines* and ensuring that items and tasks would not unfairly advantage or disadvantage one student or group of students over another. Committee members made edits and modifications to items to eliminate sources of bias and improve accessibility for all students.

3.4. Form Construction

Test form construction is the process of selecting and sequencing a set of operational and field test items for administration, which is a complex, interactive task that requires both content and psychometric expertise. Table 3.1 presents the number of test forms constructed for Spring 2024. Both ELA/L and mathematics had one core operational form and one accommodated operational form. The forms were constructed to (a) reflect the test blueprint in terms of content, item types, test length, and expected difficulty and performance along the ability continuum and (b) adhere to the following goals outlined in the test construction specifications:

- Test forms are designed to appropriately measure the assessment claims and subclaims across the full range of ability.
- Scores are comparable across forms and administrations.
- Overexposure of items is minimized.
- Parallel forms are created among the IAR forms, as possible.
- Forms are developed to industry standards for validity, reliability, and fairness (AERA et al., 2014).

Table 3.1. Number of Test Forms Constructed in Spring 2024

Assessment	#Core OP Forms	#Accommodated OP Base Forms	#FT Forms
ELA/L 3	1	1	14
ELA/L 4	1	1	14
ELA/L 5	1	1	14
ELA/L 6	1	1	14
ELA/L 7	1	1	14
ELA/L 8	1	1	12
Mathematics 3	1	1	41
Mathematics 4	1	1	45
Mathematics 5	1	1	45
Mathematics 6	1	1	48
Mathematics 7	1	1	52
Mathematics 8	1	1	51

Note. OP = operational, FT = field test

3.4.1. Operational Forms

Core forms refer to the operational forms consisting only of the items that count toward a student's score designed to facilitate psychometric equating through a common item linking strategy and to be constructed as "parallel" as possible from a content and test-taking experience. Evaluation criteria for parallelism included adherence to the blueprint; sequencing of content across the forms; statistical averages and distributions for item difficulty and discrimination; item type and cognitive complexity; and ELA/L passage characteristics including genre, topics, word count, and text complexity.

3.4.2. Field Test Forms

All students receive the same core operational items but different field test items. Field test items were either embedded in the mathematics units or administered to a select sample of students in a separate third unit for ELA/L (i.e., census field testing is conducted for mathematics, whereas a sampling plan is used for ELA/L). Mathematics forms include embedded items in Units 2 and 3 only for grades 3–5 and in each unit for all other grades.

3.4.3. Accommodated Forms

Table 3.2 presents the accommodated forms constructed based on the one accommodated operational form developed for each content area and grade, as well as the accommodations available on the operational core form. The forms are accommodated to support braille, large print, human reader/human signers, and text-to-speech (TTS). Spanish forms are provided for mathematics only.

Table 3.2. Supported Accommodations

Test Form	ELA/L	Mathematics
Accommodated Base Form (ACC1)	• Paper-Based Form • Large Print • Read Aloud • Human Reader • Human Signer • ASL • Braille • Screen Reader • Non-Screen Reader	• Paper-Based Form • Large Print • Read Aloud • Human Reader • Human Signer • ASL • Braille • Screen Reader • Non-Screen Reader • Spanish Paper • Spanish Large Print • Spanish Human Reader
Core Form (Online1)	• TTS	• TTS • Spanish Online • Spanish TTS

3.5. Data Review

Following the Spring 2024 test administration, an educator data review committee met in August 2024 to evaluate the field tested items and associated performance data in terms of appropriateness, level of difficulty, and any potential differential item functioning (DIF) for groups of interest. The committee recommended acceptance or rejection of each field tested item for inclusion in the operational item bank and made recommendations for some items to be revised and re-field tested. Items approved by the committee became eligible for use on future operational assessments.

The field tested items from the Spring 2023 administration were also reviewed in a data review meeting in July 2024 as they had not gone through a data review after the 2023 administration during the open procurement. Pearson won the development contract in August 2023, and it was decided to put the 2023 field test through data review in Summer 2024 during contract negotiations.

Section 4: Test Administration

Table 4.1 presents the Spring 2024 test administration dates. The IAR assessments are administered online, with paper accommodated forms available as needed. The online administration takes place in TestNav, Pearson's online testing platform. PearsonAccess^{next} is the student test management portal that Test Administrators use to manage student tests and registrations and order materials if needed.

Table 4.1. Test Administration Activities

Event	Dates
Administration Training	January 9–16, 2024
Receive Materials	February 20, 2024
Online Testing Window	March 4 – April 19, 2024
Paper Testing Window	March 4 – April 9, 2024
Return Materials	March 4 – April 12, 2024

To ensure a standardized administration for all students, School Test Coordinators and Test Administrators are instructed to follow the directions in the *Test Coordinator Manual* and *Test Administrator Manual* available online at https://il.mypearsonsupport.com/iar-summative-resources/. The standardization of directions, test administration conditions, and scoring procedures is necessary to support the comparability of test score interpretations both within and between administrations. When standardized procedures are not in place, differences in student performance cannot be clearly attributed to true differences in student ability because of the unknown effect of administration conditions on performance.

4.1. Accessibility Features and Accommodations

It is important to ensure that performance in the classroom and on assessments is influenced minimally, if at all, by a student's disability or linguistic/cultural characteristics that may be unrelated to the content being assessed. Through a combination of Universal Design principles and accessibility features, accessibility was considered from the initial test design through item development, field testing, and implementation of the assessments for all students, including SWDs, ELs, and ELs with disabilities. Accommodations may still be needed for some SWDs and ELs to assist in demonstrating what they know and can do, but the accessibility features available to students should minimize the need for accommodations during testing and ensure the inclusive, accessible, and fair testing of the diverse students being assessed. While all students can receive accessibility features on the assessments, four distinct groups of students may receive accommodations:

1. SWDs with an IEP
2. Students with a Section 504 plan who have a physical or mental impairment that limits one or more major life activities, have a record of such an impairment, or are regarded as having such an impairment but who do not qualify for special education services
3. Students who are ELs
4. Students who are ELs with disabilities who have an IEP or 504 plan

These students are eligible for accommodations intended for both SWDs and ELs. Testing accommodations for SWDs or students who are ELs must be documented according to the guidelines and requirements outlined in the *Accessibility Features and Accommodations Manual* available online at https://il.mypearsonsupport.com/iar-summative-resources/.

Accessibility features are tools or preferences available to all students that are either built into the online TestNav assessment system or provided externally by Test Administrators. Examples of accessibility features include the line reader, answer eliminator, magnifier, highlighter, bookmark, pop-up glossary, and notepad. Students should have the opportunity to select and practice using them prior to testing to determine which are appropriate for use on the assessment. Consideration should be given to the supports a student finds helpful and consistently uses during instruction.

Accommodations are adjustments to the testing conditions, test format, or test administration that provide equitable access during assessments for SWDs and students who are ELs. In general, the administration of the assessment should not be the first occasion on which an accommodation is introduced to the student. To the extent possible, accommodations should provide equitable access during instruction and assessments, mitigate the effects of a student’s disability, not reduce learning or performance expectations, not change the construct being assessed, and not compromise the integrity or validity of the assessment.

Accommodations are intended to reduce or eliminate the effects of a student’s disability and/or English language proficiency level, but they should never reduce learning expectations by reducing the scope, complexity, or rigor of an assessment. Accommodations must also be consistent with those provided for classroom instruction and classroom assessments. Some accommodations may be used for instruction and for formative assessments that are not allowed for the summative assessment because they impact the validity of the assessment results (e.g., allowing a student to use a thesaurus or access the internet during an assessment). There may be consequences (e.g., excluding a student’s test score) for the use of nonallowable accommodations during assessments. To the extent possible, accommodations should adhere to the following principles:

- Accommodations should enable students to participate more fully and fairly in instruction and assessments and to demonstrate their knowledge and skills.
- Accommodations should be based on an individual student’s needs rather than on the category of a student’s disability, level of English language proficiency alone, level of or access to grade-level instruction, amount of time spent in a general classroom, current program setting, or staff availability.
- Accommodations should be based on a documented need in the instruction/assessment setting and should not be provided to give the student an enhancement that could be viewed as an unfair advantage.
- Accommodations for SWDs must be described and documented in the student’s IEP or 504 plan and must be provided if they are listed.
- Accommodations for ELs should be described and documented.
- EL students with disabilities are eligible to receive accommodations for both SWDs and ELs.
- Accommodations should become part of the student’s program of daily instruction as soon as possible after completion and approval of the appropriate plan.
- Accommodations should not be introduced for the first time during the testing of a student.
- Accommodations should be monitored for effectiveness.
- Accommodations used for instruction should also be used, if allowable, on local district assessments and state assessments.

Examples of accommodations include assistive technology, a screen reader version for a student who is blind or visually impaired, a braille edition, large print edition, a paper-based edition, American Sign Language (ASL) video, human signer for test directions, and a word-to-word dictionary for ELs. If a student refuses an accommodation listed in their IEP, 504 plan, or an EL plan, the school must document in writing that the student refused the accommodation, although the accommodation must still be offered and remain available to the student during the test administration. The *Accessibility Features and Accommodations Manual* provides the full list of accessibility features and accommodations for students with disabilities and EL students.

4.2. Test Security

The IAR test administration is a secure testing event, and maintaining the security of test materials before, during, and after the test administration is crucial to obtaining valid and reliable results. All test security and administration policies are found in the *Test Coordinator Manual* and the *Test Administrator Manual*. For example, School Test Coordinators are responsible for ensuring that all personnel with authorized access to secure materials are trained in and subsequently act in accordance with all security requirements. They must implement chain-of-custody requirements for specified materials and are responsible for distributing, collecting, and returning or destroying secure test materials. School Test Coordinators must maintain a tracking log to account for the collection and destruction of test materials. Test Administrators are not to have extended access to test materials before or after administration (except for certain accessibility or accommodations purposes) and must document the receipt and return of all secure test materials (used and unused) to the School Test Coordinator immediately after testing.

The IAR test administration includes both secure and nonsecure materials that are further delineated by whether they are scorable or nonscorable depending on whether the assessments were administered online or on paper, as explained below. Students may not have access to secure test materials before testing, including printed student testing tickets.

- Secure materials must be closely monitored and tracked to prevent unauthorized access to or prohibited use or distribution of secure content such as test items, reading passages, and student work. Secure paper materials include both used and unused test booklets and used scratch paper, and secure online materials include student testing tickets, secure administration scripts (e.g., mathematics read-aloud), and used scratch paper. Nonsecure materials are any authorized testing materials that do not include secure content (e.g., items or student work), including test administration manuals, unused scratch paper, and mathematics reference sheets that have not been written on.
- Paper scorable materials consist of used test booklets (grade 3) and answer documents (grades 4+) that must be returned to Pearson to be scored. All other paper materials such as blank (i.e., unused) test booklets, test administration manuals, scratch paper, and mathematics reference sheets are deemed nonscorable. The online assessments do not have any scorable materials as student work is submitted electronically for scoring. Thus, there are limited physical materials to return (e.g., secure administration scripts for certain accommodations).

Printed mathematics reference sheets (if applicable) and scratch paper must be new and unmarked. Paper scorable secure materials provided by test administrators include test booklets (grade 3) and answer documents (grades 4+). Paper nonscorable secure materials distributed by test administrators include large print test booklets, braille test booklets, scratch paper (paper used by students to take notes and work through items), and printed mathematics reference sheets (grades 5–8 and high school).

4.3. Testing Irregularities and Security Breaches

Any action that compromises test security or score validity is prohibited and may be classified as testing irregularities or security breaches. Table 4.2 presents examples of these activities. School Test Coordinators should discuss other possible testing irregularities and security breaches with Test Administrators during training. All instances of security breaches and testing irregularities must be reported to the School Test Coordinator immediately, and the *Form to Report a Testing Irregularity or Security Breach* must be completed within two school days of the incident. If any situation occurs that could cause any part of the test administration to be compromised, schools should refer to the *Test Coordinator Manual* and follow the instructions for reporting a testing irregularity or security breach.

Table 4.2. Test Irregularity and Security Breach Examples

Topic	Examples
Electronic Devices	Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner) while secure test materials are still distributed, while students are testing, after a student turns in their test materials, or during a break *Exception*: School Test Coordinators, Technology Coordinators, and Test Administrators can use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed.
Test Supervision	• Coaching students during testing (e.g., giving students verbal or nonverbal cues, hints, suggestions, or paraphrasing or defining any part of the test) • Engaging in activities (e.g., grading papers, reading a book, newspaper, or magazine) that prevent proper student supervision while secure test materials are still distributed or while students are testing • Leaving students unattended while secure test materials are still distributed or while students are testing • Deviating from testing time procedures • Allowing cheating of any kind • Providing unauthorized persons with access to secure materials • Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore not appropriate • Allowing students to test before or after the test administration window
Test Materials	• Losing a student test booklet or answer document • Losing a student testing ticket • Leaving test materials unattended or failing to keep test materials secure at all times • Reading or viewing the passages or test items before, during, or after testing • Copying or reproducing (e.g., taking a picture of) any part of the passages or items or any secure test materials or online test forms • Revealing or discussing passages or test items with anyone, including students and school staff, through verbal exchange, email, social media, or any other form of communication • Removing secure test materials from the school's campus or removing them from locked storage for any purpose other than administering the test *Exception*: Administration of a human reader/signer accessibility feature for mathematics or accommodation for ELA/L that requires a Test Administrator to access passages or items
Testing Environment	• Allowing unauthorized visitors in the testing environment • Failing to follow administration directions exactly as specified in the *Test Administrator Manual* • Displaying testing aids in the testing environment (e.g., a bulletin board containing relevant instructional materials) during testing

Section 5: Scoring

Selected-response, technology-enabled, and technology-enhanced items are machine scored; constructed-response items are handscored using Pearson's scoring platform, ePEN2 (Electronic Performance Evaluation Network, second generation); and the ELA/L PCR items are primarily scored by Pearson's automated scoring engine known as the Intelligent Essay Assessor (IEA), with a 10% reliability second score and some outlier scoring (where the IEA score and human score differ by more than 1 point) by human scorers. To be more specific, 10% of the PCR item scores are also scored by humans in addition to IEA to compute the inter-rater agreement and monitor scoring. This is also explained in Section 5.2.3.1, in the first paragraph of Section 5.3, and in Section 5.4.

5.1. Machine Scoring

Pearson performed a key check and adjudication near the end of the test administration and before reporting to verify that the answer keys were correct for each item. If discrepancies were identified, a Pearson senior content specialist or content manager reviewed the flagged item(s) and worked to resolve the issue.

Rule-based scoring refers to item types that use various scoring models, including choice interaction that presents a set of choices where one or more choices can be selected; text entry, where the response is entered in a text box; hot spot or text interaction, where an area in a graph or text in a paragraph can be highlighted; or match interaction, where an association can be made between pairs of choices in a set. These items include the scoring rules and correct responses as part of their item XML (markup language) coding. Following the initial development of the rule-based scoring rubrics, Pearson has continued to monitor and evaluate new item development to ensure that the scoring rules are maintained within all item types as approved.

In the case of a hot spot interaction, each hot spot region was drawn as a rectangle or a circle using the background art for reference points which translate to pixels having a horizontal and a vertical position. Each hot spot was given a unique alphanumeric name. An Assessment Specialist identified all possible correct responses using these names. A second Assessment Specialist checked the item and checked the scoring. A committee chosen by ISBE reviewed the item, including the scoring. In Machine Scoring, the position where the student clicked was compared to the region drawn. Students were awarded points when the position(s) clicked occurred within the regions identified as correct responses.

5.2. Handscoring of Constructed-Response Items

Constructed-response items were handscored by human scorers who completed online training and qualification sets to demonstrate they could correctly score student responses based on the provided guidelines. Scorers who successfully completed the training and qualifying process were permitted to score student responses. All online and paper responses were scored within the ePEN2 system with monitoring conducted by Pearson. A handscoring specifications document detailed the handscoring schedule, customer requirements, quality management plans, item information, and staffing plans for each scoring administration. All Pearson employees involved in the scoring process possessed at least a four-year college degree. Roles and responsibilities were as follows:

- Scorers applied scores to student responses.
- Scoring supervisors monitored the work of a team of scorers through review of scorer statistics and backreading.

- Scoring directors managed the scoring quality of a subset of items and monitored the work of supervisors and scorers for their assigned items. Directors backread responses scored by supervisors and scorers as part of their quality-monitoring duties.
- ELA/L and mathematics content specialists managed the scoring quality and monitored the work of the scoring directors.
- The project manager documented the procedures, identified risks, and managed day-to-day administrative matters.
- A scoring manager provided oversight for the entire scoring process.

5.2.1. Scorer Training

Scorer training materials were initiated at rangefinding meetings held prior to scoring the field test items where educators and administrators interpreted the scoring rubrics and determined consensus scores for student responses. Rangefinding participants reviewed student responses and used scoring rubrics to determine consensus scores used to create the field test scorer training sets. After items were selected for operational testing, Pearson developed operational training materials for these items. When developing the scorer training materials, Pearson reviewed the detailed notes and records from the rangefinding committee meetings. Training sets were developed using the responses scored by the committees and additional suitable student response samples as needed.

During scorer training, Pearson used anchor, practice, and qualification sets, as described in Table 5.1. Two types of training sets (prototype and abbreviated) are used, as described below. The anchor and practice sets for both the prototype and abbreviated items included annotations for each student response (i.e., formal written explanations of the score).

- Prototype training sets were complete training sets consisting of the anchor, practice, and qualification sets. ELA/L had one prototype training set per task type (Research Simulation Task, Literary Analysis Task, and Narrative Writing Task) at each grade level. A mathematics prototype training set was built for a grouping of similar items for a total of 3–4 prototype sets per grade. The prototype training approach promoted consistency in scoring, as each subsequent abbreviated training set for the ELA/L task type or mathematics item grouping was based on the prototype. Once a prototype was chosen, full training materials were developed for that item, and scorers were trained to score a particular item type using the prototype training materials for that type.
- Abbreviated training sets were prepared for all items not selected for prototype training sets. The abbreviated training sets included an anchor set and two practice sets so scorers could internalize the scoring standards for these new items, which were similar to the prototype items they had previously scored.

Table 5.1. Scoring Training Materials

Training Material	Description	Specifications
Anchor Sets	Anchor sets consist of responses that are clear examples of student performance at each score point and are the primary reference for scorers as they internalize the rubric. The responses selected are representative of typical approaches to the task and are arranged to reflect a continuum of performance. All scorers have access to the anchor set when they are training and scoring and are directed to refer to it regularly.	The mathematics prototype anchor set includes three annotated responses per score point, whereas the abbreviated anchor set includes 1–3 annotated responses per score point. The ELA/L prototype anchor sets include three annotated responses per score point, including separate complete anchor sets for each scoring trait (Reading Comprehension and Written Expression and Conventions for Research Simulation and Literary Analysis Tasks, Written Expression for Narrative Writing Tasks, and Knowledge of Language and Conventions for all task types).
Practice Sets	Practice sets are used to help scorers practice applying the scoring guidelines. Scorers review the anchor sets, score the practice sets, and then compare their assigned scores for the practice sets to the actual assigned scores to help them learn. Some of these responses clearly reinforce the scoring guidelines presented in the anchor set, whereas others are more difficult to evaluate, fall near the boundary between two score categories, or represent unusual approaches to the task to provide guidance and practice in defining the line between score categories and applying the scoring criteria to a wider range of response types.	The mathematics prototype and abbreviated practice sets include 2–3 sets of 10 annotated responses. The ELA/L prototype practice sets include two sets of five annotated responses and two sets of 10 annotated responses, whereas the abbreviated practice sets include two sets of 10 annotated responses.
Qualification Sets	Qualification sets consist of student responses that are clear examples of score points to reinforce the application of the scoring criteria illustrated in the anchor set. These sets are used to confirm that scorers understand how to score the responses accurately. Scorers are required to meet specified agreement percentages on qualification sets to score student responses.	The mathematics and ELA/L prototype qualification sets include three sets of 10 responses each (not annotated). The subsequent abbreviated items do not include qualification sets.

5.2.2. Scorer Qualification

To demonstrate that they could accurately apply the scoring methodology, scorers applied scores to three qualification sets consisting of 10 responses each. ELA/L scorers applied a score for each trait on each response in the qualification sets[1], and mathematics scorers applied a score for each part of an item that was a constructed response ranging from 1–4 parts. Scorers were required to match the approved score at a certain percentage to qualify. For ELA/L qualification, scorers were required to meet the following conditions:

1. On at least one of the three qualifying sets, at least 70% of the ratings on each of the two scoring traits (considered separately) must agree exactly with the approved scores.
2. On at least two of the three qualifying sets, at least 70% of the ratings (combined across the three scoring traits) must agree exactly with the approved scores.
3. Combining over the three qualifying sets and across the two scoring traits, at least 96% of the ratings must be within one point of the approved scores.

The qualification requirements for mathematics were based on the item types and score point ranges. Because mathematics items can have one or more scoring traits, a scorer needed to achieve the requirements in Table 5.2 separately for each scoring trait. On at least two of the three qualifying sets, a scorer was required to meet the "perfect agreement" percentage for each category. Perfect agreement was achieved when the scores applied exactly matched the approved scores. Over the three qualifying sets, a scorer was required to meet the "within 1 point" percentage indicated for each category. The average is exclusive to each trait, so an item with multiple scoring traits would have multiple trait rating averages within 1 point of the approved score.

Table 5.2. Mathematics Scorer Qualification Requirements

Category	Score Point Range	Perfect Agreement	Within 1 Point
2	0–1	90%	100%
3	0–2	80%	96%
4	0–3	70%	96%
5	0–4	70%	95%
6	0–5	70%	95%
7	0–6	70%	95%

5.2.3. Scorer Monitoring

Score monitoring consisted of second scoring of at least 10% of the responses, backreading, the use of validity responses and calibration sets, and inter-rater reliability (see Section 5.4).

5.2.3.1. Second Scoring

During scoring, the ePEN2 scoring system automatically and randomly distributed a minimum of 10% of student responses for second scoring. Scorers had no indication whether a response had been scored previously. Humans applied the second score for all mathematics items, whereas second scoring for ELA/L was performed either by human scorers or the IEA automated scoring engine. If the first and second scores were nonadjacent, a third and occasionally fourth score were assigned to resolve scorer disagreements. When a resolution score (i.e., third score) was nonadjacent to one or both of the first two scores, the content specialist or scoring director would apply an adjudication score (fourth score).

[1] The Literary Analysis and Research Simulation tasks each had two traits (Reading Comprehension & Written Expression and Conventions), and the Narrative Writing task had two traits (Written Expression and Conventions).

5.2.3.2. Backreading

Backreading required the scoring supervisor to review the scores applied by scorers to help them provide additional coaching or instruction and guard against scorer drift, where scorers score responses in comparison to one another instead of in comparison to the training responses. Scoring supervisors used the ePEN2 backreading tool to review scores assigned to individual student responses by any given scorer to confirm that the scores were correctly assigned and to give feedback and remediation to individual scorers. Pearson backread approximately 5% of the handscored responses. Backreading scores did not override the original score but were used to monitor scorer performance.

5.2.3.3. Validity Responses

Prescored validity responses were strategically interspersed in the pool of live responses and indistinguishable from any other responses so that scorers were unaware they were scoring validity responses rather than live responses to help ensure that scorers were applying the same standards throughout the project. Scorers had to meet the required validity agreement requirements in Table 5.3 to continue working on the project. Scorers who did not maintain the expected agreement statistics were given a series of interventions culminating in a targeted calibration set. Scorers who did not pass targeted calibration were removed from scoring the item, and all the scores they assigned were deleted.

Table 5.3. Scoring Validity Agreement Requirements

Content Area	Score Point Range	Perfect Agreement	Within 1 Point*
ELA/L	Multi-trait	65%	96%
Mathematics	0–1	90%	96%
	0–2	80%	96%
	0–3	70%	96%
	0–4	65%	95%
	0–5	65%	95%
	0–6	65%	95%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point

In addition to the prescored validity responses, validity was at times shared with scorers in a process known as “validity as review” that provided scorers automated, immediate feedback, giving them a chance to review responses they mis-scored, with reference to the correct score and a brief explanation of that score. One validity response was sent to scorers for every 25 “live” responses scored.

5.2.3.4. Calibration Sets

Calibration sets were created by scoring directors to reinforce rangefinding standards, introduce scoring decisions, or address scoring issues and trends to help train scorers on areas of concern or focus. Calibration was used either to correct a scoring issue or trend or to continue scorer training by introducing a scoring decision. Calibration was administered regularly throughout scoring.

5.3. Automated Scoring of Prose Constructed-Response Items

Automated scoring performed by Pearson's IEA automated scoring engine was the default option for scoring the summative assessment's online PCR tasks. Under the default option, it was assumed that operational scores for approximately 90% of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. Human scoring was applied to responses that were scored while IEA was being trained, as well as to additional responses routed to human scoring when there was uncertainty about the automated scores. 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. When IEA provided the first score of record, the second reliability score was a human score.

Continuous Flow scoring facilitates the training of IEA using human scores assigned to operational online data collected early in the administration. With Continuous Flow, responses flow between the engine and human scorers so the engine can learn from humans in real time. Once IEA obtains sufficient data to train or complete a scoring model (all score points can be scored), it can be used as the primary source of scoring (although human scoring continues for the 10% reliability sample and other responses that may be routed accordingly).

When the engine is less confident in scoring a response, the response is marked with a low confidence flag that automatically routes it to human scorers (known as Smart Routing). 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. Smart Routing can be applied prompt-by-prompt to the extent needed to meet scoring quality criteria for automated scoring. It was assumed for the spring administration that operational scores for approximately 95% of the online PCR responses would be assigned by IEA, while the operational scores for the remaining online responses were assigned by human scorers.

5.3.1. Sampling Responses Used for Training IEA

The performance of human scoring was closely monitored to verify that an appropriate set of data, which would meet the criteria below, would be available for training IEA. Several characteristics of the human scoring data were monitored:

- Exact agreement between human scorers (the goal was for this to be at least 65% for each trait)
- Exact agreement between human scores at each score point (the goal was for this to be at least 50% for each trait)
- The number of responses at each score point (the goal was to have at least 40 responses at the highest score points in the training samples used by IEA)
- The number of responses with two human scores assigned (note that IEA "ordered" additional scoring of responses during the sampling period as needed)

Although the desired characteristics of the training data were easily achieved for some tasks, they were more challenging to achieve for others. For some tasks, a subset of scores were reset and clarifying directions were provided to improve human-human agreement. For other tasks, special sampling approaches (i.e., over-sampling was conducted to ensure enough responses at the top scores for PCR items that were difficult and hence had relatively few responses at top scores) were used to increase the number of responses that received top scores. A healthy percentage of responses were also backread during the sampling period, and these scores as well as double human scores were all part of the data used to train IEA.

5.3.2. Quality Criteria for Evaluating IEA Performance

The primary evaluation criterion for IEA was based on responses to validity papers with “known” scores assigned by experts. For each PCR item scored, a set of validity papers is used to monitor the human-scoring process over time. Validity papers are seeded into human scoring throughout the administration, and the expectation is that IEA can score validity papers at least as accurately as humans can.

Additional measures of inter-rater agreement for evaluating automated scoring are used, including the Pearson correlation (*r*), kappa, quadratic weighted kappa (QWK), exact agreement, and standardized mean difference (SMD). These measures are computed between pairs of human scores and between IEA and humans to evaluate how performance was the same or different. Criteria for evaluating the training of IEA given these measures include the following:

- Pearson correlation (*r*) between IEA-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- QWK between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25% of human-human.
- SMD between IEA-human should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria:

- Primary Criteria based on responses to validity papers: With Smart Routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.
- Contingent Primary Criteria based on the training responses if validity responses are not available: In these cases, IEA was evaluated based on IEA-human exact agreement for each trait score and compared to agreement based on responses that were double-scored by humans. The IEA-human exact agreement criterion is within 5.25% of human-human exact agreement.
- Secondary Criteria based on the training responses: With Smarter Routing applied as needed, IEA-human differences on statistical measures for each trait score are within the Williamson et al. (2012) tolerances for subgroups with at least 50 responses.

5.4. Inter-Rater Agreement

For 10% of all responses, a second reliability score was assigned to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. Inter-rater agreement is the agreement between the first and second scores assigned to student responses. Pearson used inter-rater agreement statistics as one factor in determining the needs for continuing training and intervention on both individual and group levels. During handscoring, the ePEN2 system included comprehensive inter-rater agreement reports that allowed supervisory personnel to monitor both individual and group performance. Based on reviews of these reports, scoring experts targeted individuals for increased backreading and feedback, and if necessary, retraining. Inter-rater agreement was also calculated for the operational online ELA/L PCR tasks scored by IEA.

In addition to perfect agreement and the Pearson correlation, two common indices used to gauge rater agreement are Cohen’s kappa and quadrative weighted kappa. Cohen’s kappa (κ) measures the agreement between two raters while accounting for the agreement expected by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o (the observed agreement) is the proportion of instances where the raters agree and P_e (the expected agreement) is the proportion of agreement expected by chance, calculated as shown below.

$$P_e = \sum_i (\frac{\sum_j n_{ij}}{N} . \frac{\sum_j n_{ji}}{N})$$

where n_{ij} is the number of items where Rater 1 assigned score *i* and Rater 2 assigned score *j* given the total number of items (*N*).

Quadratic weighted kappa (QWK) considers the agreement between raters while penalizing disagreements based on their squared difference:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} . O_{ij}}{\sum_{i,j} w_{ij} . E_{ij}}$$

where O_{ij} and E_{ij} are the observed and expected agreement matrices respectively and the weight (w_{ij}) assigned to the disagreement between scores *i* and *j* given the total number of scores available (*k*) is calculated as shown below.

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

While O_{ij} is the actual count of ratings, E_{ij} is calculated based on the marginal totals of the observed ratings:

$$E_{ij} = \frac{\sum_m O_{im} \cdot \sum_t O_{tj}}{N}$$

where O_{im} is the total number of items assigned to score *i* by one rater and O_{tj} is the total number of items assigned to score *j* by the other rater.

Table 5.4 presents the inter-rater agreement expectations and results for the constructed-response items from the Spring 2024 administration across all grades based on human scoring, and Table 5.5 presents the average agreement across the PCRs for each grade by trait from the automated scoring process, including the number of tasks included in the analyses, perfect agreement, kappa, QWK, and Pearson correlation (*r*). PCR items are scored on two traits: Reading Comprehension & Written Expression and Conventions for the Literary Analysis and Research Simulation tasks, and Written Expression and Conventions for the Narrative Writing task. For the ELA/L PCR traits, the expectation for agreement is an inter-rater agreement of 65% or higher between two scorers. When IEA provided the first score of record, the second reliability score was a human score. For a subset of responses, the first and second score were both human scores.

Table 5.4. Inter-Rater Agreement Expectations and Spring 2024 Results

Content Area	#Items	Score Point Range	Perfect Agreement Expectation	Perfect Agreement 2024 Result	Within 1 Point Expectation*	Within 1 Point 2024 Result
ELA/L	15	Multi-trait	65%	93%	96%	100%
Mathematics	10	0–2	80%	100%	96%	100%
	25	0–3	70%	100%	96%	100%
	11	0–4	65%	100%	95%	100%
	3	0–5	65%	100%	95%	100%
	2	0–6	65%	100%	95%	100%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point.

Table 5.5. ELA/L PCR Average Agreement Indices

			Written Expression				Writing Knowledge Language and Conventions			
Grade	#PCRs	#Tasks	Perfect	Kappa	QWK	*r*	Perfect	Kappa	QWK	*r*
3	2	2	75%	0.54	0.73	0.74	76%	0.58	0.75	0.76
4	3	3	80%	0.66	0.84	0.84	76%	0.61	0.80	0.81
5	2	2	74%	0.61	0.84	0.84	75%	0.61	0.83	0.84
6	2	2	75%	0.64	0.85	0.86	74%	0.63	0.84	0.85
7	3	3	82%	0.73	0.90	0.90	78%	0.66	0.86	0.86
8	3	3	77%	0.67	0.88	0.88	78%	0.67	0.88	0.88

5.5. Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the hierarchy rules in Table 5.6 determined which score was reported as the final operational score.

Table 5.6. Scoring Hierarchy Rules

Score Type	Rank	Final Score Calculation
Adjudication (fourth score)	1	If an adjudication score is assigned, this is the final score.
Resolution (third score)	2	If no adjudication score is assigned, this is the final score.
Backreading score	3	If no adjudication or resolution score is assigned, the latest backreading score is the final score.
Human first score	4	If no adjudication, resolution, or backreading score is assigned, this is the final score.
Human second score	5	If no adjudication, resolution, backreading, or human first score is assigned, this is the final score.
IEA score	6	If no human score is assigned, this is the final score.

Section 6: Reporting

6.1. Available Reports

The following reports are available for the IAR assessments. Student performance is reported on the Individual Student Report (ISR) using scale scores, performance levels, and subclaim performance indicators, as described in the *IAR Score Interpretation Guide* available online at https://il.mypearsonsupport.com/training-resources/. State, district, and school average results are included to help understand how a student's performance compares to that of other students.

- Individual Student Report (ISR)
- Student Roster Report
- District Summary of School Report
- District/School Performance Level Summary Report
- District/School Evidence Statement Analysis Report
- School Content Standards Roster Report

6.2. Interpretation of Test Scores

6.2.1. Total Scale Scores and Performance Levels

The IAR student results are expressed as total scale scores ranging from 650 to 850 for all tests, along with associated performance levels to describe how well students met the academic standards for their grade level. Not all students respond to the same set of test items, so each student's raw score (actual points earned on the test) is converted onto a common scale through the process of scaling to account for the differences in difficulty among the various forms and administrations of the test. The resulting scale score allows for an accurate comparison across test forms and administration years within a grade and content area. For example, a student who receives a raw score of 50 on one form of a mathematics test, meaning they answered 50 points correctly, might receive a scale score of 750. This scale score can then be compared to a different test form of the same test where a raw score of 55 translates into a scale score of 750. The scale scores, not the raw scores, reflect the same ability and knowledge levels.

Based on a student's total score, an inference is drawn about how much knowledge and skill in the content area the student has acquired. The overall scale scores also determine a student's performance level that classifies a student's competency based on their test performance as reflected by their test results, as described in Table 6.1. Each performance level is defined by a range of overall scale scores for the assessment established during the standard setting (see Section 7 for more details). Students classified as either Level 4 or Level 5 are meeting or exceeding the grade-level expectations. The table presents the general policy descriptions that define the high-level expectations of student achievement within each performance level across grades, as well as the expectations specific to grades 3–8. The full PLDs for the IAR assessments are available online at www.isbe.net/iar.

Table 6.1. Performance Levels

Performance Level	General Policy Description	Grades 3–8
Level 5: *Exceeded Expectations*	Students performing at this level **exceed academic expectations** for the knowledge, skills, and practices contained in the ELA/L or mathematics standards assessed at their grade level.	Students are **academically well prepared** to engage successfully in further studies in this content area.
Level 4: *Met Expectations*	Students performing at this level **meet academic expectations** for the knowledge, skills, and practices contained in the ELA/L or mathematics standards assessed at their grade level.	Students are **academically prepared** to engage successfully in further studies in this content area.
Level 3: *Approached Expectations*	Students performing at this level **approach academic expectations** for the knowledge, skills, and practices contained in the ELA/L or mathematics standards assessed at their grade level.	Students are **likely prepared** to engage successfully in further studies in this content area.
Level 2: *Partially Met Expectations*	Students performing at this level **partially meet academic expectations** for the knowledge, skills, and practices contained in the ELA/L or mathematics standards assessed at their grade level.	Students **will likely need academic support** to engage successfully in further studies in this content area.
Level 1: *Did Not Yet Meet Expectations*	Students performing at this level **do not yet meet academic expectations** for the knowledge, skills, and practices contained in the ELA/L or mathematics standards assessed at their grade level.	Students **will need academic support** to engage successfully in further studies in this content area.

6.2.2. Claim and Subclaim Scores

The ISR for the ELA/L assessments provide separate claim scale scores for both Reading and Writing. The claim scale scores and the summative scale score are on different scales, so the sum of the scale scores for each claim will not equal the summative scale score. Reading scale scores range from 10 to 90, and Writing scale scores range from 10 to 60. The claim scores can be interpreted by comparing a student's claim scale score to the average performance for the school, district, and state. The ISR provides the student scale score results and the average scale score results for the school, district, and state.

Within each reporting category are specific skill sets (subclaims) students demonstrate on the IAR. Each subclaim category includes the header identifying the subclaim, an explanatory icon representing the student's performance, and an explanation of whether the student has met the expectations of the subclaim. Subclaim indicators represent how well students performed in a subclaim category. Performance in the Level 1–2 range of that scale is categorized as "Lower level readiness" represented by the letter L, performance in the Level 3 range is categorized as "Middle level readiness" represented by the letter M, and performance in the Level 4–5 range is categorized as "Higher level readiness" represented by the letter H.

6.2.3. Additional Measures

The ISR also includes Lexile® and Quantile measures that represent both a student's reading ability and the difficulty of a text and both a student's mathematical achievement and the difficulty of a mathematical skill or concept, respectively. Student growth percentiles (SGPs) are also provided that estimate individual student progress by tracking student scores from one year to the next. The first year a student tests in Illinois is their baseline year. (See Section 15 for more information on SGPs.)

Section 7: Standard Setting

Cut scores, also known as performance standards, relate levels of performance on an assessment directly to what students are expected to learn by separating an assessment's score scale into performance levels. Standard setting, also known as performance level setting, is the process of establishing the cut scores that define the performance levels for an assessment. This section summarizes the 2015 PARCC standard setting, with the full details about the process provided in the standard setting report (Davis & Moyer, 2015).

A main objective of the assessment system is to provide information to students, parents, educators, and administrators as to whether students are on track in their learning for success after high school, defined as college and career readiness. To set performance levels associated with this objective, the evidence-based standard setting (EBSS) method (Beimers et al., 2012) was used during the standard setting meetings conducted in one-week sessions, as shown in Table 7.1.

Table 7.1. 2015 Standard Setting Meetings

Dates	Committees
August 17–21, 2015	Grades 7–8 Mathematics Grades 7–8 ELA/L
August 24–28, 2015	Grades 3–4 Mathematics Grades 5–6 Mathematics Grades 3–4 ELA/L Grades 5–6 ELA/L

7.1. Standard Setting Process

The EBSS method is a systematic method for combining various considerations into the process for setting performance levels, including policy considerations, content standards, educator judgment about what students should know and be able to demonstrate, and research to support policy goals related to college and career readiness. A defined multistep process was used to allow a diverse set of stakeholders to consider the interaction of these elements in recommending performance level threshold scores for each assessment. The following steps of the EBSS process were followed to establish cut scores for the summative assessments:

1. Define the outcomes of interest and policy goals.
2. Develop research, data collection, and analysis plans.
3. Synthesize the research results.
4. Conduct pre-policy meeting.
5. Conduct standard setting meetings with panels.
6. Conduct a reasonableness review with post-policy panel.
7. Continue to gather evidence in support of the standards.

During the standard setting meetings, committees recommended four cut scores that would define the five performance levels for each assessment. PARCC participating states and agencies solicited panelist nominations from all states that administered the assessments in 2014–2015. Nominations were solicited both from state departments of public education (K–12) and higher education (primarily for participation on the high school panels). When selecting panelists, an emphasis was placed on educators with content knowledge and experience with a variety of student groups and a balance in terms of state representation.

An Extended Modified Yes/No Angoff method (Plake et al., 2005) was used to collect judgments on the items. This method asked panelists to review each item on a reference form of the assessment and to make the following judgment: "*How many points would a borderline student at each performance level likely earn if they answered the question?*" This allowed for incorporation of the multi-point items by asking educators to evaluate whether a borderline student would earn the maximum number of points on an item, a lesser number of points on an item, or no points on the item. For single-point or multiple-choice items, this task simplified to the standard Yes/No method.

After receiving training on the standard setting procedure, panelists participated in three rounds of judgments. Each panelist made judgments for the Level 2 performance level, followed by judgments for Level 3, Level 4, and then Level 5. The individual judgments were summed across items for a committee to create an estimated total score on the reference form for each cut score. Feedback data relative to panelist agreement, student performance on the items, and student performance on the overall test were provided in between each judgment round.

The cut scores recommended by the standard setting committees were then reviewed by the Advisory Committee on College Readiness as part of a post-policy reasonableness review. Members of the original standard setting committees were recruited to participate in this process. This group reviewed both the median cut score recommendations from each committee and the variability in the cut scores as represented by the standard error of judgment of the committee. Adjustments to the median cut scores that were within two standard errors of judgment were considered consistent with the standard setting panels' recommendation.

In addition to voting to adopt the cut scores based on the committees' recommendations, this group also voted to conduct a shift in the performance levels to better meet the intended inferences about student performance. Holding the college- and career-ready (or on track) expectations (i.e., the Level 4) constant, performance levels above this expectation were combined and performance levels below this expectation were expanded to create the final system of performance levels with three below and two above the college- and career-ready (or on track) expectation. The shift in performance levels was accomplished using a scale anchoring process that involved two primary steps:

- Combine the top two performance levels, above college and career ready (or on track), into a single performance level and create an additional performance level below college and career ready (or on track) by empirically determining the midpoint between the existing two levels.
- Update the PLDs using items that discriminated student performance well at this level to create a PLD aligned with the new empirically determined performance level and review the PLDs for all performance levels for consistency and continuity.

7.2. Results

Table 7.2 presents the resulting IAR scale score cut scores (i.e., the minimum score students must receive to be classified into a certain performance level), as shown in bold.

Table 7.2. Scale Score Ranges

Assessment	Level 1	Level 2	Level 3	Level 4	Level 5
ELA/L 3	**650**–699	**700**–724	**725**–749	**750**–809	**810**–850
ELA/L 4	**650**–699	**700**–724	**725**–749	**750**–789	**790**–850
ELA/L 5	**650**–699	**700**–724	**725**–749	**750**–798	**799**–850
ELA/L 6	**650**–699	**700**–724	**725**–749	**750**–789	**790**–850
ELA/L 7	**650**–699	**700**–724	**725**–749	**750**–784	**785**–850
ELA/L 8	**650**–699	**700**–724	**725**–749	**750**–793	**794**–850
Mathematics 3	**650**–699	**700**–724	**725**–749	**750**–789	**790**–850
Mathematics 4	**650**–699	**700**–724	**725**–749	**750**–795	**796**–850
Mathematics 5	**650**–699	**700**–724	**725**–749	**750**–789	**790**–850
Mathematics 6	**650**–699	**700**–724	**725**–749	**750**–787	**788**–850
Mathematics 7	**650**–699	**700**–724	**725**–749	**750**–785	**786**–850
Mathematics 8	**650**–699	**700**–724	**725**–749	**750**–800	**801**–850

Section 8: Student Characteristics and Test Results

8.1. Student Participation

Table 8.1 presents the number and percentage of students who took the IAR assessments by administration mode (online vs. paper). The results include students taking the accommodated forms.

Table 8.1. Student Participation by Administration Mode

		Online	Online	Paper	Paper
Assessment	#Valid Cases	N	% of Grade	N	% of grade
ELA/L 3	129,997	129,503	99.6	494	0.4
ELA/L 4	129,858	129,545	99.8	313	0.2
ELA/L 5	129,335	129,085	99.8	250	0.2
ELA/L 6	130,763	130,540	99.8	223	0.2
ELA/L 7	133,542	133,353	99.9	189	0.1
ELA/L 8	134,873	134,651	99.8	222	0.2
ELA/L Total	788,368	786,677	99.8	1,691	0.2
Mathematics 3	130,057	129,482	99.6	575	0.4
Mathematics 4	129,924	129,446	99.6	478	0.4
Mathematics 5	129,432	129,064	99.7	368	0.3
Mathematics 6	130,694	130,323	99.7	371	0.3
Mathematics 7	133,394	133,049	99.7	345	0.3
Mathematics 8	134,715	134,345	99.7	370	0.3
Mathematics Total	788,216	785,709	99.7	2,507	0.3

Table 8.2 and Table 8.3 present the number of students with valid scores by demographic subgroup as captured in PearsonAccess$^{\text{next}}$ by means of a student data upload. The demographic data were verified by Illinois prior to score reporting. Students missing information on one or more of the demographic variables were omitted from the subgroup analyses.

Table 8.2. Student Participation by Demographic Subgroup—ELA/L

Demographic	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Economically Disadvantaged	67,550	66,970	66,712	66,562	67,218	67,986
Student with Disabilities (SWD)	22,675	23,960	24,060	24,123	24,257	24,452
English Learner (EL)	28,251	25,844	21,644	19,490	20,701	20,058
Male	66,022	66,235	65,954	66,552	68,216	69,345
Female	63,961	63,603	63,355	64,178	65,293	65,479
American Indian/Alaska Native	347	332	308	293	275	281
Asian	7,556	7,457	7,538	7,592	7,573	7,611
Black/African American	21,124	21,032	20,792	20,836	21,295	22,079
Hispanic/Latino	36,016	35,768	35,756	36,643	37,880	38,349
Native Hawaiian or Other Pacific Islander	107	107	106	142	132	104
White/Caucasian	58,290	58,921	58,788	59,178	60,523	60,736
Two or More Races Reported	6,379	6,081	5,850	5,885	5,698	5,531
Unknown	178	160	197	194	166	182

Table 8.3. Student Participation by Demographic Subgroup—Mathematics

Demographic	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Economically Disadvantaged	67,404	66,831	66,620	66,324	66,934	67,710
Student with Disabilities (SWD)	22,608	23,927	24,056	24,025	24,188	24,369
English Learner (EL)	28,203	25,783	21,610	19,422	20,593	19,980
Male	66,023	66,240	66,003	66,486	68,110	69,248
Female	64,019	63,664	63,403	64,174	65,250	65,420
American Indian/Alaska Native	346	330	309	290	275	277
Asian	7,543	7,445	7,543	7,572	7,547	7,593
Black/African American	21,061	20,996	20,779	20,754	21,198	21,975
Hispanic/Latino	35,957	35,677	35,698	36,537	37,742	38,227
White/Caucasian	107	107	106	141	134	103
Native Hawaiian or Other Pacific Islander	58,237	58,865	58,725	59,089	60,413	60,610
Two or More Races Reported	6,370	6,077	5,848	5,867	5,679	5,511
Unknown	436	427	424	444	406	419

8.2. Scale Score Distributions

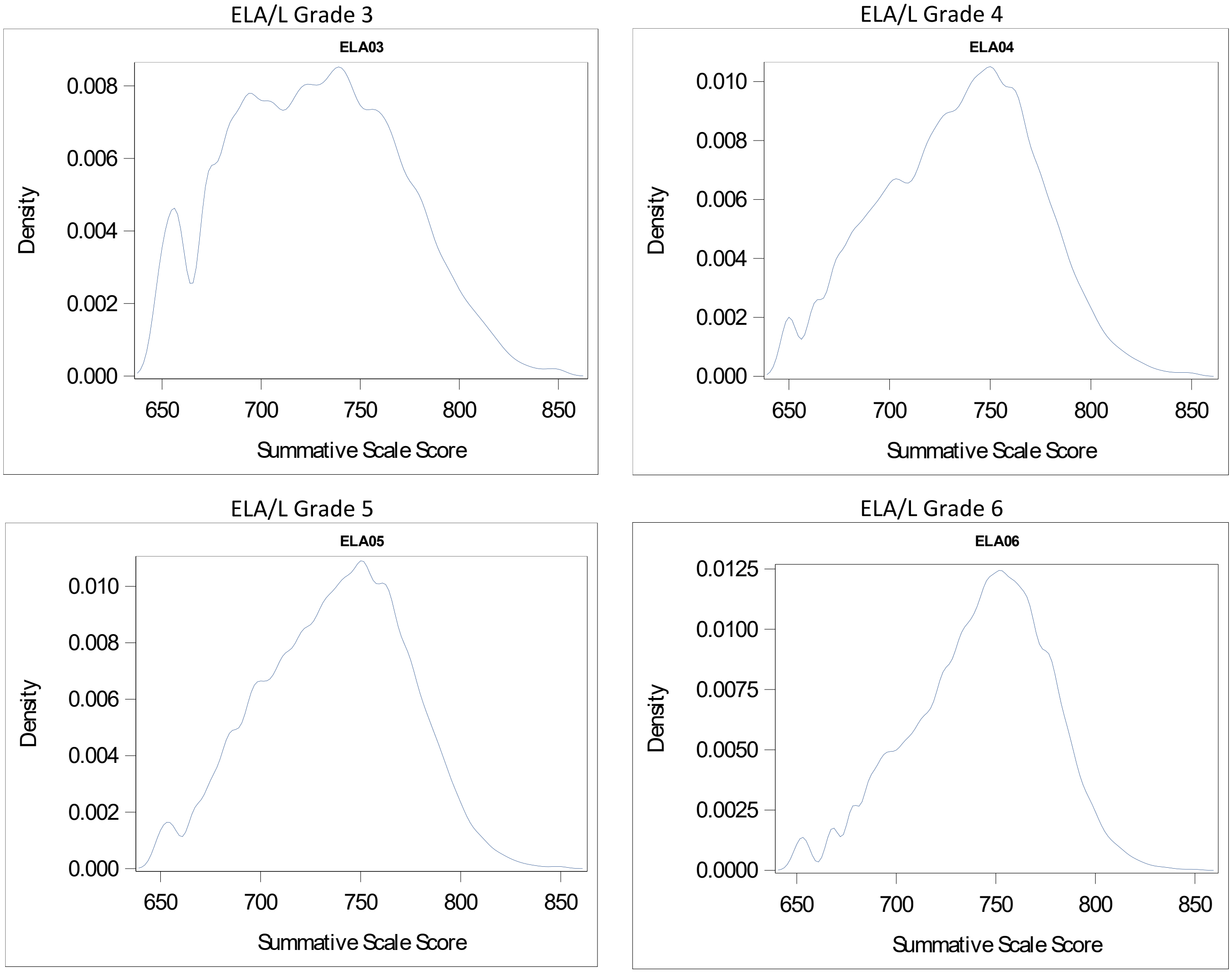
Figure 8.1 – Figure 8.4 present the Spring 2024 IAR scale score distributions. The vertical *y*-axis labeled "Density" represents the proportion of students earning the scale score point indicated along the horizontal *x*-axis. The overall score scale ranges from 650 to 850, the Reading score scale ranges from 10 to 90, and the Writing score scale ranges from 10 to 60. Appendix A presents the cumulative frequency distribution for the overall scale scores, and Appendix B presents the subgroup statistics for the summative, Reading, and Writing scale scores.

Scale score distributions for mathematics peaked between approximately 700, and the distributions of the ELA/L overall scale scores were centered around the Level 2 cut score (700) or slightly below. Reading scale scores tended to be centered around or slightly below the Level 2 cut score of 30 and were slightly more irregular than the summative scale scores.

The Writing scale score distributions were less smooth than the Reading or ELA/L summative distributions due to peaks related to the weighting of the Written Expression portion of the PCR tasks and a noticeable proportion of students at the LOSS. Due to the weighting of the Written Expression trait, multiple Writing scale score values are not likely to be obtained resulting in multiple peaks across the range of the Writing scale score. A noticeable proportion of students earned the LOSS of 10 in Writing across all grades. Students with 0 raw score points on the written portion of the assessment are automatically assigned the LOSS value of a scale. Writing items are embedded exclusively in PCR tasks, which tended to be difficult. The Written Expression trait also tended to be the most difficult of the PCR traits.

Across the ELA/L grades, few students are between 11 and 20, depending on the grade.[2] The LOSS is 10, which was selected to be consistent with the Reading LOSS and reduce truncation at the lower ends of the scale. However, the scale is defined by the theta values associated with the Level 2 and Level 4 performance levels. All other scale score values are identified through a theta-to-scale score linear transformation applying the scaling constants (Table 11.3). For Writing, the lowest theta estimate associated with raw scores ranging from one to two are linearly transformed to scale score values between 15 and 20, meaning that there may be multiple scale scores between 11 and 20 that are not assigned to a raw score. In contrast, the Reading lowest theta estimates associated with raw scores ranging from one to two are linearly transformed to scale score values closer to the LOSS. The gap in the proportion of students at the scale scores between the LOSS value of 10 and the scale score values around 17 to 19 is an artifact of the scale score task force selecting the LOSS value of 10.

Figure 8.1. Scale Score Distributions—ELA/L

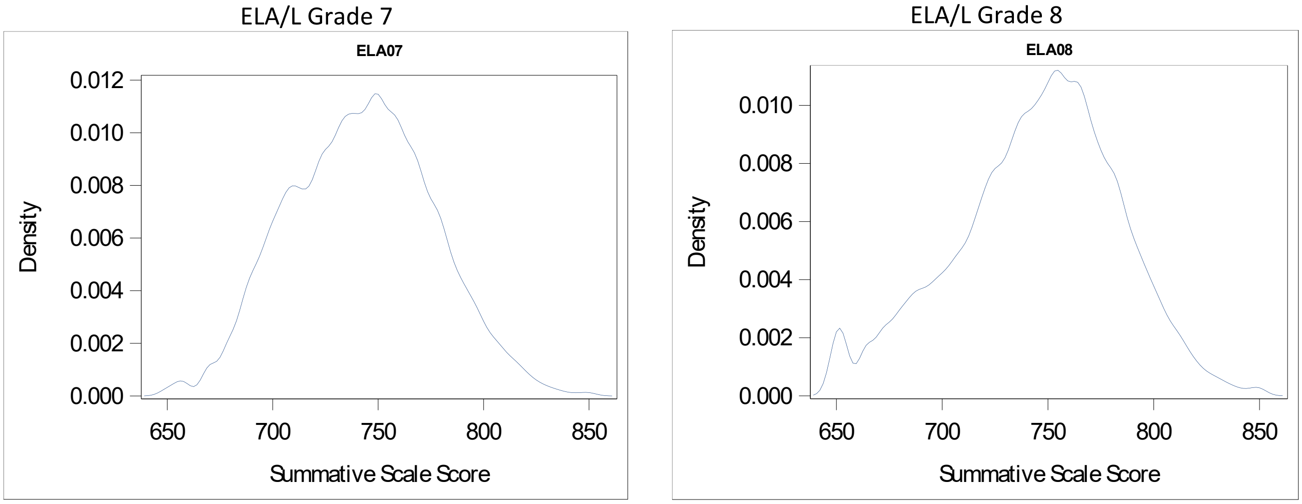

ELA/L Grade 3

ELA/L Grade 4

ELA/L Grade 5

ELA/L Grade 6

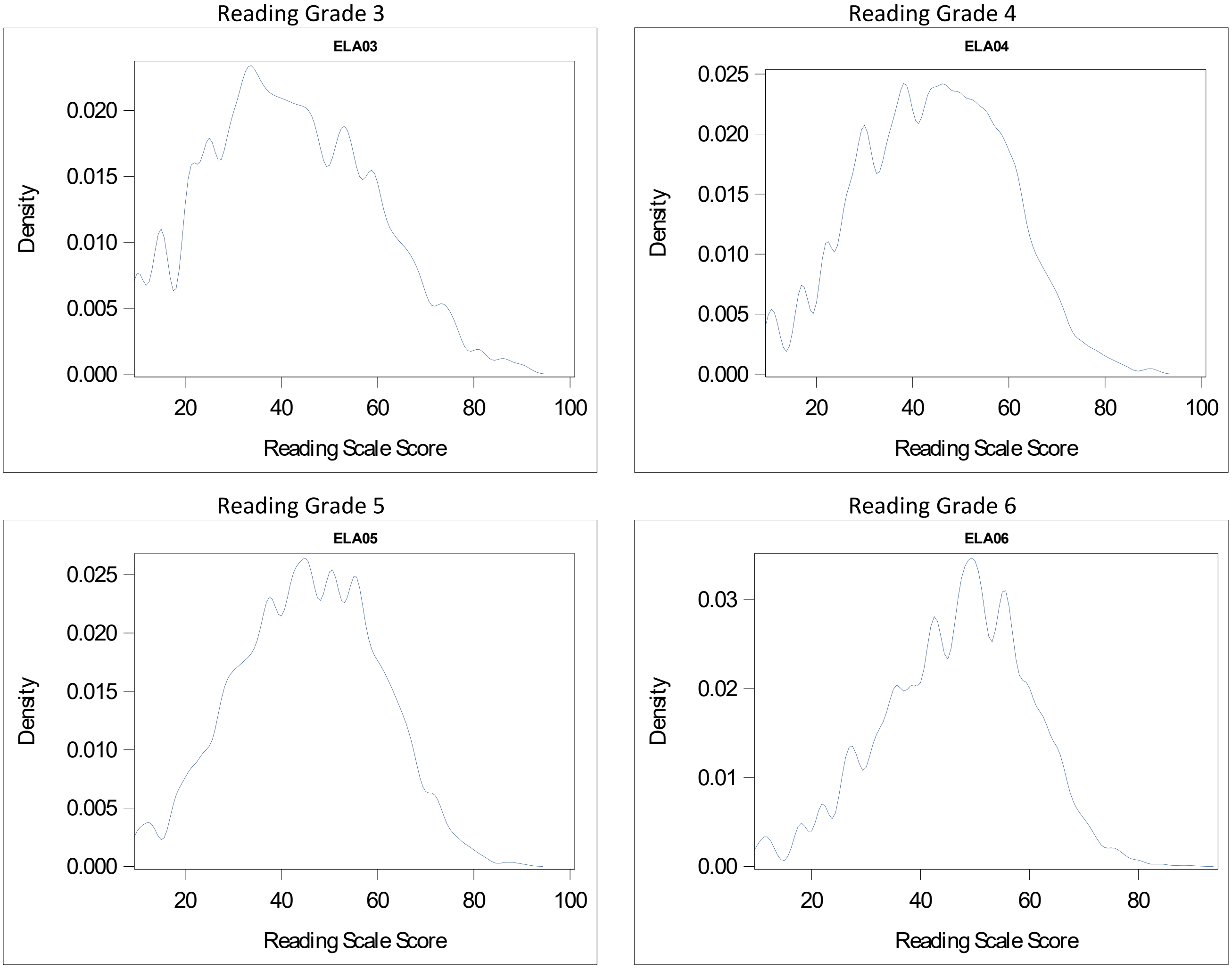
[2] Due to smoothing of the kernel density function, in some figures, particularly those with small sample sizes, the line representing the distribution may appear to remain above zero near the region.



ELA/L Grade 7

ELA/L Grade 8

Figure 8.2. Scale Score Distributions—Reading

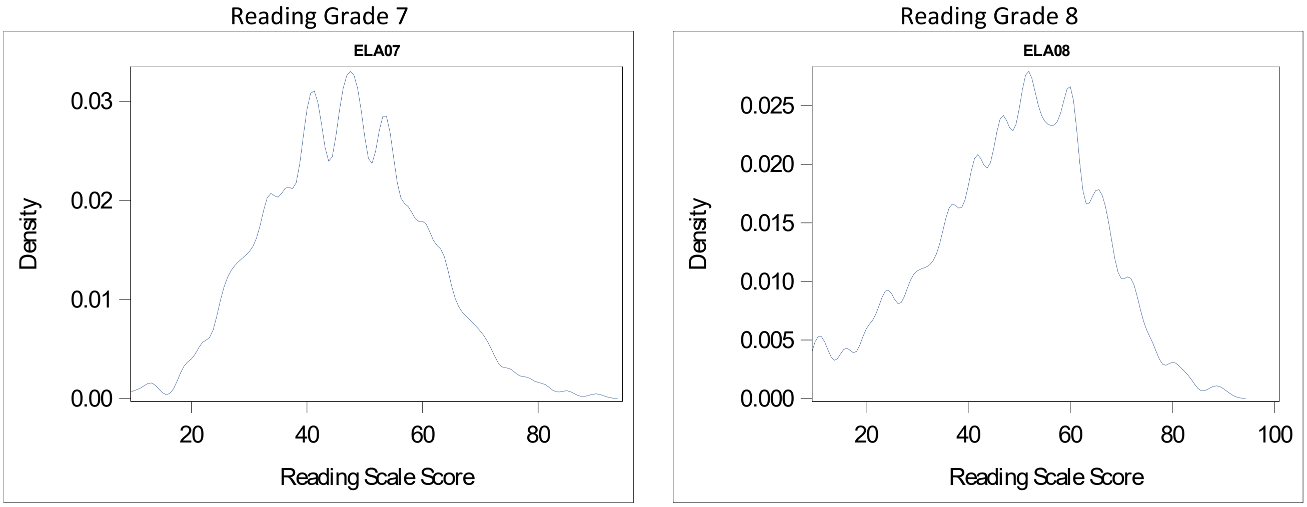


Reading Grade 3

Reading Grade 4

Reading Grade 5

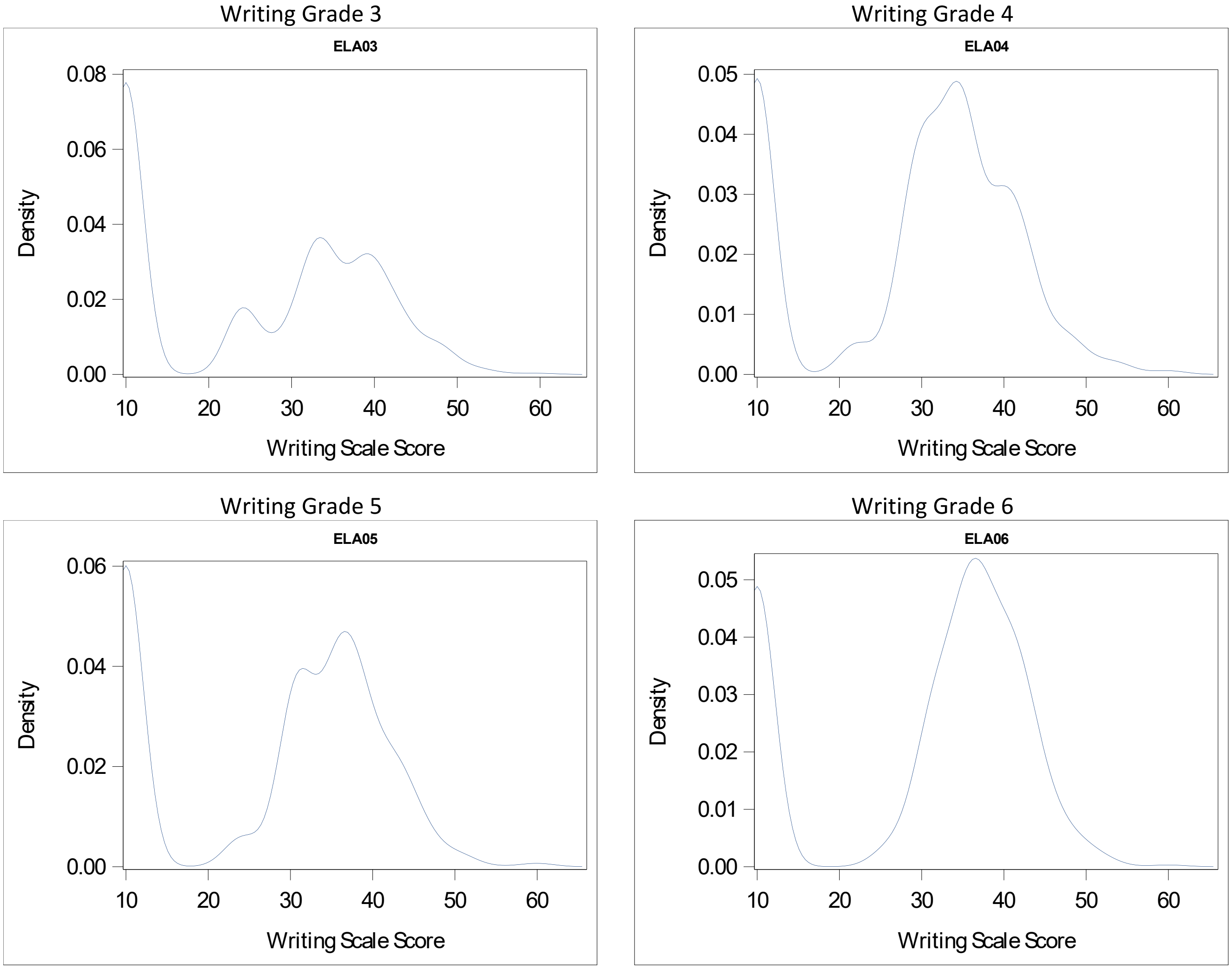
Reading Grade 6



Reading Grade 7

Reading Grade 8

Figure 8.3. Scale Score Distributions—Writing

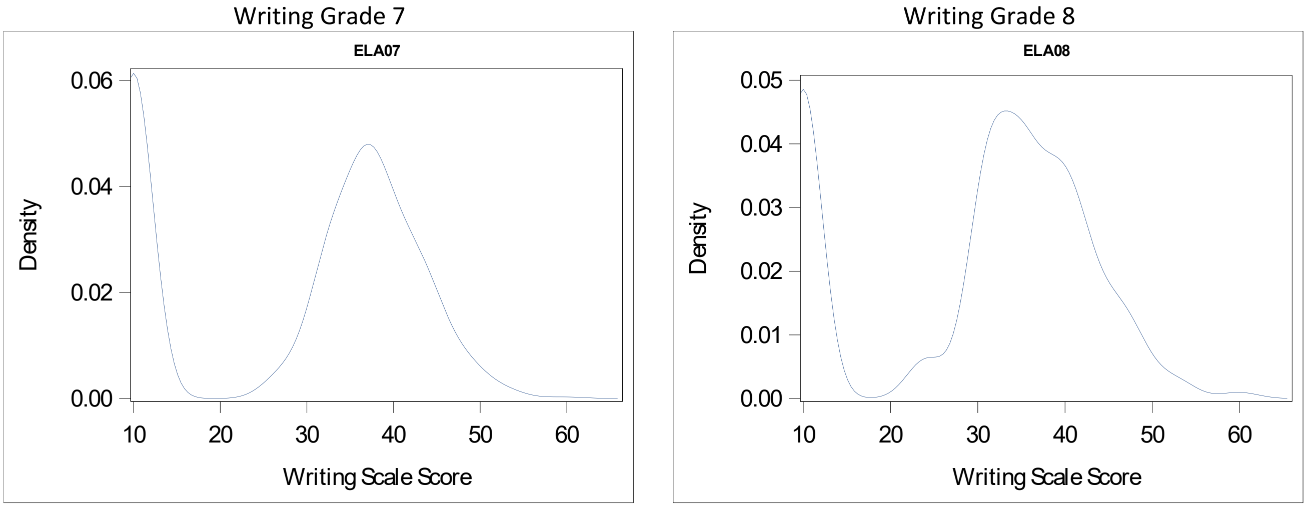


Writing Grade 3

Writing Grade 4

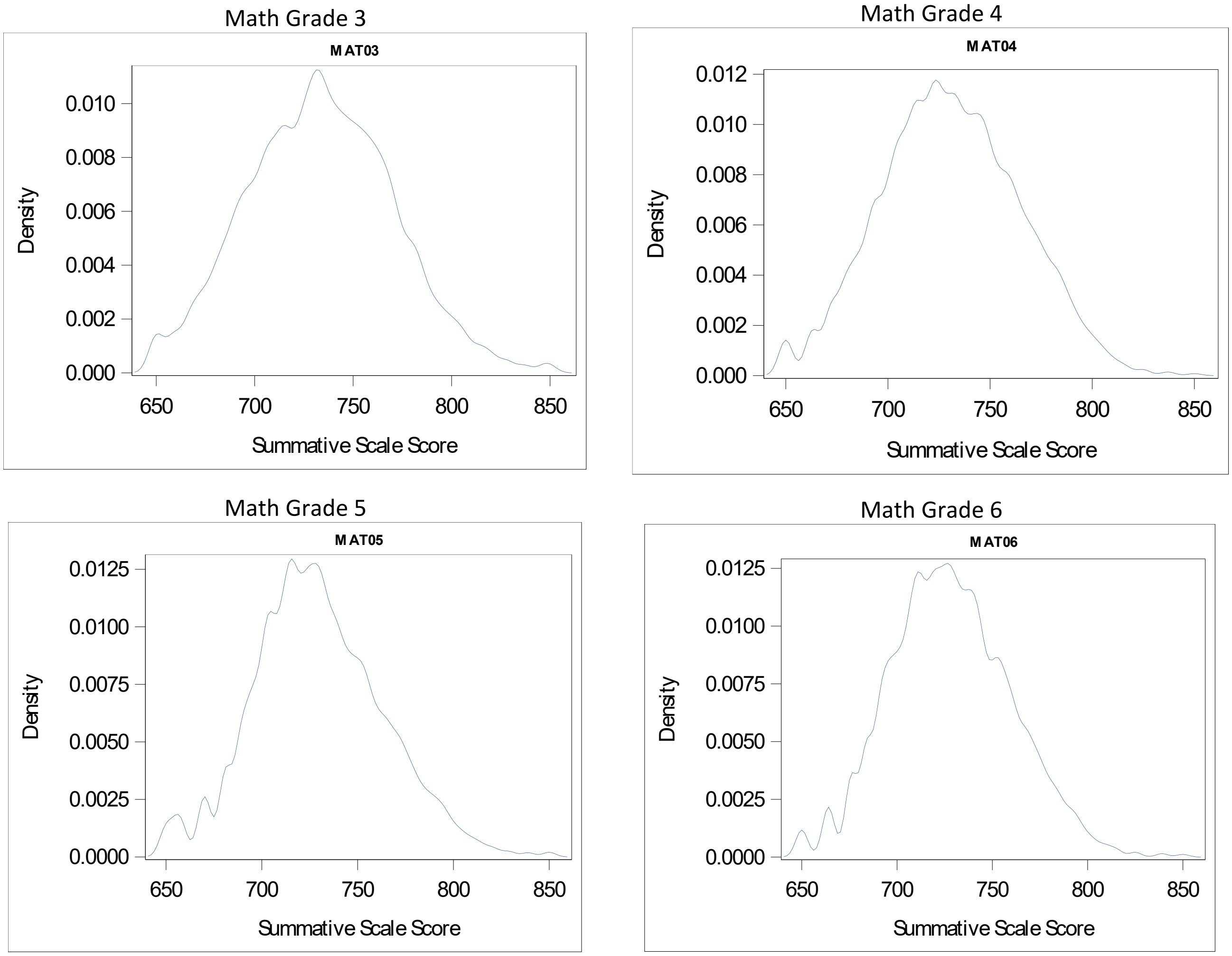
Writing Grade 5

Writing Grade 6


Writing Grade 7

Writing Grade 8

Figure 8.4. Scale Score Distributions—Mathematics

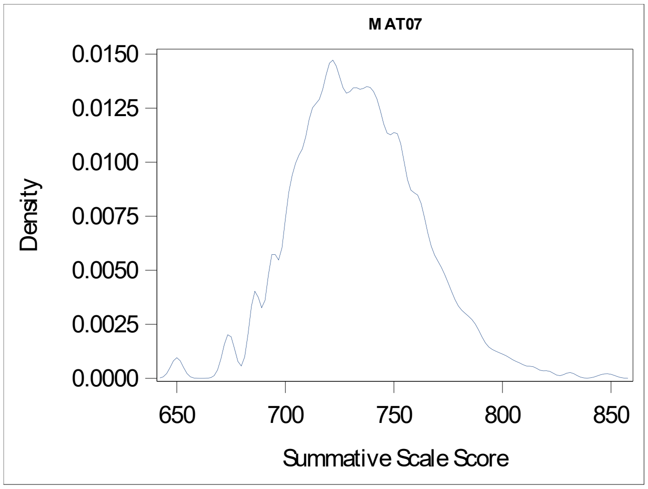

Math Grade 3

Math Grade 4

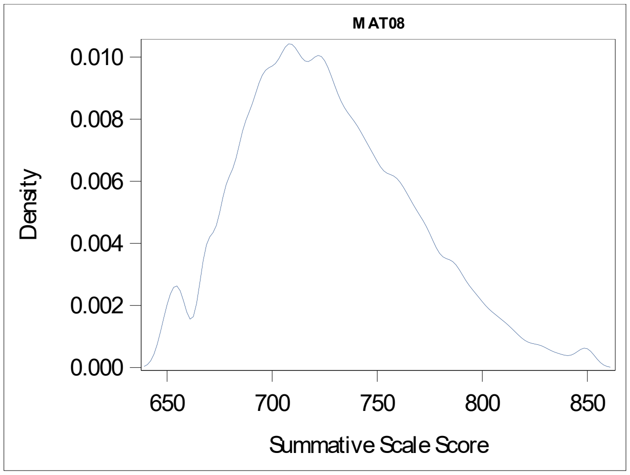
Math Grade 5

Math Grade 6

Math Grade 7



Math Grade 8



Section 9: Classical Item Analysis

This section presents item analysis results for the operational items included on the Spring 2024 test forms. All assessments were pre-equated, meaning the scoring was based on item parameters estimated using data from earlier administrations. As a result, the item analysis results are from prior operational administrations that were used to make decisions during the test construction process and for score reporting.

9.1. Data Preparation

In preparation for item analysis, student response files were processed to verify that the data were free of errors. Pearson Customer Data Quality staff ran predefined checks on all data files and verified that all fields and data needed to perform the statistical analyses were present and within expected ranges. Next, to produce higher-quality (albeit slightly smaller) datasets, Pearson psychometricians established the following criteria for including students in the operational analyses to determine which, if any, student records should be removed prior to conducting the analysis:

- Exclude all records with an invalid form number.
- Exclude all records flagged as “void.”
- Exclude all records where the student attempted fewer than 25% of items.
- For students with more than one valid record, choose the record with the higher raw score.
- Exclude records for students with administration issues or anomalies.

The following factors were also considered during the analyses:

- An operational item may appear on multiple test forms. The item analysis results present unique item counts for an assessment, and the reported item statistics may be based on student responses across multiple occurrences of an item.
- Spoiled or “do not score” items were excluded from the total test score in the item analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, or multiple/no correct answers.

9.2. Item Analyses

The following item-level analyses were calculated for the IAR assessments. Item difficulty and discrimination results are presented in this technical report, whereas the remaining analyses were conducted during key check and adjudication after the IAR test window.

- Item difficulty (*p*-value)
- Item discrimination (item-total correlation)
- Distractor-total correlation for the selected-response items
- Percentage of students choosing each answer option for the selected-response items
- Percentage of students omitting or not reaching each item
- Distribution of item scores

9.2.1. Item Difficulty (P-value)

When constructing tests, a wide range of item difficulties is desired (from easy to hard items) so that students of all ability levels can be assessed with precision. Item difficulty is measured by the *p*-value statistic bounded by 0.0 and 1.0 that indicates how easy or hard an item is for students. The *p*-value for dichotomous 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 polytomous 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. Items were flagged for review if the *p*-value was above 0.95 (i.e., too easy) or below 0.25 (i.e., too difficult).

Table 9.1 presents the *p*-value summary statistics for the operational items. The average *p*-values varied across grades, and neither subject had a clear trend of average and median *p*-value change across grades.

Table 9.1. Summary of *p*-Values

Assessment	#Unique Items	Mean	SD	Min.	Max.	Median
ELA/L 3	169	0.37	0.14	0.07	0.74	0.39
ELA/L 4	227	0.38	0.14	0.13	0.75	0.37
ELA/L 5	226	0.41	0.15	0.13	0.80	0.38
ELA/L 6	224	0.42	0.15	0.11	0.82	0.42
ELA/L 7	225	0.44	0.15	0.15	0.82	0.43
ELA/L 8	211	0.46	0.15	0.16	0.79	0.47
Mathematics 3	243	0.54	0.24	0.02	0.96	0.55
Mathematics 4	236	0.47	0.22	0.01	0.96	0.47
Mathematics 5	237	0.44	0.20	0.02	0.91	0.47
Mathematics 6	242	0.40	0.20	0.00	0.92	0.38
Mathematics 7	237	0.32	0.17	0.00	0.82	0.30
Mathematics 8	232	0.36	0.18	0.01	0.90	0.35

Note. SD = standard deviation, Min. = minimum, Max. = maximum

9.2.2. Item Discrimination (Item-Total Correlation)

Item discrimination is represented by the item-total correlation bounded by -1.0 and 1.0 that describes the relationship between performance on a specific item and performance on the total test and indicates how well an item discriminates, or distinguishes, between low- and high-performing students. 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 item-total correlation discriminates between low- and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that low-performing students performed better on an item than high-performing students, an indication that the item may be flawed. The item-total correlation was calculated for both dichotomous and polytomous items as an estimate of the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson et al., 1982). Item-total correlations below 0.15 were flagged for review.

Table 9.2 presents the item-total correlation summary statistics for the operational items. The average item-total correlations varied across grades, and neither subject had a clear trend of average and median item-total correlation change across grades.

Table 9.2. Summary of Item-Total Correlations

Assessment	#Unique Items	Mean	SD	Min.	Max.	Median
ELA/L 3	169	0.54	0.16	-0.15	0.80	0.54
ELA/L 4	227	0.52	0.17	-0.18	0.82	0.53
ELA/L 5	226	0.50	0.15	0.18	0.83	0.49
ELA/L 6	224	0.52	0.14	0.08	0.86	0.50
ELA/L 7	225	0.52	0.16	0.19	0.85	0.51
ELA/L 8	211	0.51	0.15	-0.02	0.86	0.50
Mathematics 3	243	0.47	0.14	0.14	0.79	0.47
Mathematics 4	236	0.48	0.15	-0.31	0.77	0.47
Mathematics 5	237	0.47	0.14	0.08	0.77	0.45
Mathematics 6	241	0.48	0.18	0.08	0.86	0.47
Mathematics 7	237	0.44	0.18	-0.01	0.80	0.42
Mathematics 8	232	0.45	0.15	-0.06	0.78	0.43

Note. SD = standard deviation, Min. = minimum, Max. = maximum

The item-total correlation was also calculated for the distractors of selected-response items to describe the relationship between selecting an incorrect response (i.e., a distractor) for an item and performance on the total test. Items with distractor-total correlations above 0.0 were flagged for review as these items may have multiple correct answers, be miskeyed, or have other content issues.

9.2.3. Percentage of Students Choosing Each Answer Option

Selected-response items refer primarily to single-select multiple-choice scored items that require the student to select a response from several answer options. The percentage of students choosing each answer option for single-select multiple-choice items is calculated, along with the percentages for the high-performing students who scored at the top 20% on the assessment. An item is flagged for review if more high-performing students chose an incorrect option than the correct response. Such a result could indicate that the item has multiple correct answers or is miskeyed.

9.2.4. Percentage of Students Omitting or Not Reaching Each Item

Calculating the percentage of students omitting or not reaching each item is useful for identifying problems with test features such as testing time and item/test layout. Typically, if students have an adequate amount of testing time, approximately 95% of students should attempt to answer each item on the test. A distinction is made between “omit” and “not reached” for items without responses: an item is considered “omit” if the student responded to subsequent items and “not reached” if the student did not respond to any subsequent items.

Patterns of high omit or not-reached rates for items located near the end of a test section may indicate that students did not have adequate time. Omit rates for polytomous items tend to be higher than for dichotomous items. Therefore, the omit rate for flagging individual items was 5% for dichotomous items and 15% for polytomous items. If a student omitted an item, they received a score of 0 for that item and was included in the n-count for that item. However, if an item was near the end of the test and classified as “not reached,” the student did not receive a score and was not included in the n-count for that item.

9.2.5. Distribution of Item Scores

For constructed-response items, examination of the distribution of scores is helpful to identify how well the item is functioning. If no student responses are assigned the highest possible score point, this may indicate that the item is not functioning as expected (e.g., the item could be confusing, poorly worded, or unexpectedly difficult), the scoring rubric is flawed, and/or students did not have an opportunity to learn the content. If all or most students score at the extreme ends of the distribution (e.g., 0 and 2 for a three-category item), this may indicate that there are problems with the item or the rubric so that students can receive either full credit or no credit at all, but not partial credit.

The raw score frequency distributions for constructed-response items were computed to identify items with few or no observations at any score points. Items with no observations or a low percentage (i.e., less than 3%) of students obtaining any score point were flagged. Constructed-response items were also flagged if they had U-shaped distributions, with high frequencies for extreme scores and low frequencies for middle score categories.

9.3. Flagging Criteria

Items were flagged for review if the item analysis yielded any of the following results. Pearson's psychometrics team reviewed any flagged items and submitted them to the content team to decide if the items were problematic and should be excluded from scoring.

- *P*-values below 0.25 or above 0.95 that indicates too easy or difficult items
- Item-total correlations below 0.15 that indicate poorly discriminating items
- Distractor-total correlations above 0.0 as these items may have multiple correct answers, be miskeyed, or have other content issues
- Greater number of high-performing students (top 20%) choosing a distractor than the keyed response, which indicates that the item may have multiple correct answers or is miskeyed
- High omit and not-reached rates above 5% for dichotomous items and above 15% for polytomous items, which may indicate that students did not have adequate time if patterns of high omit or not-reached rates for items are located near the end of a test section
- Polytomous items with a score value obtained by less than 3% of responses (i.e., there should be at least 3% of students at each score point)

Section 10: Differential Item Functioning (DIF)

Differential item functioning (DIF) is a statistical procedure used to flag items for potential bias when students from different demographic groups with the same overall ability have a different probability of getting an item correct (e.g., an item that seems easy for female students but not for male students). This section presents DIF results for the operational items included on the Spring 2024 test forms. All assessments were pre-equated, meaning that the scoring was based on item parameters estimated using data from earlier administrations. As a result, the DIF results are from prior operational administrations.

It is important to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify *potential* item bias only. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

10.1. DIF Methods

DIF analyses were conducted for the operational items using the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) for selected-response and dichotomously scored constructed-response items and the standardization DIF procedure for polytomously scored constructed-response items (Dorans, 2013; Dorans & Schmitt, 1991; Zwick et al., 1997) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959). The group representing students in a specific demographic group is referred to as the focal group, and the group comprised of students from outside this group is referred to as the reference group.

In the MH method, students are classified into relevant subgroups of interest (e.g., gender or ethnicity). Using the raw score total as the criteria, students in a certain total score category in the focal group are compared with students in the same total score category in the reference group. For each item, students in the focal group are also compared to students in the reference group who performed equally well on the overall test. The common odds ratio is estimated across all categories of matched student ability using the following formula (Dorans & Holland, 1993), and the resulting estimate is interpreted as the relative likelihood of success on a particular item for members of two groups when matched on ability:

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} \frac{R_{rs}W_{fs}}{N_{ts}}}{\sum_{s=1}^{S} \frac{R_{fs}W_{rs}}{N_{ts}}}, \qquad \text{(Equation 10-1)}$$

where $S$is the number of score categories, $R_{rs}$is the number of students in the reference group who answer the item correctly, $W_{fs}$is the number of students in the focal group who answer the item incorrectly, $R_{fs}$is the number of students in the focal group who answer the item correctly, $W_{rs}$is the number of students in the reference group who answer the item incorrectly, and $N_{ts}$is the total number of students.

To facilitate the interpretation of the MH results, the common odds ratio is frequently transformed to the delta scale using the following formula (Holland & Thayer, 1988):

$$MH\ D - DIF = \text{-}2.35\ ln(\hat{\alpha}_{MH}) \qquad \text{(Equation 10-2)}$$

The standardization DIF procedure compares the item means of the two groups after adjusting for differences in the distribution of students across the values of the matching variable (i.e., total test score) and is calculated as follows:

$$STD - EISDIF = \frac{\sum_{s=1}^{S} N_{fs} \times E_f(Y|X=s)}{\sum_{s=1}^{S} N_{fs}} - \frac{\sum_{s=1}^{S} N_{fs} \times E_r(Y|X=s)}{\sum_{s=1}^{S} N_{fs}}, \qquad \text{(Equation 10-3)}$$

where X = the total score, Y = the item score, S = the number of score categories, N_{fs}= the number of students in the focal group in score category s, E_r= the expected item score for the reference group, and E_f= the expected item score for the focal group.

10.2. Classification

Based on the DIF statistics, items are classified into three categories (Zieky, 1993): Category A items contain negligible DIF, Category B items exhibit slight-to-moderate DIF, and Category C items possess moderate-to-large DIF values. Positive values indicate DIF in favor of the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values indicate DIF in favor of the reference group (i.e., negative DIF items are differentially easier for the reference group). Table 10.1 presents the flagging criteria for the dichotomously scored and polytomously scored constructed-response items.

Table 10.1. DIF Categories

DIF Category	Dichotomous SR And CR Items	Polytomous CR Items
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero or is less than one.	Mantel Chi-square *p*-value > 0.05 or $\lvert STD\text{-}EISDIF/SD \rvert \leq 0.17$
B (slight to moderate)	1. Absolute value of the MH D-DIF is significantly different from zero but not from one and is at least one; or 2. Absolute value of the MH D-DIF is significantly different from one but is less than 1.5. Positive values are classified as “B+” and negative values as “B−.”	Mantel Chi-square *p*-value < 0.05 and $\lvert STD\text{-}EISDIF/SD \rvert > 0.17$
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one and is at least 1.5. Positive values are classified as “C+” and negative values as “C−.”	Mantel Chi-square *p*-value < 0.05 and $\lvert STD\text{-}EISDIF/SD \rvert > 0.25$

Note. STD-EISDIF = standardized DIF, SD = total group standard deviation of item score

10.3. Comparisons

DIF analyses were conducted on each test form for designated comparison groups based on demographic variables including gender, race/ethnicity, economic disadvantage, and special instructional needs such as students with disabilities or English learners (ELs), as shown in Table 10.2. DIF analyses were conducted when the following sample size requirements were met:

- The smaller group, reference or focal, had at least 100 students.
- The combined group, reference and focal, had at least 400 students.

Table 10.2. DIF Comparison Groups

Grouping Variable	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	American Indian/Alaska Native	White
	Black or African American	White
	Hispanic/Latino	White
Special Instructional Needs	English Learner (ELY)	Non-English Learner (ELN)
	Students with Disabilities (SWDY)	Students without Disabilities (SWDN)

10.4. Results

Appendix C presents the pre-administration item DIF results for the operational items included on the Spring 2024 test forms (i.e., the DIF results are from a previous year's bank). Spoiled or "do not score" items were excluded from the total test score for each form in the DIF analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, multiple correct answers, or no correct answers. However, the tables may include items for certain grade levels that were excluded from scoring based on later analyses.

The column "DIF Comparisons" identifies the focal and reference groups for the analysis performed, and "Total #Unique Items" reports the number of unique items included in the analysis. Because DIF analysis is conducted at the parent level for the ELA/L prose constructed responses, the total number of unique items reported in the DIF analysis is smaller than the total number of items reported in the classical item analysis and the IRT summary statistics. Furthermore, "0" indicates that the DIF analysis did not classify any items in the particular DIF category, while "n/a" indicates that the DIF analysis was not performed due to insufficient sample sizes.

Section 11: Calibration, Equating, and Scaling

This section describes the item response theory (IRT) model used in this assessment program, provides descriptive statistics of the item parameters, and describes how the reporting scale was established. All IAR assessments in Spring 2024 were pre-equated.

11.1. IRT Model

The operational items used pre-equated parameters in the context of the two-parameter logistic/generalized partial-credit (2PL/GPC) model, denoted as follows:

$$p_{im}(\theta_j) = \frac{exp[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})]}{\sum_{v=0}^{M_i-1} exp[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{ik})]} \quad \text{(Equation 11-1)}$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $p_{im}(\theta_j)$ is the probability of a student with $\theta_j$getting score$m$on item i; $D$is the IRT scale constant (1.7); $a_i$is the discrimination parameter of item i; $b_i$is the item difficulty parameter of item i; $d_{ik}$is the k^{th} step deviation value for item i; $M_i$is the number of score categories of item i with possible item scores as consecutive integers from zero to $M_i - 1$; and v indexes the response categories and is iterated from 0 to$M_i - 1$.

11.2. IRT Analysis Results

Table 11.1 and Table 11.2 present the pre-equated IRT *b*- and *a*-parameter estimates for the operational items administered in Spring 2024 administration except the items on the reused forms, if applicable, for which the summary results were reported in the technical reports of the source administrations. The tables present the statistics for the Reading and Writing claim items for ELA/L and by item type for mathematics (see Section 2.3 for a description of the item types), including the total number of items and score points, mean, standard deviation (SD), minimum, and maximum.

Table 11.1. Pre-Equated IRT Parameter Estimates Summary—ELA/L

	Item			*b* Estimates Summary				*a* Estimates Summary			
Grade	Grouping	#Points	#Items	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
3	All Items	44	20	0.48	0.95	-2.12	2.02	0.59	0.23	0.28	0.99
	Reading	32	16	0.23	0.91	-2.12	2.02	0.51	0.17	0.28	0.88
	Writing	12	4	1.45	0.14	1.24	1.53	0.92	0.08	0.82	0.99
4	All Items	78	35	0.71	0.65	-0.60	1.90	0.51	0.23	0.23	1.03
	Reading	58	29	0.59	0.63	-0.60	1.89	0.43	0.16	0.23	0.85
	Writing	20	6	1.31	0.35	0.86	1.90	0.90	0.12	0.77	1.03
5	All Items	57	26	0.46	0.92	-1.47	2.04	0.49	0.19	0.21	0.84
	Reading	44	22	0.33	0.94	-1.47	2.04	0.43	0.15	0.21	0.76
	Writing	13	4	1.18	0.26	0.97	1.56	0.79	0.05	0.73	0.84
6	All Items	58	26	0.53	0.70	-0.74	1.77	0.51	0.22	0.17	1.01
	Reading	44	22	0.38	0.64	-0.74	1.54	0.43	0.14	0.17	0.75
	Writing	14	4	1.37	0.37	0.90	1.77	0.91	0.09	0.83	1.01
7	All Items	85	38	0.33	0.75	-1.22	2.40	0.60	0.28	0.17	1.14
	Reading	64	32	0.18	0.72	-1.22	2.40	0.51	0.20	0.17	0.96
	Writing	21	6	1.11	0.25	0.79	1.37	1.08	0.06	0.99	1.14
8	All Items	79	35	0.10	0.64	-1.36	1.33	0.54	0.29	0.18	1.25
	Reading	58	29	-0.07	0.56	-1.36	1.07	0.42	0.16	0.18	0.86
	Writing	21	6	0.93	0.30	0.56	1.33	1.08	0.17	0.83	1.25

Note. SD = standard deviation, Min. = minimum, Max. = maximum

Table 11.2. Pre-Equated IRT Parameter Estimates Summary—Mathematics

Grade	Item Grouping	#Points	#Items	*b* Estimates Summary Mean	SD	Min.	Max.	*a* Estimates Summary Mean	SD	Min.	Max.
3	All Items	83	55	-0.27	1.33	-2.37	3.68	0.73	0.26	0.19	1.23
	Type I	51	46	-0.47	1.35	-2.37	3.68	0.77	0.27	0.19	1.23
	Type II	20	6	0.65	0.73	-0.29	1.49	0.55	0.14	0.38	0.73
	Type III	12	3	0.88	0.13	0.80	1.02	0.51	0.06	0.46	0.58
4	All Items	83	50	-0.09	0.94	-1.86	1.65	0.74	0.21	0.23	1.20
	Type I	51	41	-0.32	0.85	-1.86	1.22	0.77	0.22	0.23	1.20
	Type II	20	6	0.82	0.63	-0.14	1.44	0.64	0.11	0.43	0.75
	Type III	12	3	1.13	0.76	0.26	1.65	0.62	0.18	0.47	0.82
5	All Items	94	56	0.00	1.27	-6.36	2.41	0.67	0.27	0.11	1.36
	Type I	53	45	-0.19	1.32	-6.36	2.41	0.70	0.28	0.11	1.36
	Type II	20	6	0.88	0.50	0.00	1.42	0.54	0.14	0.37	0.78
	Type III	21	5	0.66	0.69	-0.32	1.29	0.55	0.15	0.46	0.81
6	All Items	71	47	0.25	1.02	-2.65	2.33	0.69	0.29	0.12	1.54
	Type I	49	41	0.20	1.04	-2.65	2.33	0.70	0.30	0.12	1.54
	Type II	10	3	0.04	0.71	-0.78	0.51	0.61	0.15	0.44	0.70
	Type III	12	3	1.16	0.57	0.82	1.82	0.58	0.14	0.46	0.73
7	All Items	77	51	0.77	0.89	-1.26	2.35	0.67	0.27	0.19	1.27
	Type I	52	44	0.76	0.95	-1.26	2.35	0.69	0.28	0.19	1.27
	Type II	10	3	1.03	0.56	0.69	1.68	0.61	0.16	0.48	0.79
	Type III	15	4	0.64	0.22	0.39	0.84	0.56	0.11	0.40	0.63
8	All Items	75	49	0.71	1.04	-1.70	2.84	0.59	0.20	0.18	0.97
	Type I	46	40	0.47	1.00	-1.70	2.84	0.60	0.22	0.18	0.97
	Type II	17	5	1.66	0.32	1.37	2.17	0.50	0.11	0.42	0.69
	Type III	12	4	1.90	0.25	1.55	2.12	0.59	0.09	0.47	0.67

Note. SD = standard deviation, Min. = minimum, Max. = maximum

11.3. Establishing the Reporting Scale

Reporting scales designate student performance into one of five performance levels, with Level 1 indicating the lowest level of performance and Level 5 indicating the highest level of performance. Threshold or cut scores associated with performance levels were initially expressed as raw scores on the standard setting forms approved by the Governing Board. A scale score task force was assembled, which made recommendations about how threshold levels would be represented on the reporting scale.

11.3.1. Summative Score Scale and Performance Levels

There are 201 defined summative scale score points for both ELA/L and mathematics, ranging from 650 to 850. The lowest obtainable scale score (LOSS) is 650, and the highest obtainable scale score (HOSS) is 850. The thresholds for summative performance levels on the scale score metric recommended by the scale score task force are Level 2 and Level 4. The cuts are the anchors for establishing the linear transformation between the theta scale and the reported scale score. A scale score of 700 is associated with minimum Level 2 performance, and a scale score of 750 is associated with minimum Level 4 performance. Not all possible scale scores may be realized in a scoring table.

For Spring 2015, scale scores were defined for each test as a linear transformation of the theta(θ_{2015})scale. The theta values associated with the Level 2 and Level 4 performance levels were identified using the test characteristic curve (TCC) associated with the standard setting form. With Levels 2 and 4 scale scores fixed at 700 and 750, respectively, the relationship between theta(θ_{2015})and scale scores($ScaleScore_{2015}$)was established as follows:

$$ScaleScore_{2015} = A_{2015} \times \theta_{2015} + B_{2015} \tag{11-2}$$

where$A_{2015}$is the slope and$B_{2015}$is the intercept. The slope and intercept were established as follows:

$$A_{2015} = \frac{750-700}{\theta_{2015_{Level4}} - \theta_{2015_{Level2}}} \tag{11-3}$$

and

$$B_{2015} = 750 - A_{2015} \times \theta_{2015_{Level4}} \tag{11-4}$$

As indicated by these formulas, the slope and intercept for the summative scale scores were based on the theta scale, and by default the item response theory (IRT) parameter scale, established in 2015. Because the Spring 2016 IRT parameter scale is the base scale for the IRT parameters, the scaling constantsA_{2015}andB_{2015}were updated in order to continue reporting performance levels, summative scale scores, claim scores, and subclaim performance levels on the same scale as 2015. Maintaining the 2015 scale allows for prior year scores to be compared to current and future scores, and it maintains the performance levels cut scores.

New scaling constants for the summative scale score were needed for the linear transformation of the theta scale$\theta_{2016}$to the 2015 reporting scale ($ScaleScore_{2015}$):

$$ScaleScore_{2015} = SA_{2016} \times \theta_{2016} + SB_{2016} \tag{11-5}$$

The slope($slope_{2015\_to\_2016}$)and intercept($intercept_{2015\_to\_2016}$)generated during the year-to-year linking defined the linear relationship between the 2015 theta scale(θ_{2015})and the 2016 theta scale(θ_{2016}). These values were included in the scale score formula, and the formulas were used to solve for the slope(SA_{2016})and(SB_{2016})intercept for 2016. The slope(A_{2016})was updated using the following formula:

$$SA_{2016} = \frac{A_{2015}}{slope_{2015\_to\_2016}} \tag{11-6}$$

where$A_{2015}$is the current scale score multiplicative constant, $slope_{2015\_to\_2016}$ is the multiplicative coefficient from the year-to-year linking, and SA_{2016} is the scale score slope constant for 2016 and beyond. The intercept(B_{2016})was updated using the following formula:

$$SB_{2016} = B_{2015} - A_{2016} \times intercept_{2015\_to\_2016} \tag{11-7}$$

where$B_{2015}$is the current scale score additive constant,$A_{2016}$is the updated scale score slope, and(SB_{2016})is the scale score intercept constant for 2016 and beyond.

In addition, new scaling constants for the Reading and Writing claim scales were needed. The same formulas were applied by replacing the slope(A_{2015})and intercept(B_{2015})with the Reading claim slope and intercept and the Writing claim slope and intercept.

11.3.2. Reading and Writing Claim Scale

There are 81 defined scale score points possible for Reading, ranging from 10 to 90. The threshold Reading and Writing performance levels on the scale score metric recommended by the scale score task force are Level 2 and Level 4. A scale score of 30 is associated with minimum Level 2 performance, and a scale score of 50 is associated with minimum Level 4 performance. There are 51 defined scale score points possible for Writing, ranging from 10 to 60. A scale score of 25 is associated with minimum Level 2 performance, and a scale score of 35 is associated with minimum Level 4 performance. Not all possible scale scores may be realized in a scoring table.

As with the summative scale scores, scale scores for Reading and Writing were defined for each test as a linear transformation of the IRT theta (θ) scale. The same IRT theta scale was used for Reading and Writing as was used for the ELA/L summative scores. The theta values associated with the Level 2 and Level 4 performance levels were identified using the TCC associated with the standard setting form. As with the summative scores, the relationship between theta and scale scores was established with Level 2 and Level 4 theta scores and the corresponding predefined scale scores. Table 11.3 presents the formulas used for this.

Table 11.3. Calculating Scaling Constants for Reading and Writing Claim Scores

Reading	Writing
$Scale = A_R \times \theta + B_R$	$Scale = A_W \times \theta + B_W$
$A_R = \dfrac{50 - 30}{\theta_{Level4} - \theta_{Level2}}$	$A_W = \dfrac{35 - 25}{\theta_{Level4} - \theta_{Level2}}$
$B_R = 50 - A \times \theta_{Level4}$	$B_W = 35 - A \times \theta_{Level4}$

11.3.3. Subclaims Scale

The Level 4 cut is defined as *Meets or Exceeds Expectations* because high school students at Level 4 or above are likely to have the skills and knowledge to meet the definition of college and career readiness. The Level 3 cut is defined as *Nearly Meets Expectations*. Subclaim outcomes center on the Level 3 and Level 4 performance levels and are reported at three levels: *Below Expectations*, *Nearly Meets Expectations*, and *Meets or Exceeds Expectations*.

The subclaim performance levels are designated through the IRT theta (θ) scale for the items associated with a particular subclaim. The theta values and corresponding raw scores associated with the Level 3 and Level 4 performance levels were identified using the TCC. Students earning a raw subclaim score equal to or greater than the Level 4 threshold were designated as *Meets or Exceeds Expectations*. Students not earning a raw subclaim score equal to or greater than the Level 3 threshold were designated as *Below Expectations*. Students whose raw subclaim score fell between the Level 3 and 4 thresholds were designated as *Nearly Meets Expectations*.

11.3.4. Conversion Tables

A conversion table relates the number of points earned by a student on an assessment to the corresponding scale score for the test form administered to that student. An IRT inverse TCC approach is used to develop the relationship between point scores and theta, θ_s (IRT ability estimates). In conducting the calculations, estimates of item parameters and thetas are substituted for parameters in the formulas in each step.

Step 1: Calculate the expected item score (i.e., estimated item true score) for every theta in the selected range (between −15 and +15, in 0.0001 increments) based on the generalized partial credit model for both dichotomous and polytomous items:

$$s_i(\theta_j) = \sum_{m=0}^{M_i-1} mp_{im}(\theta_j) \tag{11-8}$$

$$p_{im}(\theta_j) = \frac{exp[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})]}{\sum_{v=0}^{M_i-1} exp[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{iv})]} \tag{11-9}$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $s_i(\theta_j)$ is the expected item score for item i on theta, θ_j; $p_{im}(\theta_j)$ is the probability of a student, j, with$\theta_j$getting score m on item i; $m_i$is the number of score categories of item i; with possible item scores as consecutive integers from 0 to $m_i - 1$; D is the IRT scale constant (1.7); a_i is a slope parameter; b_i is a location parameter reflecting overall item difficulty; d_{ik} is a location parameter incrementing the overall item difficulty to reflect the difficulty of earning score category k; and v is the number of score categories.

Step 2: Calculate the expected (weighted) test score for every theta in the selected range:

$$T_j = \sum_{i=1}^{I} w_i \, s_i(\theta_j) \tag{11-10}$$

where$T_j$is the expected (weighted) test score on theta,θ_j; $w_i$is the item weight for item i (e.g., with w_i = 2, a dichotomous item is scored as 0 or 2, and a three-category item is scored as 0, 2, or 4); and$I$is the total number of items in a test form.

Step 3: Calculate the estimated conditional standard error of measurement (CSEM) for each theta in the selected range:

$$CSEM_j = \sqrt{\frac{1}{\sum_{i=1}^{I} L_i(\theta_j)}} \tag{11-11}$$

$$L_i(\theta_j) = (Da_i)^2[s_{i2}(\theta_j) - s_i^2(\theta_j)] \tag{11-12}$$

$$s_{i2}(\theta_j) = \sum_{m=0}^{M_i-1} m^2 p_{im}(\theta_j) \tag{11-13}$$

where $L_i(\theta_j)$ is the estimated item information function for item$i$on theta, θ_j.

Step 4: Match every raw score with a theta.$\theta_j$is the theta for a raw scorer_h, if $T_j - r_h$is minimum across allT_j.

Step 5: Calculate the reported scale score. Using the$A$and$B$scaling constants, convert each theta value to a scale score and each theta CSEM to a scale score CSEM:

$$ScaleScore = A \times \theta + B \tag{11-14}$$

$$CSEM = CSEM_\theta \times A \tag{11-15}$$

The scale scores are rounded to the nearest whole number, and CSEMs are rounded to the tenths place. Furthermore, the scale scores are truncated with the lowest obtainable scale score (LOSS) of 650 and highest obtainable scale score (HOSS) of 850.

Appendix D presents the TCCs, estimated CSEM curves, and estimated information (INF) curves for each content area and grade. The curves are based on IRT parameters from a prior operational or field test administration. The curves in each figure are for the regular core and accommodated forms and are reported on the theta scale. The vertical dotted lines indicate the performance level cuts on the theta scale.

11.3.5. Scaling Constants

Table 11.4, Table 11.5, and Table 11.6 present the *A* and *B* values resulting and the theta values associated with the performance level scale score cut scores.

Table 11.4. Cut Scores and Scaling Constants—ELA/L

Grade	Cut	Theta	Scale Score	A	B
3	Level 2	-0.9648	700	36.7227	735.4297
	Level 3	-0.2840	725		
	Level 4	0.3968	750		
	Level 5	2.0360	810		
4	Level 2	-1.3004	700	31.5462	741.0214
	Level 3	-0.5079	725		
	Level 4	0.2846	750		
	Level 5	1.5578	790		
5	Level 2	-1.3411	700	29.4580	739.5050
	Level 3	-0.4924	725		
	Level 4	0.3563	750		
	Level 5	2.0224	799		
6	Level 2	-1.3656	700	28.3160	738.6673
	Level 3	-0.4827	725		
	Level 4	0.4002	750		
	Level 5	1.8133	790		
7	Level 2	-1.2488	700	33.9161	742.3542
	Level 3	-0.5117	725		
	Level 4	0.2254	750		
	Level 5	1.2614	785		
8	Level 2	-1.2730	700	34.1183	743.4330
	Level 3	-0.5402	725		
	Level 4	0.1925	750		
	Level 5	1.4696	794		

Table 11.5. Cut Scores and Scaling Constants—Reading and Writing

	Reading		Writing	
Grade	A	B	A	B
3	14.6891	44.1719	7.3445	32.0859
4	12.6184	46.4086	6.3093	33.2043
5	11.7832	45.8019	5.8916	32.9010
6	11.3264	45.4669	5.6632	32.7335
7	13.5664	46.9416	6.7832	33.4708
8	13.6472	47.3732	6.8237	33.6866

Table 11.6. Cut Scores and Scaling Constants—Mathematics

Grade	Cut	Theta	Scale Score	A	B
3	Level 2	-1.4141	700	32.1135	745.4119
	Level 3	-0.6356	725		
	Level 4	0.1429	750		
	Level 5	1.3931	790		
4	Level 2	-1.3840	700	29.9167	741.4049
	Level 3	-0.5484	725		
	Level 4	0.2873	750		
	Level 5	1.8323	796		
5	Level 2	-1.4571	700	29.0301	742.2997
	Level 3	-0.5959	725		
	Level 4	0.2653	750		
	Level 5	1.6262	790		
6	Level 2	-1.3829	700	28.1465	738.9252
	Level 3	-0.4948	725		
	Level 4	0.3935	750		
	Level 5	1.7567	788		
7	Level 2	-1.4464	700	25.1033	736.3102
	Level 3	-0.4505	725		
	Level 4	0.5453	750		
	Level 5	1.9919	786		
8	Level 2	-0.8851	700	32.9505	729.1640
	Level 3	-0.1264	725		
	Level 4	0.6323	750		
	Level 5	2.1896	801		

Section 12: Quality Control Procedures

Quality control in a testing program is a comprehensive and ongoing process. This section describes procedures put into place to monitor the quality of the item bank, test form, and ancillary material development. The quality checks for scanning, image editing, scoring, and data screening during psychometric analyses are also outlined. Additional quality information can be found in the Program Quality Plan document.

12.1. Quality Control of the Item Bank

The IAR item bank consists of test passages and items, their metadata, and status (e.g., operational ready, field test ready, released). The items were developed by Pearson and their partners and put in the item bank once created. Pearson's Assessment Banking for Building and Interoperability (ABBI) bank houses the passages and items, art, associated metadata, rubrics, alternate text for use on accommodated forms, and text complexity documentation. It provides an item previewer that allows items to be viewed and interacted with in the same way students see and interact with them, and it manages versioning of items with a date/time stamp. Reviewers cab vote on item acceptance and record and retain their review notes for later reconciliation and reference. Item and passage review participants conduct their review in the item banking system and also view the items as the student would, voting to edit, accept, or reject the item and record their comments in the system.

12.2. Quality Control of Test Form Development

The operational test forms were built based on targets and the established blueprints set, and items were pulled into forms based on the criteria approved in the test specifications. The forms then went through an internal review process to ensure content accuracy, completeness, style guide conformity, and tools function. Revisions were incorporated into the forms before final review and approval. The forms quality assurance was performed by Pearson's Assessment and Information Quality (AIQ) organization. AIQ completed a comprehensive review of all online forms for the administration cycle. This group is part of Pearson's larger Organizational Quality group and operates exclusively to validate form operability. The group verifies that the functionality of every online form is working to specifications. The overall functionality and maneuverability of each form is checked, and the behavior of each item within the form is verified.

The items within each form were tested to verify that they operated as expected for students. As a further aspect of the testing process, AIQ confirmed that forms were loaded correctly and that the audio was correct when compared to text. Sections and overviews were reviewed. Technology-enhanced items also were tested as an additional measure. As enumerated in the *Technology Guidelines for Assessments*, user interfaces were compatible with a range of common computer devices, operating systems, and browsers.

Pearson also performed quality control tests to verify that a standard set of responses was outputted to XML as expected after the final version of the form was approved. These responses were based on the keys provided in the test map or a standard open-ended responses string that contained a valid range of characters. As part of these tests, the test maps also were validated against the form layout and item types for correctness. Pearson conducted a multifaceted validation of all item layout, rendering, and functionality. Reviewers conducted comparisons between the approved item and the item as it appeared in the field test form or how it previously appeared; verified that tools and functions in the test delivery system, TestNav, were accurately applied; and verified that the style and layout met all requirements. Answer keys were also validated through a formal key review process.

12.3. Quality Control of Test Materials

Pearson provided high-quality materials in a timely and efficient manner to meet the test administration needs. Because most printing work was done in-house, it was possible to fully control the production environment, press schedule, and quality process for print materials. Strict security requirements were also employed to protect secure materials production. Materials were produced according to the style guide and to the detailed specifications supplied in the materials list.

Pearson Print Service operates within the sanctions of an ISO 9001:2008 Quality Management System, and practices process improvement through Lean principles and employee involvement. Raw materials (paper and ink) used for scannable forms production were manufactured exclusively for Pearson Print Service using specifications created by Pearson Print Service. Samples of ink and paper were tested by Pearson prior to use in production. Project specialists were the point of contact for incoming production.

Purchase orders and other order information were assessed against manufacturing capabilities and assigned to the optimal production methodology. Expectations, quality requirements, and cost considerations were foremost in these decisions. Prior to release for manufacture, order information was checked against specifications, technical requirements, and other communication that includes expected outcomes. Records of these checks were maintained.

Files for image creation flow through one of two file preparation functions: digital pre-press for digital print methodology, or plateroom for offset print methodology. Both the digital prepress and plateroom functions verify content, file naming, imposition, pagination, numbering stream, registration of technical components, color mapping, workflow, and file integrity. Records of these checks are created and saved.

Offset production requires printing that uses a lithographic process. Offline finishing activities are required to create books and package offset output. Digital output may flow through an inkjet digital production line or a sheet-fed toner application process in the Xpress Center. A battery of quality checks was performed in these areas. The checks included color match, correct file selection, content match to proof, litho-code to serial number synchronization, registration of technical components, ink density controlled by densitometry, inspection for print flaws, perforations, punching, pagination, scanning requirements, and any unique features specified for the order. Records of these checks and samples pulled from planned production points were maintained. Offline finishing included cutting, shrink-wrapping, folding, and collating. The collation process has three robust inline detection systems that inspected each book for the following:

- Caliper validation that detects too few or too many pages. This detector will stop the collator if an incorrect caliper reading is registered.
- An optical reader that will only accept one sheet. Two or zero sheets will result in a collator stoppage.
- The correct bar code for the signature being assembled. An incorrect or upside down signature will be rejected by the bar code scanner and will result in a collator stoppage.

Pearson's Quality Assurance department personnel inspected print output prior to collation and shipment. Quality Assurance also supported process improvement, work area documentation, audited process adherence, and established training programs for employees.

12.4. Quality Control of Scoring

12.4.1. Quality Control of Scanning

Establishing and maintaining the accuracy of scanning, editing, and imaging processes is a cornerstone of the Pearson scoring process. While the scanners are designed to perform with great precision, Pearson implements other quality assurance processes to confirm that the data captured from scan processing produces a complete and accurate map to the expected results.

Pearson pioneered optical mark reading and image scanning and continues to improve in-house scanners for this purpose. Software programs drive the capture of student demographic data and student responses from the test materials during scan processing. Routinely scheduled maintenance and adjustments to the scanner components (e.g., camera) maintain scanner calibration. Test sheets inserted into every batch test scanner accuracy and calibration. Controlled processes for developing and testing software specifications included a series of validation and verification procedures to confirm the captured data can be mapped accurately and completely to the expected results and that editing application rules are properly applied.

12.4.2. Quality Control of Image Editing

The final step in producing accurate data for scoring is the editing process. Once information from the documents was captured in the scanning process, the scan program file was executed, comparing the data captured from the student documents to the project specifications. The result of the comparison was a report (or edit listing) of documents needing corrections or validation. Image Editing Services performed the tasks necessary to correct and verify the student data prior to scoring. Using the report, editors verified that all unscanned documents were scanned, or the data were imported into the system through some other method such as flatbed scan or key entry. Documents with missing or suspect data were pulled and verified, and corrections or additional data were entered. Standard edits included

- Incorrect or double gridding
- Incorrect dates (including birth year)
- Mismatches between pre-ID label and gridded information
- Incomplete names

When all edits were resolved, corrections were incorporated into the document file containing student records. Additional quality checks were also performed, including student n-count checks to ensure that

- students were placed under the correct header,
- all sheets belonged to the appropriate document,
- documents were not scanned twice, and
- no blank documents existed.

Finally, accuracy checks were performed by checking random documents against scanned data to verify the accuracy of the scanning process. Once all corrections were made, the scan program was tested a second time to verify all data were valid. When the resulting output showed that no fields were flagged as suspect, the file was considered clean and scoring began. Once all scanning was completed, the right/wrong response data were securely handed off.

12.4.3. Quality Control of Answer Document and Data

Quality control of answer document processing and scoring involves all aspects of the scoring procedures, including key-based and rule-based machine scoring and handscoring for constructed-response items and performance tasks. Based on lessons learned from previous administrations, the following quality steps were implemented:

- Raw score validation (e.g., score key validation; evidence statement, field test nonscore; double-grid combinations; possible correct combination, if applicable; out-of-range/negative test cases)
- Matching (e.g., validation of high-confidence criteria, low-confidence criteria, cross document, external or forced matching by customer; prior to and after data updates; extract file of matched and unmatched documents)
- Demographic update tests (e.g., verification of data extract against corresponding layout; valid values for updatable fields; invalid values for updatable/nonupdatable fields; negative test for nonexisting record or empty file)

The following components were also included in the quality control process:

- XML Validation: A combination of automated validation against 100% of item XMLs and human inspection of XML from selected difficult item types or composite items
- Administration/End-to-End Data Validation: An automated generation of response data from approved test maps that have known conditions against the operational scoring systems and data generation systems to verify scoring accuracy
- Psychometric Validation: Verification of data integrity using criteria typically used in psychometric processes (e.g., statistical keychecks) and categorization of identified issues to help inform investigation by other groups
- Content Validation: An examination, by subject matter experts, of all items using a combination of automated tools to generate response and scoring data

The following quality control process for answer keys and scoring was also implemented:

- Pearson’s psychometrics team conducted empirical analyses based on preliminary data files and flagged items based on statistical criteria.
- Pearson content team reviewed the flagged items and provided feedback on the accuracy of content, answer keys, and scoring.
- Items potentially requiring changes were added to the product validation log for further investigation by other Pearson teams.
- Staff was notified of items for which keys or scoring changes were recommended.
- Illinois approved/rejected scoring changes.
- All approved scoring changes were implemented and validated prior to the generation of the data files used for psychometric processing.

12.5. Quality Control of Psychometric Processes

High-quality psychometric work for the operational administrations was necessary to provide accurate and reliable results of student performance. The psychometric analyses were all conducted according to well-defined specifications, and data cleaning rules were clearly articulated and applied consistently throughout the process. Results from all analyses underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were used by members of the team for each statistical procedure.

Quality control steps performed at different stages of the psychometric analyses including data screening, classical item analysis, and the creation of conversion tables. Data screening is an important first step to ensure quality data input for meaningful analysis. The Pearson Customer Data Quality team validated all student data files used in the operational psychometric analyses. The data validation for the student data files and item response files included the following steps:

1. Validated variables in the data file for values in acceptable ranges
2. Validated that the test form ID, unique item numbers, and item sequence on the data file were consistent with the test form values on the corresponding test map
3. Computed the composite raw score, claim raw scores, and subclaim raw scores, given the item scores in the student data file
4. Compared computed raw scores to the raw scores in the student data file
5. Compared the student item response block to the item scores
6. Flagged student records with inconsistencies for further investigation

All classical item analysis results were reviewed by Pearson psychometricians, and items flagged for unusual statistical properties were reviewed by the content team. Refer to Section 9.3 for the classical item analysis item flagging criteria.

Finally, conversion tables are used to generate reported scores for students and must be accurate. Comprehensive records were maintained on item-level decisions, and thorough checks were made to ensure that the correct items were included in the final score. Pre-equated conversion tables were developed independently by two psychometricians and completely matched. A reasonableness check was also conducted by psychometricians for each content and grade level to make sure the results were in alignment with observations during the analyses prior to conversion table creation. Refer to Section 11.3.4 for the procedure to create the conversion tables.

Section 13: Reliability

Reliability focuses on the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations due to chance. Thus, reliability measures the consistency of the scores across conditions that can be assumed to differ at random. In statistical terms, the variance in the distribution of test scores (i.e., the differences among individuals) is partly due to real differences in the knowledge, skill, or ability being tested (true variance) and partly due to random errors in the measurement process (error variance). Reliability is an estimate of the proportion of the total variance that is true variance. Reliability for the IAR assessments was evaluated based on the following analyses for both raw and scale scores:

- Internal consistency
- Standard error of measurement (SEM)
- Decision accuracy and consistency
- Inter-rater agreement (see Section 5.4)

13.1. Internal Consistency and SEM

Reliability coefficients for both raw and scale scores range from 0.0 to 1.0. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain similar scores upon repeated testing occasions if the students do not change in their level of the knowledge or skills measured by the test. The reliability estimates attempt to answer the question, “How consistent would the scores of these students be over replications of the entire testing process?” Raw score reliability estimate reported for the assessment is an internal-consistency measure derived from analysis of the consistency of the performance of individuals across items within a test. It serves as a good estimate of alternate forms reliability but does not consider form-to-form variation due to lack of test form parallelism, nor is it responsive to day-to-day variation due to, for example, the student’s state of health or the testing environment. The scale score reliability results use a modified measure of internal consistency that accounts for the conversions between raw scores and scale scores.

The SEM quantifies the amount of error in the test scores. SEM is the extent by which students’ scores tend to differ from the scores they would receive if the test were perfectly reliable. As the SEM increases, the variability of students’ observed scores is likely to increase across repeated testing. Observed scores with large SEMs pose a challenge to the valid interpretation of a single test score.

Reliability estimates are influenced by test length, test characteristics, and sample characteristics (Lord & Novick, 1968; Tavakol & Dennick, 2011; Cortina, 1993). As test length decreases and samples become smaller and more homogeneous, lower estimates of alpha are obtained (Tavakol & Dennick, 2011; Pike & Hudson, 1998). Moderate to acceptable ranges of reliability tend to exceed 0.5 (Cortina, 1993; Schmitt, 1996). Estimates lower than 0.5 may indicate a lack of internal consistency. Additional analyses investigate whether lower estimates of alpha are due to a restriction in range of the sample. In these cases, the alpha estimates are not appropriate measures of internal consistency. As a result, sample-free reliability estimates are also provided, such as scale score reliability (Kolen et al., 1996).

13.1.1. Raw Score Estimation

Coefficient alpha (Cronbach, 1951), the most used measure of reliability, is an internal consistency measure derived from analysis of the consistency of the performance of students across items within a test. It is estimated by substituting sample estimates for the parameters as follows:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right] \qquad \text{(13-1)}$$

where *n* is the number of items, σ_i^2 is the variance of scores on the *i*th item, and σ_X^2 is the variance of the total score (sum of scores on the individual items).

However, because the test forms have mixed item types (dichotomous and polytomous items), it is more appropriate to report stratified alpha (Feldt & Brennan, 1989), which is a weighted average of coefficient alphas for item sets with different maximum score points or "strata." Stratified alpha is a reliability estimate computed by dividing the test into parts (strata), computing alpha separately for each part, and using the results to estimate a reliability coefficient for the total score. Stratified alpha is used here because different parts of the test consist of different item types and may measure different skills. The formula for the stratified alpha is as follows:

$$\rho_{strata} = 1 - \frac{\sum_{h=1}^{H}\sigma_{X_h}^2(1-\alpha_h)}{\sigma_X^2} \qquad \text{(13-2)}$$

where $\sigma_{X_h}^2$ is the variance for part h of the test, σ_X^2 is the variance of the total scores, and α_h is coefficient alpha for part h of the test. Estimates of stratified alpha are computed by substituting sample estimates for the parameters in the formula. The average stratified alpha is a weighted average of the stratified alphas across the test forms. The formula for the SEM is as follows:

$$\sigma_E = \sigma_X\sqrt{1 - \rho_{XX'}} \qquad \text{(13-3)}$$

where σ_X is the standard deviation of the test raw score, and $\rho_{xx'}$ is the reliability estimated by substitution of appropriate statistics for the parameters.

13.1.2. Scale Score Estimation

Like the stratified alpha coefficients, scale score reliability coefficients range from 0.0 to 1.0. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain similar scores upon repeated testing occasions if they do not change in their level of the knowledge or skills measured by the test. Because the scale scores are computed from a total score and do not have an item-level component, a stratified alpha coefficient cannot be computed for scale scores. Instead, Kolen et al.'s (1996) method for scale score reliability was used. The general formula for a reliability coefficient,

$$\rho = 1 - \frac{\sigma^2(E)}{\sigma^2(X)'} \qquad \text{(13-4)}$$

involves the error variance,$\sigma^2(E)$, and the total score variance,$\sigma^2(X)$. Using Kolen et al.'s (1996) method, conditional raw score distributions are estimated using Lord and Wingersky's (1984) recursion formula. The conditional raw score distributions are transformed into conditional scale score distributions. Denote$X$as the raw sum score ranging from 0 to$X$, and$s$as a resulting scale score after transformation. The conditional distribution of scale scores is written as$P(X = x|\theta)$. The mean and variance,$\sigma^2[s(X)]$, of this distribution can be computed using these scores and their associated probabilities. The average error variance of the scale scores is computed as follows:

$$\sigma^2(Error_{scale}) = \int_\theta \sigma^2(s(X)|\theta)\, g(\theta)d\theta \qquad (13\text{-}5)$$

where$g(\theta)$is the ability distribution. The square root of the error variance is the conditional standard error of measurement of the scale scores.

Just as the reliability of raw scores is one minus the ratio of error variance to total variance, the reliability of scale scores is one minus the ratio of the average variance of measurement error for scale scores to the total variance of scale scores:

$$\rho_{scale} = 1 - \frac{\sigma^2(Error_{scale})}{\sigma^2[s(X)]} \qquad (13\text{-}6)$$

The Windows program POLYCSEM (Kolen, 2004) was used to estimate scale score error variance and reliability.

13.1.3. Results

Reliability results are presented at the overall, subgroup, and subclaim levels. Table 13.1 and Table 13.2 present the raw and scale score test reliability estimates for the total testing group, including the average reliability that is estimated by averaging the internal consistency estimates computed for all the individual forms of the test. The Spring 2024 administration had two forms: one online core form (Online1) and one accommodated form (ACC1) taken by a small number of students. The tables present the average reliability across both forms and by form.

The average raw score reliability estimates for ELA/L range from 0.87 to 0.90, and the average raw score SEM is consistently between 3 and 4 points. The average reliability estimates for mathematics range from 0.88 to 0.91, and the raw score SEM was consistently about 3 points. Average scale score reliabilities for ELA/L range from 0.86 to 0.91, and the average SEM ranges from 10.74 to 14.33. Average scale score reliability estimates range from 0.87 to 0.91 for mathematics, and the average scale score SEM ranges from 9.30 to 12.57.

Table 13.1. Summary of Raw Score Test Reliability for Total Group

		Max. Possible	Avg. Raw	Avg.	ACC1		Online1	
Assessment	#Forms	Score	Score SEM	Reliability	N	Alpha	N	Alpha
ELA/L 3	2	54	3.06	0.87	812	0.85	128,697	0.90
ELA/L 4	2	67	3.70	0.89	856	0.87	128,699	0.91
ELA/L 5	2	67	3.83	0.88	804	0.86	128,296	0.90
ELA/L 6	2	74	3.92	0.90	745	0.88	129,822	0.92
ELA/L 7	2	74	3.90	0.90	565	0.88	132,798	0.92
ELA/L 8	2	70	4.17	0.89	496	0.88	134,182	0.90
Mathematics 3	2	52	2.98	0.91	656	0.90	129,200	0.92
Mathematics 4	2	52	3.18	0.90	577	0.89	129,290	0.92
Mathematics 5	2	52	3.06	0.90	379	0.88	129,009	0.92
Mathematics 6	2	50	3.08	0.90	324	0.89	130,561	0.92
Mathematics 7	2	52	2.93	0.91	306	0.89	133,568	0.92
Mathematics 8	2	50	2.70	0.88	234	0.84	134,969	0.91

Table 13.2. Summary of Scale Score Test Reliability for Total Group

Assessment	#Forms	Avg. Scale Score SEM	Avg. Reliability	ACC1 Reliability	Online1 Reliability
ELA/L 3	2	14.33	0.87	0.86	0.87
ELA/L 4	2	12.37	0.87	0.86	0.88
ELA/L 5	2	11.81	0.86	0.86	0.86
ELA/L 6	2	10.77	0.88	0.88	0.88
ELA/L 7	2	10.74	0.91	0.90	0.92
ELA/L 8	2	11.84	0.89	0.90	0.89
Mathematics 3	2	10.71	0.90	0.90	0.91
Mathematics 4	2	9.30	0.91	0.91	0.91
Mathematics 5	2	10.01	0.90	0.90	0.89
Mathematics 6	2	9.51	0.90	0.90	0.91
Mathematics 7	2	9.53	0.88	0.88	0.88
Mathematics 8	2	12.57	0.87	0.87	0.87

Appendix E presents the raw score reliability and SEM for various demographic subgroups with sufficiently large sample sizes (i.e., 100 or more for a given test form). Reliability estimates depend on score variance, and subgroups with smaller variance are likely to have lower reliability estimates than the total group. Overall, the reliability estimates for the subgroups of interest were close to the reliability estimates of the total group.

Table 13.3 and Table 13.4 present the reliability estimates for each major claim and subclaim. Subclaims with greater numbers of points tend to have greater reliability estimates. Across grades, the average reliabilities range from 0.47 to 0.85 for the Reading claim and 0.69 to 0.86 for the Writing claim. The average reliabilities across all subclaims for mathematics range from 0.43 to 0.83 across grades.

Table 13.3. Average Reliability Estimates by Subclaim—ELA/L

	Reading: Total		Reading: Literature		Reading: Information		Reading: Vocabulary		Writing: Total		Writing Expression		Writing: Knowledge Language and Conventions	
Grade	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability
3	30–30	0.81	11–11	0.73	11–11	0.55	8–8	0.47	24–24	0.76	18–18	0.69	6–6	0.78
4	40–40	0.85	16–16	0.73	16–16	0.66	8–8	0.49	27–27	0.75	21–21	0.70	6–6	0.74
5	40–40	0.83	16–16	0.65	14–16	0.62	8–10	0.56	27–27	0.77	21–21	0.72	6–6	0.78
6	44–44	0.84	18–18	0.72	16–16	0.59	10–10	0.59	30–30	0.83	24–24	0.83	6–6	0.83
7	44–44	0.84	18–20	0.70	16–16	0.65	8–10	0.58	30–30	0.84	24–24	0.86	6–6	0.86
8	40–40	0.84	16–16	0.70	16–16	0.61	8–8	0.59	30–30	0.80	24–24	0.77	6–6	0.82

Note. RS = raw score, Avg. = average

Table 13.4. Average Reliability Estimates by Subclaim—Mathematics

	Major Content		Additional & Supporting Content		Mathematics Reasoning		Modeling Practice	
Grade	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability	RS Range	Avg. Reliability
3	20–20	0.82	10–10	0.68	10–10	0.62	12–12	0.67
4	21–21	0.83	9–9	0.71	10–10	0.65	12–12	0.43
5	20–20	0.76	10–10	0.70	10–10	0.55	12–12	0.68
6	18–18	0.78	10–10	0.57	10–10	0.74	12–12	0.62
7	20–20	0.78	10–10	0.54	10–10	0.70	12–12	0.75
8	19–20	0.75	8–9	0.53	10–10	0.61	12–12	0.63

Note. RS = raw score, Avg. = average

13.2. Decision Accuracy and Consistency

The reliability of the classifications for the students was calculated using the computer program BB-CLASS (Brennan, 2004), which operationalizes a statistical method developed by Livingston and Lewis (1993, 1995). As Livingston and Lewis (1993, 1995) explain, this method uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient) to estimate two kinds of statistics, decision accuracy and decision consistency. Decision accuracy refers to the extent to which the classifications of students based on their scores on the test form agree with the classifications made based on the classifications that would be made if the test scores were perfectly reliable. Decision consistency refers to the agreement between these classifications based on two nonoverlapping, equally difficult forms of the test hypothetically taken by the same group of students. The idea of decision consistency is conceptual as in real world, students rarely take more than one test form under exactly the same testing condition, and BB-CLASS computes the decision consistency by comparing the actual observed score distribution with observed score distribution based on a hypothetical test form predicted from the model.

Decision consistency values are always lower than the corresponding decision accuracy values because both classifications are subject to measurement error in decision consistency. In decision accuracy, only one of the classifications is based on a score that contains an error(s). It is not possible to know which students were accurately classified, but it is possible to estimate the proportion of the students who were accurately classified. Similarly, it is not possible to know which students would be consistently classified if they were retested with another form, but it is possible to estimate the proportion of the students who would be consistently classified.

Table 13.5 presents decision accuracy and consistency results based on the summative scale. “Exact Level” presents the estimates of the indices based on classifications of students into one of the five performance levels, and “Level 4 or Higher vs. 3 or Lower” presents the estimates of the indices based on classifications of students as being either in one of the upper two levels (Levels 4 and 5) or in one of the lower three levels (Levels 1, 2, and 3). Level 4 is considered the college and career readiness standard on the IAR assessments. These results are specific to the Illinois student population and should not be compared to previous PARCC results that had much higher sample sizes.

Table 13.5. Decision Accuracy and Consistency Summary

	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
Assessment	Exact Level	Level 4 or Higher vs. 3 or Lower	Exact Level	Level 4 or Higher vs. 3 or Lower
ELA/L 3	0.70	0.90	0.61	0.86
ELA/L 4	0.69	0.90	0.59	0.86
ELA/L 5	0.72	0.90	0.62	0.85
ELA/L 6	0.73	0.90	0.63	0.86
ELA/L 7	0.71	0.90	0.61	0.86
ELA/L 8	0.70	0.90	0.60	0.86
Mathematics 3	0.72	0.92	0.63	0.88
Mathematics 4	0.73	0.91	0.63	0.88
Mathematics 5	0.72	0.92	0.62	0.89
Mathematics 6	0.73	0.92	0.63	0.89
Mathematics 7	0.75	0.92	0.66	0.89
Mathematics 8	0.69	0.92	0.59	0.88

Appendix F provides more detailed information about the accuracy and the consistency of the classification of students into performance levels by grade. Each cell in the 5-by-5 tables shows the estimated proportion of students who would be classified into a particular combination of performance levels. The sum of the five bold values on the diagonal is approximately equal to the level of decision accuracy or consistency presented in Table 13.5. For "Level 4 and Higher vs. 3 and Lower" in the summary tables, the sum of the shaded values in Appendix F is approximately equal to the level of decision accuracy or consistency presented in Table 13.5. The sums based on values in Appendix F may not match exactly to the values in the summary tables due to truncation and rounding.

Section 14: Validity

As stated in 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 for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p. 11). The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, and psychometric characteristics. This chapter summarizes the evidence based on test content and the internal structure of the tests.

14.1. Evidence Based on Test Content

Content validity addresses whether the test adequately samples the relevant material it purports to cover. Evidence based on content of achievement tests is supported by the degree of correspondence between test items and content standards. The degree to which the test measures what it claims to measure is known as construct validity. The summative assessments adhere to the principles of evidence-centered design, in which the standards to be measured (the Illinois Learning Standards) are identified, and the performance a student needs to achieve to meet those standards is delineated in the evidence statements. Test items are reviewed for adherence to universal design principles, which maximize the participation of the widest possible range of students. Accommodations were also made available based on individual student need documented in the student's approved Individualized Education Program (IEP), 504 Plan, or an EL Plan.

Content is also aligned through the articulation of performance in the performance level descriptors (PLDs). At the policy level, the PLDs include policy claims about the educational achievement of students who attain a particular performance level, and a broad description of the grade-level knowledge, skills, and practices students performing at a particular achievement level are able to demonstrate. Those policy descriptions are the foundation for the content area- and grade-specific PLDs, which, along with the evidence frameworks, guide the development of the operational items and tasks.

The college- and career-ready determinations in ELA/L and mathematics describe the academic knowledge, skills, and practices students must demonstrate to show readiness for success in entry-level, credit-bearing college courses and relevant technical courses. The states and agencies determined that this level means graduating from high school and having at least a 75% likelihood of earning a grade of "C" or better in credit-bearing courses without the need for remedial coursework. After reviewing the standards and test design, the PARCC Governing Board (made up of the K–12 education chiefs in participating states and agencies) in conjunction with the Advisory Committee on College Readiness (composed of higher education chiefs in the participating states or agencies), determined that students who achieve at Levels 4 and 5 on the final high school assessments are likely to have acquired the skills and knowledge to meet the definition of college and career readiness. To validate the determinations, a postsecondary educator judgment study and a benchmark study of the SAT, ACT, National Assessment of Educational Progress, Trends in International Mathematics and Science Study, Programme of International Student Assessment, and Progress in International Reading Literacy Study tests were conducted (McClarty et al., 2015).

Gathering construct validity evidence for the assessments is embedded in the process by which the test content is developed and validated. See Section 2 for an overview of the content development process. The items and tasks were then field tested prior to their operational use. During the initial field test administration in 2014, PARCC participating states and agencies collected feedback from students, test administrators, test coordinators, and classroom teachers on their experience with the assessments, including the quality of test items and student experience. Information pertaining to this process can be found at https://resources.newmeridiancorp.org/research/. The feedback from that survey was used to inform test directions, test timing, and the function of online task interactions. Performance data from the field test also informed the future development of additional items and tasks.

Finally, an important consideration when constructing test forms is recognition of items that may introduce construct-irrelevant variance. Such items should not be included on test forms to help ensure fairness to all subgroups of students.

14.2. Evidence Based on Internal Structure

Internal structure refers to “the degree to which the relationships among test items and test components conform to the construct on which the proposed test interpretations are based” (AERA et al., 2014, p. 16). If an item has poor internal structure, it may not be measuring the intended construct accurately, which can lead to invalid or unreliable results. Evidence for the summative assessments includes (1) intercorrelations between an assessment’s subclaims to examine how they relate to each other and verify the unidimensionality of the assessment (i.e., measuring only one construct); (2) reliability correlation coefficients that measure a test’s internal consistency, or the extent to which the items in an assessment are measuring the same underlying construct; and (3) local item independence, an assumption under the IRT model that assumes any item pair is uncorrelated, conditioned on the latent trait an instrument is intended to measure (e.g., mathematics proficiency).

14.2.1. Intercorrelations

The ELA/L summative assessments have two claim scores (Reading and Writing) and five subclaim scores: Reading Literature (RL), Reading Information (RI), Reading Vocabulary (RV), Writing Written Expression (WE), and Writing Knowledge Language and Conventions (WKL). The Reading claim score is a composite of RL, RI, and RV. The Writing claim score is a composite of WE and WKL and comprises only PCR items that are the same in each subclaim. The mathematics summative tests have four subclaim scores: Major Content (MC), Mathematical Reasoning (MR), Modeling Practice (MP), and Additional and Supporting Content (ASC). These analyses were conducted between the ELA/L Reading and Writing claim scores and subclaims (RL, RI, RV, WE, and WKL) and between the mathematics subclaims.

Table 14.1 and Table 14.2 present the weighted average Pearson intercorrelations between subclaims by averaging the intercorrelations computed for all the core operational forms of each assessment. The shaded values along the diagonal are the reliabilities from Section 13.1**Error! Reference source not found.**. The average intercorrelations are provided in the lower portion of the tables, and the total sample sizes are provided in the upper portion of the tables. Results are as follows:

- The WR, WE, and WKL scores tended to be highly correlated. RL, RI, and RV (all subclaims of Reading) are moderately to highly correlated. The WR claim and the WE and WKL subclaims are also moderately correlated with RD subclaims (of RL, RI, and RV). These moderate-to-high ELA/L intercorrelations among the subclaims are sufficiently high to provide evidence that the ELA/L tests are unidimensional. The moderate intercorrelations among the subclaims and claims suggest the claims may be sufficient for individual student reporting.

- The mathematics intercorrelations are moderate. The main observable pattern in the mathematics intercorrelations is that the MC subclaim has slightly higher correlations with the ASC, MR, and MP subclaims; the intercorrelations among the ASC, MR, and MP subclaims are usually slightly lower. The mathematics intercorrelations are sufficiently high to suggest that the mathematics tests are likely to be unidimensional with some minor secondary dimensions.

Table 14.1. Average Interrelations and Reliability between Subclaims—ELA/L

Grade	Subclaim	RD	RL	RI	RV	WR	WE	WKL
3	RD	0.81	128,697	128,697	128,697	128,697	128,697	128,697
	RL	0.91	0.73	128,697	128,697	128,697	128,697	128,697
	RI	0.87	0.70	0.55	128,697	128,697	128,697	128,697
	RV	0.83	0.63	0.59	0.47	128,697	128,697	128,697
	WR	0.75	0.71	0.72	0.53	0.76	128,697	128,697
	WE	0.73	0.69	0.71	0.51	0.99	0.69	128,697
	WKL	0.69	0.65	0.65	0.51	0.89	0.80	0.78
4	RD	0.85	128,699	128,699	128,699	128,699	128,699	128,699
	RL	0.92	0.73	128,699	128,699	128,699	128,699	128,699
	RI	0.89	0.70	0.66	128,699	128,699	128,699	128,699
	RV	0.79	0.62	0.59	0.49	128,699	128,699	128,699
	WR	0.74	0.64	0.74	0.53	0.75	128,699	128,699
	WE	0.72	0.62	0.73	0.52	0.99	0.70	128,699
	WKL	0.72	0.62	0.71	0.52	0.94	0.89	0.74
5	RD	0.83	128,296	128,296	128,296	128,296	128,296	128,296
	RL	0.91	0.65	128,296	128,296	128,296	128,296	128,296
	RI	0.88	0.67	0.62	128,296	128,296	128,296	128,296
	RV	0.79	0.61	0.59	0.56	128,296	128,296	128,296
	WR	0.70	0.59	0.72	0.50	0.77	128,296	128,296
	WE	0.70	0.58	0.71	0.50	0.99	0.72	128,296
	WKL	0.68	0.57	0.68	0.49	0.95	0.91	0.78
6	RD	0.84	129,822	129,822	129,822	129,822	129,822	129,822
	RL	0.92	0.72	129,822	129,822	129,822	129,822	129,822
	RI	0.88	0.70	0.59	129,822	129,822	129,822	129,822
	RV	0.82	0.64	0.60	0.59	129,822	129,822	129,822
	WR	0.74	0.68	0.72	0.49	0.83	129,822	129,822
	WE	0.73	0.68	0.72	0.49	1.00	0.83	129,822
	WKL	0.71	0.66	0.70	0.48	0.96	0.94	0.83
7	RD	0.84	132,798	132,798	132,798	132,798	132,798	132,798
	RL	0.92	0.70	132,798	132,798	132,798	132,798	132,798
	RI	0.89	0.72	0.65	132,798	132,798	132,798	132,798
	RV	0.84	0.69	0.62	0.58	132,798	132,798	132,798
	WR	0.75	0.69	0.75	0.53	0.84	132,798	132,798
	WE	0.75	0.69	0.74	0.52	1.00	0.86	132,798
	WKL	0.74	0.68	0.74	0.53	0.97	0.94	0.86

Grade	Subclaim	RD	RL	RI	RV	WR	WE	WKL
8	RD	0.84	134,182	134,182	134,182	134,182	134,182	134,182
	RL	0.89	0.70	134,182	134,182	134,182	134,182	134,182
	RI	0.87	0.62	0.61	134,182	134,182	134,182	134,182
	RV	0.81	0.62	0.60	0.59	134,182	134,182	134,182
	WR	0.69	0.55	0.72	0.49	0.80	134,182	134,182
	WE	0.68	0.53	0.71	0.48	1.00	0.77	134,182
	WKL	0.69	0.55	0.71	0.49	0.96	0.93	0.82

Note. RD = Reading, RL = Reading Literature, RI = Reading Information, RV = Reading Vocabulary, WR = Writing, WE = Written Expression, WKL = Writing Knowledge and Conventions

Table 14.2. Average Interrelations and Reliability between Subclaims—Mathematics

Grade	Subclaim	MC	ASC	MR	MP
3	MC	0.82	129,200	129,200	129,200
	ASC	0.79	0.68	129,200	129,200
	MR	0.75	0.67	0.62	129,200
	MP	0.70	0.64	0.66	0.67
4	MC	0.83	129,290	129,290	129,290
	ASC	0.77	0.71	129,290	129,290
	MR	0.74	0.67	0.65	129,290
	MP	0.73	0.67	0.72	0.43
5	MC	0.76	129,009	129,009	129,009
	ASC	0.75	0.70	129,009	129,009
	MR	0.74	0.69	0.55	129,009
	MP	0.70	0.68	0.66	0.68
6	MC	0.78	130,561	130,561	130,561
	ASC	0.73	0.57	130,561	130,561
	MR	0.79	0.68	0.74	130,561
	MP	0.77	0.69	0.74	0.62
7	MC	0.78	133,568	133,568	133,568
	ASC	0.73	0.54	133,568	133,568
	MR	0.77	0.69	0.70	133,568
	MP	0.75	0.67	0.74	0.75
8	MC	0.75	134,969	134,969	134,969
	ASC	0.75	0.53	134,969	134,969
	MR	0.71	0.64	0.61	134,969
	MP	0.73	0.67	0.69	0.63

Note. MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, MP = Modeling Practice, n/r = not reported due to low n-count (no subclaim reliability could be calculated)

14.2.2. Reliability

Internal consistency is typically measured via correlations among the items on an assessment and provides an indication of how much the items measure the same general construct. As shown in Section 13.1, the reliability estimates computed using coefficient alpha (Cronbach, 1951) indicate an acceptable level of reliability for ELA/L and mathematics. Appendix E summarizes the test reliability for groups of interest. Overall, the reliability estimates indicate that the items within each assessment measure a similar construct.

14.2.3. Local Item Independence

Local item independence is a primary assumption of IRT that states the probability of success on one item is not influenced by performance on other items when controlling for ability level. This implies that ability or theta accounts for the associations among the observed items. Local item dependence (LID), when present, overstates the amount of information predicted by the IRT model. It can exert other undesirable psychometric effects and represents a threat to validity since other factors besides the construct of interest are present. Classical statistics are also affected when LID is present because estimates of test reliability like IRT information can be inflated (Zenisky et al., 2003). The LID issue affects the choice of item scoring in IRT calibrations. If evidence suggests these items indeed have LID, it might be preferable to sum the item scores into clusters or testlets as a method of minimizing it. However, if these items do not appear to have strong LID, retaining the scores as individual item scores in an IRT calibration is preferred because more information concerning item properties is retained.

Local item independence was evaluated in prior studies. Please refer to the previous technical report for details (New Meridian, 2023).

14.3. Evidence Based on Relationships to Other Variables

Empirical results concerning the relationships between test scores and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, and demographic characteristics of students that are expected to be related and unrelated to test performance. For example, when a test’s scores are highly correlated with scores from a different, external assessment, it provides evidence that the tests measure the same or similar construct.

The relationship of the scores across the IAR assessments were evaluated Pearson correlations between the ELA/L and the mathematics scores and between the ELA/L subclaims, as shown in Table 14.3. Students must have a valid test score from Spring 2024 for both ELA/L and mathematics at the same grade level to be included in the tables, and only correlations for pairings with total sample sizes of at least 100 are shown in the tables (blank cells indicate pairings with sample sizes less than 100). The correlation is presented in the lower triangle, and the sample size is presented in the upper triangle.

ELA/L, Reading, and Writing are low to moderately correlated with mathematics, with correlations ranging from 0.65 to 0.96. These correlations suggest that the ELA/L and mathematics tests are assessing different content. The higher intercorrelations between the ELA/L, Reading, and Writing scores suggest stronger internal relationships when compared to the correlations with mathematics.

Table 14.3. Correlations between ELA/L and Mathematics

Grade	Content Area	ELA/L	Reading	Writing	Mathematics
3	ELA/L	–	128,697	128,697	128,697
	Reading	0.96	–	128,697	128,697
	Writing	0.91	0.75	–	128,697
	Mathematics	0.75	0.75	0.65	–
4	ELA/L	–	128,699	128,699	128,699
	Reading	0.95	–	128,699	128,699
	Writing	0.91	0.74	–	128,699
	Mathematics	0.79	0.78	0.68	–
5	ELA/L	–	128,296	128,296	128,296
	Reading	0.94	–	128,296	128,296
	Writing	0.90	0.71	–	128,296
	Mathematics	0.74	0.72	0.65	–
6	ELA/L	–	129,822	129,822	129,822
	Reading	0.95	–	129,822	129,822
	Writing	0.91	0.74	–	129,822
	Mathematics	0.76	0.75	0.65	–
7	ELA/L	–	132,798	132,798	132,798
	Reading	0.95	–	132,798	132,798
	Writing	0.92	0.75	–	132,798
	Mathematics	0.77	0.76	0.67	–
8	ELA/L	–	134,182	134,182	134,182
	Reading	0.93	–	134,182	134,182
	Writing	0.91	0.69	–	134,182
	Mathematics	0.73	0.70	0.65	–

14.4. Evidence from Special Studies

Several research studies have been conducted to provide additional validity evidence for the assessment's goals of assessing more rigorous academic expectations, helping to prepare students for college and careers, and providing information back to teachers and parents about their students' progress toward college and career readiness:

- Content alignment studies (Doorey & Polikoff, 2016; Schultz et al., 2017)
- Benchmarking study (McClarty et al., 2015)
- Mode and device comparability studies
- Alternate blueprint study

14.4.1. Content Alignment Studies

In 2016, the grades 5 and 8 assessments were evaluated by the Fordham Institute to determine how well the assessments were aligned to the CCSS (Doorey & Polikoff, 2016). To conduct the study, content experts judged how well the items aligned to the CCSS, the depth of knowledge of the items, and the accessibility of the items to all students, including SWDs and ELs and students with disabilities. The content experts reviewing the assessments were required to be familiar with the CCSS but could not be employed by participating organizations or be the writers of the CCSS. Therefore, an effort was made to eliminate any potential conflicts of interest.

To conduct the study, individual content experts reviewed and rated each item, followed by the group of content experts reaching consensus on the final ratings for the content alignment, depth, and accessibility to all students. The content experts also provided an explanation of their ratings, which were then used by the full group to provide narrative comments regarding the overall ratings and to provide feedback and recommendation about the assessments.

Each assessment was rated as Excellent Match for ELA/L content and depth and Good Match for mathematics content and depth for grades 5 and 8, although the content study did note some weaknesses and strengths of the assessments. For example, the ELA/L assessments include complex texts, a range of cognitive demands, and have a variety of item types and “require close reading, assess writing to sources, research, and inquiry, and emphasize vocabulary and language skills.” A weakness of the ELA/L assessments is the lack of a listening and speaking component, and they could be enhanced by the inclusion of a research task that requires the use of two or more sources of information.

A strength for mathematics is that the assessments are aligned to the major work for each grade level. While the grade 5 assessment includes a range of cognitive demand items, the grade 8 assessment includes several higher-demand items and may not fully assess the standards at the lowest level of cognitive demand. It was suggested that the grade 5 assessment could include more focus on the major work and the grade 8 assessment could include items at the lowest cognitive demand level. The reviewers also noted that some of the mathematics items should be reviewed for editorial and mathematical accuracy.

In 2017, HumRRO conducted a study to evaluate the quality and alignment of ELA/L and mathematics assessments for grades 3, 4, 6, and 7 (Schultz et al., 2017) following a similar methodology as the 2016 study. An item’s cognitive complexity was defined a measure of the rigor of an individual item based on the amount of text a student must process from the corresponding passage to answer the item correctly, the way in which students are expected to interact with the item’s functionality, and the linguistic demands and reading load that exists within the components of the item itself. Reviewers determined the extent to which items were aligned to the CCSS, using “fully,” “partially,” or “not aligned” as the rating categories. Ratings were averaged to determine overall alignment. For ELA/L, 99.6% of grades 3 and 4 items, 95.5% of grade 6 items, and 94.6% of grade 7 items were fully aligned. For mathematics, 92.0% of grade 3 items, 91.1% of grade 4 items, 83.1% of grade 6 items, and 94.0% of grade 7 items were fully aligned. Most items that did not fall into fully aligned were considered partially aligned to the standards.

The CCSS are designed to be measured by multiple items, so items that aligned to multiple CCSS received a partially aligned rating. The overall item-to-CCSS alignment was captured by a holistic alignment rating that indicated if an item captured the identified standards as a set. Holistic ratings (either yes or no) were found by averaging review ratings across clusters for items that included more than one standard. For ELA/L, for all four grades, at least 93% of items had a holistic alignment rating of yes to indicate that the identified standards captured the skills or knowledge required. For mathematics, grade 6 had the lowest percentage for the holistic alignment rating of yes (84.8%), and grade 7 had the highest (96.3%). Overall, the alignment study suggests that the identified CCSS capture the knowledge and skills required in the items.

In addition to the alignment study, HumRRO also evaluated the CCSSO criteria for content and depth for ELA/L and mathematics grades 3, 4, 6, and 7 (Schultz et al., 2017). ELA/L content has five criteria: close reading, writing, vocabulary and language skills, research and inquiry, and speaking and listening. Reviewers rated the content as Excellent, Good, Limited/Uneven, or Weak Match. For grades 3, 4, 6, and 7, the ELA/L assessments received a composite rating of Excellent Match for assessing the content needed for college and career readiness. ELA/L depth has four criteria: text quality and types, complexity of texts, cognitive demand, and high-quality items and item variety. All grades received a composite rating of Good Match for depth. For mathematics content, the composite rating is based on two criteria: focus and concepts, procedures and applications. Grades 3, 4, and 6 received a composite content rating of Good Match, and grade 7 received a composite content rating of Excellent Match. The mathematics composite depth rating is based on three criteria: connecting practice to content, cognitive demand, and high-quality items and item variety. All grades were rated as Excellent Match at assessing the depth needed to successfully meet college and career readiness.

Finally, the 2017 HumRRO study looked at cognitive complexity of the ELA/L and mathematics items at grades 3, 4, 6, and 7 (Schultz et al., 2017). Reviewers indicated their agreement with the intended cognitive complexity ratings of low, medium, or high. The results indicated that the reviewers generally agreed with the distribution of complexity levels. There were differences in agreements in ELA/L language cluster and a few exceptions to agreement in mathematics, particularly at grade 6, where there was disagreement in the ratings at the medium complexity level for two domains and the high complexity level for one domain. Grade 7 had agreement across low, medium, and high in all domains.

14.4.2. Benchmarking Study

The purpose of the benchmarking study was to provide information that would inform the standard setting process (McClarty et al., 2015). An evidence-based standard setting approach (EBSS; McClarty et al., 2013) was used to establish the performance levels. In EBSS, the threshold scores for performance levels are set based on a combination of empirical research evidence and expert judgment. This benchmarking study provided one source of empirical evidence to inform the college- and career-readiness performance level (i.e., Level 4). The study findings were provided to a pre-policy standard setting committee who suggested a reasonable range for the percentage of students meeting or exceeding the Level 4 threshold score and therefore considered college and career ready.

For the benchmarking study, external information was analyzed to provide information about the Level 4 cut scores for the grades 4 and 8 ELA/L and mathematics assessments. The assessments and Level 4 expectations were compared with comparable assessments and expectations for the Programme of International Student Assessment, Trends in International Mathematics and Science Study, Progress in International Reading Literacy Study, National Assessment of Educational Progress, ACT, SAT, the Michigan Merit Exam, and the Virginia end-of-course exams. For each external assessment, the best-matched performance level was determined and the percentage of students reaching that level across the nation and in the PARCC participating states and agencies was determined. Across grades and subjects, the data indicated that 25% to 50% of students were college and career ready or on track to readiness based on the Level 4 expectations.

14.4.3. Mode and Device Comparability Studies

A two-pronged study consisted of a mode comparability analysis and a device comparability analysis. The mode comparability analysis compared scores from the paper and online administrations, and the device comparability analysis compared the online scores from tests administered using a tablet and tests administered from any other type of electronic administration where a tablet was not present (i.e., laptops, desktops, Chromebooks).

The goal of this study was threefold: (a) to investigate whether test items were of similar difficulty across the levels of conditions for each analysis (i.e., paper or online for the mode comparability analysis and tablet and non-tablet for the device comparability analysis), (b) to determine whether the psychometric properties of test scores were similar across the levels of conditions for each analysis, and (c) to determine whether overall test performance was similar across the levels of conditions for each analysis. This study examined performance on 12 assessments, split evenly between mathematics and ELA/L. Students were matched on demographic variables and on the score from the summative assessment in the same content area in the prior year, creating comparable samples that allowed for an unbiased comparison of performance across different conditions.

The mode comparability analysis results were mixed and found to be consistent with prior research. The item means suggested that items were of similar difficulty on the paper and online modes. Only two items were flagged for mode effects, both of which were on the mathematics assessments. C-level DIF was present in both analyses. All the items flagged for C-level DIF in the mathematics assessments favored the online students, whereas most items flagged for C-level DIF in the ELA/L assessments favored the paper students. None of the test forms were flagged for mode effects with respect to test reliability. The test-level adjustment analysis and the change of the PBT students' performance levels after the adjustment constants were applied to the paper students' scores indicated that more scale scores were adjusted downward than were adjusted upward on the PBT test form for each assessment except grades 5 and 7 mathematics. However, all adjustments were less than the minimum standard error of theta. Therefore, the adjustments are within measurement precision for each assessment.

The results of the device comparability study revealed consistent evidence supporting the comparability between the tablet condition (TC) and the non-tablet condition (NTC). The item means suggested that items were similarly difficult for the TC and NTC, and none of the items were flagged for device effects. The DIF analysis revealed that none of the items had C-level DIF. Consistent with the findings at the item level, an examination of test reliability indicated that the TC and NTC test forms were similarly reliable and that none of the test forms were flagged for device effects. Furthermore, the test-level adjustment analysis and the change of the students' performance levels after the adjustment constants were applied did not indicate evidence of device effects.

The generalizability of the findings from this study may be limited due to the small sample size of both the PBT students (for mode comparability) and the tablet students (for device comparability) at the high school grades. However, high-quality matching supports the internal validity of this study's findings. For mode and device comparability, few to no items were flagged for mode or device effects, the psychometric properties of test scores were similar across assessment conditions, and any adjustments to student performance for the PBT or tablet condition were within measurement precision.

14.4.4. Alternate Blueprint Study

New Meridian, in coordination with multiple states and vendors, developed an alternate form of the summative assessment to meet the provision for shorter testing times requested by several states. The shorter version of the blueprint is referred to as the alternate assessment and the original blueprint is referred to as the original assessment. Research conducted using 2017 (Boyd et al., 2018) and 2018 (Minchen et al., 2018) student data evaluated the effects of removing items from the original assessments to determine if scores arising from the two versions would be comparable. Research was conducted in several steps: (a) subject matter experts identified item subsets from the original forms that maintained the integrity of the assessment and were approximately 65% to 80% of the original test length; (b) students were rescored on the item subsets, producing a set of hypothetical scores, as if the students had only taken the subset of items; and (c) a series of analyses were conducted.

Through extensive research, the alternate blueprint was available in Spring 2019 in addition to the original blueprint with the option to administer either blueprint at the state or agency level. Because some states administered the alternate blueprint and some states administered the original blueprint, this study evaluated the comparability between the two blueprints with respect to scale score comparability and performance level comparability.

The goal was to determine additional evidence to support scale score comparability and performance level comparability according to the guidelines in the *Quality Testing Standards* (Center for Assessment, 2018). Scale score comparability is defined by the Center for Assessment (2018) as follows: *If a student taking the alternate assessments with New Meridian content took the original assessment, would the student obtain a similar scale score*? Performance level comparability is defined by the Center for Assessment (2018) as follows: *If a student taking the alternate assessment with New Meridian content took the original assessment, would the student receive a similar designation in terms of college and career readiness or Level 4 on the original blueprint*? For the Spring 2019 assessments, the mathematics items on the alternate forms also appeared on the corresponding original forms, whereas a small number of ELA/L items were unique to the alternate forms. The scale scores were reported on the same scale regardless of the form and used the same performance level cut scores.

Three sets of analyses were conducted. Most of the analyses were conducted on a set of matched samples from the 2019 alternate and original forms, allowing for direct comparisons of assessment characteristics and outcomes to be made. Such samples were obtained through coarsened exact matching (Iacus et al., 2012), which used demographic information and prior achievement scores, where possible. Prior achievement scores were grouped into bands within each performance level, and students taking the alternate forms were matched with students who took the original forms who had identical information on all demographic and prior achievement variables.

Table 14.4 presents the prior assessments used in the matching process, and Table 14.5 presents the sample sizes before and after the matching process. For grade 3, only demographic information is used in the matching process due to the lack of prior assessment data.

Table 14.4. 2019 Alternate Blueprint Study: Prior Grades used in Matching

Content Area	Current Grade	Prior Grade	Prior Test Year
ELA/L	Grade 3	N/A	N/A
	Grade 4	Grade 3	2018
	Grade 5	Grade 4	2018
	Grade 6	Grade 5	2018
	Grade 7	Grade 6	2018
	Grade 8	Grade 7	2018
Mathematics	Grade 3	N/A	N/A
	Grade 4	Grade 3	2018
	Grade 5	Grade 4	2018
	Grade 6	Grade 5	2018
	Grade 7	Grade 6	2018
	Grade 8	Grade 7	2018

Table 14.5. 2019 Alternate Blueprint Study: Matching Sample Size Results

		Unmatched		Matched	
Assessment	Form	#Forms	Original #Forms	#Forms	Original #Forms
ELA/L 3	1	105,482	32,034	31,481	31,481
	2	105,309	31,861	31,272	31,272
ELA/L 4	1	105,826	28,153	27,695	27,695
	2	126,875	34,071	33,444	33,444
ELA/L 5	1	136,148	36,313	35,742	35,742
	2	101,869	27,272	26,721	26,721
ELA/L 6	1	119,838	31,031	30,667	30,667
	2	120,218	30,802	30,506	30,506
ELA/L 7	1	116,933	29,877	29,544	29,544
	2	117,757	29,835	29,593	29,593
ELA/L 8	1	118,198	29,638	29,312	29,312
	2	119,059	29,248	28,898	28,898
Mathematics 3	1	88,858	26,531	25,970	25,970
	2	88,919	26,595	25,987	25,987
Mathematics 4	1	87,291	25,941	25,070	25,070
	2	87,488	26,192	25,207	25,207
Mathematics 5	1	91,136	27,333	26,377	26,377
	2	91,739	27,611	26,754	26,754
Mathematics 6	1	95,174	28,514	27,677	27,677
	2	94,800	28,342	27,665	27,665
Mathematics 7	1	93,777	24,547	23,855	23,855
	2	93,265	24,141	23,485	23,485
Mathematics 8	1	83,289	15,293	14,962	14,962
	2	76,135	13,973	13,695	13,695

The remaining analyses were conducted on assessment data from 2018 and 2019 rather than the matched samples. The second set of analyses was conducted at the grade level, using all available data from both 2018 and 2019, examining grade-level statistics over the course of two years, ensuring state participation was similar within each grade for both years. Finally, the last set of analyses used two-year student cohorts examining students' scores over two years. Only students who completed assessments in both 2018 and 2019 were included, so grade 3 student data from 2019 were not included. The following analyses were conducted, which demonstrated that there appears to be broad comparability between the alternate and original scale scores and performance levels and that the alternate forms have less measurement precision than the original forms:

- Scale score comparability: item-level analysis (p-values, polyserial correlations, and DIF)
- Scale score comparability: test-level analysis (analyzing reliability, scale score distributions, ELA/L claim score distributions, and subclaim distributions)
- Scale score comparability: longitudinal analysis
- Performance level comparability: test-level analysis (performance level distributions)
- Performance level comparability: classification analyses
- Performance level comparability: longitudinal analysis

14.5. Evidence Based on Response Processes

Additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that students are using the intended response processes when responding to the items in a test (AERA et al., 2014). This type of evidence may be gathered from interacting with students to understand what processes underlie their item responses. Evidence may also be derived from feedback provided by test administrators and teachers involved in the administration of the test and scorers involved in the scoring of constructed-response items. Evidence may also be gathered by evaluating the correct and incorrect responses to short constructed-response items (e.g., items requiring a few words to respond) or by evaluating the response patterns to multi-part items.

Several studies have been conducted to investigate the quality of the items, tasks, and stimuli, focusing on whether students interact with items/tasks as intended, whether they were given enough time to complete the assessments, and the degree to which scoring rubrics allow accurate and reliable scoring. Accessibility for SWDs and ELs was also examined based on students' understanding of the format of the assessments and the use of technology.

The first two studies (Brandt et al., 2015a; Brandt et al., 2015b) focused on evaluating the usability of the tool itself both in the general population and among students with low vision and fine motor impairment disabilities. During these studies, detailed information regarding the functionality of the tool was collected, and it was determined that the items should be tested operationally. The third and fourth studies (Minchen et al., 2018b; Steedle & LaSalle, 2016) involved evaluating the effect of the tool in the context of the operational assessments. The third study was conducted in grade 3, and the fourth study was conducted in grades 4 and 5. To evaluate the drawing tool in context, a set of items was studied by field testing the items with and without the drawing tool. The drawing tool version of each item was randomly assigned to students so that comparisons could be made. The goal was to explore the impact of the drawing tool on item performance. In general, the results showed that the drawing tool usually did not have a significant impact on performance or item statistics. However, items that included access to the drawing tool did show longer response times for grades 4 and 5, prompting a limitation to be placed on the number of drawing tool items in each unit.

14.6. Evidence Based on the Consequences to Testing

The consequences of testing should also be investigated to support the validity evidence for the use of the summative assessments as tests are usually administered “with the expectation that some benefit will be realized from the intended use of the scores” (AERA et al., 2014). When this is the case, evidence that the expected benefits accrue will provide support for the intended use of the scores. Evidence of the consequence of testing will also accrue with the continued implementation of the Illinois Learning Standards and the continued administration of the assessments.

14.7. Summary

The goal of providing validity evidence is to demonstrate that the assessment is accurately measuring the intended construct. The item development process involved educators, assessment experts, and bias and sensitivity experts in review of text, items, and tasks for accuracy, appropriateness, and freedom from bias. Several studies were conducted during the item development process to evaluate the item development process (e.g., technological functionalities, answer time required, and student experiences). Items were then field tested prior to the initial operational administration, and data and feedback from students, test administrators, and classroom teachers were used to improve the operational administration of the items and to inform future item development. The multiple item and form reviews conducted by educators and studies to evaluate item administration help to ensure the integrity of the assessments.

Psychometric analyses further provided evidence that the assessments measure what is intended. For example, the intercorrelations of the subclaims and the reliability analyses indicate that the summative assessments are both unidimensional, and the correlations between ELA/L and mathematics indicate that the two assessments are measuring different content. Several studies have also been conducted, including the content alignment studies, the benchmarking study conducted in support of the standard setting meeting, and the mode and device comparability studies.

In addition to the validity information presented in this section of the technical report, other information in support of the uses and interpretations of the scores appear in the following sections:

- Section 8 presents information regarding student characteristics and test results for the spring administration.
- Section 9 provides information concerning the test characteristics based on classical test theory.
- Section 10 provides information regarding the DIF analyses.
- Section 13 provides information on the test reliability (total test score and for subclaims).

Section 15: Student Growth Measures

Student growth percentiles (SGPs) are normative measures of annual progress that are useful in answering questions like “How does my academic progress compare with the academic progress of my peers?” In contrast to criterion-referenced measures of growth that describe academic growth toward a particular goal, norm-referenced measures of growth describe students’ growth relative to that of students who performed similarly in the past (Betebenner, 2009). SGPs measure individual student progress by tracking student scores from one year to the next and compare a student’s performance to that of their academic peers, defined as students in the norm group who took the same assessment as the student in prior years and achieved a similar score.

The SGP describes a student’s location in the distribution of current test scores for all students who performed similarly in the past and indicates the percentage of academic peers above whom the student scored. With a range of 1 to 99, higher numbers represent higher growth and lower numbers represent lower growth. For example, an SGP of 60 on grade 7 ELA/L means that the student scored better than 60% of the students who took the grade 7 ELA/L assessment in Spring 2019 *and* who had achieved a similar score as this student on the grade 6 ELA/L assessment in Spring 2018 and the grade 5 ELA/L assessment in Spring 2017.[3] An SGP of 50 represents typical (median) student growth.

15.1. Norm Groups

The norm groups consisted of students with the same prior scores based on grade progressions (academic peers). In the SGP analysis for the Spring 2024 administration, SGPs were based on up to one year of prior test scores from the Spring 2023 administration, as shown in Table 15.1 that presents the grade progressions required for SGPs based on one prior test score for both ELA/L and mathematics. Students who did not have a previous test score, which included any new students and all grade 3 students, did not receive an SGP. The sample size threshold of conducting SGP analysis was 1,000, so it was not conducted for the high school assessments due to low sample sizes.

Table 15.1. SGP Grade-Level Progressions for One-Year Prior Scores

One Prior Year Test Score	Current Year Test Score
N/A	Grade 3*
Grade 3	Grade 4
Grade 4	Grade 5
Grade 5	Grade 6
Grade 6	Grade 7
Grade 7	Grade 8

*SGP was not calculated for grade 3 because there are no prior scores.

[3] Because regression modeling is used to establish the relationship between prior and current scores, the SGP is for students with the exact same prior scores. This can lead to confusion among nontechnical stakeholders who often ask, “How many students are there with exactly the same prior scores?” To avoid explaining regression to nontechnical stakeholders, the “similar scores” is often used to finesse the idea of regression without mentioning it.

15.2. SGP Estimation

SGPs are calculated using quantile regression that describes the conditional distribution of the response variable with greater precision than traditional linear regression, which describes only the conditional mean (Betebenner, 2009). This application of quantile regression uses B-spline smoothing to fit a curvilinear relationship between a norm group's prior and current scores. Cubic B-spline basis functions are used when calculating SGPs to better model the heteroscedasticity, nonlinearity, and skewness in the test data.

For each group, the quantile regression fits 100 relationships (one for each percentile) between students' prior and current scores. The result is a single coefficient matrix that relates students' prior achievement to their current achievement at each percentile. The National Center for the Improvement of Educational Assessment performed the analyses using Betebenner's (2009) nonlinear quantile-regression based SGP. The analysis was done in the SGP package in R (Betebenner et al., 2017). For details on SGPs, see Betebenner's *A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories* (2011).

Betebenner's (2009) SGP model uses Koenker's (2005) quantile regression approach to estimate the conditional density associated with a student's score at administration *t* conditioned on the student's prior score(s). Quantile regression functions represent the solution to a loss function much in the way that least squares regression represents the solution to a minimization of squared deviations. The conditional quantile functions are parametrized as a linear combination of B-spline basis functions (Wei & He, 2006) to smooth irregularities found in the data. For scores from administration t (where *t* ≥2), the τth quantile function for Y_t conditional on prior scores $(Y_{t-1}, \ldots, Y_1)$ is

$$Q_{Yt}(\tau | Y_{t-1}, \ldots, Y_1) = \sum_{u=1}^{t-1} \sum_{j=1}^{n} \phi_{ju}(Y_u) \beta_{ju}(\tau) \text{ (15-1)}$$

where ϕ_{ju} (j =1, 2,…, n students; u =1, …, $t - 1$ administrations) represent the B-spline basis functions. The SGP of each student i is the midpoint between the two consecutive τ whose quantile scores capture the student's current score, multiplied by 100. For example, a student with a current score that lies between the fitted value for $\tau = .595$and$\tau = .605$would receive an SGP of 60.

SGPs are assumed to be uniformly distributed and uncorrelated with prior achievement. Scale score conditional standard errors of measurement were incorporated for calculation of SGP standard errors of measurement. Goodness of fit results were checked (i.e., uniform distribution of SGPs by prior achievement) for indications of ceiling/floor effects for each SGP norm-group analysis.

15.3. SGP Results

The estimation of SGPs was conducted for each student who had at least one prior score. Each analysis is defined by the norm cohort group (grade/sequence). A goodness of fit plot is produced for each analysis run, and a ceiling/floor effects test identifies potential problems at the HOSS and LOSS. Other fit plots compare the observed conditional density of SGP estimates with the theoretical uniform density. If there is perfect model fit, 10% of the estimated SGPs are expected within each decile band. A Q-Q plot compares the observed distribution with the theoretical distribution; ideally, the step function lines do not deviate much from the ideal line of perfect fit.

15.3.1. Summary for Total Group

Table 15.2 summarizes the SGP estimates for the total testing group from the Spring 2024 IAR administration for ELA/L and mathematics. Median SGPs were all close to 50. If the model is a perfect fit, the median is expected to be 50 with norm-referenced data. The average standard error for the SGPs is within expectations for these models. In general, SGPs can be divided into three categories: (a) an SGP below 30, indicating that a student is not meeting a year's worth of growth; (b) an SGP of 30–70, indicating that a student did achieve a year's worth of growth; and (c) an SGP over 70, indicating that the student surpassed a year's worth of growth. It is important to note that definitions such as these are not inherent to the SGP method but rather require expert judgment (Betebenner, 2009). The observed standard errors, ranging from 13.44 to 16.34 across content areas and grades, support these interpretations (Betebenner et al., 2017).

Table 15.2. Summary of SGP Estimates for Total Group

Assessment	Sample Size	Average SGP	Average Standard Error	Median SGP
ELA/L 4	122,276	49.89	13.46	50
ELA/L 5	122,078	49.97	15.12	50
ELA/L 6	123,415	49.94	14.43	50
ELA/L 7	126,591	49.92	14.51	50
ELA/L 8	127,832	49.84	14.97	50
Mathematics 4	119,647	50.02	13.90	50
Mathematics 5	119,760	50.13	15.64	50
Mathematics 6	121,251	50.09	15.77	50
Mathematics 7	125,043	50.02	16.13	50
Mathematics 8	126,338	49.99	16.34	50

15.3.2. Subgroups of Interest

Appendix G presents the SGP results for subgroups of interest from the Spring 2024 IAR administration. With norm-referenced data, the median of all SGPs is expected to be close to 50. Median subgroup SGPs below 50 represent growth lower than the median, and median SGPs above 50 represent growth higher than the median. As shown in the appendix, the median SGPs for subgroups of interest fell within the band of 30–70, which is considered adequate growth. Results by subgroup are as follows:

Gender:

- ELA/L: The median SGPs for females tend to be higher than the median SGPs for males. The median SGP ranges from 50 to 54 for females and 46 to 49 for males. The standard error for males and females is comparable to the total group.
- Mathematics: The median SGPs for females tend to be higher than the median SGPs for males except grade 4. The median SGP ranges from 48 to 52 for both females and males. The standard errors for both are similar to the total group.

Ethnicity:

- ELA/L: American Indian/Alaska Native students had median SGPs ranging from 40 to 48. For all ethnicity groups, standard errors are similar to that of the total group.
- Mathematics: American Indian/Alaska Native had median SGPs ranging from 46 to 52. For all ethnicities, the standard errors for all groups are under 20 points.

Special Instructional Needs:

- ELA/L: Students with disabilities had an observed median SGP of 40 to 44, whereas the median SGP ranges from 50 to 52 for students without disabilities. The standard errors for special instructional needs subgroups are similar to those observed for the total group.
- Mathematics: The median SGP ranges from 40 to 46 for students with disabilities and 51 to 52 for students without disabilities. The standard errors for special education students are similar to the total group.

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.

Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012). *Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments*. Pearson.

Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. National Center for the Improvement of Educational Assessment.

Betebenner, D. W., Van Iwaarden, A., Domingue, B., & Shang, Y. (2017). *SGP: Student growth percentiles & percentile growth trajectories* (R package version, 1-7 [Computer software].

Boyd, A., Minchen, N., & McBride, M. (2018). *Alternative blueprinting options research report*. Pearson.

Brandt, R., Bercovitz, E., McNally, S., & Zimmerman, L. (2015a). *Drawing response interaction usability study for PARCC*. Partnership for Assessment of Readiness for College and Careers. https://files.eric.ed.gov/fulltext/ED599260.pdf

Brandt, R., Bercovitz, E., & Zimmerman, L. (2015b). *Drawing response interaction usability study for PARCC*. Pearson. https://eric.ed.gov/?id=ED599261

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0). (CASMA Research Report No. 9). Center for Advanced Studies in Measurement, University of Iowa.

Center for Assessment. (2018). *PARCC comparability review guidelines*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Davis, L. L., & Moyer, E. L. (2015, December). *PARCC* performance *level setting technical report*. https://eric.ed.gov/?q=source%3a%22Partnership+for+Assessment+of+Readiness+for+College+and+Careers%22&pg=2&id=ED599257

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute.

Dorans, N. J. (2013). *ETS contributions to the quantitative assessment of item, test and score fairness* (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04). Educational Testing Service.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. RR-91-47). Educational Testing Service.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.

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

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Kolen, M. J. (2004). *POLYCSEM windows console version* [Computer software]. The Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129–140.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Livingston, S. A., & Lewis, C. (1993). *Estimating the consistency and accuracy of classifications based on test scores* (ETS Research Report No. RR-93-48). Educational Testing Service.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*(4), 453–461.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748.

McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: establishing a validity framework for cut scores. *Educational Researcher, 42*(2), 78–88.

Minchen, N., Boyd, A., & McBride, M. (2018a). *Alternative blueprinting options 2018 research report*. Pearson.

Minchen, N. LaSalle, A., & Boyd, A. (2018b). *Operational study 4: Accessibility of new items/functionality component 4 report*. Pearson.

New Meridian. (2023). *Technical report 2021–2022, alternate blueprint*. https://www.isbe.net/Documents/PARCC-Spring-2023-Tech-Manual.pdf

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Biometrika, 47*, 337–347.

Pike, C. K., & Hudson, W. W. (1998). Reliability and measurement error in the presence of homogeneity. *Journal of Social Service Research, 24*(1–2), 149–163.

Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting multiple performance standards using the Yes/No method: An alternative item mapping method* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Schultz, S. R., Norman Dvorak, R., & Chen, J. (2017). *Evaluating the quality and alignment of PARCC ELA/literacy and mathematics assessments: Grades 3, 4, 6, and 7*. (HumRRO Report 2017 No. 040). Human Resources Research Organization.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of *Educational Measurement, 28*(3), 237–247.

Steedle, J., & LaSalle, A. (2016). *Operational study 4: Accessibility of new items/functionality component 3 report*. Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

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

Wainer, H., & Thissen, D. (2001). *Test scoring*. Lawrence Erlbaum.

Wei, Y., & He, X. (2006). Conditional growth charts. *Annals of Statistics, 34*(5), 2069–2097.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices, 31*(1), 2–13.

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.

Zenisky, A. L., Hambleton, R. K., & Sireci, S.C. (2003). *Effects of local dependence on the validity of IRT item test, and ability statistics* (Technical Report). American College Admissions Test.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Lawrence Erlbaum

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (ETS Research Report RR-97-05). Educational Testing Service.

Appendix A: Scale Score Cumulative Frequencies

Table A.1. Scale Score Cumulative Frequencies—ELA/L Grade 3

Score Band	N	%	Cumulative N	Cumulative %
650–654	4,081	3.14	4,081	3.14
655 –659	5,754	4.43	9,835	7.57
660–664	–	–	–	–
665–669	130	0.10	9,965	7.67
670–674	7,593	5.84	17,558	13.51
675–679	–	–	–	–
680–684	7,898	6.08	25,456	19.58
685–689	125	0.10	25,581	19.68
690–694	6,854	5.27	32,435	24.95
695–699	6,010	4.62	38,445	29.57
700–704	4,824	3.71	43,269	33.28
705–709	4,499	3.46	47,768	36.75
710–714	4,042	3.11	51,810	39.85
715–719	3,878	2.98	55,688	42.84
720–724	7,483	5.76	63,171	48.59
725–729	3,583	2.76	66,754	51.35
730–734	7,225	5.56	73,979	56.91
735–739	3,533	2.72	77,512	59.63
740–744	6,706	5.16	84,218	64.78
745–749	6,405	4.93	90,623	69.71
750–754	3,198	2.46	93,821	72.17
755–759	5,890	4.53	99,711	76.70
760–764	2,786	2.14	102,497	78.85
765–769	5,371	4.13	107,868	82.98
770–774	2,499	1.92	110,367	84.90
775–779	4,408	3.39	114,775	88.29
780–784	3,718	2.86	118,493	91.15
785–789	1,660	1.28	120,153	92.43
790–794	2,700	2.08	122,853	94.50
795–799	1,142	0.88	123,995	95.38
800–804	1,949	1.50	125,944	96.88
805–809	790	0.61	126,734	97.49
810–814	731	0.56	127,465	98.05
815–819	1,081	0.83	128,546	98.88
820–824	370	0.28	128,916	99.17
825–829	307	0.24	129,223	99.40
830–834	184	0.14	129,407	99.55
835–839	172	0.13	129,579	99.68
840–844	126	0.10	129,705	99.78
845–849	82	0.06	129,787	99.84
850	210	0.16	129,997	100.00

Table A.2. Scale Score Cumulative Frequencies—ELA/L Grade 4

Score Band	N	%	Cumulative N	Cumulative %
650–654	2,656	2.05	2,656	2.05
655 –659	–	–	–	–
660–664	3,151	2.43	5,807	4.47
665–669	46	0.04	5,853	4.51
670–674	4,330	3.33	10,183	7.84
675–679	75	0.06	10,258	7.90
680–684	4,631	3.57	14,889	11.47
685–689	4,643	3.58	19,532	15.04
690–694	4,207	3.24	23,739	18.28
695–699	3,643	2.81	27,382	21.09
700–704	3,632	2.80	31,014	23.88
705–709	3,503	2.70	34,517	26.58
710–714	3,138	2.42	37,655	29.00
715–719	6,412	4.94	44,067	33.93
720–724	6,392	4.92	50,459	38.86
725–729	6,235	4.80	56,694	43.66
730–734	3,116	2.40	59,810	46.06
735–739	6,166	4.75	65,976	50.81
740–744	9,222	7.10	75,198	57.91
745–749	6,121	4.71	81,319	62.62
750–754	5,885	4.53	87,204	67.15
755–759	5,758	4.43	92,962	71.59
760–764	8,096	6.23	101,058	77.82
765–769	4,887	3.76	105,945	81.59
770–774	4,354	3.35	110,299	84.94
775–779	3,862	2.97	114,161	87.91
780–784	4,802	3.70	118,963	91.61
785–789	2,643	2.04	121,606	93.65
790–794	2,158	1.66	123,764	95.31
795–799	1,719	1.32	125,483	96.63
800–804	1,375	1.06	126,858	97.69
805–809	553	0.43	127,411	98.12
810–814	822	0.63	128,233	98.75
815–819	618	0.48	128,851	99.22
820–824	262	0.20	129,113	99.43
825–829	202	0.16	129,315	99.58
830–834	258	0.20	129,573	99.78
835–839	79	0.06	129,652	99.84
840–844	84	0.06	129,736	99.91
845–849	–	–	–	–
850	122	0.09	129,858	100.00

Table A.3. Scale Score Cumulative Frequencies—ELA/L Grade 5

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,261	0.97	1,261	0.97
655 –659	1,506	1.16	2,767	2.14
660–664	47	0.04	2,814	2.18
665–669	2,290	1.77	5,104	3.95
670–674	52	0.04	5,156	3.99
675–679	2,867	2.22	8,023	6.20
680–684	3,401	2.63	11,424	8.83
685–689	3,489	2.70	14,913	11.53
690–694	3,510	2.71	18,423	14.24
695–699	3,581	2.77	22,004	17.01
700–704	3,452	2.67	25,456	19.68
705–709	6,518	5.04	31,974	24.72
710–714	3,249	2.51	35,223	27.23
715–719	6,359	4.92	41,582	32.15
720–724	6,251	4.83	47,833	36.98
725–729	3,200	2.47	51,033	39.46
730–734	6,343	4.90	57,376	44.36
735–739	6,326	4.89	63,702	49.25
740–744	9,260	7.16	72,962	56.41
745–749	6,239	4.82	79,201	61.24
750–754	6,141	4.75	85,342	65.99
755–759	5,917	4.57	91,259	70.56
760–764	8,272	6.40	99,531	76.96
765–769	5,228	4.04	104,759	81.00
770–774	4,673	3.61	109,432	84.61
775–779	4,080	3.15	113,512	87.77
780–784	5,169	4.00	118,681	91.76
785–789	2,772	2.14	121,453	93.91
790–794	2,278	1.76	123,731	95.67
795–799	1,798	1.39	125,529	97.06
800–804	1,296	1.00	126,825	98.06
805–809	466	0.36	127,291	98.42
810–814	828	0.64	128,119	99.06
815–819	514	0.40	128,633	99.46
820–824	188	0.15	128,821	99.60
825–829	165	0.13	128,986	99.73
830–834	110	0.09	129,096	99.82
835–839	88	0.07	129,184	99.88
840–844	52	0.04	129,236	99.92
845–849	39	0.03	129,275	99.95
850	60	0.05	129,335	100.00

Table A.4. Scale Score Cumulative Frequencies—ELA/L Grade 6

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,781	1.36	1,781	1.36
655 –659	–	–	–	–
660–664	–	–	–	–
665–669	1,953	1.49	3,734	2.86
670–674	–	–	–	–
675–679	2,720	2.08	6,454	4.94
680–684	–	–	–	–
685–689	3,218	2.46	9,672	7.40
690–694	3,327	2.54	12,999	9.94
695–699	3,195	2.44	16,194	12.38
700–704	3,044	2.33	19,238	14.71
705–709	2,860	2.19	22,098	16.90
710–714	5,224	4.00	27,322	20.89
715–719	5,042	3.86	32,364	24.75
720–724	5,141	3.93	37,505	28.68
725–729	5,326	4.07	42,831	32.75
730–734	5,263	4.02	48,094	36.78
735–739	8,109	6.20	56,203	42.98
740–744	5,508	4.21	61,711	47.19
745–749	8,489	6.49	70,200	53.68
750–754	8,447	6.46	78,647	60.14
755–759	8,439	6.45	87,086	66.60
760–764	8,185	6.26	95,271	72.86
765–769	7,577	5.79	102,848	78.65
770–774	4,618	3.53	107,466	82.18
775–779	6,324	4.84	113,790	87.02
780–784	5,391	4.12	119,181	91.14
785–789	2,983	2.28	122,164	93.42
790–794	2,390	1.83	124,554	95.25
795–799	2,538	1.94	127,092	97.19
800–804	1,132	0.87	128,224	98.06
805–809	854	0.65	129,078	98.71
810–814	648	0.50	129,726	99.21
815–819	445	0.34	130,171	99.55
820–824	159	0.12	130,330	99.67
825–829	186	0.14	130,516	99.81
830–834	85	0.07	130,601	99.88
835–839	65	0.05	130,666	99.93
840–844	35	0.03	130,701	99.95
845–849	25	0.02	130,726	99.97
850	37	0.03	130,763	100.00

Table A.5. Scale Score Cumulative Frequencies—ELA/L Grade 7

Score Band	N	%	Cumulative N	Cumulative %
650–654	286	0.21	286	0.21
655 –659	663	0.50	949	0.71
660–664	21	0.02	970	0.73
665–669	–	–	–	–
670–674	1,366	1.02	2,336	1.75
675–679	2,152	1.61	4,488	3.36
680–684	–	–	–	–
685–689	2,810	2.10	7,298	5.46
690–694	3,294	2.47	10,592	7.93
695–699	3,461	2.59	14,053	10.52
700–704	6,889	5.16	20,942	15.68
705–709	3,332	2.50	24,274	18.18
710–714	6,416	4.80	30,690	22.98
715–719	6,033	4.52	36,723	27.50
720–724	6,040	4.52	42,763	32.02
725–729	5,880	4.40	48,643	36.43
730–734	5,796	4.34	54,439	40.77
735–739	8,640	6.47	63,079	47.24
740–744	5,650	4.23	68,729	51.47
745–749	8,386	6.28	77,115	57.75
750–754	8,229	6.16	85,344	63.91
755–759	7,632	5.72	92,976	69.62
760–764	4,823	3.61	97,799	73.23
765–769	6,774	5.07	104,573	78.31
770–774	6,014	4.50	110,587	82.81
775–779	5,154	3.86	115,741	86.67
780–784	3,085	2.31	118,826	88.98
785–789	3,793	2.84	122,619	91.82
790–794	2,193	1.64	124,812	93.46
795–799	2,740	2.05	127,552	95.51
800–804	1,499	1.12	129,051	96.64
805–809	1,259	0.94	130,310	97.58
810–814	1,107	0.83	131,417	98.41
815–819	826	0.62	132,243	99.03
820–824	360	0.27	132,603	99.30
825–829	429	0.32	133,032	99.62
830–834	157	0.12	133,189	99.74
835–839	120	0.09	133,309	99.83
840–844	82	0.06	133,391	99.89
845–849	65	0.05	133,456	99.94
850	86	0.06	133,542	100.00

Table A.6. Scale Score Cumulative Frequencies—ELA/L Grade 8

Score Band	N	%	Cumulative N	Cumulative %
650–654	3,266	2.42	3,266	2.42
655 –659	–	–	–	–
660–664	1,961	1.45	5,227	3.88
665–669	–	–	–	–
670–674	2,320	1.72	7,547	5.60
675–679	2,485	1.84	10,032	7.44
680–684	39	0.03	10,071	7.47
685–689	2,527	1.87	12,598	9.34
690–694	2,551	1.89	15,149	11.23
695–699	4,542	3.37	19,691	14.60
700–704	2,292	1.70	21,983	16.30
705–709	2,420	1.79	24,403	18.09
710–714	4,922	3.65	29,325	21.74
715–719	2,681	1.99	32,006	23.73
720–724	5,564	4.13	37,570	27.86
725–729	5,699	4.23	43,269	32.08
730–734	5,849	4.34	49,118	36.42
735–739	6,127	4.54	55,245	40.96
740–744	6,179	4.58	61,424	45.54
745–749	6,466	4.79	67,890	50.34
750–754	9,803	7.27	77,693	57.60
755–759	6,631	4.92	84,324	62.52
760–764	6,579	4.88	90,903	67.40
765–769	9,140	6.78	100,043	74.18
770–774	5,575	4.13	105,618	78.31
775–779	5,058	3.75	110,676	82.06
780–784	4,547	3.37	115,223	85.43
785–789	4,043	3.00	119,266	88.43
790–794	3,585	2.66	122,851	91.09
795–799	2,906	2.15	125,757	93.24
800–804	2,392	1.77	128,149	95.01
805–809	1,885	1.40	130,034	96.41
810–814	1,551	1.15	131,585	97.56
815–819	1,149	0.85	132,734	98.41
820–824	436	0.32	133,170	98.74
825–829	701	0.52	133,871	99.26
830–834	265	0.20	134,136	99.45
835–839	225	0.17	134,361	99.62
840–844	151	0.11	134,512	99.73
845–849	153	0.11	134,665	99.85
850	208	0.15	134,873	100.00

Table A.7. Scale Score Cumulative Frequencies—Mathematics Grade 3

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,686	1.30	1,686	1.30
655 –659	1,578	1.21	3,264	2.51
660–664	–	–	–	–
665–669	2,039	1.57	5,303	4.08
670–674	2,392	1.84	7,695	5.92
675–679	2,701	2.08	10,396	7.99
680–684	3,155	2.43	13,551	10.42
685–689	3,316	2.55	16,867	12.97
690–694	3,560	2.74	20,427	15.71
695–699	3,602	2.77	24,029	18.48
700–704	3,710	2.85	27,739	21.33
705–709	7,761	5.97	35,500	27.30
710–714	4,023	3.09	39,523	30.39
715–719	7,864	6.05	47,387	36.44
720–724	4,063	3.12	51,450	39.56
725–729	7,985	6.14	59,435	45.70
730–734	7,982	6.14	67,417	51.84
735–739	7,860	6.04	75,277	57.88
740–744	3,880	2.98	79,157	60.86
745–749	7,412	5.70	86,569	66.56
750–754	7,226	5.56	93,795	72.12
755–759	3,437	2.64	97,232	74.76
760–764	6,682	5.14	103,914	79.90
765–769	5,976	4.59	109,890	84.49
770–774	2,705	2.08	112,595	86.57
775–779	2,576	1.98	115,171	88.55
780–784	4,375	3.36	119,546	91.92
785–789	1,951	1.50	121,497	93.42
790–794	1,774	1.36	123,271	94.78
795–799	1,452	1.12	124,723	95.90
800–804	1,310	1.01	126,033	96.91
805–809	1,127	0.87	127,160	97.77
810–814	877	0.67	128,037	98.45
815–819	2	0.00	128,039	98.45
820–824	685	0.53	128,724	98.98
825–829	510	0.39	129,234	99.37
830–834	–	–	–	–
835–839	339	0.26	129,573	99.63
840–844	–	–	–	–
845–849	253	0.19	129,826	99.82
850	231	0.18	130,057	100.00

Table A.8. Scale Score Cumulative Frequencies—Mathematics Grade 4

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,609	1.24	1,609	1.24
655 –659	24	0.02	1,633	1.26
660–664	2,001	1.54	3,634	2.80
665–669	–	–	–	–
670–674	2,728	2.10	6,362	4.90
675–679	3,340	2.57	9,702	7.47
680–684	–	–	–	–
685–689	3,753	2.89	13,455	10.36
690–694	4,050	3.12	17,505	13.47
695–699	4,217	3.25	21,722	16.72
700–704	8,708	6.70	30,430	23.42
705–709	4,397	3.38	34,827	26.81
710–714	9,137	7.03	43,964	33.84
715–719	4,634	3.57	48,598	37.40
720–724	8,823	6.79	57,421	44.20
725–729	8,373	6.44	65,794	50.64
730–734	7,746	5.96	73,540	56.60
735–739	3,757	2.89	77,297	59.49
740–744	6,863	5.28	84,160	64.78
745–749	9,298	7.16	93,458	71.93
750–754	5,411	4.16	98,869	76.10
755–759	4,957	3.82	103,826	79.91
760–764	4,656	3.58	108,482	83.50
765–769	4,089	3.15	112,571	86.64
770–774	3,724	2.87	116,295	89.51
775–779	3,310	2.55	119,605	92.06
780–784	2,931	2.26	122,536	94.31
785–789	1,308	1.01	123,844	95.32
790–794	2,209	1.70	126,053	97.02
795–799	993	0.76	127,046	97.78
800–804	794	0.61	127,840	98.40
805–809	670	0.52	128,510	98.91
810–814	513	0.39	129,023	99.31
815–819	376	0.29	129,399	99.60
820–824	–	–	–	–
825–829	267	0.21	129,666	99.80
830–834	–	–	–	–
835–839	168	0.13	129,834	99.93
840–844	–	–	–	–
845–849	–	–	–	–
850	90	0.07	129,924	100.00

Table A.9. Scale Score Cumulative Frequencies—Mathematics Grade 5

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,369	1.06	1,369	1.06
655 –659	1,891	1.46	3,260	2.52
660–664	-	-	-	-
665–669	-	-	-	-
670–674	2,937	2.27	6,197	4.79
675–679	-	-	-	-
680–684	4,005	3.09	10,202	7.88
685–689	4,855	3.75	15,057	11.63
690–694	43	0.03	15,100	11.67
695–699	5,782	4.47	20,882	16.13
700–704	6,110	4.72	26,992	20.85
705–709	6,331	4.89	33,323	25.75
710–714	12,257	9.47	45,580	35.22
715–719	5,792	4.47	51,372	39.69
720–724	10,422	8.05	61,794	47.74
725–729	4,852	3.75	66,646	51.49
730–734	8,492	6.56	75,138	58.05
735–739	7,403	5.72	82,541	63.77
740–744	6,707	5.18	89,248	68.95
745–749	5,791	4.47	95,039	73.43
750–754	5,187	4.01	100,226	77.44
755–759	4,602	3.56	104,828	80.99
760–764	3,988	3.08	108,816	84.07
765–769	3,552	2.74	112,368	86.82
770–774	4,692	3.63	117,060	90.44
775–779	2,709	2.09	119,769	92.53
780–784	1,262	0.98	121,031	93.51
785–789	2,222	1.72	123,253	95.23
790–794	1,955	1.51	125,208	96.74
795–799	844	0.65	126,052	97.39
800–804	723	0.56	126,775	97.95
805–809	628	0.49	127,403	98.43
810–814	577	0.45	127,980	98.88
815–819	402	0.31	128,382	99.19
820–824	363	0.28	128,745	99.47
825–829	259	0.20	129,004	99.67
830–834	–	–	–	–
835–839	202	0.16	129,206	99.83
840–844	–	–	–	–
845–849	–	–	–	–
850	226	0.17	129,432	100.00

Table A.10. Scale Score Cumulative Frequencies—Mathematics Grade 6

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,228	0.94	1,228	0.94
655 –659	–	–	–	–
660–664	2,274	1.74	3,502	2.68
665–669	–	–	–	–
670–674	43	0.03	3,545	2.71
675–679	3,596	2.75	7,141	5.46
680–684	4,705	3.60	11,846	9.06
685–689	–	–	–	–
690–694	5,371	4.11	17,217	13.17
695–699	5,672	4.34	22,889	17.51
700–704	5,875	4.50	28,764	22.01
705–709	6,017	4.60	34,781	26.61
710–714	11,474	8.78	46,255	35.39
715–719	5,337	4.08	51,592	39.48
720–724	9,817	7.51	61,409	46.99
725–729	8,502	6.51	69,911	53.49
730–734	7,720	5.91	77,631	59.40
735–739	6,808	5.21	84,439	64.61
740–744	8,847	6.77	93,286	71.38
745–749	5,190	3.97	98,476	75.35
750–754	4,733	3.62	103,209	78.97
755–759	6,275	4.80	109,484	83.77
760–764	3,729	2.85	113,213	86.62
765–769	3,478	2.66	116,691	89.29
770–774	3,181	2.43	119,872	91.72
775–779	2,822	2.16	122,694	93.88
780–784	2,499	1.91	125,193	95.79
785–789	1,075	0.82	126,268	96.61
790–794	1,844	1.41	128,112	98.02
795–799	742	0.57	128,854	98.59
800–804	562	0.43	129,416	99.02
805–809	449	0.34	129,865	99.37
810–814	–	–	–	–
815–819	334	0.26	130,199	99.62
820–824	–	–	–	–
825–829	217	0.17	130,416	99.79
830–834	–	–	–	–
835–839	156	0.12	130,572	99.91
840–844	–	–	–	–
845–849	–	–	–	–
850	122	0.09	130,694	100.00

Table A.11. Scale Score Cumulative Frequencies—Mathematics Grade 7

Score Band	N	%	Cumulative N	Cumulative %
650–654	876	0.66	876	0.66
655 –659	–	–	–	–
660–664	–	–	–	–
665–669	–	–	–	–
670–674	1,963	1.47	2,839	2.13
675–679	–	–	–	–
680–684	28	0.02	2,867	2.15
685–689	3,615	2.71	6,482	4.86
690–694	5,067	3.80	11,549	8.66
695–699	–	–	–	–
700–704	6,154	4.61	17,703	13.27
705–709	6,904	5.18	24,607	18.45
710–714	7,074	5.30	31,681	23.75
715–719	13,523	10.14	45,204	33.89
720–724	6,088	4.56	51,292	38.45
725–729	10,471	7.85	61,763	46.30
730–734	9,124	6.84	70,887	53.14
735–739	8,139	6.10	79,026	59.24
740–744	10,381	7.78	89,407	67.02
745–749	5,931	4.45	95,338	71.47
750–754	7,923	5.94	103,261	77.41
755–759	6,662	4.99	109,923	82.40
760–764	5,675	4.25	115,598	86.66
765–769	3,267	2.45	118,865	89.11
770–774	4,025	3.02	122,890	92.13
775–779	2,271	1.70	125,161	93.83
780–784	2,035	1.53	127,196	95.35
785–789	1,759	1.32	128,955	96.67
790–794	792	0.59	129,747	97.27
795–799	1,329	1.00	131,076	98.26
800–804	557	0.42	131,633	98.68
805–809	504	0.38	132,137	99.06
810–814	437	0.33	132,574	99.39
815–819	–	–	–	–
820–824	311	0.23	132,885	99.62
825–829	–	–	–	–
830–834	251	0.19	133,136	99.81
835–839	–	–	–	–
840–844	–	–	–	–
845–849	136	0.10	133,272	99.91
850	122	0.09	133,394	100.00

Table A.12. Scale Score Cumulative Frequencies—Mathematics Grade 8

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,826	1.36	1,826	1.36
655 –659	2,816	2.09	4,642	3.45
660–664	–	–	–	–
665–669	4,968	3.69	9,610	7.13
670–674	–	–	–	–
675–679	6,521	4.84	16,131	11.97
680–684	61	0.05	16,192	12.02
685–689	7,567	5.62	23,759	17.64
690–694	7,964	5.91	31,723	23.55
695–699	7,872	5.84	39,595	29.39
700–704	44	0.03	39,639	29.42
705–709	7,435	5.52	47,074	34.94
710–714	7,243	5.38	54,317	40.32
715–719	6,515	4.84	60,832	45.16
720–724	11,612	8.62	72,444	53.78
725–729	5,006	3.72	77,450	57.49
730–734	4,653	3.45	82,103	60.95
735–739	4,340	3.22	86,443	64.17
740–744	7,322	5.44	93,765	69.60
745–749	3,300	2.45	97,065	72.05
750–754	5,891	4.37	102,956	76.43
755–759	2,640	1.96	105,596	78.38
760–764	4,827	3.58	110,423	81.97
765–769	4,132	3.07	114,555	85.04
770–774	1,963	1.46	116,518	86.49
775–779	3,481	2.58	119,999	89.08
780–784	1,546	1.15	121,545	90.22
785–789	2,942	2.18	124,487	92.41
790–794	1,281	0.95	125,768	93.36
795–799	2,293	1.70	128,061	95.06
800–804	1,010	0.75	129,071	95.81
805–809	948	0.70	130,019	96.51
810–814	1,556	1.16	131,575	97.67
815–819	658	0.49	132,233	98.16
820–824	–	–	–	–
825–829	549	0.41	132,782	98.57
830–834	510	0.38	133,292	98.94
835–839	433	0.32	133,725	99.27
840–844	–	–	–	–
845–849	307	0.23	134,032	99.49
850	683	0.51	134,715	100.00

Appendix B: Scale Score Performance by Demographic Subgroup

Table B.1. Scale Score Performance by Demographic Subgroup—ELA/L Grade 3

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**129,997**	**727.19**	**42.14**	**650**	**850**
Female	63,961	730.66	42.90	650	850
Male	66,022	723.84	41.10	650	850
American Indian/Alaska Native	347	722.04	39.09	650	822
Asian	7,556	751.02	43.26	650	850
Black or African American	21,124	709.23	37.74	650	850
Hispanic/Latino	36,016	713.12	39.29	650	850
Native Hawaiian or Pacific Islander	107	732.93	38.05	650	818
Two or More Races	6,379	731.74	42.80	650	850
White	58,290	738.89	39.73	650	850
Not Economically Disadvantaged	61,817	742.97	40.40	650	850
Economically Disadvantaged	67,550	712.97	38.41	650	850
Non-English Learner (EL)	101,493	733.06	41.74	650	850
English Learner (EL)	28,251	706.24	36.55	650	850
Students without Disabilities	106,089	732.23	41.54	650	850
Student with Disability (SWD)	22,675	704.99	37.21	650	850
Reading Claim Score	**129,997**	**41.59**	**16.98**	**10**	**90**
Female	63,961	42.73	17.24	10	90
Male	66,022	40.48	16.66	10	90
American Indian/Alaska Native	347	39.27	15.49	10	81
Asian	7,556	50.61	17.58	10	90
Black or African American	21,124	34.56	14.98	10	90
Hispanic/Latino	36,016	35.89	15.53	10	90
Native Hawaiian or Pacific Islander	107	44.54	15.16	10	81
Two or More Races	6,379	43.62	17.40	10	90
White	58,290	46.30	16.31	10	90
Not Economically Disadvantaged	61,817	47.93	16.61	10	90
Economically Disadvantaged	67,550	35.87	15.18	10	90
Non-English Learner	101,493	44.04	16.95	10	90
English Learner	28,251	32.82	13.92	10	90
Students without Disabilities	106,089	43.52	16.79	10	90
Student with Disability (SWD)	22,675	33.06	15.04	10	90
Writing Claim Score	**129,997**	**25.56**	**13.64**	**10**	**60**
Female	63,961	26.87	13.71	10	60
Male	66,022	24.30	13.45	10	60
American Indian/Alaska Native	347	24.33	13.16	10	53
Asian	7,556	32.64	13.23	10	60
Black or African American	21,124	20.22	12.56	10	60
Hispanic/Latino	36,016	21.72	13.01	10	60
Native Hawaiian or Pacific Islander	107	25.96	13.36	10	49
Two or More Races	6,379	26.50	13.81	10	60
White	58,290	28.87	13.07	10	60

Subgroup	N	Mean	SD	Min.	Max.
Not Economically Disadvantaged	61,817	30.04	13.05	10	60
Economically Disadvantaged	67,550	21.52	12.87	10	60
Non-English Learner	101,493	27.05	13.57	10	60
English Learner	28,251	20.23	12.52	10	60
Students without Disabilities	106,089	27.05	13.54	10	60
Student with Disability (SWD)	22,675	18.97	12.02	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.2. Scale Score Performance by Demographic Subgroup—ELA/L Grade 4

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**129,858**	**734.73**	**38.11**	**650**	**850**
Female	63,603	738.00	38.11	650	850
Male	66,235	731.58	37.85	650	850
American Indian/Alaska Native	332	721.64	35.42	650	834
Asian	7,457	754.93	37.94	650	850
Black or African American	21,032	716.74	34.51	650	844
Hispanic/Latino	35,768	721.92	36.44	650	850
Native Hawaiian or Pacific Islander	107	738.59	39.03	650	819
Two or More Races	6,081	738.87	38.86	650	850
White	58,921	746.07	34.93	650	850
Not Economically Disadvantaged	62,170	749.52	35.36	650	850
Economically Disadvantaged	66,970	721.24	35.39	650	850
Non-English Learner (EL)	103,725	740.24	37.17	650	850
English Learner (EL)	25,844	712.83	33.70	650	850
Students without Disabilities	104,775	740.27	36.48	650	850
Student with Disability (SWD)	23,960	711.69	35.99	650	850
Reading Claim Score	**129,858**	**44.39**	**15.10**	**10**	**90**
Female	63,603	45.00	14.88	10	90
Male	66,235	43.80	15.29	10	90
American Indian/Alaska Native	332	39.45	14.22	10	78
Asian	7,457	52.06	15.05	10	90
Black or African American	21,032	37.70	13.59	10	90
Hispanic/Latino	35,768	39.38	14.24	10	90
Native Hawaiian or Pacific Islander	107	45.84	15.47	11	84
Two or More Races	6,081	46.23	15.49	10	90
White	58,921	48.71	14.15	10	90
Not Economically Disadvantaged	62,170	50.14	14.30	10	90
Economically Disadvantaged	66,970	39.14	13.84	10	90
Non-English Learner	103,725	46.59	14.83	10	90
English Learner	25,844	35.64	12.87	10	84
Students without Disabilities	104,775	46.41	14.56	10	90
Student with Disability (SWD)	23,960	35.97	14.41	10	90

Subgroup	N	Mean	SD	Min.	Max.
Writing Claim Score	**129,858**	**28.88**	**12.41**	**10**	**60**
Female	63,603	30.51	12.23	10	60
Male	66,235	27.33	12.39	10	60
American Indian/Alaska Native	332	25.01	12.12	10	60
Asian	7,457	34.55	11.49	10	60
Black or African American	21,032	23.48	12.03	10	60
Hispanic/Latino	35,768	25.38	12.40	10	60
Native Hawaiian or Pacific Islander	107	30.19	12.21	10	52
Two or More Races	6,081	29.71	12.52	10	60
White	58,921	32.18	11.27	10	60
Not Economically Disadvantaged	62,170	33.02	11.21	10	60
Economically Disadvantaged	66,970	25.13	12.25	10	60
Non-English Learner	103,725	30.34	12.08	10	60
English Learner	25,844	23.13	12.05	10	60
Students without Disabilities	104,775	30.66	11.78	10	60
Student with Disability (SWD)	23,960	21.48	12.20	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.3. Scale Score Performance by Demographic Subgroup—ELA/L Grade 5

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**129,335**	**736.77**	**36.27**	**650**	**850**
Female	63,355	740.84	36.47	650	850
Male	65,954	732.85	35.63	650	850
American Indian/Alaska Native	308	734.00	36.67	650	830
Asian	7,538	758.32	35.97	650	850
Black or African American	20,792	719.52	32.85	650	850
Hispanic/Latino	35,756	724.57	34.81	650	850
Native Hawaiian or Pacific Islander	106	746.01	37.24	650	835
Two or More Races	5,850	740.35	36.66	650	850
White	58,788	747.23	33.09	650	850
Not Economically Disadvantaged	62,002	750.79	33.47	650	850
Economically Disadvantaged	66,712	723.96	33.87	650	850
Non-English Learner (EL)	107,412	742.48	34.73	650	850
English Learner (EL)	21,644	708.64	30.04	650	830
Students without Disabilities	104,276	742.60	34.26	650	850
Student with Disability (SWD)	24,060	712.56	34.27	650	850
Reading Claim Score	**129,335**	**45.40**	**14.73**	**10**	**90**
Female	63,355	46.23	14.53	10	90
Male	65,954	44.60	14.88	10	90

Subgroup	N	Mean	SD	Min.	Max.
American Indian/Alaska Native	308	44.40	15.13	10	87
Asian	7,538	53.53	14.66	10	90
Black or African American	20,792	38.99	13.58	10	90
Hispanic/Latino	35,756	40.63	14.15	10	90
Native Hawaiian or Pacific Islander	106	49.13	15.22	10	87
Two or More Races	5,850	47.01	14.95	10	90
White	58,788	49.39	13.61	10	90
Not Economically Disadvantaged	62,002	50.87	13.77	10	90
Economically Disadvantaged	66,712	40.41	13.76	10	90
Non-English Learner	107,412	47.71	14.17	10	90
English Learner	21,644	34.01	11.88	10	90
Students without Disabilities	104,276	47.62	13.98	10	90
Student with Disability (SWD)	24,060	36.19	14.19	10	90
Writing Claim Score	**129,335**	**27.95**	**13.09**	**10**	**60**
Female	63,355	30.04	12.83	10	60
Male	65,954	25.94	13.01	10	60
American Indian/Alaska Native	308	27.05	13.13	10	51
Asian	7,538	34.92	11.57	10	60
Black or African American	20,792	22.10	12.43	10	60
Hispanic/Latino	35,756	24.28	12.84	10	60
Native Hawaiian or Pacific Islander	106	30.92	12.63	10	56
Two or More Races	5,850	28.66	13.20	10	60
White	58,788	31.30	12.14	10	60
Not Economically Disadvantaged	62,002	32.37	11.91	10	60
Economically Disadvantaged	66,712	23.91	12.79	10	60
Non-English Learner	107,412	29.61	12.72	10	60
English Learner	21,644	19.79	11.73	10	60
Students without Disabilities	104,276	29.89	12.52	10	60
Student with Disability (SWD)	24,060	19.90	12.28	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.4. Scale Score Performance by Demographic Subgroup—ELA/L Grade 6

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**130,763**	**742.56**	**33.87**	**650**	**850**
Female	64,178	746.85	33.42	650	850
Male	66,552	738.42	33.78	650	850
American Indian/Alaska Native	293	736.79	33.91	650	836
Asian	7,592	762.56	32.20	650	850
Black or African American	20,836	726.62	32.21	650	836
Hispanic/Latino	36,643	732.00	33.28	650	845
Native Hawaiian or Pacific Islander	142	751.24	32.09	650	832
Two or More Races	5,885	745.27	33.93	650	850
White	59,178	751.95	30.40	650	850

Subgroup	N	Mean	SD	Min.	Max.
Not Economically Disadvantaged	63,469	755.23	30.47	650	850
Economically Disadvantaged	66,562	730.71	32.52	650	850
Non-English Learner (EL)	110,995	747.65	32.10	650	850
English Learner (EL)	19,490	713.72	28.83	650	850
Students without Disabilities	105,679	748.00	31.48	650	850
Student with Disability (SWD)	24,123	719.72	34.03	650	845
Reading Claim Score	**130,763**	**46.52**	**13.21**	**10**	**90**
Female	64,178	47.50	12.98	10	90
Male	66,552	45.57	13.36	10	90
American Indian/Alaska Native	293	44.04	13.24	10	84
Asian	7,592	54.01	12.84	10	90
Black or African American	20,836	40.63	12.63	10	84
Hispanic/Latino	36,643	42.27	12.85	10	90
Native Hawaiian or Pacific Islander	142	49.71	12.03	10	80
Two or More Races	5,885	47.79	13.25	10	90
White	59,178	50.16	11.95	10	90
Not Economically Disadvantaged	63,469	51.44	12.00	10	90
Economically Disadvantaged	66,562	41.91	12.60	10	90
Non-English Learner	110,995	48.55	12.54	10	90
English Learner	19,490	35.01	10.79	10	90
Students without Disabilities	105,679	48.50	12.35	10	90
Student with Disability (SWD)	24,123	38.23	13.49	10	84
Writing Claim Score	**130,763**	**30.25**	**12.91**	**10**	**60**
Female	64,178	32.59	12.13	10	60
Male	66,552	28.00	13.24	10	60
American Indian/Alaska Native	293	28.72	13.15	10	52
Asian	7,592	36.70	10.61	10	60
Black or African American	20,836	24.82	13.07	10	60
Hispanic/Latino	36,643	27.30	13.12	10	60
Native Hawaiian or Pacific Islander	142	33.03	12.16	10	52
Two or More Races	5,885	30.66	12.99	10	60
White	59,178	33.16	11.79	10	60
Not Economically Disadvantaged	63,469	34.20	11.42	10	60
Economically Disadvantaged	66,562	26.58	13.13	10	60
Non-English Learner	110,995	31.75	12.39	10	60
English Learner	19,490	21.79	12.55	10	60
Students without Disabilities	105,679	32.22	12.06	10	60
Student with Disability (SWD)	24,123	21.97	13.14	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.5. Scale Score Performance by Demographic Subgroup—ELA/L Grade 7

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**133,542**	**742.09**	**34.11**	**650**	**850**
Female	65,293	747.20	34.25	650	850
Male	68,216	737.20	33.26	650	850
American Indian/Alaska Native	275	733.60	36.04	660	837
Asian	7,573	766.35	35.29	650	850
Black or African American	21,295	727.80	30.13	650	849
Hispanic/Latino	37,880	731.38	32.33	650	850
Native Hawaiian or Pacific Islander	132	747.95	33.21	670	814
Two or More Races	5,698	744.15	34.01	650	850
White	60,523	750.66	32.04	650	850
Not Economically Disadvantaged	65,622	754.30	32.57	650	850
Economically Disadvantaged	67,218	730.37	31.31	650	850
Non-English Learner (EL)	112,546	747.09	32.97	650	850
English Learner (EL)	20,701	715.06	26.79	650	849
Students without Disabilities	108,347	747.27	32.58	650	850
Student with Disability (SWD)	24,257	719.83	31.49	650	850
Reading Claim Score	**133,542**	**46.52**	**13.30**	**10**	**90**
Female	65,293	47.99	13.32	10	90
Male	68,216	45.12	13.13	10	90
American Indian/Alaska Native	275	43.25	13.83	14	81
Asian	7,573	55.81	14.17	10	90
Black or African American	21,295	41.38	11.83	10	90
Hispanic/Latino	37,880	42.33	12.47	10	90
Native Hawaiian or Pacific Islander	132	49.02	13.17	19	81
Two or More Races	5,698	47.66	13.35	10	90
White	60,523	49.71	12.60	10	90
Not Economically Disadvantaged	65,622	51.21	12.86	10	90
Economically Disadvantaged	67,218	42.03	12.10	10	90
Non-English Learner	112,546	48.50	12.89	10	90
English Learner	20,701	35.84	10.05	10	85
Students without Disabilities	108,347	48.44	12.74	10	90
Student with Disability (SWD)	24,257	38.31	12.51	10	90
Writing Claim Score	**133,542**	**28.49**	**14.03**	**10**	**60**
Female	65,293	31.10	13.50	10	60
Male	68,216	25.99	14.07	10	60
American Indian/Alaska Native	275	25.42	14.53	10	60
Asian	7,573	36.83	11.89	10	60
Black or African American	21,295	22.87	13.45	10	60
Hispanic/Latino	37,880	25.03	13.85	10	60
Native Hawaiian or Pacific Islander	132	30.29	13.64	10	49
Two or More Races	5,698	28.45	14.19	10	60
White	60,523	31.63	13.21	10	60
Not Economically Disadvantaged	65,622	32.77	12.98	10	60
Economically Disadvantaged	67,218	24.39	13.76	10	60

Subgroup	N	Mean	SD	Min.	Max.
Non-English Learner	112,546	30.12	13.69	10	60
English Learner	20,701	19.73	12.53	10	60
Students without Disabilities	108,347	30.48	13.49	10	60
Student with Disability (SWD)	24,257	19.94	13.07	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.6. Scale Score Performance by Demographic Subgroup—ELA/L Grade 8

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**134,873**	**744.60**	**38.91**	**650**	**850**
Female	65,479	751.05	38.76	650	850
Male	69,345	738.50	38.05	650	850
American Indian/Alaska Native	281	735.59	42.90	650	848
Asian	7,611	768.21	36.65	650	850
Black or African American	22,079	728.30	35.30	650	850
Hispanic/Latino	38,349	733.00	38.42	650	850
Native Hawaiian or Pacific Islander	104	747.82	41.05	650	816
Two or More Races	5,531	746.56	38.53	650	850
White	60,736	754.81	36.04	650	850
Not Economically Disadvantaged	66,288	758.39	35.67	650	850
Economically Disadvantaged	67,986	731.37	37.19	650	850
Non-English Learner (EL)	114,543	750.31	36.84	650	850
English Learner (EL)	20,058	712.14	34.12	650	842
Students without Disabilities	109,597	750.85	36.31	650	850
Student with Disability (SWD)	24,452	717.52	38.18	650	850
Reading Claim Score	**134,873**	**48.77**	**15.95**	**10**	**90**
Female	65,479	50.38	15.66	10	90
Male	69,345	47.24	16.08	10	90
American Indian/Alaska Native	281	45.22	17.61	10	90
Asian	7,611	57.69	15.08	10	90
Black or African American	22,079	42.80	14.84	10	90
Hispanic/Latino	38,349	44.30	15.94	10	90
Native Hawaiian or Pacific Islander	104	49.52	16.26	10	88
Two or More Races	5,531	49.97	15.92	10	90
White	60,736	52.57	14.84	10	90
Not Economically Disadvantaged	66,288	54.14	14.67	10	90
Economically Disadvantaged	67,986	43.61	15.39	10	90
Non-English Learner	114,543	51.10	15.12	10	90
English Learner	20,058	35.51	13.98	10	88
Students without Disabilities	109,597	51.22	14.89	10	90
Student with Disability (SWD)	24,452	38.14	16.04	10	90
Writing Claim Score	**134,873**	**29.96**	**13.10**	**10**	**60**
Female	65,479	32.75	12.52	10	60
Male	69,345	27.32	13.09	10	60

Subgroup	N	Mean	SD	Min.	Max.
American Indian/Alaska Native	281	27.16	14.02	10	60
Asian	7,611	37.27	10.92	10	60
Black or African American	22,079	24.42	12.76	10	60
Hispanic/Latino	38,349	26.63	13.01	10	60
Native Hawaiian or Pacific Islander	104	31.42	13.17	10	53
Two or More Races	5,531	29.99	13.21	10	60
White	60,736	33.18	12.14	10	60
Not Economically Disadvantaged	66,288	34.13	11.84	10	60
Economically Disadvantaged	67,986	25.96	13.00	10	60
Non-English Learner	114,543	31.50	12.68	10	60
English Learner	20,058	21.20	11.95	10	60
Students without Disabilities	109,597	31.86	12.42	10	60
Student with Disability (SWD)	24,452	21.71	12.81	10	60

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.7. Scale Score Performance by Demographic Subgroup—Mathematics Grade 3

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**130,057**	**733.52**	**37.02**	**650**	**850**
Female	64,019	731.63	35.70	650	850
Male	66,023	735.36	38.17	650	850
American Indian/Alaska Native	346	729.45	36.26	650	828
Asian	7,543	758.70	37.67	650	850
Black or African American	21,061	712.88	32.51	650	850
Hispanic/Latino	35,957	721.44	33.04	650	850
Native Hawaiian or Pacific Islander	107	737.65	34.98	650	807
Two or More Races	6,370	735.98	38.18	650	850
White	58,237	745.09	34.36	650	850
Not Economically Disadvantaged	61,763	748.62	35.17	650	850
Economically Disadvantaged	67,404	719.97	33.16	650	850
Non-English Learner (EL)	101,343	737.76	37.18	650	850
English Learner (EL)	28,203	718.60	32.32	650	850
Students without Disabilities	105,961	737.79	35.79	650	850
Student with Disability (SWD)	22,608	714.95	36.69	650	850
Spanish	5,271	708.19	31.14	650	847

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.8. Scale Score Performance by Demographic Subgroup—Mathematics Grade 4

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**129,924**	**730.82**	**33.94**	**650**	**850**
Female	63,664	728.78	32.56	650	850
Male	66,240	732.78	35.10	650	850
American Indian/Alaska Native	330	722.21	31.56	650	816
Asian	7,445	755.15	34.42	650	850
Black or African American	20,996	710.75	29.15	650	850
Hispanic/Latino	35,677	718.81	29.93	650	850
Native Hawaiian or Pacific Islander	107	730.91	32.64	663	797
Two or More Races	6,077	733.90	34.79	650	850
White	58,865	742.06	31.18	650	850
Not Economically Disadvantaged	62,112	745.22	32.15	650	850
Economically Disadvantaged	66,831	717.72	29.95	650	850
Non-English Learner (EL)	103,590	734.94	33.98	650	850
English Learner (EL)	25,783	714.66	28.50	650	825
Students without Disabilities	104,626	735.21	32.68	650	850
Student with Disability (SWD)	23,927	712.80	33.14	650	850
Spanish	3,817	702.32	26.45	650	816

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.9. Scale Score Performance by Demographic Subgroup—Mathematics Grade 5

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**129,432**	**729.66**	**33.77**	**650**	**850**
Female	63,403	728.88	31.77	650	850
Male	66,003	730.41	35.58	650	850
American Indian/Alaska Native	309	728.71	35.12	657	850
Asian	7,543	757.93	36.64	650	850
Black or African American	20,779	710.32	27.68	650	850
Hispanic/Latino	35,698	718.46	29.04	650	850
Native Hawaiian or Pacific Islander	106	736.80	37.61	657	829
Two or More Races	5,848	732.30	35.02	650	850
White	58,725	739.51	31.78	650	850
Not Economically Disadvantaged	61,964	743.66	33.00	650	850
Economically Disadvantaged	66,620	716.87	29.01	650	850
Non-English Learner (EL)	107,318	733.76	33.81	650	850
English Learner (EL)	21,610	709.58	25.40	650	850
Students without Disabilities	104,170	734.03	32.78	650	850
Student with Disability (SWD)	24,056	711.75	31.90	650	850
Spanish	3,539	702.24	26.86	650	810

Note. SD = standard deviation. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.10. Scale Score Performance by Demographic Subgroup—Mathematics Grade 6

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**130,694**	**728.37**	**31.76**	**650**	**850**
Female	64,174	727.76	30.35	650	850
Male	66,486	728.95	33.05	650	850
American Indian/Alaska Native	290	723.16	30.49	650	825
Asian	7,572	755.47	34.54	650	850
Black or African American	20,754	709.87	26.17	650	850
Hispanic/Latino	36,537	717.87	27.56	650	850
Native Hawaiian or Pacific Islander	141	736.74	32.60	650	850
Two or More Races	5,867	729.79	32.80	650	850
White	59,089	737.88	29.53	650	850
Not Economically Disadvantaged	63,393	741.56	30.70	650	850
Economically Disadvantaged	66,324	716.01	27.41	650	850
Non-English Learner (EL)	110,745	732.38	31.42	650	850
English Learner (EL)	19,422	705.87	23.08	650	825
Students without Disabilities	105,482	732.84	30.52	650	850
Student with Disability (SWD)	24,025	709.72	30.08	650	850
Spanish	3,149	701.70	24.34	650	781

Note. SD = standard deviation, n/r = not reported due to n<20, n/a = not applicable. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.11. Scale Score Performance by Demographic Subgroup—Mathematics Grade 7

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**133,394**	**733.81**	**28.81**	**650**	**850**
Female	65,250	733.60	27.96	650	850
Male	68,110	734.02	29.60	650	850
American Indian/Alaska Native	275	728.91	29.70	650	846
Asian	7,547	760.15	32.59	650	850
Black or African American	21,198	717.43	22.98	650	850
Hispanic/Latino	37,742	724.76	24.74	650	850
Native Hawaiian or Pacific Islander	134	736.62	28.41	674	814
Two or More Races	5,679	735.25	29.76	650	850
White	60,413	741.87	27.24	650	850
Not Economically Disadvantaged	65,517	745.04	28.57	650	850
Economically Disadvantaged	66,934	723.02	24.57	650	850
Non-English Learner (EL)	112,266	737.35	28.64	650	850
English Learner (EL)	20,593	714.77	21.45	650	831
Students without Disabilities	108,051	737.98	27.55	650	850
Student with Disability (SWD)	24,188	715.89	27.43	650	850
Spanish	2,355	706.38	19.88	650	778

Note. SD = standard deviation, n/r = not reported due to n<20, n/a = not applicable. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.12. Scale Score Performance by Demographic Subgroup—Mathematics Grade 8

Subgroup	N	Mean	SD	Min.	Max.
Overall Score	**134,715**	**727.00**	**40.16**	**650**	**850**
Female	65,420	727.01	39.25	650	850
Male	69,248	726.98	41.01	650	850
American Indian/Alaska Native	277	720.07	39.30	650	850
Asian	7,593	764.39	44.93	650	850
Black or African American	21,975	705.54	31.08	650	850
Hispanic/Latino	38,227	714.54	34.56	650	850
Native Hawaiian or Pacific Islander	103	739.83	43.91	650	850
Two or More Races	5,511	728.92	41.19	650	850
White	60,610	737.93	38.66	650	850
Not Economically Disadvantaged	66,175	742.47	40.25	650	850
Economically Disadvantaged	67,710	712.15	33.95	650	850
Non-English Learner (EL)	114,228	731.78	40.20	650	850
English Learner (EL)	19,980	700.08	27.36	650	830
Students without Disabilities	109,305	732.57	39.26	650	850
Student with Disability (SWD)	24,369	703.15	35.07	650	850
Spanish	2,307	689.36	23.70	650	791

Note. SD = standard deviation, n/r = not reported due to n<20, n/a = not applicable. Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Appendix C: Differential Item Functioning (DIF) Results

This appendix presents the number of items in each DIF category, along with the percentage out of the total number of items. The abbreviations are as follows:

- AI/AN = American Indian/Alaska Native
- NH/PI = Native Hawaiian or Pacific Islander
- Multiracial = multiple races selected
- NED = not economically disadvantaged
- ED = economically disadvantaged
- ELN = not an English learner
- ELY = English learner
- SWDN = not student with disability
- SWDY = student with disability

Table C.1. Pre-Administration DIF Results—ELA/L Grade 3

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	113					113	100				
White vs. Black/African American	113	4	4	7	6	102	90				
White vs. Hispanic/Latino	113			7	6	106	94				
White vs. Asian	113					110	97	3	3		
White vs. AI/AN	113					113	100				
White vs. NH/PI	113					113	100				
White vs. Multiracial	113					113	100				
NED vs. ED	113					113	100				
ELN vs. ELY	113			8	7	105	93				
SWDN vs. SWDY	113			4	4	109	96				

Table C.2. Pre-Administration DIF Results—ELA/L Grade 4

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	157			1	1	152	97	4	3		
White vs. Black/African American	157			6	4	151	96				
White vs. Hispanic/Latino	157			7	4	150	96				
White vs. Asian	157					156	99	1	1		
White vs. AI/AN	157					157	100				
White vs. NH/PI	157					157	100				
White vs. Multiracial	157					157	100				
NED vs. ED	157					157	100				
ELN vs. ELY	157			17	11	140	89				
SWDN vs. SWDY	157			4	3	152	97	1	1		

Table C.3. Pre-Administration DIF Results—ELA/L Grade 5

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	148	2	1	5	3	139	94	2	1		
White vs. Black/African American	148			10	7	138	93				
White vs. Hispanic/Latino	148	1	1	6	4	141	95				
White vs. Asian	148			1	1	147	99				
White vs. AI/AN	148					148	100				
White vs. NH/PI	148					148	100				
White vs. Multiracial	148					148	100				
NED vs. ED	148					148	100				
ELN vs. ELY	148	4	3	10	7	134	91				
SWDN vs. SWDY	148	1	1	4	3	143	97				

Table C.4. Pre-Administration DIF Results—ELA/L Grade 6

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	149	1	1	5	3	139	93	3	2	1	1
White vs. Black/African American	149			11	7	138	93				
White vs. Hispanic/Latino	149	2	1	7	5	140	94				
White vs. Asian	149					149	100				
White vs. AI/AN	149					149	100				
White vs. NH/PI	149					149	100				
White vs. Multiracial	149					147	99	2	1		
NED vs. ED	149					149	100				
ELN vs. ELY	149	9	6	19	13	121	81				
SWDN vs. SWDY	149			8	5	141	95				

Table C.5. Pre-Administration DIF Results—ELA/L Grade 7

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	162	3	2	8	5	142	88	9	6		
White vs. Black/African American	162	3	2	8	5	151	93				
White vs. Hispanic/Latino	162	1	1	12	7	149	92				
White vs. Asian	162					159	98	3	2		
White vs. AI/AN	162					162	100				
White vs. NH/PI	162					162	100				
White vs. Multiracial	162					162	100				
NED vs. ED	162					162	100				
ELN vs. ELY	162	5	3	18	11	139	86				
SWDN vs. SWDY	162			5	3	157	97				

Table C.6. Pre-Administration DIF Results—ELA/L Grade 8

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	143	1	1	12	8	129	90	1	1		
White vs. Black/African American	143	2	1	4	3	134	94	3	2		
White vs. Hispanic/Latino	143	1	1	6	4	136	95				
White vs. Asian	143			1	1	140	98	2	1		
White vs. AI/AN	143					143	100				
White vs. NH/PI	143					143	100				
White vs. Multiracial	143					143	100				
NED vs. ED	143					143	100				
ELN vs. ELY	143	11	8	26	18	106	74				
SWDN vs. SWDY	143			4	3	139	97				

Table C.7. Pre-Administration DIF Results—Mathematics Grade 3

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	156	1	1	6	4	144	92	5	3		
White vs. Black/African American	156			21	13	133	85	2	1		
White vs. Hispanic/Latino	156			5	3	151	97				
White vs. Asian	156					152	97	4	3		
White vs. AI/AN	156					156	100				
White vs. NH/PI	156					156	100				
White vs. Multiracial	156					155	99	1	1		
NED vs. ED	156					156	100				
ELN vs. ELY	156			5	3	151	97				
SWDN vs. SWDY	156			5	3	148	95	3	2		

Table C.8. Pre-Administration DIF Results—Mathematics Grade 4

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	153	1	1	7	5	144	94	1	1		
White vs. Black/African American	153	1	1	8	5	141	92	3	2		
White vs. Hispanic/Latino	153			4	3	148	97	1	1		
White vs. Asian	153					151	99	1	1	1	1
White vs. AI/AN	153					153	100				
White vs. NH/PI	153					153	100				
White vs. Multiracial	153					153	100				
NED vs. ED	153					153	100				
ELN vs. ELY	153	1	1	5	3	146	95	1	1		
SWDN vs. SWDY	153			4	3	145	95	4	3		

Table C.9. Pre-Administration DIF Results—Mathematics Grade 5

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	160	1	1	6	4	153	96				
White vs. Black/African American	160	1	1	17	11	140	88	2	1		
White vs. Hispanic/Latino	160			3	2	157	98				
White vs. Asian	160					150	94	8	5	2	1
White vs. AI/AN	160					160	100				
White vs. NH/PI	160					160	100				
White vs. Multiracial	160					160	100				
NED vs. ED	160					160	100				
ELN vs. ELY	160			10	6	149	93	1	1		
SWDN vs. SWDY	160			2	1	158	99				

Table C.10. Pre-Administration DIF Results—Mathematics Grade 6

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	164	2	1	7	4	155	95				
White vs. Black/African American	164	1	1	18	11	143	87	2	1		
White vs. Hispanic/Latino	164	1	1	9	5	153	93	1	1		
White vs. Asian	164					157	96	7	4		
White vs. AI/AN	164					164	100				
White vs. NH/PI	164					164	100				
White vs. Multiracial	164					164	100				
NED vs. ED	164					164	100				
ELN vs. ELY	164	2	1	6	4	155	95			1	1
SWDN vs. SWDY	164			4	2	160	98				

Table C.11. Pre-Administration DIF Results—Mathematics Grade 7

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	155	3	2	8	5	144	93				
White vs. Black/African American	155			10	6	145	94				
White vs. Hispanic/Latino	155			10	6	145	94				
White vs. Asian	155					145	94	6	4	4	3
White vs. AI/AN	155					155	100				
White vs. NH/PI	155					155	100				
White vs. Multiracial	155					155	100				
NED vs. ED	155					155	100				
ELN vs. ELY	155			10	6	144	93	1	1		
SWDN vs. SWDY	155			2	1	149	96	3	2	1	1

Table C.12. Pre-Administration DIF Results—Mathematics Grade 8

DIF Comparison	Total #Unique Items	C-		B-		A		B+		C+	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	148	2	1	4	3	141	95	1	1		
White vs. Black/African American	148	1	1	14	9	132	89	1	1		
White vs. Hispanic/Latino	148			2	1	146	99				
White vs. Asian	148					138	93	6	4	4	3
White vs. AI/AN	148					148	100				
White vs. NH/PI	148					148	100				
White vs. Multiracial	148					148	100				
NED vs. ED	148					148	100				
ELN vs. ELY	148	1	1	10	7	135	91	2	1		
SWDN vs. SWDY	148			2	1	143	97	3	2		

Appendix D: TCCs, CSEM Curves, and TIF Curves

This appendix presents the pre-equated IRT test characteristic curves (TCCs), conditional standard error of measurement (CSEM) curves, and test information function (TIF) curves by content area and grade.

Figure D.1. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 3

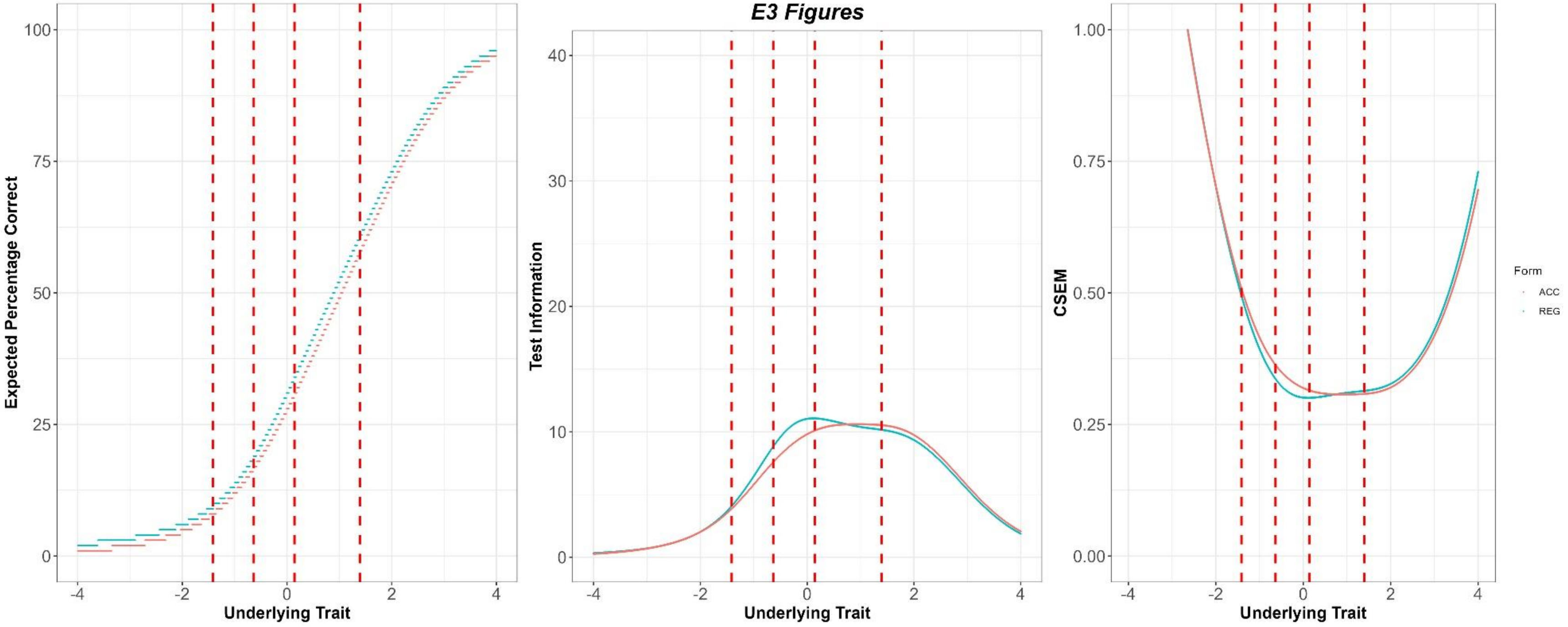


Figure D.2. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 4

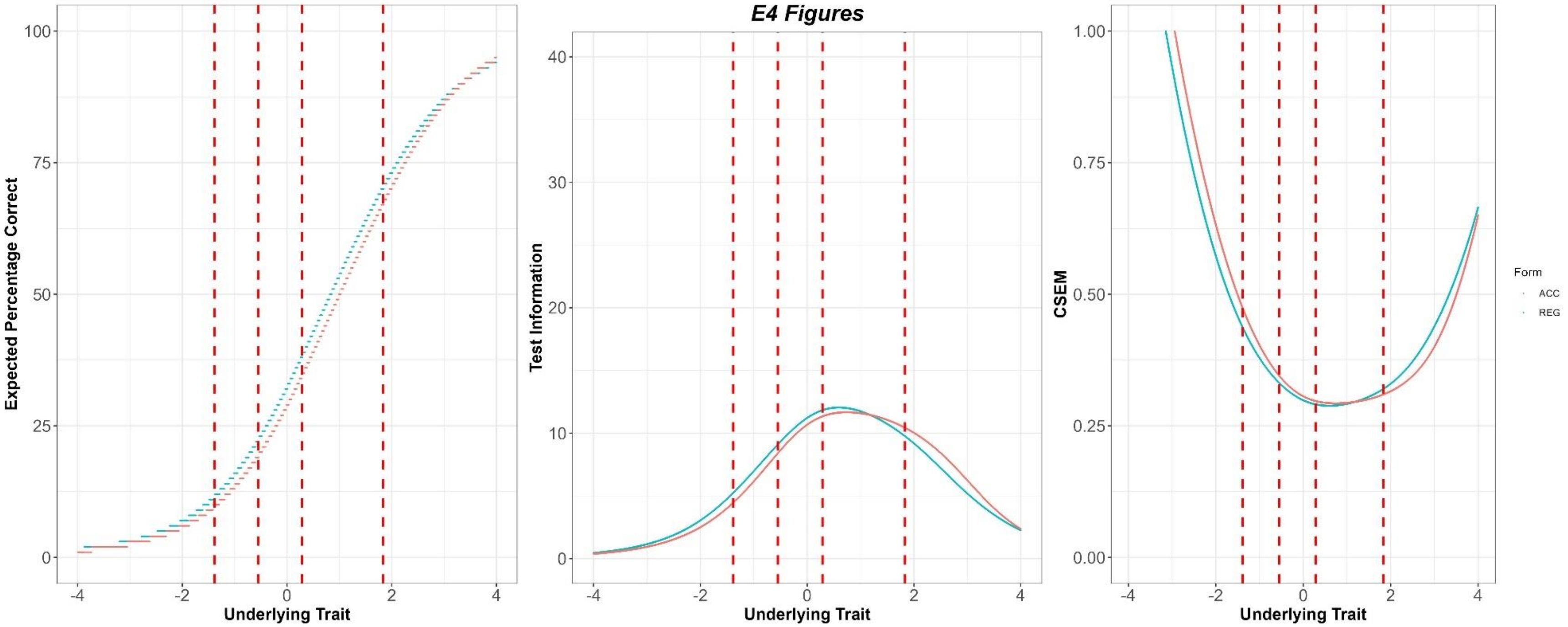


Figure D.3. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 5

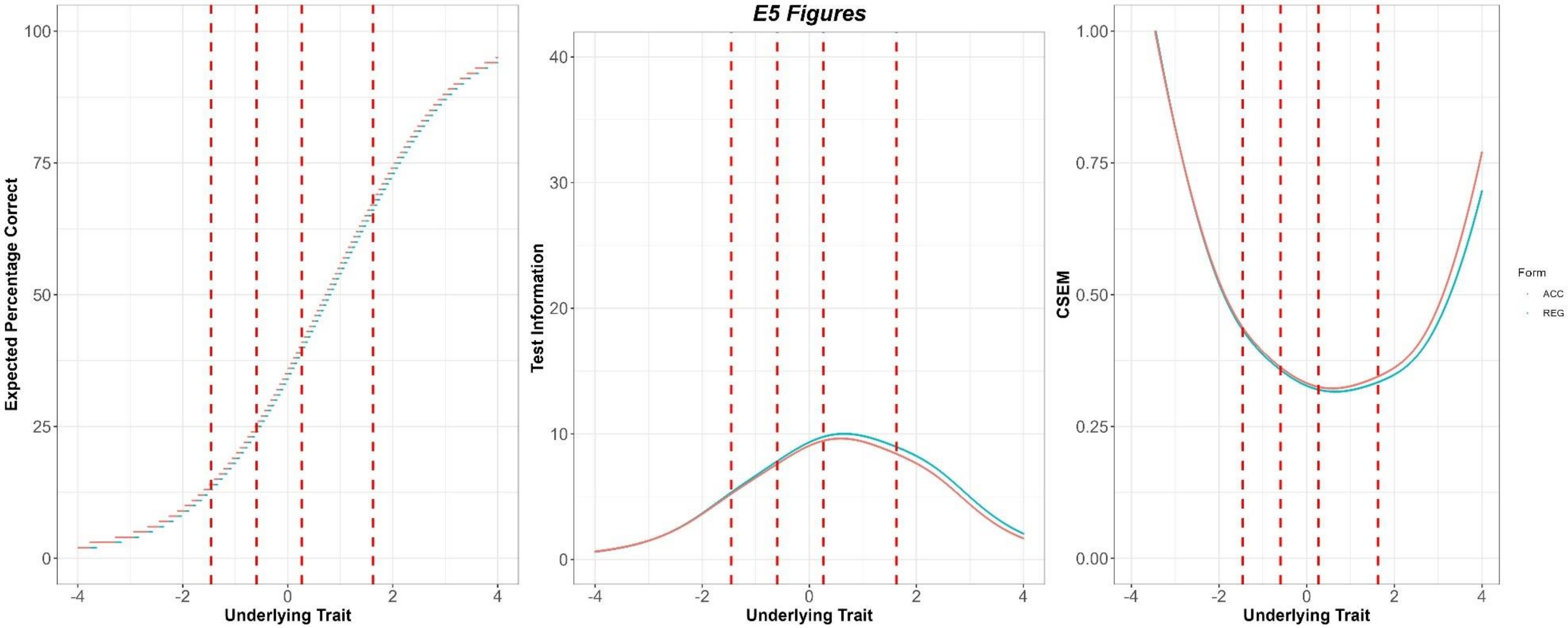


Figure D.4. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 6

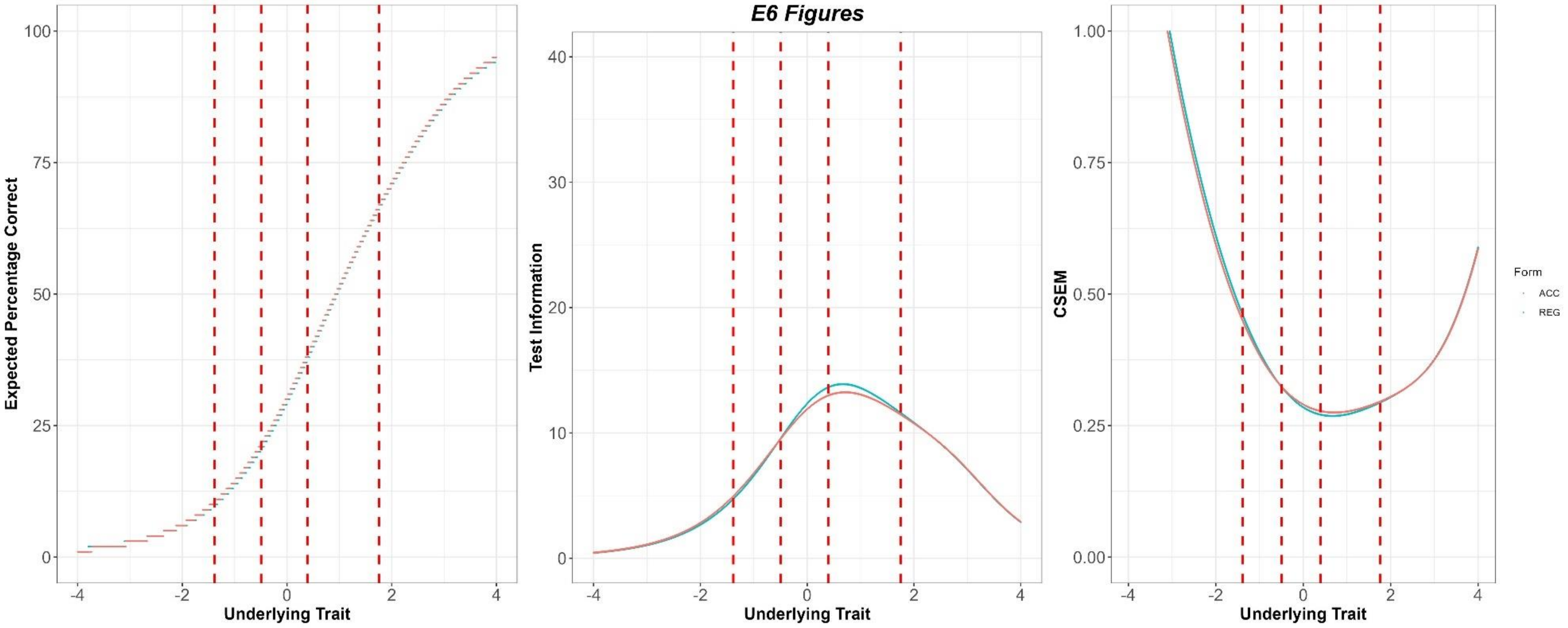


Figure D.5. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 7

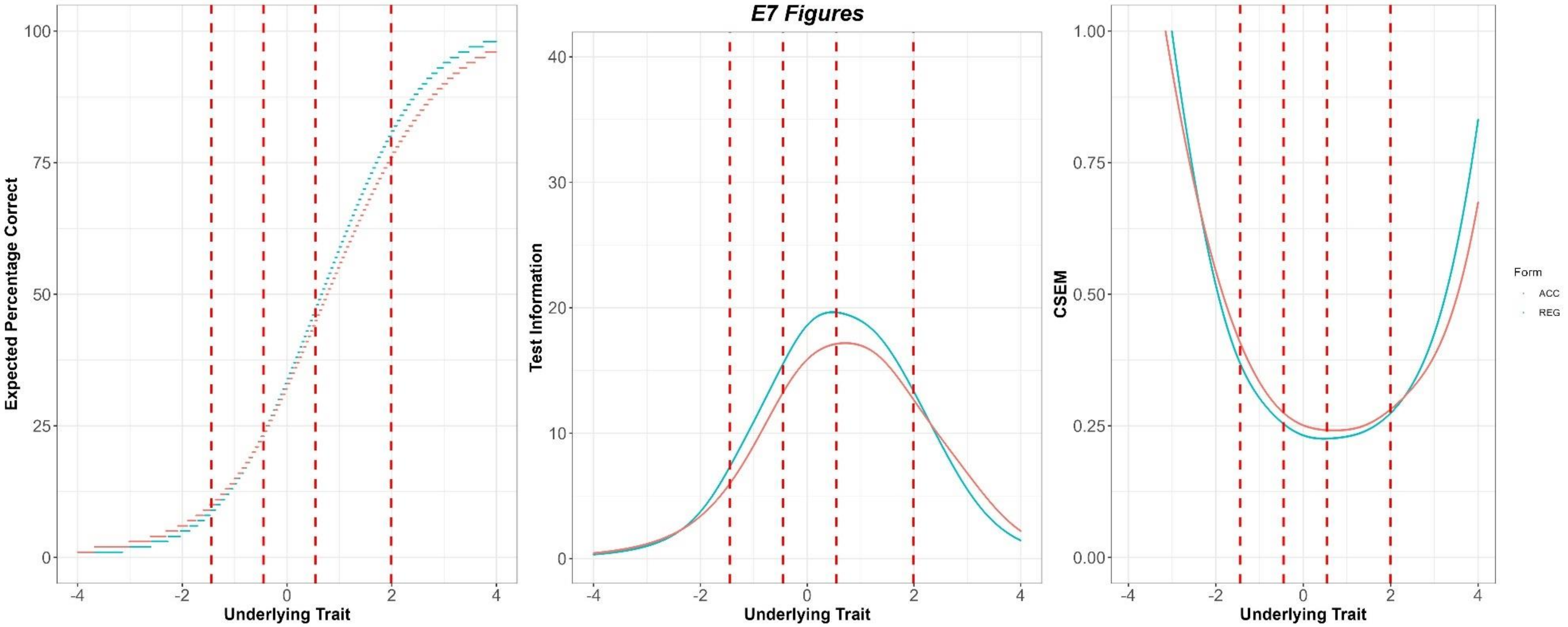


Figure D.6. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 8

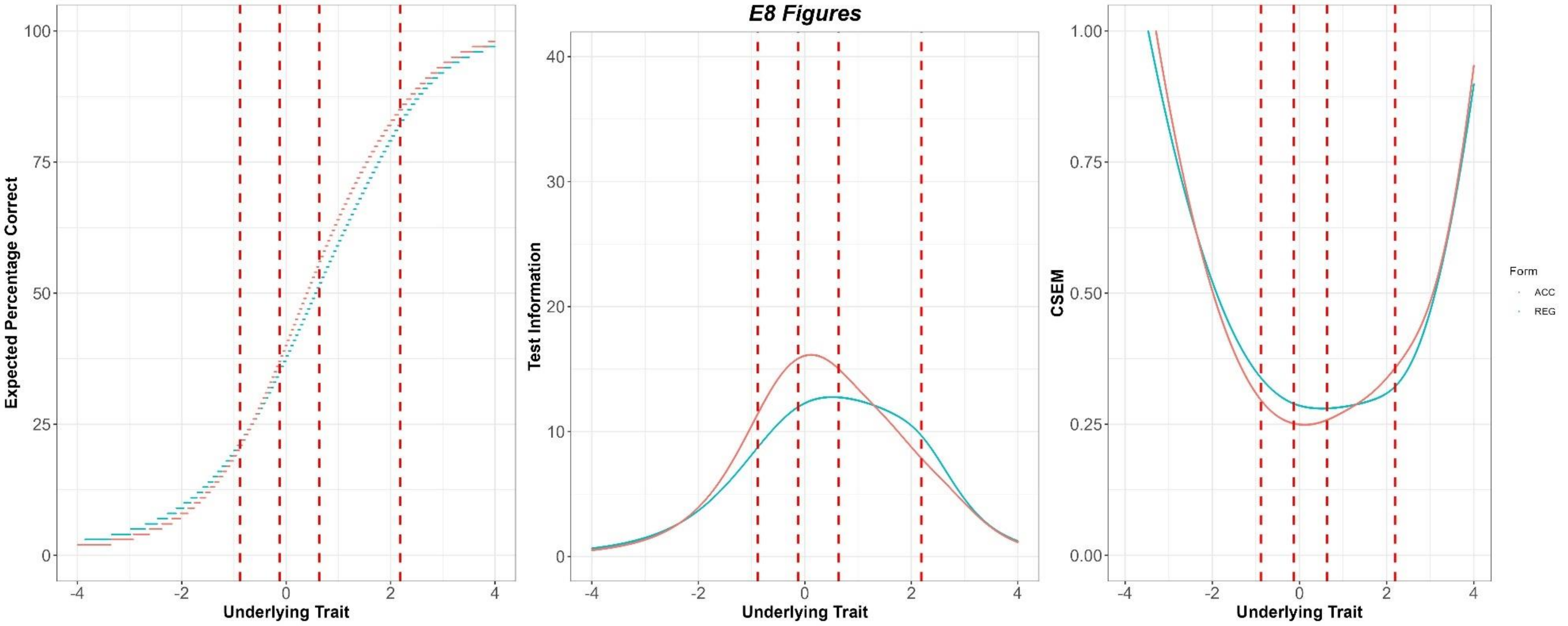


Figure D.7. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 3

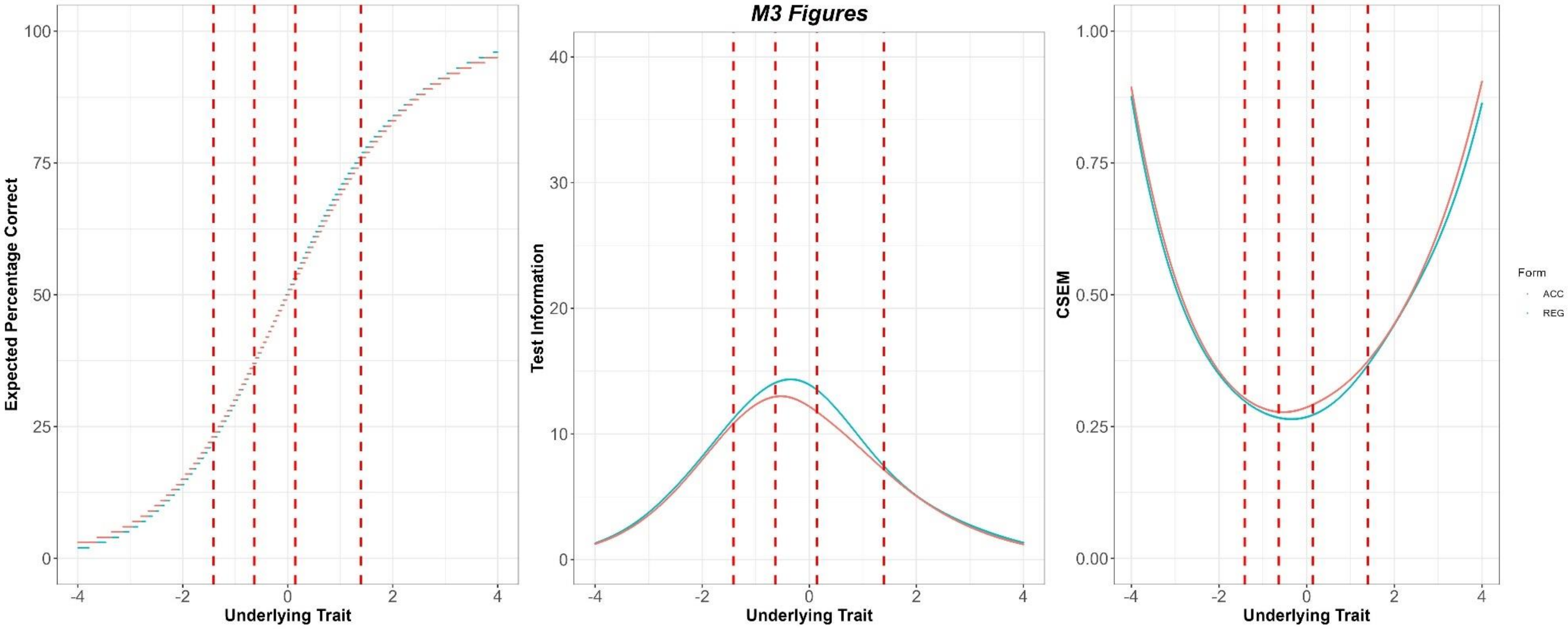


Figure D.8. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 4

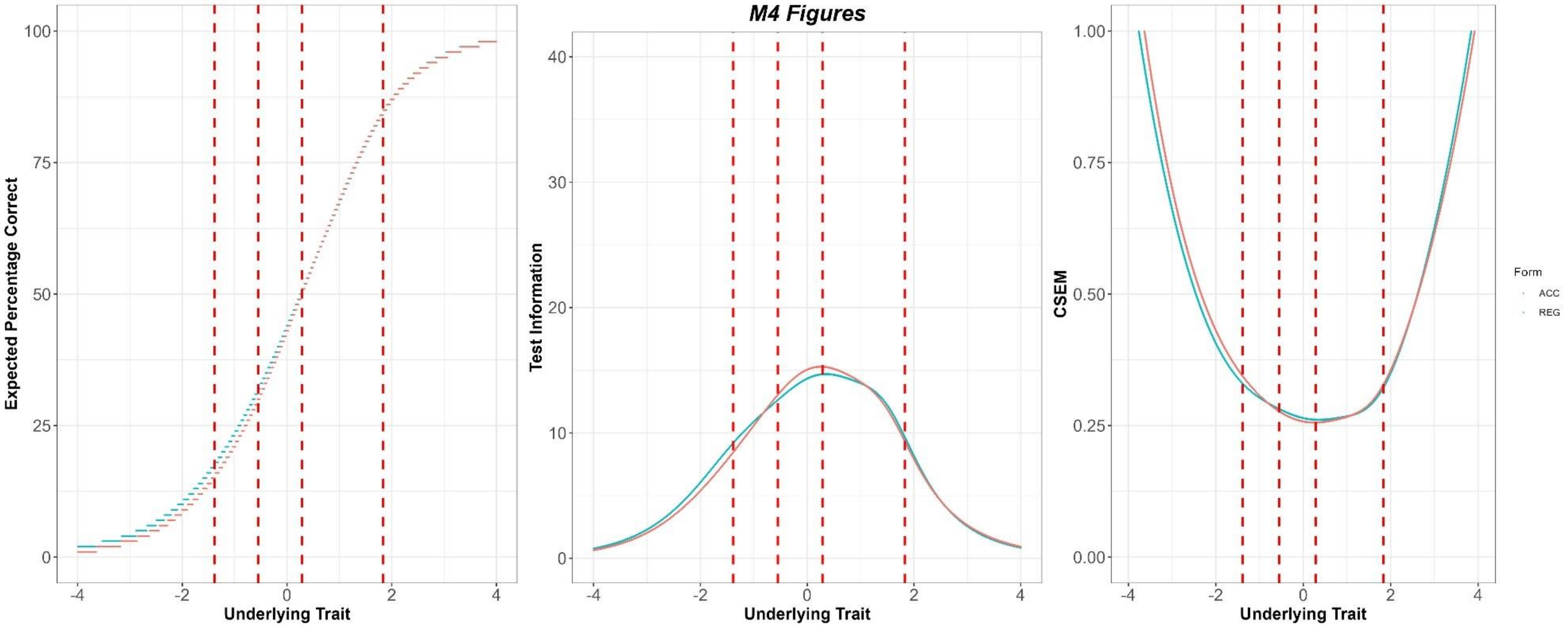


Figure D.9. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 5

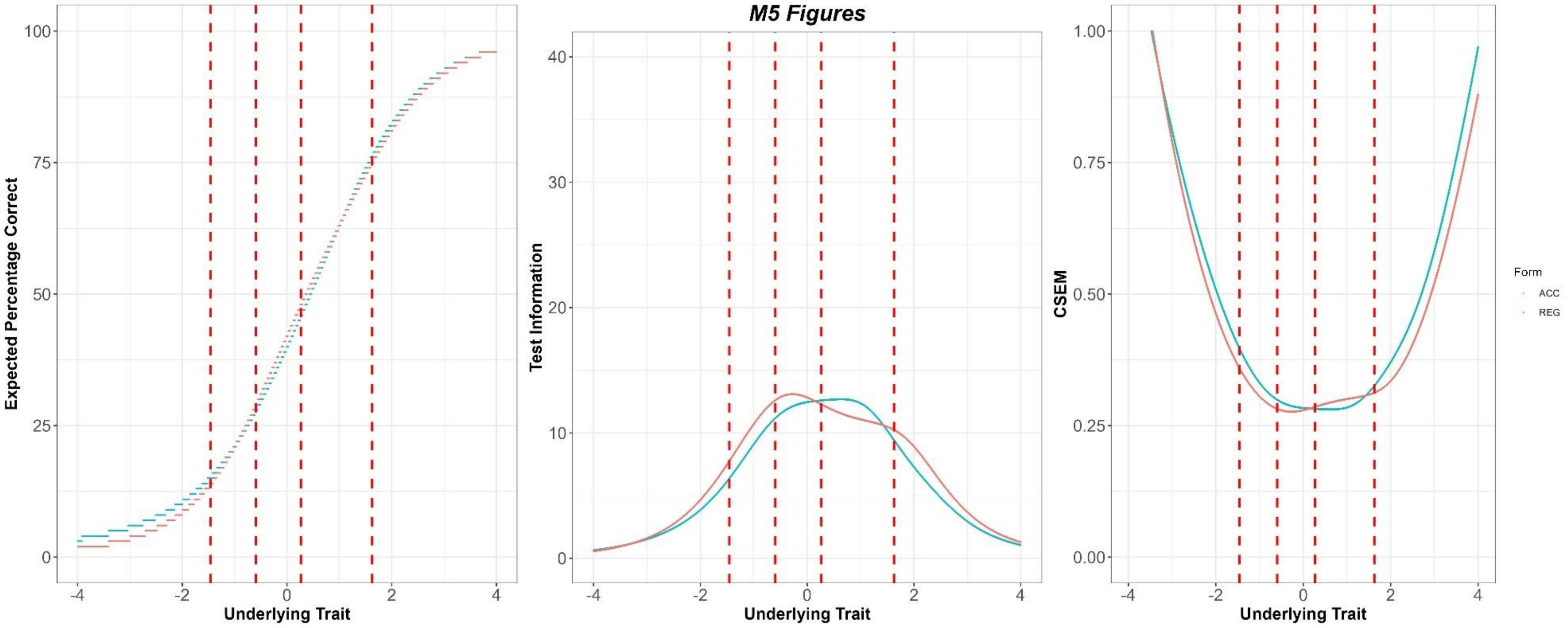


Figure D.10. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 6

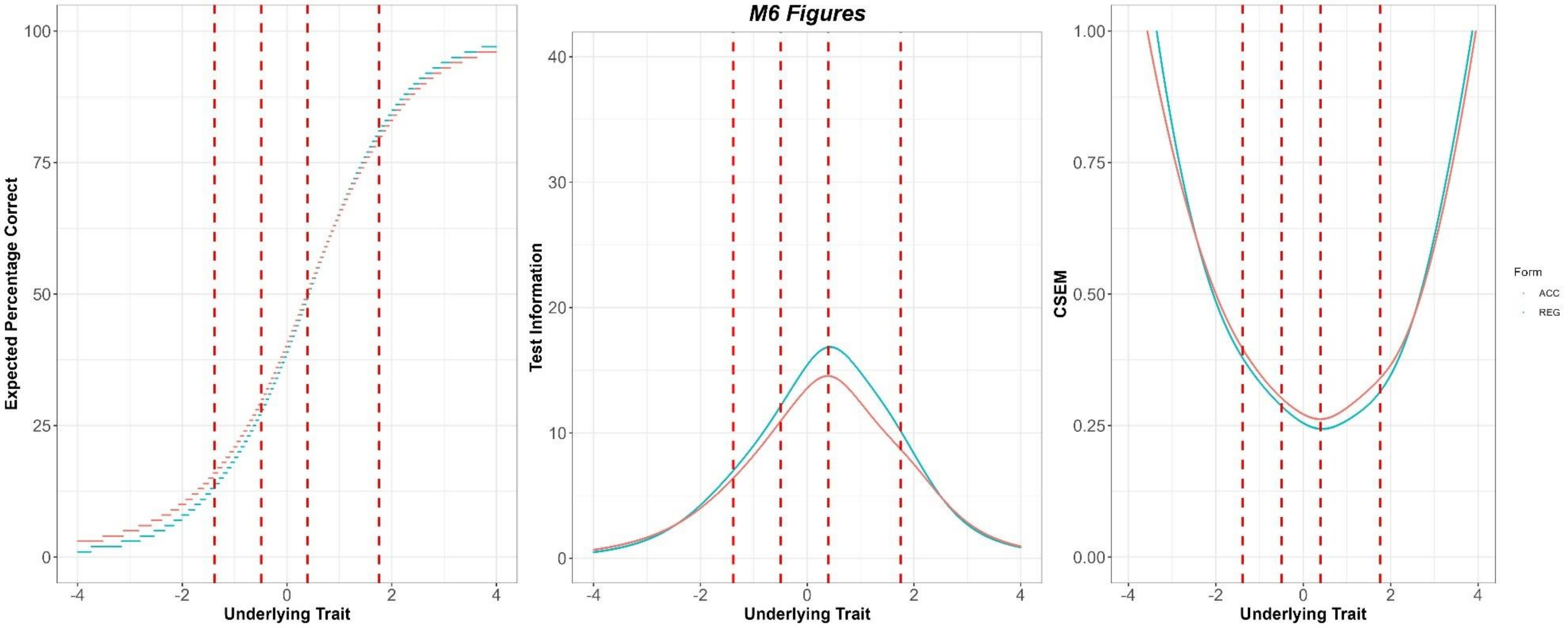


Figure D.11. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 7

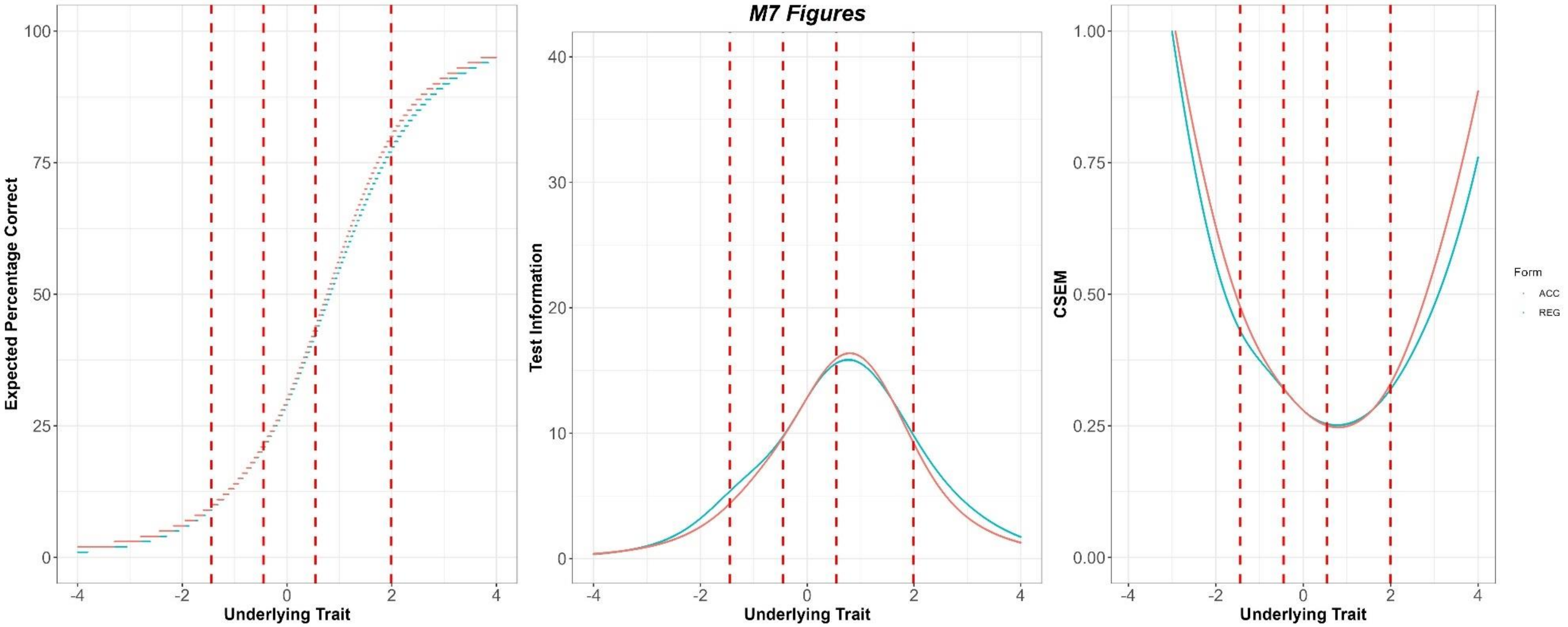
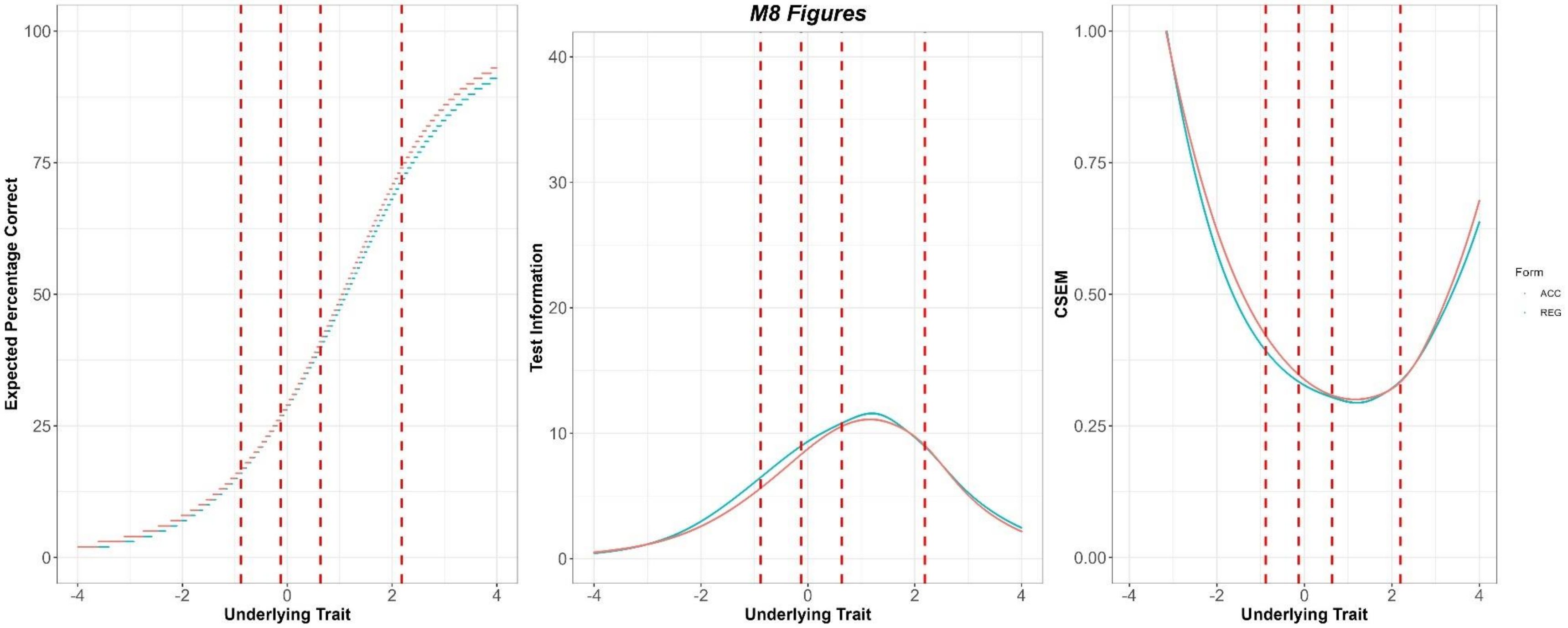


Figure D.12. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 8



Appendix E: Reliability by Subgroup

Table E.1. Test Reliability Estimates by Subgroup—ELA/L Grade 3

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1 N	ACC1 Alpha	Online1 N	Online1 Alpha
Total Group	54	3.06	0.87	812	0.85	128,697	0.90
Male	54	3.01	0.87	482	0.85	65,245	0.90
Female	54	3.11	0.87	330	0.84	63,438	0.90
Black/African American	54	n/a	n/a	n/a	n/a	20,847	0.89
Asian/Pacific Islander	54	n/a	n/a	n/a	n/a	7,625	0.90
Hispanic/Latino	54	2.83	0.83	236	0.78	35,645	0.89
American Indian/Alaska Native	54	n/a	n/a	n/a	n/a	336	0.89
Multiple	54	n/a	n/a	n/a	n/a	6,301	0.91
White	54	3.22	0.87	397	0.85	57,772	0.89
Economically Disadvantaged	54	2.87	0.85	446	0.81	66,784	0.89
Not Economically Disadvantaged	54	3.24	0.88	358	0.87	61,300	0.89
English Learner (EL)	54	2.74	0.79	199	0.71	27,944	0.87
Non-EL	54	3.14	0.88	612	0.86	100,506	0.90
Students with Disabilities (SWD)	54	2.76	0.86	650	0.83	21,633	0.89
Students without Disabilities	54	3.28	0.87	161	0.85	105,840	0.90
American Sign Language (ASL)	54	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	54	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	54	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	54	n/a	n/a	n/a	n/a	3,179	0.75

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.2. Test Reliability Estimates by Subgroup—ELA/L Grade 4

	Max. Raw	Avg. Raw	Avg.	ACC1		Online1	
Subgroup	Score	Score SEM	Reliability	N	Alpha	N	Alpha
Total Group	67	3.70	0.89	856	0.87	128,699	0.91
Male	67	3.58	0.89	521	0.87	65,552	0.91
Female	67	3.84	0.89	334	0.87	63,128	0.91
Black/African American	67	3.27	0.82	122	0.76	20,816	0.88
Asian/Pacific Islander	67	n/a	n/a	n/a	n/a	7,519	0.90
Hispanic/Latino	67	3.45	0.86	236	0.82	35,447	0.90
American Indian/Alaska Native	67	n/a	n/a	n/a	n/a	331	0.89
Multiple	67	n/a	n/a	n/a	n/a	6,035	0.91
White	67	3.89	0.89	425	0.88	58,396	0.89
Economically Disadvantaged	67	3.47	0.86	474	0.83	66,318	0.89
Not Economically Disadvantaged	67	3.91	0.89	371	0.88	61,678	0.90
English Learner (EL)	67	3.26	0.80	182	0.73	25,609	0.87
Non-EL	67	3.79	0.89	670	0.88	102,809	0.90
Students with Disabilities (SWD)	67	3.38	0.87	717	0.84	23,038	0.90
Students without Disabilities	67	3.92	0.90	137	0.90	104,548	0.90
American Sign Language (ASL)	67	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	67	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	67	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	67	n/a	n/a	n/a	n/a	3,673	0.78

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.3. Test Reliability Estimates by Subgroup—ELA/L Grade 5

	Max. Raw	Avg. Raw	Avg.	ACC1		Online1	
Subgroup	Score	Score SEM	Reliability	N	Alpha	N	Alpha
Total Group	67	3.83	0.88	804	0.86	128,296	0.90
Male	67	3.71	0.88	499	0.86	65,336	0.90
Female	67	3.96	0.88	305	0.86	62,934	0.90
Black/African American	67	3.38	0.83	121	0.78	20,607	0.88
Asian/Pacific Islander	67	n/a	n/a	n/a	n/a	7,606	0.90
Hispanic/Latino	67	3.63	0.84	269	0.80	35,423	0.89
American Indian/Alaska Native	67	n/a	n/a	n/a	n/a	307	0.91
Multiple	67	n/a	n/a	n/a	n/a	5,804	0.90
White	67	4.00	0.88	352	0.87	58,363	0.89
Economically Disadvantaged	67	3.58	0.83	427	0.78	66,175	0.89
Not Economically Disadvantaged	67	4.04	0.88	366	0.88	61,521	0.89
English Learner (EL)	67	3.38	0.80	225	0.75	21,380	0.84
Non-EL	67	3.91	0.88	575	0.87	106,650	0.89
Students with Disabilities (SWD)	67	3.53	0.87	717	0.85	23,207	0.90
Students without Disabilities	67	n/a	n/a	n/a	n/a	104,100	0.89
American Sign Language (ASL)	67	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	67	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	67	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	67	n/a	n/a	n/a	n/a	3,791	0.76

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.4. Test Reliability Estimates by Subgroup—ELA/L Grade 6

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	74	3.92	0.90	745	0.88	129,822	0.92
Male	74	3.78	0.89	471	0.87	65,970	0.92
Female	74	4.08	0.90	273	0.89	63,820	0.91
Black/African American	74	3.58	0.86	136	0.82	20,654	0.90
Asian/Pacific Islander	74	n/a	n/a	n/a	n/a	7,700	0.91
Hispanic/Latino	74	3.71	0.88	234	0.85	36,364	0.91
American Indian/Alaska Native	74	n/a	n/a	n/a	n/a	289	0.91
Multiple	74	n/a	n/a	n/a	n/a	5,838	0.92
White	74	4.09	0.89	315	0.88	58,794	0.90
Economically Disadvantaged	74	3.75	0.88	418	0.86	66,052	0.91
Not Economically Disadvantaged	74	4.05	0.90	318	0.89	63,055	0.90
English Learner (EL)	74	3.47	0.84	207	0.81	19,259	0.86
Non-EL	74	4.00	0.90	536	0.89	110,295	0.91
Students with Disabilities (SWD)	74	3.66	0.89	695	0.87	23,321	0.92
Students without Disabilities	74	n/a	n/a	n/a	n/a	105,551	0.91
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	n/a	n/a	3,643	0.84

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.5. Test Reliability Estimates by Subgroup—ELA/L Grade 7

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	74	3.90	0.90	565	0.88	132,798	0.92
Male	74	3.82	0.90	359	0.88	67,767	0.92
Female	74	3.97	0.90	206	0.88	64,998	0.92
Black/African American	74	3.59	0.87	100	0.84	21,164	0.91
Asian/Pacific Islander	74	n/a	n/a	n/a	n/a	7,675	0.93
Hispanic/Latino	74	3.64	0.88	171	0.84	37,661	0.92
American Indian/Alaska Native	74	n/a	n/a	n/a	n/a	272	0.93
Multiple	74	n/a	n/a	n/a	n/a	5,660	0.92
White	74	4.07	0.90	244	0.88	60,208	0.92
Economically Disadvantaged	74	3.69	0.89	334	0.87	66,793	0.91
Not Economically Disadvantaged	74	4.10	0.90	224	0.88	65,319	0.92
English Learner (EL)	74	3.38	0.83	135	0.79	20,546	0.87
Non-EL	74	3.99	0.90	427	0.88	111,969	0.92
Students with Disabilities (SWD)	74	3.60	0.89	496	0.86	23,679	0.92
Students without Disabilities	74	n/a	n/a	n/a	n/a	108,190	0.92
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	n/a	n/a	3,747	0.82

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.6. Test Reliability Estimates by Subgroup—ELA/L Grade 8

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	70	4.17	0.89	496	0.88	134,182	0.90
Male	70	3.99	0.88	312	0.87	68,928	0.90
Female	70	4.35	0.89	184	0.89	65,206	0.89
Black/African American	70	n/a	n/a	n/a	n/a	21,925	0.88
Asian/Pacific Islander	70	n/a	n/a	n/a	n/a	7,695	0.88
Hispanic/Latino	70	3.98	0.87	172	0.84	38,123	0.90
American Indian/Alaska Native	70	n/a	n/a	n/a	n/a	279	0.92
Multiple	70	n/a	n/a	n/a	n/a	5,495	0.90
White	70	4.37	0.89	196	0.90	60,494	0.88
Economically Disadvantaged	70	3.94	0.86	278	0.82	67,589	0.89
Not Economically Disadvantaged	70	4.34	0.89	215	0.90	66,009	0.88
English Learner (EL)	70	3.66	0.82	128	0.77	19,892	0.87
Non-EL	70	4.24	0.89	367	0.89	114,031	0.89
Students with Disabilities (SWD)	70	3.82	0.89	443	0.87	23,889	0.90
Students without Disabilities	70	n/a	n/a	n/a	n/a	109,479	0.89
American Sign Language (ASL)	70	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	70	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	70	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	70	n/a	n/a	n/a	n/a	3,647	0.81

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.7. Test Reliability Estimates by Subgroup—Mathematics Grade 3

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	52	2.98	0.91	656	0.90	129,200	0.92
Male	52	2.97	0.92	383	0.91	65,543	0.92
Female	52	2.99	0.89	273	0.87	63,642	0.92
Black/African American	52	2.78	0.89	110	0.87	20,935	0.91
Asian/Pacific Islander	52	n/a	n/a	n/a	n/a	7,640	0.92
Hispanic/Latino	52	2.85	0.89	140	0.87	35,799	0.91
American Indian/Alaska Native	52	n/a	n/a	n/a	n/a	342	0.92
Multiple	52	n/a	n/a	n/a	n/a	6,335	0.92
White	52	3.06	0.90	359	0.90	57,865	0.91
Economically Disadvantaged	52	2.89	0.90	395	0.88	67,006	0.91
Not Economically Disadvantaged	52	3.06	0.91	259	0.91	61,460	0.91
English Learner (EL)	52	2.80	0.90	128	0.89	28,072	0.90
Non-EL	52	3.02	0.91	527	0.90	100,768	0.92
Students with Disabilities (SWD)	52	2.87	0.91	421	0.91	21,972	0.92
Students without Disabilities	52	3.03	0.89	232	0.86	105,888	0.91
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	n/a	n/a	48,380	0.92

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.8. Test Reliability Estimates by Subgroup—Mathematics Grade 4

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1 N	ACC1 Alpha	Online1 N	Online1 Alpha
Total Group	52	3.18	0.90	577	0.89	129,290	0.92
Male	52	3.17	0.91	332	0.89	65,898	0.92
Female	52	3.18	0.89	245	0.88	63,372	0.91
Black/African American	52	n/a	n/a	n/a	n/a	20,934	0.89
Asian/Pacific Islander	52	n/a	n/a	n/a	n/a	7,547	0.92
Hispanic/Latino	52	n/a	n/a	n/a	n/a	35,637	0.89
American Indian/Alaska Native	52	n/a	n/a	n/a	n/a	330	0.91
Multiple	52	n/a	n/a	n/a	n/a	6,075	0.92
White	52	3.26	0.89	369	0.88	58,516	0.91
Economically Disadvantaged	52	3.01	0.89	337	0.88	66,598	0.89
Not Economically Disadvantaged	52	3.31	0.90	234	0.89	61,889	0.91
English Learner (EL)	52	n/a	n/a	n/a	n/a	25,759	0.88
Non-EL	52	3.23	0.90	494	0.88	103,156	0.92
Students with Disabilities (SWD)	52	2.97	0.89	367	0.87	23,495	0.91
Students without Disabilities	52	3.24	0.90	206	0.88	104,599	0.91
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	n/a	n/a	49,110	0.91

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.9. Test Reliability Estimates by Subgroup—Mathematics Grade 5

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1 N	ACC1 Alpha	Online1 N	Online1 Alpha
Total Group	52	3.06	0.90	379	0.88	129,009	0.92
Male	52	3.06	0.90	230	0.89	65,763	0.92
Female	52	3.05	0.88	149	0.86	63,220	0.91
Black/African American	52	n/a	n/a	n/a	n/a	20,767	0.86
Asian/Pacific Islander	52	n/a	n/a	n/a	n/a	7,626	0.93
Hispanic/Latino	52	n/a	n/a	n/a	n/a	35,660	0.88
American Indian/Alaska Native	52	n/a	n/a	n/a	n/a	308	0.92
Multiple	52	n/a	n/a	n/a	n/a	5,836	0.92
White	52	3.20	0.89	229	0.88	58,521	0.91
Economically Disadvantaged	52	2.92	0.87	198	0.85	66,521	0.88
Not Economically Disadvantaged	52	3.19	0.90	174	0.89	61,774	0.91
English Learner (EL)	52	n/a	n/a	n/a	n/a	21,563	0.82
Non-EL	52	3.13	0.90	298	0.88	107,073	0.92
Students with Disabilities (SWD)	52	2.82	0.87	300	0.84	23,710	0.90
Students without Disabilities	52	n/a	n/a	n/a	n/a	104,221	0.91
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	n/a	n/a	47,392	0.91

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.10. Test Reliability Estimates by Subgroup—Mathematics Grade 6

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	50	3.08	0.90	324	0.89	130,561	0.92
Male	50	3.08	0.91	193	0.90	66,426	0.92
Female	50	3.07	0.89	131	0.87	64,102	0.91
Black/African American	50	n/a	n/a	n/a	n/a	20,833	0.87
Asian/Pacific Islander	50	n/a	n/a	n/a	n/a	7,712	0.93
Hispanic/Latino	50	n/a	n/a	n/a	n/a	36,596	0.89
American Indian/Alaska Native	50	n/a	n/a	n/a	n/a	292	0.91
Multiple	50	n/a	n/a	n/a	n/a	5,866	0.92
White	50	3.20	0.90	204	0.89	58,974	0.91
Economically Disadvantaged	50	2.94	0.89	198	0.89	66,414	0.89
Not Economically Disadvantaged	50	3.20	0.90	123	0.90	63,320	0.91
English Learner (EL)	50	n/a	n/a	n/a	n/a	19,451	0.82
Non-EL	50	3.13	0.90	274	0.89	110,737	0.92
Students with Disabilities (SWD)	50	2.83	0.89	243	0.88	23,829	0.91
Students without Disabilities	50	n/a	n/a	n/a	n/a	105,688	0.91
American Sign Language (ASL)	50	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	50	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	50	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	50	n/a	n/a	n/a	n/a	45,768	0.91

Note. AI/AN = American Indian/Alaska Native, SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.11. Test Reliability Estimates by Subgroup—Mathematics Grade 7

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	52	2.93	0.91	306	0.89	133,568	0.92
Male	52	2.87	0.90	170	0.88	68,214	0.93
Female	52	2.98	0.91	136	0.90	65,320	0.92
Black/African American	52	n/a	n/a	n/a	n/a	21,352	0.87
Asian/Pacific Islander	52	n/a	n/a	n/a	n/a	7,689	0.93
Hispanic/Latino	52	n/a	n/a	n/a	n/a	37,903	0.89
American Indian/Alaska Native	52	n/a	n/a	n/a	n/a	275	0.92
Multiple	52	n/a	n/a	n/a	n/a	5,703	0.93
White	52	3.08	0.90	198	0.90	60,387	0.91
Economically Disadvantaged	52	2.73	0.88	182	0.87	67,205	0.89
Not Economically Disadvantaged	52	3.13	0.91	119	0.91	65,570	0.92
English Learner (EL)	52	n/a	n/a	n/a	n/a	20,708	0.83
Non-EL	52	3.01	0.90	258	0.89	112,476	0.92
Students with Disabilities (SWD)	52	2.57	0.88	225	0.85	24,065	0.91
Students without Disabilities	52	n/a	n/a	n/a	n/a	108,490	0.92
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	n/a	n/a	45,827	0.92

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table E.12. Test Reliability Estimates by Subgroup—Mathematics Grade 8

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	ACC1		Online1	
				N	Alpha	N	Alpha
Total Group	50	2.70	0.88	234	0.84	134,969	0.91
Male	50	2.63	0.87	140	0.82	69,364	0.92
Female	50	n/a	n/a	n/a	n/a	65,556	0.91
Black/African American	50	n/a	n/a	n/a	n/a	22,082	0.86
Asian/Pacific Islander	50	n/a	n/a	n/a	n/a	7,715	0.93
Hispanic/Latino	50	n/a	n/a	n/a	n/a	38,397	0.88
American Indian/Alaska Native	50	n/a	n/a	n/a	n/a	278	0.91
Multiple	50	n/a	n/a	n/a	n/a	5,523	0.92
White	50	2.83	0.88	131	0.85	60,695	0.91
Economically Disadvantaged	50	2.57	0.80	128	0.72	67,991	0.88
Not Economically Disadvantaged	50	2.82	0.90	104	0.89	66,286	0.91
English Learner (EL)	50	n/a	n/a	n/a	n/a	20,081	0.80
Non-EL	50	2.76	0.88	198	0.85	114,523	0.91
Students with Disabilities (SWD)	50	2.42	0.85	173	0.80	24,278	0.89
Students without Disabilities	50	n/a	n/a	n/a	n/a	109,780	0.91
American Sign Language (ASL)	50	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	50	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	50	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	50	n/a	n/a	n/a	n/a	45,377	0.91

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Appendix F: Decision Accuracy and Consistency by Performance Level

Table F.1. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 3

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.25**	0.04	0.00	0.00	0.00	0.28
	700–724	0.05	**0.11**	0.04	0.00	0.00	0.21
	725–749	0.00	0.05	**0.10**	0.05	0.00	0.20
	750–809	0.00	0.00	0.04	**0.24**	0.02	0.31
	810–850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision Consistency	650–699	**0.23**	0.05	0.01	0.00	0.00	0.30
	700–724	0.05	**0.08**	0.05	0.01	0.00	0.19
	725–749	0.01	0.05	**0.07**	0.06	0.00	0.19
	750–809	0.00	0.01	0.06	**0.21**	0.01	0.30
	810–850	0.00	0.00	0.00	0.01	**0.00**	0.02

Table F.2. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 4

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.17**	0.03	0.00	0.00	0.00	0.20
	700–724	0.04	**0.12**	0.04	0.00	0.00	0.20
	725–749	0.00	0.05	**0.13**	0.05	0.00	0.23
	750–789	0.00	0.00	0.05	**0.24**	0.04	0.33
	790–850	0.00	0.00	0.00	0.02	**0.03**	0.05
Decision Consistency	650–699	**0.16**	0.04	0.01	0.00	0.00	0.21
	700–724	0.04	**0.09**	0.05	0.01	0.00	0.19
	725–749	0.01	0.05	**0.10**	0.06	0.00	0.22
	750–789	0.00	0.01	0.06	**0.20**	0.04	0.31
	790–850	0.00	0.00	0.00	0.04	**0.03**	0.07

Table F.3. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 5

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.14**	0.02	0.00	0.00	0.00	0.17
	700–724	0.04	**0.12**	0.04	0.00	0.00	0.20
	725–749	0.00	0.05	**0.14**	0.05	0.00	0.24
	750–798	0.00	0.00	0.05	**0.31**	0.03	0.39
	799–850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision Consistency	650–699	**0.14**	0.04	0.00	0.00	0.00	0.18
	700–724	0.04	**0.10**	0.05	0.01	0.00	0.20
	725–749	0.00	0.05	**0.11**	0.06	0.00	0.23
	750–798	0.00	0.01	0.06	**0.27**	0.02	0.36
	799–850	0.00	0.00	0.00	0.02	**0.01**	0.03

Table F.4. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 6

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.10**	0.02	0.00	0.00	0.00	0.12
	700–724	0.03	**0.12**	0.03	0.00	0.00	0.18
	725–749	0.00	0.04	**0.16**	0.05	0.00	0.24
	750–789	0.00	0.00	0.05	**0.33**	0.04	0.43
	790–850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650–699	**0.10**	0.03	0.00	0.00	0.00	0.13
	700–724	0.03	**0.10**	0.05	0.00	0.00	0.18
	725–749	0.00	0.05	**0.13**	0.06	0.00	0.24
	750–789	0.00	0.00	0.06	**0.29**	0.04	0.39
	790–850	0.00	0.00	0.00	0.04	**0.02**	0.06

Table F.5. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 7

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.09**	0.02	0.00	0.00	0.00	0.10
	700–724	0.03	**0.14**	0.04	0.00	0.00	0.21
	725–749	0.00	0.05	**0.18**	0.05	0.00	0.28
	750–784	0.00	0.00	0.05	**0.23**	0.03	0.31
	785–850	0.00	0.00	0.00	0.02	**0.08**	0.10
Decision Consistency	650–699	**0.08**	0.03	0.00	0.00	0.00	0.12
	700–724	0.03	**0.11**	0.06	0.00	0.00	0.21
	725–749	0.00	0.06	**0.14**	0.06	0.00	0.26
	750–784	0.00	0.00	0.06	**0.19**	0.04	0.30
	785–850	0.00	0.00	0.00	0.04	**0.08**	0.12

Table F.6. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 8

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.12**	0.02	0.00	0.00	0.00	0.14
	700–724	0.03	**0.10**	0.03	0.00	0.00	0.16
	725–749	0.00	0.04	**0.12**	0.05	0.00	0.21
	750–793	0.00	0.00	0.05	**0.32**	0.05	0.42
	794–850	0.00	0.00	0.00	0.03	**0.05**	0.07
Decision Consistency	650–699	**0.12**	0.03	0.00	0.00	0.00	0.15
	700–724	0.03	**0.08**	0.04	0.01	0.00	0.16
	725–749	0.00	0.04	**0.09**	0.06	0.00	0.20
	750–793	0.00	0.01	0.06	**0.27**	0.05	0.39
	794–850	0.00	0.00	0.00	0.05	**0.05**	0.10

Table F.7. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 3

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.17**	0.03	0.00	0.00	0.00	0.19
	700–724	0.04	**0.15**	0.04	0.00	0.00	0.23
	725–749	0.00	0.05	**0.15**	0.04	0.00	0.24
	750–789	0.00	0.00	0.04	**0.20**	0.02	0.27
	790–850	0.00	0.00	0.00	0.01	**0.05**	0.06
Decision Consistency	650–699	**0.16**	0.04	0.00	0.00	0.00	0.21
	700–724	0.04	**0.12**	0.06	0.00	0.00	0.22
	725–749	0.00	0.06	**0.12**	0.05	0.00	0.23
	750–789	0.00	0.00	0.05	**0.18**	0.03	0.26
	790–850	0.00	0.00	0.00	0.03	**0.05**	0.07

Table F.8. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 4

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.16**	0.03	0.00	0.00	0.00	0.19
	700–724	0.04	**0.17**	0.04	0.00	0.00	0.25
	725–749	0.00	0.05	**0.17**	0.05	0.00	0.27
	750–795	0.00	0.00	0.04	**0.22**	0.01	0.27
	796–850	0.00	0.00	0.00	0.00	**0.01**	0.02
Decision Consistency	650–699	**0.16**	0.05	0.00	0.00	0.00	0.20
	700–724	0.05	**0.13**	0.06	0.00	0.00	0.24
	725–749	0.00	0.06	**0.13**	0.06	0.00	0.25
	750–795	0.00	0.00	0.06	**0.20**	0.01	0.27
	796–850	0.00	0.00	0.00	0.01	**0.01**	0.03

Table F.9. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 5

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.17**	0.03	0.00	0.00	0.00	0.20
	700–724	0.04	**0.18**	0.05	0.00	0.00	0.27
	725–749	0.00	0.05	**0.17**	0.05	0.00	0.26
	750–789	0.00	0.00	0.04	**0.18**	0.02	0.23
	790–850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650–699	**0.16**	0.05	0.00	0.00	0.00	0.21
	700–724	0.05	**0.15**	0.06	0.00	0.00	0.26
	725–749	0.00	0.06	**0.13**	0.05	0.00	0.25
	750–789	0.00	0.00	0.05	**0.16**	0.02	0.23
	790–850	0.00	0.00	0.00	0.02	**0.03**	0.05

Table F.10. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 6

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.16**	0.03	0.00	0.00	0.00	0.19
	700–724	0.04	**0.20**	0.05	0.00	0.00	0.29
	725–749	0.00	0.05	**0.18**	0.04	0.00	0.28
	750–787	0.00	0.00	0.03	**0.17**	0.01	0.21
	788–850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650–699	**0.15**	0.05	0.00	0.00	0.00	0.20
	700–724	0.05	**0.16**	0.06	0.00	0.00	0.28
	725–749	0.00	0.06	**0.15**	0.05	0.00	0.27
	750–787	0.00	0.00	0.05	**0.15**	0.01	0.22
	788–850	0.00	0.00	0.00	0.01	**0.02**	0.04

Table F.11. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 7

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.09**	0.02	0.00	0.00	0.00	0.11
	700–724	0.04	**0.22**	0.05	0.00	0.00	0.30
	725–749	0.00	0.05	**0.22**	0.04	0.00	0.32
	750–785	0.00	0.00	0.03	**0.19**	0.01	0.24
	786–850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650–699	**0.08**	0.04	0.00	0.00	0.00	0.12
	700–724	0.04	**0.18**	0.06	0.00	0.00	0.29
	725–749	0.00	0.06	**0.19**	0.05	0.00	0.30
	750–785	0.00	0.00	0.05	**0.17**	0.01	0.24
	786–850	0.00	0.00	0.00	0.01	**0.03**	0.05

Table F.12. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 8

	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650–699	**0.24**	0.05	0.00	0.00	0.00	0.29
	700–724	0.06	**0.13**	0.05	0.00	0.00	0.24
	725–749	0.00	0.05	**0.11**	0.04	0.00	0.20
	750–800	0.00	0.00	0.03	**0.17**	0.02	0.23
	801–850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650–699	**0.23**	0.06	0.01	0.00	0.00	0.30
	700–724	0.06	**0.10**	0.05	0.01	0.00	0.23
	725–749	0.01	0.05	**0.08**	0.05	0.00	0.19
	750–800	0.00	0.01	0.05	**0.15**	0.02	0.23
	801–850	0.00	0.00	0.00	0.02	**0.03**	0.05

Appendix G: Student Growth Percentile (SGP) Estimates by Subgroup

Table G.1. SGP Estimates by Subgroup—ELA/L Grade 4

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	62,343	49.66	13.44	50
Female	59,921	50.13	13.48	50
White	56,819	52.60	13.41	54
African American	19,619	43.61	13.58	41
Asian/Pacific Islander	6,907	56.01	13.43	58
American Indian/Alaska Native	312	46.31	13.77	43.5
Hispanic	32,802	47.65	13.50	47
Multiple	5,734	50.46	13.37	51
Economically Disadvantaged	62,355	46.37	13.50	45
Not Economically Disadvantaged	59,921	53.56	13.42	55
English Learner (EL)	22,749	46.67	13.44	45
Non-EL	99,527	50.63	13.46	51
Students with Disabilities (SWD)	23,071	42.63	13.53	39
Students without Disabilities	99,205	51.58	13.44	52

Table G.2. SGP Estimates by Subgroup—ELA/L Grade 5

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	62,272	48.71	15.19	48
Female	59,784	51.27	15.04	52
White	56,753	51.46	15.07	52
African American	19,397	46.13	15.21	45
Asian/Pacific Islander	7,071	55.97	15.27	59
American Indian/Alaska Native	288	50.97	15.47	48.5
Hispanic	32,909	48.43	15.12	48
Multiple	5,564	49.64	15.09	49
Economically Disadvantaged	62,274	47.94	15.09	47
Not Economically Disadvantaged	59,804	52.09	15.15	53
English Learner (EL)	18,776	47.14	15.13	46
Non-EL	103,302	50.48	15.12	51
Students with Disabilities (SWD)	23,205	44.35	15.40	42
Students without Disabilities	98,873	51.29	15.05	52

Table G.3. SGP Estimates by Subgroup—ELA/L Grade 6

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	62,854	47.38	14.47	46
Female	60,531	52.61	14.39	54
White	56,938	50.50	14.52	51
African American	19,427	47.49	14.30	46
Asian/Pacific Islander	7,176	55.33	14.84	58
American Indian/Alaska Native	275	50.04	13.90	50
Hispanic	33,952	49.50	14.27	49
Multiple	5,552	48.72	14.44	48
Economically Disadvantaged	62,167	48.52	14.22	48
Not Economically Disadvantaged	61,248	51.39	14.64	52
English Learner (EL)	16,847	48.01	14.20	47
Non-EL	106,568	50.25	14.46	50
Students with Disabilities (SWD)	23,282	44.86	14.53	42
Students without Disabilities	100,133	51.13	14.40	52

Table G.4. SGP Estimates by Subgroup—ELA/L Grade 7

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	64,661	48.44	14.68	48
Female	61,902	51.47	14.34	52
White	58,385	49.66	14.21	49
African American	19,965	49.42	15.11	49
Asian/Pacific Islander	7,184	58.42	14.15	62
American Indian/Alaska Native	264	44.66	14.99	42
Hispanic	35,305	49.20	14.75	49
Multiple	5,410	48.24	14.47	47
Economically Disadvantaged	63,100	48.61	14.83	48
Not Economically Disadvantaged	63,491	51.23	14.20	52
English Learner (EL)	18,243	48.06	15.42	47
Non-EL	108,348	50.23	14.36	50
Students with Disabilities (SWD)	23,396	45.64	15.74	44
Students without Disabilities	103,195	50.89	14.23	51

Table G.5. SGP Estimates by Subgroup—ELA/L Grade 8

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	65,810	47.45	14.91	46
Female	61,979	52.37	15.03	54
White	58,631	50.66	15.15	51
African American	20,635	47.91	14.73	47
Asian/Pacific Islander	7,246	54.58	15.44	57
American Indian/Alaska Native	265	47.86	14.41	47
Hispanic	35,734	48.90	14.72	49
Multiple	5,222	48.50	14.96	48
Economically Disadvantaged	63,557	48.10	14.67	47
Not Economically Disadvantaged	64,275	51.56	15.26	52
English Learner (EL)	17,646	46.86	14.16	45
Non-EL	110,186	50.32	15.10	50
Students with Disabilities (SWD)	23,492	44.57	14.38	42
Students without Disabilities	104,340	51.03	15.10	52

Table G.6. SGP Estimates by Subgroup—Mathematics Grade 4

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,058	51.33	13.83	52
Female	58,578	48.66	13.98	48
White	56,694	52.61	13.57	54
African American	19,547	43.51	14.52	41
Asian/Pacific Islander	6,907	57.74	13.65	60
American Indian/Alaska Native	279	46.77	14.16	46
Hispanic	30,424	47.50	14.19	46
Multiple	5,713	50.88	13.89	51
Economically Disadvantaged	60,172	46.55	14.20	45
Not Economically Disadvantaged	59,475	53.53	13.61	55
English Learner (EL)	20,391	47.83	14.24	47
Non-EL	99,256	50.47	13.83	51
Students with Disabilities (SWD)	22,704	45.78	14.29	44
Students without Disabilities	96,943	51.01	13.81	51
Spanish Language Form	2,351	43.70	14.63	41

Table G.7. SGP Estimates by Subgroup—Mathematics Grade 5

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,122	49.04	15.60	48
Female	58,616	51.28	15.68	52
White	56,657	50.36	15.20	51
African American	19,369	47.50	16.54	47
Asian/Pacific Islander	7,060	58.09	14.91	61
American Indian/Alaska Native	275	52.86	15.87	52
Hispanic	30,756	49.56	16.04	49
Multiple	5,549	49.85	15.73	50
Economically Disadvantaged	60,435	48.47	16.10	48
Not Economically Disadvantaged	59,325	51.83	15.18	53
English Learner (EL)	16,813	49.82	16.53	49
Non-EL	102,947	50.18	15.50	50
Students with Disabilities (SWD)	22,934	47.19	16.22	46
Students without Disabilities	96,826	50.83	15.51	51
Spanish Language Form	2,133	46.26	16.16	45

Table G.8. SGP Estimates by Subgroup—Mathematics Grade 6

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,767	49.42	15.71	49
Female	59,454	50.79	15.84	51
White	56,801	50.30	15.24	51
African American	19,347	48.14	16.89	47
Asian/Pacific Islander	7,158	56.85	15.04	60
American Indian/Alaska Native	257	49.47	16.19	48
Hispanic	32,066	49.62	16.23	49
Multiple	5,528	48.88	15.67	49
Economically Disadvantaged	60,560	48.63	16.35	48
Not Economically Disadvantaged	60,691	51.55	15.20	52
English Learner (EL)	15,248	47.12	17.02	46
Non-EL	106,003	50.52	15.60	51
Students with Disabilities (SWD)	22,971	44.62	16.75	42
Students without Disabilities	98,280	51.37	15.55	52
Spanish Language Form	1,849	46.59	16.93	45

Table G.9. SGP Estimates by Subgroup—Mathematics Grade 7

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	63,848	49.44	16.12	49
Female	61,167	50.62	16.14	51
White	58,265	49.94	15.71	50
African American	19,856	47.64	17.12	47
Asian/Pacific Islander	7,173	55.41	15.26	57
American Indian/Alaska Native	258	49.75	16.72	49
Hispanic	34,044	50.60	16.45	51
Multiple	5,374	48.93	16.07	49
Economically Disadvantaged	61,881	48.88	16.63	48
Not Economically Disadvantaged	63,162	51.13	15.63	52
English Learner (EL)	17,049	48.33	16.96	48
Non-EL	107,994	50.28	16.00	50
Students with Disabilities (SWD)	23,167	42.88	17.01	40
Students without Disabilities	101,876	51.64	15.93	52
Spanish Language Form	1,179	42.17	16.92	38

Table G.10. SGP Estimates by Subgroup—Mathematics Grade 8

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	65,020	48.54	16.36	48
Female	61,276	51.52	16.31	52
White	58,487	49.96	15.68	50
African American	20,513	48.90	17.70	48
Asian/Pacific Islander	7,238	56.77	14.69	59
American Indian/Alaska Native	255	50.27	17.32	52
Hispanic	34,539	49.38	16.96	49
Multiple	5,206	49.25	16.41	49
Economically Disadvantaged	62,376	48.85	17.14	48
Not Economically Disadvantaged	63,962	51.10	15.56	52
English Learner (EL)	16,492	48.28	18.04	47.5
Non-EL	109,846	50.24	16.08	50
Students with Disabilities (SWD)	23,243	44.65	18.08	43
Students without Disabilities	103,095	51.19	15.95	52
Spanish Language Form	1,091	45.38	18.68	44