New Meridian

# Technical Report 2022–2023 Illinois Assessment of Readiness

November 4, 2024

# Table of Contents

# List of Tables

# List of Figures

# Executive Summary

The purpose of this report is to describe the technical qualities of the 2022–2023 operational administration of the English language arts/literacy (ELA/L) and mathematics assessments in grades 3 through 8 for the state of Illinois.

Committees of educators and state education agency staff led the work in the development of the Illinois Assessment of Readiness that is aligned to the Illinois Learning Standards and is intended to measure more complex skills like critical thinking, persuasive writing, and problem-solving. New Meridian assumes the responsibility for management of the Illinois Assessment of Readiness, as well as item development and forms construction. New Meridian, built forms in accordance with the Illinois Assessment of Readiness blueprints.

Over three years, New Meridian used a phase-in approach to new items for Grades 3–8 ELA/L and mathematics to Illinois specifications using a selected set of licensed items from its bank of PARCC items. New Meridian calibrated the newly developed items to the Illinois assessment scale; and developed operational forms for school year 2022–2023 that contain a mixture of items from custom IAR development and licensed from the New Meridian bank.

The ELA/L assessments focus on reading and comprehending a range of sufficiently complex texts independently and writing effectively when analyzing text. The ELA/L assessments contain literary and informational texts; each passage set has four to eight brief comprehension and vocabulary questions. ELA/L constructed-response items include three types of tasks: literary analysis, narrative writing, and research simulation. For each task, students are instructed to read one or more texts, answer several brief questions, and then write an essay based on the material they read.

The mathematics assessments contain tasks that measure a combination of conceptual understanding, applications, skills, and procedures. 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 ability to apply concepts by means of selected-response or technology-enhanced items. In addition, students are required to demonstrate their skills and knowledge by answering innovative selected-response and short-answer questions that measure concepts and skills.

In both content areas, students also demonstrate their acquired skills and knowledge by answering selected-response items and fill-in-the-blank questions. Each assessment consists of multiple units, and additionally, one of the mathematics units is split into two sections: a non-calculator section and a calculator section.

The summative assessments are designed to achieve several purposes. First, the tests are intended to provide evidence to determine whether students are on track for college- and career-readiness. Second, the tests are structured to assess the full range of CCSS and measure the total breadth of student performance. Finally, the tests are designed to provide data to help inform classroom instruction, student interventions, and professional development.

This technical report includes the following topics:

- Background and purpose of the assessments
- Test development of items and forms
- Test administration, security, and scoring
- Student characteristics
- Classical item analyses and differential item functioning
- Reliability and validity of scores
- Item response theory (IRT) calibration and scaling
- Performance level setting
- Development of the score reporting scales and student performance
- Student growth measures
- Quality control procedures

The information provided in this technical report 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 *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014).

<This page intentionally left blank>

# Section 1: Introduction

## 1.1 Background

The Illinois Assessment of Readiness (IAR) is a custom adaptation of the prior 3–8 assessment, New Meridian Affiliate/PARCC. The IAR was created via a three-year plan to support the development of an Illinois owned bank of items while providing continued access to New Meridian's bank of high-quality assessment content to ensure comparability and validity of the IAR with Affiliate/PARCC historical data during this transition period. States and agencies associated with the Partnership for Assessment of Readiness for College and Careers (PARCC) came together in early 2010 with a shared vision of ensuring that all students—regardless of income, family background, or geography—have equal access to a world-class education that will prepare them for success after high school in college and/or careers. The goal was to develop new assessments that tie into more rigorous academic expectations and help prepare students for success in college and the workforce, as well as to provide information back to teachers and parents about where students are on their path to success. Calling on the expertise of thousands of teachers, higher education faculty, and other educators in multiple states, the resulting assessment system is a high-quality set of summative assessments, diagnostic assessments, formative tasks, and other support materials for teachers that include professional development and communications tools.

The partnership developed and administered next-generation assessments that, compared to traditional K–12 assessments, more accurately measured student progress toward college and career readiness. The assessments were aligned to the Common Core State Standards (CCSS) and included both English language arts/literacy (ELA/L) assessments (grades 3 through 11) and mathematics assessments (grades 3 through 8 and high school). Compared to traditional standardized tests, these assessments were intended to measure more complex skills like critical thinking, persuasive writing, and problem-solving.

In 2013, the PARCC Governing Board launched PARCC Inc., a nonprofit organization designed to support the successful delivery of the tests in 2014–2017 and the long-term success of the multi-state partnership. States continued to govern decisions about the assessment system; the nonprofit organization was their "agent" for overseeing the many vendors involved in the assessment system, coordinating the multiple work groups and committees (including Governing Board meetings), managing the intellectual property, overseeing the research agenda and the Technical Advisory Committee, and developing and launching the multiple non-summative tools.

Summative assessments for the first operational administration were constructed in 2014. Eleven states including the District of Columbia participated in the first administration of the summative assessments during the 2014–2015 school year. Six states, the Bureau of Indian Education (BIE), and District of Columbia participated in the second administration in school year 2015–2016. Five states, the Bureau of Indian Education, the Department of Defense Education Activity, and the District of Columbia participated in the third administration in school year 2016–2017. Four states, the Bureau of Indian Education, the Department of Defense Education Activity, and the District of Columbia participated in the fourth administration in school year 2017–2018.

Following the PARCC, Inc. contract ending in June 2017, participating states and agencies released the intellectual property (IP) of the contract to the Council of Chief State School Officers (CCSSO) and contracted with New Meridian to manage the intellectual property and provide item development, forms construction, and governance. Starting in August 2017, New Meridian oversaw item development, data review for field test items, and test construction activities.

In 2017, New Meridian, in coordination with multiple states and vendors, developed an alternate form of the summative assessment to meet the needs for shorter testing times desired by several states. Through extensive research and guidance from the Technical Advisory Committee, the alternate blueprint and the original blueprint were available in spring 2019. New Meridian's state-centric solution to educational assessment allowed states the flexibility of selecting the assessment solution that best fits their specific needs. For the academic year 2018–2019, participating states and agencies included the Bureau of Indian Education, the District of Columbia, Illinois, New Jersey, and New Mexico. For the academic years 2019–2020 and 2020–2021, participating states and agencies included the Bureau of Indian Education, the District of Columbia, the Department of Defense Education Activity, Illinois, and New Jersey. Most testing in spring 2020 was cancelled due to the COVID-19 global pandemic, except for a small number of students in IL who tested prior to the closure of schools. Some states further cancelled administration in spring 2021. For the academic year 2021–2022, participating states and agencies included the District of Columbia, the Department of Defense Education Activity, Illinois, and New Jersey. For the academic year 2022–2023, the assessment was built using the Illinois Assessment of Readiness blueprints using 50% IAR and 50% Affiliate/PARCC items

The purpose of this technical report is to describe the operational administration of the custom Illinois summative assessments in the 2022–2023 academic year, including test form construction, test administration, item scoring, student characteristics, classical item analysis results, reliability results, evidence of validity, item response theory (IRT) calibrations and scaling, performance level setting procedure, growth measures, and quality control procedures.

## 1.2 Purpose of the Operational Tests

The Illinois Assessment of Readiness is designed to achieve several purposes. First, the assessments are intended to provide evidence to determine whether students are on track for college- and career-readiness. Second, the assessments are structured to access the full range of CCSS and measure the total breadth of student performance. Finally, the assessments are designed to provide data to help inform classroom instruction, student interventions, and professional development.

## 1.3 Composition of Operational Tests

Each operational test form is constructed to reflect the test blueprint in terms of content, standards measured, and item types. Sets of common items, included to provide data to support horizontal linking across test forms within a grade and content area, are proportionally representative of the operational test blueprint. The summative assessment is a mixed-format test. The current summative assessments are administered in either computer-based (CBT) or paper-based (PBT) format. The paper-based format is offered only as an accommodation for students who cannot take the computer-based assessment.

The ELA/L assessments focus on reading and comprehending a range of sufficiently complex texts independently and writing effectively when analyzing text. The ELA/L assessments contain literary and informational texts; each passage set has four to eight brief comprehension and vocabulary questions. ELA/L constructed-response items include three types of tasks: literary analysis, narrative writing, and research simulation. For each task, students are instructed to read one or more texts, answer several brief questions, and then write an essay based on the material they read.

The mathematics assessments contain tasks that measure a combination of conceptual understanding, applications, skills, and procedures. 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 ability to apply concepts by means of selected-response or technology-enhanced items. In addition, students are required to demonstrate their skills and knowledge by answering innovative selected-response and short-answer questions that measure concepts and skills.

In both content areas, students also demonstrate their acquired skills and knowledge by answering selected-response items and fill-in-the-blank questions. Each assessment consists of three units. Unit one consists of two sections, non-calculator and calculator.

## 1.4 Intended Population

The tests are intended for students taking ELA/L and Mathematics in grades 3 through 8. For these students, the tests measure whether students are meeting state academic standards and mastering the knowledge and skills needed to progress in their K–12 education and beyond.

Pearson serves as the primary vendor for the assessment and is responsible for producing all testing materials, packaging and distribution, receiving and scanning of materials, scoring, and program management and customer service. Pearson Psychometrics is responsible for all psychometric analyses of the operational test data. This includes classical item analyses, differential item functioning (DIF) analysis, item calibrations based on item response theory (IRT), and scaling. For the 2023 test administration, New Meridian Psychometrics was responsible for development of all score conversion tables.

## 1.6 Overview of the Technical Report

This report begins by providing explanations of the test form construction process, test administration, and scoring of the test items. Subsequent sections of the report present descriptions of student characteristics, results of classical item analyses, item response theory (IRT) calibrations and scaling, performance level setting procedures, quality control procedures, results of students' scale score analyses, results of reliability analyses, evidence of validity, and measures of student growth. The technical report contains the following sections:

## Section 2 – Test Development

Test design and the procedures followed during the development of operational test forms.

## Section 3 – Test Administration

Operational administration schedule, information regarding test security and confidentiality, accessibility features and accommodations, testing irregularities, and security breaches.

## Section 4 – Item Scoring

Key-based and rule-based processes for machine-scored items, as well as the training and monitoring processes for human-scored items.

## Section 5 – Classical Item Analysis

Classical item-level statistics calculated for the operational test data, the flagging criteria used to identify items that performed differently than expected, and the results of these analyses.

## Section 6 – Differential Item Functioning

Methods for conducting differential item functioning analyses as well as corresponding flagging criteria. This is followed by definitions of the comparison groups and subsequent results for the comparison groups.

## Section 7 – IRT Model and Parameters

Information related to the IRT models used and the descriptive statistics of the item parameters. Note that all tests employed a pre-equated model, in which previously estimated item parameters are used to generate scoring tables.

## Section 8 – Performance Level Setting

Performance levels and policy definitions, as well as the processes followed to establish performance level thresholds.

## Section 9 – Quality Control Procedures

Quality control, including quality assurance of item banking, test form construction, and testing materials, quality control of scanning, image editing, and scoring. This is followed by a detailed description of the steps taken to ensure that all psychometric analyses were of the highest quality.

## Section 10 – Operational Test Forms

Operational test forms including high level blueprints for the assessments.

## Section 11 – Student Characteristics

Composition of test forms, rules for inclusion of students in analyses, distributions of students by grade, mode, and gender, and distributions of demographic variables of interest.

## Section 12 – Scale Scores

Overview of the claims and subclaims, describes the development of reporting scales and conversion tables, and presents scale score distributions. Finally, information regarding the interpretation of claim scores and subclaim scores is presented.

## Section 13 – Reliability

Results of scale score reliability, internal consistency reliability analyses and corresponding standard errors of measurement for each grade, content area, and mode (CBT or PBT) for all students, and for subgroups of interest. This is followed by reliability results for subscores and reliability of classification (i.e., decision accuracy and decision consistency). Finally, expectations and results for inter-rater agreement for hand-scored items are summarized.

## Section 14 – Validity

Validity evidence based on analyses of the internal structure of the tests is provided in this section. Correlations between subscores are reported by grade, content area, and mode (CBT or PBT) for all students.

## Section 15 – Student Growth Measures

Details on student growth percentiles (SGP). Information about the model, model fit, and SGP averages at the overall level for all students, and for subgroups of interest, are provided in this section.

## References

## Appendices

To facilitate utility, tables in the appendices are numbered sequentially according to the section represented by the tables. For example, the first appendix table for Section 6 is numbered A.6.1, the second appendix table for Section 6 is numbered A.6.2, etc.

# Section 2: Test Development

## 2.1 Overview of the Summative Assessments, Claims, and Design

For the school year 2022–2023, IAR forms were constructed with a mix of custom Illinois-owned items and items from the Affiliate/PARCC banks to the IAR blueprints.

Aligned to the Common Core State Standards (CCSS) as articulated in the Model Content Frameworks, the Affiliate/PARCC summative assessments are designed to determine whether students are on track to be college- and career-ready, assess the full range of the CCSS, measure the full range of student performance, and provide data to help inform instruction, interventions, and professional development. Test development involved educators, researchers, psychometricians, subject matter professionals, and assessment experts who participated in the development of the test design and its underlying foundational documents; developed and reviewed passages and items used to build the summative assessments; monitored the program for quality, accessibility, and fairness for all students; and constructed, reviewed, and scored the assessments.

The summative assessments include both English language arts/literacy (ELA/L) and mathematics assessments in grades 3 through 8 and high school. Assessments contain selected response, brief and extended constructed response, technology-enabled and technology-enhanced items (TEI), as well as performance tasks. Technology-enabled items are single-response or constructed-response items that involve some type of digital stimulus or open-ended response box with which the students engage in answering questions.

Technology-enhanced items involve specialized student interactions for collecting performance data. In other words, the act of performing the task is the way in which data is 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 TEI is an interaction in which students are asked to drag response options onto a Venn diagram to show the relationship among ideas.

The summative assessments offer a wide range of accessibility features for all students and accommodations for students with disabilities (e.g., screen reader, assistive technology, braille, large print [LP], text-to-speech [TTS], and American Sign Language [ASL] video versions of the test, as well as response accommodations that allow students to respond to test items using different formats). For English learners who are native Spanish speakers, participating states and agencies offer the mathematics assessments in Spanish, and both LP and TTS versions of the test in Spanish (refer to the Accessibility Features and Accommodations Manual for in-depth information).

### 2.1.1 English Language Arts/Literacy (ELA/L) Assessments—Claims and Subclaims

The ELA/L summative assessment at each grade level consists of three task types: literary analysis, research simulation, and narrative writing. For each literary analysis and research simulation task, students are asked to read or view one or more texts, answer comprehension and vocabulary questions, and write an extended response that requires them to draw evidence from the texts. For each narrative writing task, students are asked to read or view one text, answer comprehension questions, and write an extended narrative response based on the text.

The claim structure, grounded in the CCSS, undergirds the design and development of the ELA/L summative assessments.

**Master Claim**. The master claim is the overall performance goal for the ELA/L Summative Assessment System—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, and 2) writing effectively when using and/or analyzing sources.

**Subclaims:** The subclaims further explicate what is measured on the summative assessments and include claims about student performance on the standards and evidence outlined in the evidence tables for reading and writing (refer to the test specifications documents). The claims and evidence are grouped into the following categories:

1. Vocabulary Interpretation and Use

2. Reading Literature
3. Reading Informational Text
4. Written Expression
5. Knowledge of Language and Conventions

### 2.1.2 Mathematics Assessments Claims and Subclaims

The summative mathematics assessment at each grade level includes both short- and extended-response questions focused on applying skills and concepts to solve problems that require demonstration of the mathematical practices from the CCSS, with a focus on modeling and reasoning with precision. The assessments also include performance-based short-answer questions focused on conceptual understanding, procedural skills, and application.

The claim structure, grounded in the CCSS, undergirds the design and development of the summative assessments.

**Master Claim**. The degree to which a student is college- or career-ready or on track to being ready in mathematics. The student solves grade-level/course-level problems aligned to the Standards for Mathematical Content with connections to the Standards for Mathematical Practice.

**Subclaims:** The subclaims further explicate what is measured on the summative assessments and include claims about student performance on the standards and evidences outlined in the evidence statement tables for mathematics (refer to the test specifications documents). 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 Development Activities

Test development activities began with the standards and model content frameworks. These documents include the College- and Career-Ready Determinations and Performance-Level Descriptions, Claim Structure, Evidence Statement Tables, blueprints, informational guides, Passage Selection Guidelines, Mathematics Sequencing Guidelines, Task Generation Models, Fairness and Sensitivity Guidelines, Text Selection Guidelines, and the style guide. Refer to the New Meridian website for further information about these documents.

### 2.2.1 Item Development Process

Affiliate/PARCC test and item development activities were conducted by Pearson under the guidance and oversight of the K–12 state leads, the Higher Education Leadership Team, the Technical Advisory Committee, the Operational Working Group (OWG) members from each of the member states and agencies, the Text and Content Item Review Committees, and staff members from New Meridian, the project manager.

Developing high quality assessment content with authentic stimuli for computer-based tests (CBT) and paper-based tests (PBT) measuring rigorous standards was a complex process involving the services of many experts including assessment designers, psychometricians, managers, trainers, content providers, content experts, editors, artists, programmers, technicians, human scorers, advisors, and members of the OWGs.

### Bank Analysis and Item Development Plan

The summative item bank houses passages and items at each assessed grade level and subject. The bank supports the administration of the assessments, along with item release and practice tests. Items are developed and field tested annually. As the first step in annual item development cycle, the item development teams, in conjunction with members of the OWGs for ELA/L and mathematics, evaluated the strengths of the bank and considered the needs for future tests to establish an item development plan.

### Text Selection for ELA/L

Using the Passage Selection Guidelines, English language arts subject matter experts were trained to search for appropriate passages to support an annual pool of passages for consideration. Guided by the test specifications documents, contracted subject matter experts worked to deliver the number of texts specified in the annual asset development plan. The Passage Selection Guidelines provided a text complexity framework and guidance on selecting a variety of text types and passages that allowed for a range of standards and evidences to be demonstrated to meet the assessment claims. ELA/L tests are based on authentic texts, including multi-media stimuli. Authentic texts are grade-appropriate texts 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. Staff content experts reviewed the passages for adherence to the Passage Selection Guidelines to meet the annual asset development plan described above in the number and distribution of genres and topics prior to review and consideration by the Text Review Committee. ELA/L item development was not conducted until after texts were approved by the Text Review Committee.

### Item Development

Item writers were recruited, trained, and managed to develop the number of items specified in the annual asset development plan. Prior to committee reviews, staff reviewed the items for content accuracy, alignment to the standards, range of difficulty, adherence to universal design principles (which maximize the participation of the widest possible range of students), bias and sensitivity, and copy editing to enable the accurate measurement of the standards.

## 2.2.2 Item and Text Review Committees

Members of the OWGs for ELA/L and mathematics, state-level experts, local educators, post-secondary faculty, and community members conducted rigorous reviews of every item and passage being developed for the summative assessment system to ensure all test items are of the highest quality, aligned to the standards, and fair for all student populations. All reviewers were nominated by their state education agency. The purpose of the educator reviews was to provide feedback on the quality, accuracy, alignment, and appropriateness of the test passages and items developed annually for the summative assessments. 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 of grade/subject working committees where additional training was provided.

### Text Review

The Text Review Committee meets to review and approve the texts eligible for item development. Participants reviewed and provided feedback to about the grade-level appropriateness, content, and potential bias concerns, and reached consensus about which texts would move forward for development. The Text Review Committee was made up of members of both Content Item Review and Bias and Sensitivity Review Committees.

### Content Item Review

During Content Item Review, committees reviewed and edited test items for adherence to the foundational documents, basic universal design principles, Accessibility Guidelines, associated item metadata, and the style guide. Committees accessed the item content within the Pearson Assessment Banking for Building and Interoperability (ABBI) system that previews how the passages and items will be displayed in an operational online environment. Committees also verified that the appropriate scoring rule had been applied to each item. The Content Item Review Committees were made up of OWG members and educators nominated by participating states.

### Bias and Sensitivity Review

Educators and community members made up the committee that reviewed items and tasks to confirm that there were no bias or sensitivity issues that would interfere with a student's ability to achieve his or her best performance. The committee reviewed items and tasks to evaluate adherence to the Fairness and Sensitivity Guidelines and to ensure that items and tasks do not unfairly advantage or disadvantage one student or group of students over another. Bias and Sensitivity committee members made edits and modifications to items and passages to eliminate sources of bias and improve accessibility for all students.

### Editorial Review

The Editorial Review Committee consisted of editors who reviewed up to 10 percent of the items and tasks. The committee reviewed the items for grammar, punctuation, clarity, and adherence to the style guide.

### Data Review

Following field tests, educator and bias committee members met to evaluate test items and associated performance data regarding appropriateness, level of difficulty, and potential gender, ethnic, or other bias, then recommended acceptance or rejection of each field-test item for inclusion on an operational assessment. The Data Review Committee also made recommendations that items be revised and re-field tested. Items that were approved by the committee are eligible for use on operational summative assessments.

## 2.2.3 Operational Test Construction

Under the guidance in the operational test form creation specifications, New Meridian constructed the operational forms to adhere to the test blueprints and the assessment goals outlined in the form creation specifications. These goals were as follows:

- Test forms designed to measure well across the full range of student ability
- Scores that are comparable among forms and across test administrations
- Scales that support classification of students into performance levels
- Maximization of the number of parallel forms
- Minimization of overexposure of items
- Adherence to standards for validity, reliability, and fairness (*Standards for Educational and Psychological Testing,* AERA, APA, & NCME, 2014)

Each content-area and grade-level assessment was based on a specific test blueprint that guided how each test was built. Test blueprints determined the range and distribution of content, and the distribution of points across the subclaims and task types.

Multiple core forms were constructed for a given assessment to enhance test security and to support the opportunity for item release. Core forms were the operational test forms, consisting of only those items that counted toward a student's score. These forms were designed to facilitate psychometric equating through a common item linking strategy.

Additionally, appropriate forms were identified as accessibility and accommodated forms. These forms are accommodated to support braille, large print, human reader/human signers, assistive technology, text-to-speech, closed captioning, and Spanish. Human reader/human signers and Spanish are provided for mathematics assessments only. Closed captioning is provided for ELA/L assessments only.

### Test Construction Activities

After the data review meetings and prior to the test form verification meetings, New Meridian content specialists and psychometric staff constructed initial versions of all the core forms. Initial core forms were based on the support documents and specific processes to achieve fair forms. The following steps were taken to construct the operational core forms taken to the test form verification meeting for review.

1. Constructed the online forms to match the blueprint and test construction specifications

2. Constructed the paper forms to match the blueprint and test construction specifications
3. Constructed accommodated and accessibility forms to match the blueprint, test construction specifications, and Accessibility, Accommodations, and Fairness (AAF) constraints

The test construction process included iterative steps between content specialists and psychometricians. Custom test construction reports generated by the psychometric team provided information on adherence to blueprint and statistical averages/distributions of item difficulty and discrimination describing the forms and allowing comparison of the forms. These reports facilitated content changes to better achieve the test construction goals. Equating across operational forms within an administration was accomplished by repeating core items across forms. Linking across administrations for operational forms was accomplished by including prior operational items on the current operational test forms.

New Meridian assessment specialists identified forms for each grade/subject suitable for use as the accommodated forms. Psychometrics reviewed the psychometric properties of each of the accommodated forms with respect to the required criteria. The content of these forms was also reviewed by accessibility specialists allowing for content changes prior to the Test Construction Committee meetings.

These test construction activities provided significant inputs to commence the meetings, including

- The proposed items for the initial operational core forms and the accommodated forms described above.
- Reports describing each form and comparing parallel forms.
- Recommended accommodated forms.

## Test Form Verification Meeting to Review Test Construction Inputs

Members of the Content Item Review Committees and the AAF OWG participated in the building of operational core forms that met the summative assessment requirements. In that process, they met in an in-person meeting to review and make recommendations for changes so that test forms conformed to both the content and psychometric requirements of the assessment.

## Accommodated Form Review Process

In addition to participating in many of the development activities including the Text Review and the Bias and Sensitivity Review meetings, the AAF OWG reviewed the proposed accommodated forms at the Test Construction Committee meeting for accessibility to make sure that the content can be accommodated for students with disabilities and English learners without changing the underlying measured construct.

Forms were identified to support the following accommodations:

Accommodated Base 1 (A1)

- Spanish paper (also serves Spanish LP, Spanish human reader paper)
- Spanish human reader/human signer online
- Base accommodated paper (serves braille, LP, human reader paper)
- Human reader/human signer online
- Assistive technology screen reader
- Assistive technology non-screen reader
- American Sign Language (ASL)

Accommodated Base 2 (A2)

- Closed captioning, first form
- Text-to-speech, first form
- Spanish online

- Spanish text-to-speech

Accommodated Base 3 (A3 – mathematics only)

- Closed captioning, first form
- Text-to-speech, second form

Spanish is mathematics only. Closed captioning is ELA/L only.

At the conclusion of the meetings, all test forms were constructed to meet test blueprints and requirements, and if necessary, reflect the operational linking design. Each test form reflected the test blueprint in terms of content, item types, and test length, as well as *expected* difficulty and performance along the ability continuum. Linking sets were proportionally representative of the operational test blueprint. The operational core forms, linking set forms, and field-test forms were reviewed by the Forms Review Committees and approved prior to the test administration.

### Spanish-Language Assessments for Mathematics

For English learners, the mathematics assessments are offered in Spanish, as well as in Spanish-language large print and text-to-speech (TTS) versions. Once the operational form was approved, the form was sent to Pearson's subcontractor, Teneo, for transadaptation of the items. Transadaptation differs from translation in that it takes into consideration the grade-level appropriateness of the words, as well as the linguistic and cultural differences that exist between speakers of two different languages. Accounting for these differences allows the item to measure the achievement of Spanish language speakers in the same way that the original version of the item does for native speakers of English. The Spanish glossary provided guidance to the translator conducting the transadaptation in grade-level and culturally appropriate ways of transadapting the items. For the Spanish language TTS form, the alternate text (used for description and/or text in art and graphics) was transadapted from the alternate text for the English language version of the TTS form. Phonetic mark-up, which guides how the TTS reader pronounces content-specific words and phrases, was also applied in this process.

In addition to the expert review of potential content for all accommodated forms conducted by the AAF OWG with assistance from content experts at the test construction meetings, the transadapted forms underwent additional quality checks: a Pearson Spanish copy edit services review and approval, and an AAF OWG review and approval.

## 2.2.4 Linking Design of the Operational Test

To support the goal of score comparability within and across administrations and years, a hybrid approach was implemented that incorporated the strengths of common item linking and randomly equivalent groups. The use of repeated operational core items was leveraged for common item linking. In addition, all forms were available throughout the operational administration, with spiraling at the student level, leveraged to support linking through randomly equivalent groups.

The operational test forms involved various types of linking; horizontal linking and across-administration linking. Horizontal linking consisted of linking items, or common items, included in both forms in a single administration, which was the case for mathematics forms and some ELA/L forms. Across-administration linking, or year-to-year linking, consisted of common items included in two different administrations, and used for all forms due to the pre-equated model. The placement of linking items across forms or administrations supports the development of comparable scores.

Linking item sets can be internal or external linking sets. Internal linking sets consist of common items in operational positions such that the items contribute to the students' scores. External linking sets consist of common items in positions resulting in the items not contributing to students' scores. The current linking designs included internal linking sets.

## 2.2.5 Field Test Data Collection Overview

Field test items were embedded in the spring operational mathematics forms. Field test items for ELA/L operational forms were administered as a separate unit.

# Section 3: Test Administration

## 3.1 Test Security and Administration Policies

The administration of the summative assessment is a secure testing event. Maintaining the security of test materials before, during, and after the test administration is crucial to obtaining valid and reliable results. 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.

School Test Coordinators must implement chain-of-custody requirements for specified materials. School Test Coordinators are responsible for distributing materials to Test Administrators, collecting materials from Test Administrators, returning secure test materials, and securely destroying certain specified materials after testing.

The administration of the summative assessment includes both secure and nonsecure materials, and these materials are further delineated by whether they are “scorable” or “nonscorable,” depending on whether the assessments were administered via paper/pencil (i.e., paper-based assessments) or online (i.e., computer-based assessments). For the paper-based administration, students used paper-based answer documents (except in grade 3 where students responded directly into test booklets). Nearly all of the summative assessments administered during the 2022–2023 administration were online assessments.

### 3.1.1 Secure vs. Non-Secure Materials

Secure materials are defined as those that 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, student work, etc. For paper-based tests, secure materials include both used and unused test booklets and used scratch paper, while for computer-based tests, secure materials include student testing tickets, secure administration scripts (e.g., mathematics read-aloud) and used scratch paper. Non-secure materials are defined as any authorized testing materials that do not include secure content (e.g., test items or student work). These include test administration manuals, unused scratch paper, and mathematics reference sheets that have not been written upon, etc.

### 3.1.2 Scorable vs. Nonscorable Materials

Paper-based assessments have both scorable and nonscorable materials, while computer-based assessments have only nonscorable materials. Scorable materials for paper-based assessments consist of used (including student work) test booklets (grade 3) and answer documents (grades 4 and above) only. Scorable materials must be returned to the vendor to be scored. All other materials for paper-based testing, such as blank (i.e., unused) test booklets, test administration manuals, scratch paper, mathematics reference sheets, etc., are deemed nonscorable. For computer-based tests, there are no 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).

Students taking the computer-based test may not have access to secure test materials before testing, including printed student testing tickets. Printed mathematics reference sheets (if applicable) and scratch paper must be new and unmarked.

Students taking the paper-based test may not have access to scorable or nonscorable secure test content before or after testing. Scorable secure materials provided by Test Administrators include test booklets (grade 3) or answer documents (grades 4 through high school). Nonscorable secure materials distributed by Test Administrators to paper-based testing students 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 through 8 and high school).

School Test Coordinators are required to maintain a tracking log to account for collection and destruction of test materials, including mathematics reference sheets and scratch paper written on by students. As part of the test administration policy, schools are required to maintain the Chain-of-Custody Form or tracking log of secure materials for at least three years unless otherwise directed by state policy. Copies of the Chain-of-Custody Form for paper-based testing are included in each Local Education Agency (LEA) or school’s test materials shipment.

Test Administrators are not to have extended access to test materials before or after administration (except for certain accessibility or accommodations purposes). Test Administrators must document the receipt and return of all secure test materials (used and unused) to the School Test Coordinator immediately after testing.

All test security and administration policies are found in the *Test Coordinator Manual and the Test Administrator Manuals*. State-specific policies are included in *Appendix C* of the *Test Coordinator Manual*.

# 3.2 Accessibility Features and Accommodations

## 3.2.1 Participation Guidelines for Assessments

All students, including students with disabilities and English learners, are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for English learners in their first year in a U.S. school, and certain students with disabilities who have been identified by the Individualized Education Program (IEP) team to take their state's alternate assessment. Federal laws governing student participation in statewide assessments include the No Child Left Behind Act of 2001 (NCLB), the Individuals with Disabilities Education 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. All students can receive accessibility features on the summative assessments.

Four distinct groups of students may receive accommodations on the summative assessments:

1. Students with disabilities who have an IEP
2. Students with a Section 504 plan who have a physical or mental disability that substantially 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 English learners
4. Students who are English learners with disabilities who have an IEP or 504 plan

These students are eligible for accommodations intended for both students with disabilities and English learners. Testing accommodations for students with disabilities or students who are English learners must be documented according to the guidelines and requirements outlined in the *Accessibility Features and Accommodations Manual*.

## 3.2.2 Accessibility System

Through a combination of universal design principles and accessibility features, Illinois designed an inclusive assessment system by considering accessibility from initial design through item development, field testing, and implementation of the assessments for all students, including students with disabilities, English learners, and English learners with disabilities. Accommodations may still be needed for some students with disabilities and English learners to assist in demonstrating knowledge and abilities. However, the accessibility features available to students should minimize the need for accommodation during testing and ensure the inclusive, accessible, and fair testing of the diverse students being assessed.

## 3.2.3 What are Accessibility Features?

On computer-based assessments, accessibility features are tools or preferences that are either built into the assessment system or provided externally by Test Administrators and may be used by any student taking the summative assessments (i.e., students with and without disabilities, gifted students, English learners, and English learners with disabilities). Since accessibility features are intended for all students, they are not classified as accommodations. Students should have the opportunity to select and practice using them prior to testing to determining which are appropriate for use on the assessment. Consideration should be given to the supports a student finds helpful and consistently uses during instruction. Practice tests that include accessibility features are available for teacher and student use throughout the year.

## 3.2.4 Accommodations for Students with Disabilities and English Learners

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. For the summative assessments, accommodations are adjustments to the testing conditions, test format, or test administration that provide equitable access during assessments for students with disabilities and students who are English learners. 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.
- Not compromise the integrity or validity of the assessment.

Accommodations are intended to reduce and/or eliminate the effects of a student's disability and/or English language proficiency level; however, accommodations should never reduce learning expectations by reducing the scope, complexity, or rigor of an assessment. Moreover, accommodations provided to a student on the summative assessments must be generally consistent with those provided for classroom instruction and classroom assessments. There are some accommodations that 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—for example, 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 non-allowable accommodations during assessments. It is important for educators to become familiar with the participating state and agencies' policies regarding accommodations used for assessments.

To the extent possible, accommodations should adhere to the following principles:

- Accommodations enable students to participate more fully and fairly in instruction and assessments and to demonstrate their knowledge and skills.
- Accommodations should be based upon 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 availability of staff.
- Accommodations should be based on a documented need in the instruction/assessment setting and should not be provided for the purpose of giving the student an enhancement that could be viewed as an unfair advantage.
- Accommodations for students with disabilities must be described and documented in the student's appropriate plan (i.e., either a 504 plan or an approved IEP), and must be provided if they are listed.
- Accommodations for English learners should be described and documented.
- Students who are English learners with disabilities are eligible to receive accommodations for both students with disabilities and English learners.
- 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.

In the following scenarios, the school must follow Illinois' policies and procedures for notifying the state assessment office if:

- A student was provided a test accommodation that was not listed in his or her IEP/504 plan/documentation for an English learner, or
- A student was not provided a test accommodation that was listed in his or her IEP/504 plan/documentation for an English learner.

### 3.2.5 Unique Accommodations

A comprehensive list of accessibility features and accommodations was provided in the *Accessibility Features and Accommodations Manual* that are designed to increase access to the summative assessments and that will result in valid, comparable assessment scores. However, students with disabilities or English learners may require additional accommodations that are not already listed. Participating states and agencies individually review requests for unique accommodations in their respective states and provide a determination as to whether the accommodation would result in a valid score for the student, and if so, would approve the request.

### 3.2.6 Emergency Accommodations

Emergency accommodation may be appropriate for a student who incurs a temporary disabling condition that interferes with test performance shortly before or during the assessment window. A student, whether or not they already have an IEP or 504 plan, may require an accommodation as a result of a recently occurring accident or illness. Cases include a student who has a recently fractured limb (e.g., arm, wrist, or shoulder); a student whose only pair of eyeglasses has broken; or a student returning to school after a serious or prolonged illness or injury. Emergency accommodation should be given only if the accommodation will result in a valid score for the student (i.e., does not change the construct being measured by the test[s]). If the principal (or designee) determines that a student requires an emergency accommodation on the summative assessment, an Emergency Accommodation Form must be completed and maintained in the student's assessment file. If required by a state, the school may need to consult with the state or district assessment office for approval**.** The parent must be notified that an emergency accommodation was provided. If appropriate, the Emergency Accommodation Form may also be submitted to the District Assessment Coordinator to be retained in the student's central office file. Requests for emergency accommodations will be approved after it is determined that use of the accommodation would result in a valid score for the student.

### 3.2.7 Student Refusal Form

If a student refuses an accommodation listed in his or her IEP, 504 plan, or (if required by the member state) an English learner plan, the school should document in writing that the student refused the accommodation, and the accommodation must be offered and remain available to the student during testing. This form must be completed and placed in the student's file, and a copy sent to the parent on the day of refusal. Principals (or designee) should work with Test Administrators to determine who, if any others, should be informed when a student refuses an accommodation documented in an IEP, 504, or (if required by the member state) English learner plan.

## 3.3 Testing Irregularities and Security Breaches

Any action that compromises test security or score validity is prohibited. These may be classified as testing irregularities or security breaches. Below are examples of activities that compromise test security or score validity (note that these lists are not exhaustive). It is highly recommended that School Test Coordinators discuss other possible testing irregularities and security breaches with Test Administrators during training.

Examples of test security breaches and irregularities include but are not limited to the following:

#### Electronic Devices

- Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner) is not permitted while secure test materials are distributed, while students are testing, after a student turns in his or her test materials, or during a break.
- Exception: Test Coordinators, Technology Coordinators, Test Administrators, and Proctors are permitted to use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed. LEAs may set additional restrictions on allowable devices as needed.

#### Test Supervision

- Coaching students during testing, including 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 at all times while secure test materials are distributed or while students are testing.
- Leaving students unattended for any period of time while secure test materials are distributed or while students are testing.
- Deviating from testing time procedures.
- Allowing cheating of any kind.
- Providing unauthorized persons with access to secure materials.
- Unlocking a test in PearsonAccess$^{next}$ during non-testing times.

- 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 state's 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.
- Exception: Administration of a human reader/signer accessibility feature for mathematics or accommodation for English language arts/literacy, which requires a Test Administrator to access passages or test items.
- Copying or reproducing (e.g., taking a picture of) any part of the passages or test 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.

### 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.

All instances of security breaches and testing irregularities must be reported to the School Test Coordinator immediately. The Form to Report a Testing Irregularity or Security Breach must be completed within two school days of the incident.

If any situation occurred that could cause any part of the test administration to be compromised, schools should refer to the *Test Coordinator Manual* for each state's policy and immediately follow those steps. Instructions for the School Test Coordinator or LEA Test Coordinator to report a testing irregularity or security breach is available in the *Test Coordinator Manual*.

## 3.4 Data Forensics Analyses

Maintaining the validity of test scores is essential in any high-stakes assessment program, and misconduct represents a serious threat to test score validity. When used appropriately, data forensic analyses can serve as an integral component of a wider test security protocol. The results of these data forensic analyses may be instrumental in identifying potential cases of misconduct for further follow-up and investigation.

The following data forensics analyses were conducted on the operational assessments:

- Response Change Analysis
- Aberrant Response Analysis
- Plagiarism Analysis
- Longitudinal Performance Modeling
- Internet and Social Media Monitoring
- Off-Hours Testing Monitoring

An overview of each data forensics analysis method is provided next.

### 3.4.1 Response Change Analysis

Response change analysis looks at how often student answers are changed, focusing specifically on an excessive number of wrong answers changed to right answers. In traditional paper-based, multiple-choice testing programs, this is sometimes referred to as "erasure analysis."[1] The rationale for erasure analysis is that a teacher or administrator who is intent on improving classroom performance might be motivated to change student responses after the answer sheets are collected. A clustered number of student answer documents from the same school or classroom with unusually high numbers of answers changed from wrong to right might provide evidence to support follow-up investigation. The response change analysis extended the traditional erasure method to account for issues specific to computer-based testing as well as the variety of item types on the summative assessments, such as partial-credit, multi-part, and multiple-select items.

### 3.4.2 Aberrant Response Analysis

Aberrant response pattern detection analysis looks at the unusualness of student responses compared with what would be expected. Most simply, this can be thought of as quantifying the extent to which higher-scoring students miss easy questions and lower-scoring students answer difficult questions correctly. While it would be difficult to draw a definitive inference about a single student flagged as having an aberrant response pattern, a cluster of students with aberrant response patterns within a classroom or school might warrant further investigation.

### 3.4.3 Plagiarism Analysis

Plagiarism analysis compares the responses given for a group of written composition items, looking for high degrees of similarity. For the summative assessments, the primary item type of interest was the Prose Constructed-Response (PCR) tasks in the English language arts/literacy (ELA/L) content area. This analysis was conducted for PCR tasks administered online using some of the same artificial intelligence (AI) techniques that are applied in automated essay scoring. Specifically, this method was based on Latent Semantic Analysis (LSA) technology to detect possible plagiarism. Using LSA, the content of each constructed-response was compared against the content of every other constructed-response and a measure that indicated the degrees of similarity was generated for each pair of response comparison. Because LSA provided a semantic representation of language, rather than a syntactic or word-based representation, it allowed the detection of potential copying behaviors, even when students or administrators substituted synonymous words or phrases.

### 3.4.4 Longitudinal Performance Monitoring

Longitudinal performance modeling evaluates the performance on the summative assessments across test administrations and identifies unusual performance gains in the unit of interest (e.g., school or district). Weighted Least Squares (WLS) regression methodology was evaluated and recommended by the Technical Advisory Committee (TAC) for implementation starting in spring, 2017. The WLS identified unusual changes in test performance across two consecutive administrations of the assessment. In the WLS regression approach, mean current year scale scores are regressed on mean prior year scale scores, weighting by unit sample size. Standardized residuals are calculated by dividing raw residuals by their respective standard deviations. Units with a standardized residual exceeding 3.0 are flagged for unexpected performance.

### 3.4.5 Internet and Social Media Monitoring

Internet and social media monitoring were conducted by Caveon, LLC. Caveon's team monitored English-language websites and searchable forums that were publicly available for suspected proxy testing solicitations and website postings that contain, or appear to contain, infringements of protected operational test content. The internet and social media outlets monitored included popular websites (such as Facebook and Twitter), blogs, discussion forums, video archives, document archives, brain dumps, auction sites, media outlets, peer-to-peer servers, etc. Caveon's process generated regular updates that categorize identified threats by level of actual or potential risk based upon the representations made on the websites, or actual analysis of the proffered content. For example, categorizations typically ranged from "cleared" (lowest risk but bookmarked for continued monitoring) to "severe" (highest risk). Note that this process only considered potential breaches of secure item content, not violations of testing administration policies. Potential breaches were reported directly to the state(s) implicated for further action. Summary reports describing the threats were provided through notification emails.

[1] The term "erasure analysis" is sometimes objected to because it is inferential rather than descriptive. A more descriptive term is "mark discrimination analysis," which recognizes that the scanning approach makes discriminations among the darkness of selected answer choices when multiple responses to a multiple-choice item are detected during answer sheet processing.

### 3.4.6 Off-Hours Testing Monitoring

Off-hours testing monitoring checks for suspicious testing activities at test administration locations occurring outside of the set windows for computer-based testing sessions. Participating states and agencies established set start and end times for administering computer-based assessments. Based on these hours, authorized users (that is, users with the state role) were allowed to override the start and end times for a test session. The off-hours testing monitoring process tracked such occurrences and logged them in an operational report, which listed the sessions within an organization that selected to test outside the set window. States could use this report to follow up with the organizations identified in the report.

# Section 4: Item Scoring

## 4.1 Machine-Scored Items

### 4.1.1 Key-Based Items

Pearson performed a key review prior to the test administration to verify that the scoring (answer) keys were correct for each item. Once the forms were constructed and approved for publication, an independent key review was performed by an experienced third-party vendor. The vendor reviewed each item and confirmed that the key was correct. If discrepancies were identified, a Pearson senior content specialist or content manager reviewed the flagged item(s) and worked with the item developers to resolve the issue.

### 4.1.2 Rule-Based Items

Rule-based scoring refers to item types that use various scoring models. Participating states and agencies use Question and Test Interoperability (QTI) item type implementation based on scoring model rules. Examples of these item types include "choice interaction," which 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 (for example) 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.

During the initial stages of item development, Pearson staff worked closely with participating states and agencies to first delineate the rules for the scoring rubrics and then to adjust those rules based on student responses. During item studies in spring 2015, Pearson content staff received input from the staff of participating states and agencies to develop a thorough rule-based scoring process that met their needs.

Pearson worked with the item developers to review initial scoring rules created during the item development. Once the rule-based scoring process was approved, and prior to test construction, Pearson content staff worked closely with the item developers to finalize scoring rubrics for items to be scored via the rule-based scoring method. The proposed scoring rubrics were sent for review, and if any additional changes were needed or new rules added, Pearson documented and applied the requested edits.

During test construction, Pearson monitored and evaluated the scoring and updated the scoring keys/scoring rules in the item bank. After the tryout items were scored, Pearson prepared a frequency distribution of student responses for each item or task scored using a rule-based approach and compared this to the expected response based on correct answers to ensure that scoring keys and rules were appropriately applied. The content team analyzed the student response data to determine if scoring was acceptable using the item metadata and the student response file in conjunction with any potential item issues as flagged by psychometrics. These frequency distributions included an indication of right/wrong and other identifying information defined by participating states and agencies, and those items that showed a statistical anomaly, whereby the frequency distribution was outside of the expected range, were sent to content experts to verify that the items were coded with the correct key.

Following the Rule-Based Scoring Educator Committee's review, which occurred prior to year one test construction, Pearson analyzed the feedback from the committees and made recommendations about adjustments to the scoring rubrics based on the results of the reviews. Upon submission of the results, Pearson worked with the staff of participating states and agencies to discuss these findings and determine next steps prior to the completion of scoring. In subsequent years as scoring inquiries arise throughout the process of test construction, forms creation, testing, scoring, and psychometric analysis, items with scoring discrepancies are brought before the Priority Alert Task Force for resolution. This committee consists of representatives from each state as well as the content specialists at participating states and agencies and Pearson.

Following the initial development of the rule-based scoring rubrics, Pearson has continued to monitor and evaluate new item development to ensure the scoring rules established are maintained within all item types as approved.

Pearson continues to use several avenues to monitor scoring each year. Prior to testing, a third-party key review checks operational and field test items for correct keys. Any disputed items go to a second review with Pearson content experts, and anything still in question is taken before the task force for review and possible key change. During testing, Pearson creates early testing files for frequency distribution analysis whereby items for which an incorrect key receives a high distribution of responses are further evaluated for accuracy. After testing, all responses are again evaluated for the distribution of responses and potential scoring abnormalities during psychometric analysis. Any change in scoring that may be requested as a result of the psychometric analysis is

also taken before the Priority Alert Task Force for decisions. These processes are the same for both paper and online modes of testing.

## 4.2 Human or Hand-scored Items

Constructed-response items were scored by human scorers in a process referred to as hand-scoring. Online training units were used to train all scorers. The online training units included prompts (items), passages, rubrics, training sets, and qualification sets. Scorers who successfully completed the training and qualified, demonstrating they could correctly score student responses based on the guidelines in the online training units, were permitted to score student responses using the ePEN2 (Electronic Performance Evaluation Network, second generation) scoring platform. All online and paper responses were scored within the ePEN2 system. Pearson monitored quality throughout scoring.

Pearson staff 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, which is a review of responses scored by each scorer. When backreading, a supervisor sees the scores applied by scorers, which helps the supervisor provide additional coaching or instruction to the scorer being backread.
- 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.
- English language arts/literacy (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 portfolio manager provided oversight for the entire scoring process.

All Pearson employees involved in the scoring or the supervision of scoring possessed at least a four-year college degree.

### 4.2.1 Scorer Training

Key steps in the development of scorer training materials were rangefinding and rangefinder review meetings where educators and administrators from states met to interpret the scoring rubrics and determine consensus scores for student responses. Rangefinding meetings were held prior to scoring field-test items, and rangefinder review meetings were held prior to scoring operational items.

At rangefinding meetings, educators and administrators from states reviewed student responses and used scoring rubrics to determine consensus scores. Those responses scored in rangefinding were used to create field test scorer training sets. After items were selected for operational testing, educators and administrators attended rangefinder review meetings to review and approve proposed operational scorer training sets.

When developing scorer training materials, Pearson scoring directors carefully reviewed detailed notes and records from rangefinding and rangefinder review committee meetings. Training sets were developed using the responses scored by the committees and additional suitable student response samples (as needed). All scorer training sets were reviewed and approved prior to scorer training.

During training, scorers reviewed training sets of scored student responses with annotations that explained the rationale for the score assigned. The anchor set was the primary reference for scorers as they internalized the rubric during training. Each anchor set consisted of responses that were clear examples of student performance at each score point. The responses selected were representative of typical approaches to the task and arranged to reflect a continuum of performance. All scorers had access to the anchor set when they were training and scoring and were directed to refer to it regularly during scoring.

Practice sets were used in training to help trainees practice applying the scoring guidelines. Scorers reviewed the anchor sets, scored the practice sets, and then were able to compare their assigned scores for the practice sets to the actual assigned scores to help them learn.

Qualification sets were used to confirm that scorers understood how to score student responses accurately. They were composed of responses that were clear examples of score points. Scorers were required to meet specified agreement percentages on qualification sets in order to score student responses.

Pearson has developed two types of training sets to train scorers: prototype and abbreviated sets. Prototype training sets were complete training sets consisting of anchor, practice, and qualification sets (refer to 4.2.2 for information on the qualification process). In ELA/L, there was one prototype training set per task type (Research Simulation Task, Literary Analysis Task, and Narrative Writing Task) at each of the nine grade levels (grades 3 through 10). In mathematics, a prototype training set was built for a grouping of similar items for a total of approximately three to four prototype sets per grade level or course.

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 at each grade level, 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 prototype items they had previously scored.

Anchor and practice sets for both prototype and abbreviated items included annotations for each response. Annotations are formal written explanations of the score for each student response.

Table 4.1 details the composition of the anchor sets, practice sets, and qualification sets.

Table 4.1 Training Materials Used During Scoring

| Training Set Development | |
|---|---|
| **Description** | **Specification** |
| **Anchor Set** | |
| The anchor set is the primary reference for scorers as they internalize the rubric during training. All scorers have access to the anchor set when they are training and scoring and are directed to refer to it regularly.<br><br>The anchor set comprises clear examples of student performance at each score point. The responses selected may be representative of typical approaches to the task or arranged to reflect a continuum of performance. | The anchor set for mathematics prototype items comprises three annotated responses per score point.<br><br>The anchor set for subsequent abbreviated items for mathematics comprise one to three annotated responses per score point. |
| | The anchor sets for ELA/L prototype items comprise three annotated responses per score point. Anchor sets for prototype items include separate complete anchor sets for each applicable scoring trait (Reading Comprehension and Written Expression and Conventions [RCWE] for Research Simulation and Literary Analysis Tasks, Written Expression [WE] for Narrative Writing Tasks, and Knowledge of Language and Conventions for all task types). |
| **Practice Sets** | |
| Practice sets are used to help trainees develop experience in independently applying the scoring guide (the rubric) to student responses. Some of these responses clearly reinforce the scoring guidelines presented in the anchor set. Other responses are selected because they are more difficult to evaluate, fall near the boundary between two score categories, or represent unusual approaches to the task.<br><br>The practice sets provide guidance and practice for trainees in defining the line between score categories, as well as applying the scoring criteria to a wider range of types of responses. | The practice sets for mathematics prototype and abbreviated items include two to three sets of ten annotated responses. |
| | ELA/L practice sets for prototype items include two sets of five annotated responses and two sets of ten annotated responses.<br><br>The subsequent ELA/L practice sets for abbreviated items include two sets of ten annotated responses. |

| Qualification Sets | |
|---|---|
| Qualification sets are used to confirm that scorer trainees understand the scoring criteria and are able to assign scores to student responses accurately. The responses in these sets are selected to reinforce the application of the scoring criteria illustrated in the anchor set.<br><br>Scorer trainees must demonstrate acceptable performance on these sets by meeting a pre-determined standard for accuracy in order to qualify to score. Pearson scoring staff defined and documented qualifying standards in conjunction with participating states and agencies prior to scoring. | The qualification sets for mathematics prototype items include three sets of ten responses each (not annotated).<br><br>The subsequent mathematics abbreviated items for mathematics do not include qualification sets. |
| | The qualification sets for ELA/L prototype items include three sets of ten responses each (not annotated).<br><br>The subsequent ELA/L abbreviated items do not include qualification sets. |

### 4.2.2 Scorer Qualification

To score items, scorers were required to show that they were able to apply scoring methodology accurately through a qualification process. Scorers were asked to apply scores to three qualification sets consisting of ten responses each. ELA/L scorers applied a score for each trait on each response in the qualification sets. Literary Analysis and Research Simulation Tasks each had two traits: the Reading Comprehension and Written Expression trait and the Conventions trait. The Narrative Writing Task had two traits: Written Expression and Conventions. Mathematics scorers applied a score for each part of an item that was a constructed-response. The number of constructed-response parts for each mathematics item ranged from one to four. Scorers were required to match the approved score at a percentage agreed to by participating states and agencies in order to qualify.

For ELA/L qualification, scorers were required to meet the following three conditions:

1. On at least one of the three qualifying sets, at least 70 percent 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 percent 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 percent of the ratings must be within one point of the approved scores.

For mathematics qualification, the requirements 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 following requirements separately for each scoring trait (when applicable to the item):

Table 4.2 Mathematics Qualification Requirements

| Category | Score Point Range | Perfect Agreement | Within One 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% |

On at least two of the three qualifying sets, a scorer was required to meet the "perfect agreement" percentage indicated in the table above 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 one point" percentage indicated in the table above for each category. The average is exclusive to each trait, so an item with multiple scoring traits would have multiple-trait rating averages within one point of the approved score.

## 4.2.3 Managing Scoring

Pearson created a hand-scoring specifications document that detailed the hand-scoring schedule, customer requirements, rangefinding plans, quality management plans, item information, and staffing plans for each scoring administration.

## 4.2.4 Monitoring Scoring

### Second Scoring

During scoring, Pearson's ePEN2 scoring system automatically and randomly distributed a minimum of 10 percent 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. Second scoring for ELA/L was performed either by human scorers or by the Intelligent Essay Assessor. If the first and second scores applied were nonadjacent, a third and occasionally a fourth score was assigned to resolve scorer disagreements. When a resolution score (i.e., third score) was nonadjacent to one or both the first and second scores, the content specialist or scoring director would apply an adjudication score (fourth score).
Table 4.3 provides the ruleset applied to determine the final score if f a response was scored more than once.

Table 4.3 Scoring Hierarchy Rules

| Score Type | Rank | Final Score Calculation |
|---|---|---|
| Adjudication | 1 | If an adjudication score is assigned, this is the final score. |
| Resolution | 2 | If no adjudication score is assigned, this is the final score. |
| Backread | 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. |
| Intelligent Essay Assessor Score | 6 | If no human score is assigned, this is the final score. |

### Backreading

Backreading was one of the major responsibilities of Pearson Scoring Supervisors and a primary tool for proactively guarding against scorer drift, where scorers score responses in comparison to one another instead of in comparison to the training responses. Scoring supervisory staff 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 five percent of the hand-scored responses. Backreading scores did not override the original score but were used to monitor scorer performance.

## Validity

Validity responses are pre-scored responses strategically interspersed in the pool of live responses. These responses were not distinguishable from any other responses so that scorers were not aware they were scoring validity responses rather than live responses. The use of validity responses provided an objective measure that helped ensure that scorers were applying the same standards throughout the project. In addition, validity was at times shared with scorers in a process known as "validity as review." Validity as review provided scorers automated, immediate feedback: 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.

Validity agreement requirements for scorers are listed in Table 4.4. Scorers had to meet the required validity agreement percentages to continue working on the project. Scorers who did not maintain expected agreement statistics were given a series of interventions culminating in a targeted calibration set: a test of scorer knowledge. Scorers who did not pass targeted calibration were removed from scoring the item, and all the scores they assigned were deleted.

### Table 4.4 Scoring Validity Agreement Requirements

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

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

## Calibration Sets

Calibration sets are special sets created during scoring to help train scorers on particular areas of concern or focus. Scoring directors used calibration sets to reinforce rangefinding standards, introduce scoring decisions, or address scoring issues and trends. 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.

## Inter-rater Agreement

Inter-rater agreement is the agreement between the first and second scores assigned to student responses and is the measure of how often scorers agree with each other. Pearson scoring staff used inter-rater agreement statistics as one factor in determining the needs for continuing training and intervention on both individual and group levels. Inter-rater agreement expectations are shown in Table 4.5.

Table 4.5 Inter-rater Agreement Expectations and Results

| Subject | Score Point Range | Perfect Agreement Expectation | Perfect Agreement Result | Within One Point Expectation* | Within One Point Result |
|---|---|---|---|---|---|
| Mathematics | 0–1 | 90% | 98% | 96% | 100% |
| Mathematics | 0–2 | 80% | 97% | 96% | 100% |
| Mathematics | 0–3 | 70% | 96% | 96% | 99% |
| Mathematics | 0–4 | 65% | 94% | 95% | 99% |
| Mathematics | 0–5 | 65% | 91% | 95% | 98% |
| Mathematics | 0–6 | 65% | 95% | 95% | 98% |
| ELA/L | Multi-trait | 65% | 83% | 96% | 100% |

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

Pearson's ePEN2 scoring 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.

The perfect agreement rate for mathematics responses scored by two scorers ranged from 76 to 100 percent and the within one point rate ranged from 96 to 100 percent. For all ELA/L responses scored by two scorers, the perfect agreement rate ranged from 69 to 100 percent, and the within one point rate ranged from 97 to 100 percent.

The results by grade level for ELA/L are provided in Section 4.3.7: Inter-rater Agreement for Prose Constructed-Response.

## 4.3 Automated Scoring for PCRs

Automated scoring performed by Pearson's Intelligent Essay Assessor (IEA) was the default option for scoring the summative assessment's online prose constructed-response (PCR) tasks. Under the default option, it was assumed that operational scores for approximately 90 percent 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 percent of responses, a second "reliability" score was assigned. The purpose of the reliability score was 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.

### 4.3.1 Concepts Related to Automated Scoring

The text below describes concepts related to automated scoring.

#### Continuous Flow

Continuous flow scoring results in an integrated connection between human scoring and automated scoring. It refers to a system of scoring where either an automated score, a human score, or both can be assigned based on a predetermined asynchronous operational flow.

#### Training of IEA using Operational Data

Continuous flow scoring facilitates the training of IEA using human scores assigned to operational online data collected early in the administration. Once IEA obtains sufficient data to train, it can be "turned on" and becomes the primary source of scoring (although human scoring continues for the 10 percent reliability sample and other responses that may be routed accordingly).

## 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.

## Quality Criteria for Evaluating Automated Scoring

The state leads approved specific quality criteria for evaluating automated scoring. The primary evaluation criteria for IEA was based on responses to validity papers with "known" scores assigned by experts. For each prompt 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. 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 were proposed based on the research literature (Williamson et al., 2012). These measures were previously utilized in Pearson's automated scoring research and include Pearson correlation, kappa, quadratic - weighted kappa, exact agreement, and standardized mean difference. These measures are computed between pairs of human scores, as well as 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 between IEA-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- Quadratic - weighted kappa between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA-human should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria and are noted below.

**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: With smart routing applied as needed, IEA-human exact agreement is within 5.25 percent of human-human exact agreement for each trait score.

**Secondary Criteria**—Based on the training responses: With smart routing applied as needed, IEA-human differences on statistical measures for each trait score are within the Williamson et al. tolerances for subgroups with at least 50 responses.

## Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the following hierarchy determines which score was reported operationally:

- The IEA score is reported if it is the only score assigned.
- If an IEA score and a human score are assigned, the human score is reported.
- If a first human score and a second human score are assigned, the first human score is reported.
- If a backread score and human and/or IEA scores are assigned, the backread score is reported if there is no resolution or adjudication score assigned.
- If a resolution score is assigned and an adjudicated score is not assigned, the resolution score is reported (note that if nonadjacent scores are encountered, responses are automatically routed to resolution).
- If an adjudicated score is assigned, it is reported (note that if a resolution score is nonadjacent to the other scores assigned, responses are automatically routed to adjudication).

### 4.3.2 Sampling Responses Used for Training IEA

For prompts trained using 2022 operational data, the early performance of human scoring was closely monitored to verify that an appropriate set of data would be available for training IEA. In particular, several characteristics of the human scoring data were monitored, including

- Exact agreement between human scorers (the goal was for this to be at least 65 percent for each trait).
- Exact agreement between human scores conditioned on score point (the goal was for this to be at least 50 percent 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 prompts, they were more challenging to achieve for others. For some prompts, a subset of scores were reset, and clarifying directions were provided to scorers to improve human-human agreement. For other prompts, special sampling approaches were used to increase the numbers of responses that received top scores. In addition, a healthy percentage of responses were backread during the sampling period, and these scores, as well as double human scores, were all part of the data used to train IEA.

### 4.3.3 Primary Criteria for Evaluating IEA Performance

The primary criteria for evaluating IEA performance is based on evaluating validity papers and is stated as follows: With smart routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.

To operationalize the primary criteria for a given prompt, the following general steps are undertaken:

1. Determine agreement of the human scores with the validity papers for each trait.
2. Calculate agreement of the IEA scores with the validity papers for each trait.
3. Compare the IEA validity agreement with the human agreement.
4. If the IEA validity agreement is greater than or equal to the human agreement for each trait, IEA can be deployed operationally.

In addition to looking at overall validity agreement, conditional agreement was also examined. In general, it was desirable for IEA to exceed 65 percent agreement at every score point as well as be close to or exceed the human validity agreement at each score point.

### 4.3.4 Contingent Primary Criteria for Evaluating IEA Performance

For many of the prompts trained in 2023, it was not possible to utilize human-scored validity responses in evaluating IEA performance. 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. A portion of the data was held out for evaluating IEA-human exact agreement according to the following steps:

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

In addition to the overall comparison, the following performance thresholds were targeted in the test data set: 1) at least 65 percent overall IEA-human agreement; and 2) 50 percent IEA-human agreement by score point (i.e., conditioned on the human score). These targets went beyond the contingent primary criteria approved by the state leads.

### 4.3.5 Applying Smart Routing

With smart routing, the quality of automated scoring can be increased by routing responses that are more likely to disagree with a human score to receive an additional human score.

When human scorers read a paper, they typically apply integer scores based on a scoring rubric. When there is strong agreement between two independent human readers, the readers might both assign a score of 3 such that the average score over both raters is also a 3 (i.e., (3+3)/2 = 3). IEA simulates this behavior, but because its scores come from an artificial intelligence algorithm, it generates continuous (i.e., decimalized) scores. In this case, the IEA score might be a 2.9 or 3.1. When human readers disagree on the score for a paper, say one reader gives the paper a score of 3 and another reader gives the paper a score of 4, the average of the two scores would be 3.5 (i.e., 3+4=7/2=3.5). For this paper, IEA would likely provide a score between 3 and 4, say 3.4 or 3.6. Because this continuous score needs to be rounded to an integer score for reporting, it might be reported as a 3 or a 4, depending on the rounding rules. Smart routing involves routing those responses with "in between" IEA scores to additional human scoring because the nature of the responses suggests there may be less confidence in the IEA score. Since these "in between" IEA scores are based on modeling human scores, it follows that human scores may be less certain as well, and thus such responses tend to be the ones that it makes sense to have double-scored and possibly to resolve if the IEA and human scores are nonadjacent.

Smart routing was utilized as needed to help IEA achieve targeted quality metrics (e.g., validity agreement or agreement with human scorers). Smart routing involved the application of the following four steps:

1. The continuous IEA score for each of the two trait scores was rounded to the nearest score interval of 0.2, starting from zero. For example, IEA scores between 0 and 0.1 were rounded to an interval score of 0, scores between 0.1 and 0.3 were rounded to an interval score of 0.2, scores between 0.3 and 0.5 were rounded to an interval score of 0.4, and so on.
2. Within each of these intervals, the percentage of exact agreement between IEA integer scores and the human scores was calculated for each trait.
3. For each prompt, agreement rates were evaluated by rounding interval. Those intervals for which the agreement rates were below a designated threshold for either trait were identified.
4. Once IEA scoring was implemented, responses within intervals for which IEA-human agreement was below the designated threshold were routed for additional human scoring.

In training IEA, the scoring models without smart routing were evaluated first by applying either the primary validity criteria or the contingent criteria as described in Section 4.3. For those prompts that did not meet these criteria, increasing smart routing thresholds were applied in an iterative fashion to filter scores and evaluate the remaining scores against the criteria. That is, in any one iteration a particular smart routing threshold was applied such that only scores falling in intervals for which exact agreement exceeded the threshold were included in evaluating the criteria. If the primary or contingent criteria were not met with this level of smart routing, an increased smart routing threshold was applied iteratively until the primary or contingent criteria were met, or the maximum threshold reached. If the criteria were still not met after a maximum threshold was applied, different models were investigated and/or additional human scoring data utilized until an IEA scoring model was found that met the criteria.

### 4.3.6 Evaluation of Secondary Criteria for Evaluating IEA Performance

The secondary criteria for evaluating IEA performance involved comparing agreement indices for IEA-human scoring for various demographic subgroups. Because of the importance of protecting personally identifiable information (PII), student demographic data is stored and managed separately from the performance scoring data. For this reason, it was not possible to evaluate subgroup performance in real time as IEA was being trained.

For those prompts trained on early operational data, attempts were made to prioritize the data being returned from the field to include data from states or districts where more diverse populations of students were anticipated. In addition, requests for additional human scores were made to increase the likelihood that there would be sufficient numbers of responses with two human scores for most of the demographic subgroups of interest.

Once IEA was trained and deployed, scoring sets used in training were matched to demographic information so that agreement between IEA and human scorers could be evaluated across subgroups. The analysis was conducted for the ten comparison groups listed in Table 4.6:

Table 4.6 Comparison Groups

| Comparison Groups | |
|---|---|
| Gender | Female |
| | Male |
| Ethnicity | American Indian/Alaska Native |
| | Asian |
| | Black/African American |
| | Hispanic/Latino |
| | Native Hawaiian/Pacific Islander |
| | Two or More Races |
| | White |
| Special Instructional Needs | English Language Learners (EL) |
| | Students with Disabilities (SWD) |
| | Economically Disadvantaged |

IEA-human agreement indices were calculated for all cases with an IEA score and at least one human score. Human-human agreement was calculated for all cases with two human scores.

To evaluate the training of IEA for subgroups, the following criteria approved by the state leads for subgroups with at least 50 IEA-human scores and at least 50 human-human scores were applied:

- Pearson correlation between IEA-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- Quadratic - weighted kappa between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA-human should be less than ±0.15 (this criterion was applied to subgroups with at least 50 IEA-human scores).

Although it was not expected that these criteria would be met for all subgroups for all prompts, if results of the evaluation between IEA and human scoring for subgroups for any prompt indicated that IEA performance persistently failed on the criteria listed above, consideration would be given to resetting the responses scored by IEA and reverting to human scoring until such time that an alternate IEA model could be established with improved subgroup performance.

In addition to the secondary criteria approved by the State Leads, the performance of IEA was compared to the following targets on the various measures for subgroups with at least 50 responses:

- Pearson correlation between IEA-human should be 0.70 or above.
- Kappa between IEA-human should be 0.40 or above.
- Quadratic - weighted kappa between IEA-human should be 0.70 or above.
- Exact agreement between IEA-human should be 65 percent or above.

These targets were not intended to be directly applied in decisions about whether to deploy IEA operationally or not. Such targets may or may not be met by human scoring for any particular prompt and/or subgroup, and if they are not met by human scoring, they are unlikely to be met by IEA scoring. Nevertheless, comparisons to these targets provided additional information about IEA performance (and human scoring) in an absolute sense.

### 4.3.7 Inter-rater Agreement for Prose Constructed-response

This section presents the inter-rater agreement for operational results for the online Prose Constructed-Response (PCR) tasks by trait and grade level. PCR items are scored on two traits: (1) Reading Comprehension and Written Expression and (2) Knowledge of Language and Conventions for Research Simulation for Literary Analysis tasks and (1) Written Expression and (2) Knowledge of Language and Conventions for the Narrative task.

For 10 percent of responses, a second "reliability" score was assigned. The purpose of the reliability score is 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 and is the measure of how often scorers agree with each other. Pearson scoring staff used inter-rater agreement indices as one factor in determining the needs for continuing training and intervention on both individual and group levels. Inter-rater agreement expectations are provided in Table 4.5 in Section 4.2.4. For ELA/L PCR traits, the expectation for agreement is an inter-rater agreement of 65 percent 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 4.7 presents the average agreement across the PCRs for each grade level by trait. The number of prompts included in the analyses is listed for each grade level. The agreement indices (exact agreement, kappa, quadratic-weighted kappa, and Pearson correlation) were calculated separately by PCR for each trait (Reading Comprehension and Written Expression or Written Expression and Conventions). For each grade level, the agreement indices were averaged across the PCRs. Table 4.7 presents the average count and the average for the agreement indices.

The exact agreement for the PCR traits is above the criteria of a 65 percent agreement rate for all PCRs. The strength of agreement between raters is moderate to substantial agreement as defined by Landis and Koch (1977) for all PCRs. The quadratic-weighted kappa (QW Kappa) distinguishes between differences in ratings that are close to each other versus larger differences. The weighted kappa is substantial to almost perfect agreement for all grades. The Pearson correlations (*r*) ranged from 0.75 to 0.95.

Table 4.7 PCR Average Agreement Indices by Test

| | | | Written Expression | | | | Conventions | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Number of PCRs | Count | Exact | Kappa | QW Kappa | *r* | Exact | Kappa | QW Kappa | *r* |
| 3 | 4 | 26875 | 74.43 | 0.57 | 0.74 | 0.75 | 75.1 | 0.6 | 0.78 | 0.78 |
| 4 | 4 | 26078 | 73.63 | 0.61 | 0.81 | 0.81 | 73.08 | 0.6 | 0.8 | 0.8 |
| 5 | 4 | 28513 | 73.45 | 0.6 | 0.81 | 0.81 | 73.55 | 0.6 | 0.81 | 0.82 |
| 6 | 3 | 51516 | 75.93 | 0.65 | 0.86 | 0.86 | 76.5 | 0.66 | 0.85 | 0.85 |
| 7 | 3 | 74197 | 74.17 | 0.64 | 0.87 | 0.87 | 73.83 | 0.64 | 0.86 | 0.86 |
| 8 | 4 | 29737 | 75.3 | 0.66 | 0.9 | 0.9 | 75.03 | 0.66 | 0.88 | 0.88 |

# Section 5: Classical Item Analysis

## 5.1 Overview

This section describes the results of the classical item analysis conducted for data obtained from the operational test items. All ELA/L and mathematics assessments were pre-equated. The item statistics provided in this section were from prior operational administrations and reflect the statistics that were used in test construction and for score reporting. Item analysis serves two purposes: to inform item exclusion decisions for IRT analysis and to provide item statistics for the item bank. Item analysis included data from the following types of items: key-based selected-response items, rule-based machine-scored items, and hand-scored constructed-response items. For each item, the analysis produced item difficulty, item discrimination, and item response frequencies.

## 5.2 Data Screening Criteria

Item analyses were conducted by test form based on administration mode. In preparation for item analysis, student response files were processed to verify that the data were free of errors. Pearson Customer Data Quality (CDQ) 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.

Before beginning item analysis, Pearson performed the following data screening operations:

1. All records with an invalid form number were excluded.
2. All records that were flagged as “void” were excluded.
3. All records where the student attempted fewer than 25 percent of items were excluded.
4. For students with more than one valid record, the record with the higher raw score was chosen.
5. Records for students with administration issues or anomalies were excluded.

## 5.3 Description of Classical Item Analysis Statistics

A set of classical item statistics were computed for each operational item. Each statistic was designed to evaluate the performance of each item.

The following statistics and associated flagging rules were used to identify items that were not performing as expected:

### Classical item difficulty indices (p-value and average item score)

When constructing tests, a wide range of item difficulties is desired (i.e., from easy to hard items) so that students of all ability levels can be assessed with precision. At the operational stage, item difficulty statistics are used by test developers to build forms that meet desired test difficulty targets.

For dichotomously scored items, item difficulty is indicated by its p-value, which is the proportion of students who answered that item correctly. P-values range from 0 to 1, with higher values indicating easier items and lower numbers indicating more difficult items. Dichotomously scored 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).

For polytomously scored items, difficulty is indicated by the average item score (AIS). The AIS can range from 0 to the maximum total possible points for an item. To facilitate interpretation, the AIS values for polytomously scored items are often expressed as percentages of the maximum possible score, which are equivalent to the p-values of dichotomously scored items and thus range from 0 to 1. Polytomously scored items were flagged for review if the p-value was above 0.95 or below 0.25.

### The percentage of students choosing each response option

Selected-response items on the summative assessments refer primarily to single-select multiple-choice dichotomously scored items. These items require that the student select a response from a number of answer options. These statistics for single-select multiple-choice items indicate the percentage of students who select each of the answer options and the percentage that did not respond to

(omitted) the item. The percentages are also computed for the high-performing subgroup of students who scored at the top 20 percent on the assessment. Items were 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 mis-keyed.

### Item-total correlation

This statistic describes the relationship between students' performance on a specific item and their performance on the total test. For operational item analysis, the total score on the assessment was used as the total test score. The item-total correlation was calculated for both selected-response items and constructed-response 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 can range from -1 to 1. Desired values are positive and larger than 0.15. Negative item-total correlations indicate that low-ability students perform better on an item than high-ability students, an indication that the item may be potentially flawed. Item-total correlations below 0.15 were flagged for review.

### Distractor-total correlation

For selected-response items, this estimate describes the relationship between selecting an incorrect response (i.e., a distractor) for a specific item and performance on the total test. The item-total correlation is calculated for the distractors. Items with distractor-total correlations above 0 were flagged for review as these items may have multiple correct answers, be mis-keyed, or have other content issues.

### Percentage of students omitting or not reaching each item

For both selected-response and constructed-response items, this statistic 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 percent of students should attempt to answer each question 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.
- An item is considered "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 to complete the test. Items with high omit rates were flagged. Omit rates for constructed-response items tend to be higher than for selected-response items. Therefore, the omit rate for flagging individual items was 5 percent for selected-response items and 15 percent for constructed-response items. If a student omitted an item, then the student 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.

### 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 students' 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 just unexpectedly difficult), the scoring rubric is flawed, and/or students did not have an opportunity to learn the content. In addition, if all or most students score at the extreme ends of the distribution (e.g., 0 and 2 for a 3-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 percent) of students obtaining any score point were flagged. In addition, constructed-response items were flagged if they had U-shaped distributions, with high frequencies for extreme scores and very low frequencies for middle score categories.

## 5.4 Summary of Classical Item Analysis Flagging Criteria

In summary, items are flagged for review if the item analysis yielded any of the following results:

1. P-value above 0.95 for dichotomous items or polytomous items

2. P-value below 0.25 for dichotomous items or polytomous items
3. Item-total correlation below 0.15
4. Any distractor-total correlation above 0
5. Greater number of high-performing students (top 20 percent) choosing a distractor rather than the keyed response
6. High percentage of omitted responses: above 5 percent for selected-response items and above 15 percent for constructed-response items
7. High percentage that did not reach the item: above 5 percent for selected-response items and above 15 percent for constructed-response items
8. Constructed-response items with a score value obtained by less than 3 percent of responses

The procedure was for Pearson's psychometric staff to review any flagged items and submit them to the Priority Alert Task Force to decide if the items were problematic and should be excluded from scoring.

## 5.5 Classical Item Analysis Results

This section presents tables summarizing the analyses for items on the spring operational forms. All assessments were pre-equated, meaning that the scoring was based on item parameters estimated using data from earlier administrations. Item analysis results in this section are the item statistics from prior administrations that were used to make decisions during the test construction process and for scoring.

- Table 5.1 presents pre-administration p-value information by grade for the ELA/L operational items.
- Table 5.2 presents pre-administration p-value information by grade for the mathematics operational items.
- Table 5.3 presents pre-administration item-total correlations by grade for the ELA/L operational items.
- Table 5.4 presents pre-administration item-total correlations by grade for the mathematics operational items.

An operational item may appear on multiple test forms. The tables list 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 item analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, or multiple/no correct answers.

### Table 5.1 Pre-Administration P-values for ELA/L Operational Items by Grade

| Grade | N of Unique | Mean | SD | Min | Max | Median |
|---|---|---|---|---|---|---|
| 3 | 32 | 0.47 | 0.16 | 0.16 | 0.74 | 0.44 |
| 4 | 51 | 0.46 | 0.17 | 0.15 | 0.78 | 0.46 |
| 5 | 51 | 0.48 | 0.19 | 0.13 | 0.85 | 0.46 |
| 6 | 45 | 0.47 | 0.15 | 0.20 | 0.79 | 0.44 |
| 7 | 51 | 0.45 | 0.14 | 0.14 | 0.82 | 0.43 |
| 8 | 51 | 0.49 | 0.15 | 0.19 | 0.84 | 0.46 |

Table 5.2 Pre-Administration P-values for Mathematics Operational Items by Grade

| Grade | N of Unique | Mean | SD | Min | Max | Median |
|---|---|---|---|---|---|---|
| 3 | 87 | 0.54 | 0.23 | 0.07 | 0.93 | 0.55 |
| 4 | 85 | 0.49 | 0.22 | 0.07 | 0.94 | 0.46 |
| 5 | 84 | 0.47 | 0.23 | 0.09 | 0.95 | 0.46 |
| 6 | 80 | 0.38 | 0.20 | 0.06 | 0.92 | 0.38 |
| 7 | 86 | 0.39 | 0.21 | 0.05 | 0.84 | 0.34 |
| 8 | 82 | 0.33 | 0.18 | 0.05 | 0.71 | 0.29 |

Table 5.3 Pre-Administration Item-Total Correlations for ELA/L Operational Items by Grade

| Grade | N of Unique | Mean | SD | Min | Max | Median |
|---|---|---|---|---|---|---|
| 3 | 32 | 0.55 | 0.14 | 0.29 | 0.83 | 0.55 |
| 4 | 51 | 0.48 | 0.15 | 0.20 | 0.84 | 0.45 |
| 5 | 51 | 0.52 | 0.14 | 0.20 | 0.88 | 0.50 |
| 6 | 45 | 0.50 | 0.12 | 0.35 | 0.89 | 0.48 |
| 7 | 51 | 0.50 | 0.15 | 0.28 | 0.88 | 0.48 |
| 8 | 51 | 0.47 | 0.13 | 0.12 | 0.81 | 0.47 |

Table 5.4 Pre-Administration Item-Total Correlations for Mathematics Operational Items by Grade

| Grade/ | N of Unique | Mean | SD | Min | Max | Median |
|---|---|---|---|---|---|---|
| 3 | 87 | 0.57 | 0.15 | 0.19 | 0.79 | 0.59 |
| 4 | 85 | 0.54 | 0.14 | 0.25 | 0.83 | 0.56 |
| 5 | 84 | 0.54 | 0.15 | 0.16 | 0.81 | 0.55 |
| 6 | 80 | 0.55 | 0.16 | 0.15 | 0.79 | 0.58 |
| 7 | 86 | 0.54 | 0.16 | 0.20 | 0.80 | 0.56 |
| 8 | 82 | 0.50 | 0.15 | 0.18 | 0.81 | 0.51 |

# Section 6: Differential Item Functioning

## 6.1 Overview

Differential item functioning (DIF) compares two groups of performance-matched students to determine whether student performance was significantly different between the two groups. DIF analyses were conducted using the data obtained from the operational items. If an item performs differentially across identifiable subgroups (e.g., gender, ethnicity groups, etc.) when students are matched on ability, the item may be measuring something other than the intended construct (i.e., possible evidence of DIF). It is important, however, 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. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

Subgroups of interest are separated into two categories, referred to as the "focal" and "reference" groups, for comparison purposes. Historically, the reference group was defined as the group assumed to be more likely to be advantaged than the focal group (e.g., white students are typically defined as the reference group when compared to other ethnicities). However, the definition of these groups is arbitrary and the DIF procedure is not impacted by which group is defined as focal, and which is defined as reference.

In this section, the DIF statistics used at test construction to make decisions about items are provided for all online and paper mathematics and ELA/L tests.

## 6.2 DIF Procedures

### Dichotomous Items

The Mantel-Haenszel (MH) DIF statistic was calculated for dichotomously scored selected-response items and constructed-response items. In this method, students are classified into relevant subgroups of interest (e.g., males and females). Using the raw score total as the criteria, students in a certain total score category in the focal group (e.g., females) are compared with students in the same total score category in the reference group (e.g., males). For each item, students in the focal group are also compared to students in the reference group who performed equally well on the test as a whole. 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.

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} \frac{R_{rs} W_{fs}}{N_{ts}}}{\sum_{s=1}^{S} \frac{R_{fs} W_{rs}}{N_{ts}}}, \quad (6\text{-}1)$$

in which:

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

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

$$MH\ D - DIF = -2.35\ ln(\hat{\alpha}_{MH}) \quad (6\text{-}2)$$

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).

## Polytomous Items

For polytomously scored constructed-response items, the MH D-DIF statistic is not calculated. Instead, the standardized DIF (Dorans & Schmitt, 1991; Zwick et al., 1997; Dorans, 2013), in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959), is used to identify items with DIF.

The standardized DIF 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 using the following formula:

$$STD - EISDIF = \frac{\sum_{s=1}^{S} N_{fs} \times E_f(Y|X=s)}{\sum_{s=1}^{S} N_{fs}} - \frac{\sum_{s=1}^{S} N_{fs} \times E_r(Y|X=s)}{\sum_{s=1}^{S} N_{fs}}, \quad (6\text{-}3)$$

in which:

$X$ = the total score,
$Y$ = the item score,
$S$ = the number of score categories,
$N_{rs}$= the number of students in the reference group in score category s,
$N_{fs}$= 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.

A positive *STD-EISDIF* value means that, conditional on the total test score, the focal group has a higher mean item score than the reference group. In contrast, a negative *STD-EISDIF* value means that, conditional on the total test score, the focal group has a lower mean item score than the reference group.

## Classification

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C (Zieky, 1993). Category A items contain negligible difference in performance; Category B items exhibit slight to moderate differences in performance; and Category C items possess moderate to large differences in performance. Positive values indicate that, conditional on the total score, the focal group has a higher mean item score than the reference group. In contrast, negative DIF values indicate that, conditional on the total test score, the focal group has a lower mean item score than the reference group. The flagging criteria for dichotomously scored items are presented in Table 6.1; the flagging criteria for polytomously scored constructed-response items are provided in Table 6.2.

Table 6.1 DIF Categories for Dichotomous Selected-Response and Constructed-Response Items

| DIF Category | Criteria |
|---|---|
| A (negligible) | Absolute value of the MH D-DIF is not significantly different from zero or is less than one. |
| 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<br>2. Absolute value of the MH D-DIF is significantly different from one but is less than 1.5.<br>Positive values are classified as "B+" and negative values as "B-". |
| C (moderate to large) | Absolute value of the MH D-DIF is significantly different from one and is at least 1.5.<br>Positive values are classified as "C+" and negative values as "C-". |

Table 6.2 DIF Categories for Polytomous Constructed-Response Items

| DIF Category | Criteria |
|---|---|
| A (negligible) | Mantel Chi-square p-value > 0.05 or $\lvert STD\text{-}EISDIF/SD \rvert \leq 0.17$ |
| B (slight to moderate) | Mantel Chi-square p-value < 0.05 and $\lvert STD\text{-}EISDIF/SD \rvert > 0.17$ |
| C (moderate to large) | 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.

## 6.3 Operational Analysis DIF Comparison Groups

DIF analyses were conducted for designated comparison groups defined on the basis of demographic variables including gender, race/ethnicity, economic disadvantage, and special instructional needs such as students with disabilities (SWD) or English learners (EL). Student demographic information was provided by Illinois and obtained from PearsonAccess$^{next}$ by means of a student data upload. The demographic data was verified by Illinois prior to score reporting. These comparison groups are specified in Table 6.3.

Table 6.3 Traditional DIF Comparison Groups

| Grouping Variable | Focal Group | Reference Group |
|---|---|---|
| Gender | Female | Male |
| Ethnicity | American Indian/Alaska Native | White |
| | Asian | White |
| | Black/African American | White |
| | Hispanic/Latino | White |
| | Native Hawaiian/Pacific Islander | White |
| | Multiple Race Selected | White |
| Economic Status* | Economically Disadvantaged | Not Economically Disadvantaged |
| Special Instructional Needs | English Learner | Non-English Learner |
| | Students with Disabilities | Students without Disabilities |

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

DIF analyses were conducted when the following sample size requirements were met:

1. The reference and focal groups had at least 100 students each.
2. The combined group (reference and focal) had at least 400 students.

## 6.4 Operational Differential Item Functioning Results

Appendix 6 presents tables summarizing the DIF results for the spring pre-administration item DIF results that were used to inform decisions at test construction for both ELA/L and mathematics, as well as the post-administration item DIF results for ELA/L. There is one table prepared for each content and grade level (e.g., ELA/L Grade 3). Tables 6.4-6.5 below provide examples for ELA/L and mathematics grade 3.

Spoiled or "do not score" items were excluded from the total test score for each form in 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 in this section may include items for certain grade levels that were excluded from scoring based on later analyses (refer to Section 7.5 Items Excluded from Score Reporting for more information).

In the DIF results tables, the column "DIF Comparisons" identifies the focal and reference groups for the analysis performed; "Total N of Unique Items" reports the number of unique items included in the analysis. "Total N of Item Occurrences Included in DIF Analysis" reports the number of occurrences with sufficient sample sizes to be included in DIF analyses. Because DIF analysis is conducted at the parent level for PCRs in ELA/L tests, 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 (see Tables 5.1 and 5.2) and the IRT summary statistics (see Tables 7.7–7.9) for each ELA/L test.

Table 6.4 Pre-Administration Differential Item Functioning for ELA/L Grade 3

| | | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DIF Comparison | Total N of Unique Items | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 32 | | | | | 32 | 100 | | | | |
| White vs Black | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Hispanic | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Asian | 32 | | | | | 31 | 97 | 1 | 3 | | |
| White vs American Indian | 32 | | | | | 32 | 100 | | | | |
| White vs Pacific Islander | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Two or more races | 32 | | | | | 32 | 100 | | | | |
| Not vs Economically Disadvantaged | 32 | | | | | 32 | 100 | | | | |
| Non vs English Learners | 32 | 1 | 3 | 1 | 3 | 30 | 94 | | | | |
| Without vs Students with Disabilities | 32 | | | | | 32 | 100 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not English learner, Without = not student with disability.

Table 6.5 Pre-Administration Differential Item Functioning for Mathematics Grade 3

| | | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DIF Comparison | Total N of | N | % of | N | % of | N | % of | N | % of | N | % of |
| Female vs Male | 87 | | | 2 | 2 | 83 | 95 | 2 | 2 | | |
| White vs Black | 87 | 1 | 1 | 7 | 8 | 78 | 90 | 1 | 1 | | |
| White vs Hispanic | 87 | | | 1 | 1 | 85 | 98 | 1 | 1 | | |
| White vs Asian | 87 | | | 1 | 1 | 82 | 94 | 4 | 5 | | |
| White vs American Indian | 87 | 1 | 1 | | | 86 | 99 | | | | |
| White vs Pacific Islander | 87 | | | 2 | 2 | 85 | 98 | | | | |
| White vs Two or more races | 87 | | | 1 | 1 | 85 | 98 | 1 | 1 | | |
| Not vs Economically | 87 | | | 2 | 2 | 85 | 98 | | | | |
| Non vs English Learners | 87 | | | 2 | 2 | 84 | 97 | 1 | 1 | | |
| Without vs Students with Disabilities | 87 | | | 3 | 3 | 84 | 97 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not English learner, Without = not student with disability.

# Section 7: IRT Model and Parameters

## 7.1 Overview

Multiple operational core forms were administered for each grade in ELA/L and mathematics assessments. All tests in spring 2023 were pre-equated. This section describes the item response theory (IRT) model used in this assessment program and provides descriptive statistics of the item parameters.

## 7.2 Two-Parameter Logistic/Generalized Partial Credit Model

The operational items used pre-equated parameters in the context of the 2PL/GPC model, which is denoted as

$$p_{im}(\theta_j) = \frac{exp\left[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})\right]}{\sum_{v=0}^{M_i-1} exp\left[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{ik})\right]} \quad (7\text{-}1)$$

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 $M_i - 1$.

## 7.3 Summary Statistics and Distributions from IRT Analyses

Tables 7.1 through 7.4 present summary statistics for the IRT (*b*- and *a*-) parameter estimates, the standard errors (SEs) of the parameter estimates, and the IRT model fit values (chi-square and adjusted fit) for ELA/L and mathematics assessments. The summary statistics for IRT parameter estimates include all the items administered in the spring 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 information is provided by content area (ELA/L and mathematics) for all items at each grade level or course. The summary statistics shown include the total number of items and score points, along with the mean, standard deviation (SD), minimum, and maximum.

### 7.3.1 IRT Summary Statistics for English Language Arts/Literacy

Table 7.1 shows the pre-equated *b*- and *a*-parameter estimates for all ELA/L assessments. Table 7.2 shows the source year for the item statistics for each of the ELA/L assessments. IRT summary statistics are provided in Appendix 7 for ELA/L for all items, reading claim items, and writing claim items.

Table 7.1 Pre-Equated IRT Parameter Estimates Summary for All Items for ELA/L by Grade

| | | | *b* Estimates Summary | | | | *a* Estimates Summary | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Grade | No. of Score Points | No. of Items | Mean | SD | Min | Max | Mean | SD | Min | Max |
| 3 | 80 | 36 | 0.36 | 0.92 | -2.12 | 1.69 | 0.59 | 0.22 | 0.25 | 1.04 |
| 4 | 125 | 56 | 0.49 | 1.00 | -1.30 | 2.66 | 0.48 | 0.24 | 0.18 | 1.06 |
| 5 | 126 | 56 | 0.42 | 1.07 | -1.65 | 3.59 | 0.51 | 0.24 | 0.13 | 1.02 |
| 6 | 109 | 49 | 0.39 | 0.74 | -1.02 | 1.86 | 0.50 | 0.23 | 0.24 | 1.16 |
| 7 | 125 | 56 | 0.36 | 0.70 | -1.47 | 1.60 | 0.50 | 0.28 | 0.13 | 1.30 |
| 8 | 125 | 56 | 0.15 | 0.92 | -2.98 | 2.38 | 0.49 | 0.25 | 0.08 | 1.14 |

Table 7.2 Pre-Equated IRT Parameter Distribution by Year for All Items for ELA/L by Grade

| Grade | No. of Items | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2022 |
|---|---|---|---|---|---|---|---|---|
| 3 | 36 | 0 | 0 | 0 | 9 | 3 | 13 | 11 |
| 4 | 56 | 0 | 0 | 0 | 12 | 3 | 19 | 22 |
| 5 | 56 | 0 | 0 | 0 | 3 | 10 | 9 | 34 |
| 6 | 49 | 0 | 4 | 0 | 7 | 17 | 9 | 12 |
| 7 | 56 | 0 | 5 | 10 | 5 | 12 | 13 | 11 |
| 8 | 56 | 0 | 0 | 1 | 12 | 14 | 18 | 11 |

### 7.3.2 IRT Summary Statistics for Mathematics

Table 7.3 shows the *b*- and *a*-parameter estimates for the mathematics assessments. Table 7.4 shows the source year for the item statistics for each of the assessments. IRT summary statistics are provided in Appendix 7 for mathematics for all items, single-select multiple-choice items, constructed-response items, and subclaims.

Table 7.3 Pre-Equated IRT Parameter Estimates Summary for All Items for Mathematics by Grade

| Grade | No. of Score Points | No. of Items | *b* Estimates Summary | | | | *a* Estimates Summary | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | Mean | SD | Min | Max |
| 3 | 139 | 87 | -0.24 | 1.24 | -2.56 | 3.68 | 0.77 | 0.28 | 0.19 | 1.42 |
| 4 | 143 | 85 | -0.05 | 1.09 | -2.65 | 2.36 | 0.71 | 0.23 | 0.31 | 1.46 |
| 5 | 144 | 84 | 0.01 | 1.15 | -2.34 | 2.13 | 0.69 | 0.25 | 0.18 | 1.50 |
| 6 | 142 | 80 | 0.45 | 1.12 | -3.57 | 4.46 | 0.71 | 0.30 | 0.16 | 1.54 |
| 7 | 144 | 86 | 0.61 | 1.08 | -2.23 | 3.42 | 0.73 | 0.33 | 0.19 | 1.72 |
| 8 | 139 | 82 | 0.95 | 1.00 | -1.70 | 2.70 | 0.60 | 0.28 | 0.10 | 1.44 |

Table 7.4 Pre-Equated IRT Parameter Distribution by Year for All Items for Mathematics by Grade

| Grade | No. of Items | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2022 |
|---|---|---|---|---|---|---|---|---|
| 3 | 87 | 0 | 14 | 7 | 5 | 6 | 19 | 36 |
| 4 | 85 | 0 | 10 | 17 | 7 | 3 | 17 | 31 |
| 5 | 84 | 0 | 13 | 8 | 10 | 5 | 19 | 29 |
| 6 | 80 | 0 | 12 | 6 | 9 | 6 | 10 | 37 |
| 7 | 86 | 0 | 8 | 11 | 8 | 6 | 15 | 38 |
| 8 | 82 | 0 | 12 | 7 | 6 | 6 | 9 | 42 |

# Section 8: Performance Level Setting

## 8.1 Performance Standards

Performance standards relate levels of performance on an assessment directly to what students are expected to learn. This is done by establishing threshold scores that distinguish between performance levels. Performance level setting (PLS) is the process of establishing these threshold scores that define the performance levels for an assessment.

## 8.2 Performance Levels and Policy Definitions

For the summative assessments, the performance levels are:

Level 5: Exceeded expectations

Level 4: Met expectations

Level 3: Approached expectations

Level 2: Partially met expectations

Level 1: Did not yet meet expectations

Detailed descriptions of each performance level, known as policy definitions, are as follows:

### Level 5: Exceeded expectations

Students performing at this level exceed academic expectations for the knowledge, skills, and practices contained in the standards assessed at their grade level or course.

Grades 3–8: Students performing at this level exceed academic expectations for the knowledge, skills, and practices contained in the standards for English language arts/literacy (ELA/L) or mathematics assessed at their grade level. They 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 standards assessed at their grade level or course.

Grades 3–8: Students performing at this level meet academic expectations for the knowledge, skills, and practices contained in the standards for ELA/L or mathematics assessed at their grade level. They 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 standards assessed at their grade level or course.

Grades 3–8: Students performing at this level approach academic expectations for the knowledge, skills, and practices contained in the standards for ELA/L or mathematics assessed at their grade level. They 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 standards assessed at their grade level or course.

Grades 3–8: Students performing at this level partially meet academic expectations for the knowledge, skills, and practices contained in the standards for ELA/L or mathematics assessed at their grade level. They 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 standards assessed at their grade level or course.

Grades 3–8: Students performing at this level do not yet meet academic expectations for the knowledge, skills, and practices contained in the standards for ELA/L or mathematics assessed at their grade level. They will need academic support to engage successfully in further studies in this content area.

## 8.3 Performance Level Setting Process for the Assessment System

One of the main objectives of the Affiliate/PARCC 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, participating states and agencies used the evidence-based standard setting (EBSS) method (Beimers et al., 2012) for the PLS 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 seven steps of the EBSS process that were followed to establish performance standards for the summative assessments are:

1. Define 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 performance level setting (PLS) meetings with panels.
6. Conduct reasonableness review with post-policy panel.
7. Continue to gather evidence in support of standards.

A summary of key components within these steps is provided below. Additional detail about each step in the PLS process is provided in the *Performance Level Setting Technical Report*.

### 8.3.1 Research Studies

Participating states and agencies conducted two research studies in support of their policy goals—the benchmarking study and the postsecondary educators' judgment (PEJ) study. The benchmarking study included a review of the literature relative to college- and career-readiness as well as consideration of the percentage of students obtaining a level equivalent to college- and career-readiness on a set of external assessments (e.g., ACT, SAT, NAEP). The PEJ study involved a group of nearly 200 college faculty reviewing items on the Algebra II and ELA/L grade 11 assessments and making judgments about the level of performance needed on each item to be academically ready for an entry-level college-credit bearing course in mathematics or ELA/L. Additional detail[2] about the benchmarking study can be found in the *Performance Level Setting Technical Report* as well as in the *PARCC Benchmarking Study Report*. Additional detail about the PEJ study can be found in the *Performance Level Setting Technical Report* as well as in the *Postsecondary Educators' Judgment Study Final Report*.

### 8.3.2 Pre-Policy Meeting

Prior to the PLS meetings, a pre-policy meeting was convened to determine reasonable ranges that would be shown to panelists during the high school PLS meetings. Pre-policy meeting participants included representatives from both K–12 and higher education

[2] More information is available online from https://resources.newmeridiancorp.org/research/.

who served in roles such as commissioner/superintendent, deputy/assistant commissioner, state board member, director of assessment, director of academic affairs, senior policy associate, and so on. The reasonable ranges recommended by the pre-policy meeting defined the minimum and maximum percentage of students that would be expected to be classified as college- and career-ready. The pre-policy meeting participants reviewed the test purpose, how the performance standards will be used, and the results of the research studies to provide the recommendations for the reasonable ranges without viewing any student performance data.

### 8.3.3 Performance Level Setting Meetings

The task of the PLS committee was to recommend four threshold scores that would define the five performance levels for each assessment. Participating states and agencies solicited nominations from all states that had administered the assessments in 2014–2015 for panelists to serve on the PLS committees. 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 those educators who had content knowledge as well as experience with a variety of student groups and attempted to balance the panels in terms of state representation.

Participating states and agencies used an extended modified Angoff (Yes/No) method to collect educator 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 extension to the Yes/No standard setting method (Plake et al., 2005) allowed for incorporation of the multipoint items by asking educators to evaluate (Yes or No) 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. In the case of a single point or multiple-choice item, this task simplifies to the standard Yes/No method.

After receiving training on the PLS procedure, panelists participated in three rounds of judgments for each assessment. Within each round, panelists were asked to consider the items in the test form, starting with the performance-based assessment (PBA) component and then the end-of-year (EOY) component. Each panelist made a judgment for the Level 2 performance level, followed by judgments for the Level 3 performance level, the Level 4 performance level, and the Level 5 performance level, in this order. The panelists entered their item judgments for each round by completing an online item judgment survey. Educator judgments were summed across items to create an estimated total score on the reference form for each performance level threshold. Feedback data relative to panelist agreement, student performance on the items, and student performance on the test as a whole were provided in between each of the three rounds of judgment. Panelists were shown the pre-policy reasonable ranges prior to making their Round 1 judgments and again as feedback data following each round of judgment.

A dry run of the PLS meeting process was held for grade 11 ELA/L and Algebra II in order to evaluate the implementation of the PLS method with the innovative characteristics of the summative assessments. These content areas were selected because they combined all the various aspects of the assessments, including the various types of items, scoring rules, and performance level decisions. The dry-run PLS meetings provided the opportunity to implement and evaluate multiple aspects of the operational plan for the actual PLS meeting, including pre-work, meeting materials, data analysis and feedback, and staff and panelist functions. The results of the dry-run PLS meeting were used to implement improvements in the process for the operational PLS meetings. Additional information about the methods and results of the dry-run PLS meeting is available in the Performance Level Setting Dry-Run meeting report.

The PLS meetings for the summative assessments were conducted in three one-week sessions. The dates of the twelve PLS committee meetings are shown in Table 8.1.

Additional information about the methods and results of the PLS meetings is available in the *Performance Level Setting Technical Report*.

### 8.3.4 Post-Policy Reasonableness Review

Performance standards for all summative assessments were recommended by PLS committees and reviewed by the Governing Board and (for the Algebra II, Integrated Mathematics III, and ELA/L grade 11 assessments) the Advisory Committee on College Readiness as part of a post-policy reasonableness review. This group reviewed both the median threshold score recommendations from each committee and the variability in the threshold scores as represented by the standard error of judgment (SEJ) of the committee. Adjustments to the median threshold scores that were within 2 SEJ were considered to be consistent with the PLS panels' recommendation.

Table 8.1 Performance Level Setting Committee Meetings and Dates

| Dates | Committees by Subjects and Grades |
|---|---|
| **July 27–31, 2015** | Algebra I/Integrated Mathematics I |
| | Geometry/Integrated Mathematics II |
| | Algebra II/Integrated Mathematics III |
| | Grade 9 English Language Arts/Literacy |
| | Grade 10 English Language Arts/Literacy |
| | Grade 11 English Language Arts/Literacy |
| **August 17–21, 2015** | Grades 7 & 8 Mathematics |
| | Grades 7 & 8 English Language Arts/Literacy |
| **August 24–28, 2015** | Grades 3 & 4 Mathematics |
| | Grades 5 & 6 Mathematics |
| | Grades 3 & 4 English Language Arts/Literacy |
| | Grades 5 & 6 English Language Arts/Literacy |

In addition to voting to adopt the performance standards based on the committee's 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 current 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. In the first step, the top two performance levels, above college- and career-ready (or on-track), were combined into a single performance level and an additional performance level below college- and career-ready (or on-track) was created by empirically determining the midpoint between the existing two levels. In the second step, the performance level descriptors (PLDs) were updated using items that discriminated student performance well at this level to create a PLD aligned with the new empirically determined performance level. At this same time, PLDs for all performance levels were reviewed for consistency and continuity. Members of the original PLS committees were recruited to participate in this process. Additional information about this process can be found in the *Performance Level Setting Technical Report*.

# Section 9: 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.

## 9.1 Quality Control of the Item Bank

The summative item bank consists of test passages and items, their associated metadata, and status (e.g., operational-ready, field-test ready, released, etc.). The Affiliate/PARCC banked items on the assessments were developed by Pearson and West Ed and put in the item bank once created. Custom Illinois items were created by New Meridian subject matter experts and were reviewed by Illinois educators.

Pearson's 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 items and tools and manages versioning of items with a date/time stamp. It allows reviewers to vote on item acceptance, and to record and retain their review notes for later reconciliation and reference. Item and passage review committee participants conducted their review in the item banking system. The committee members viewed the items as the student would, and could vote to alter the item, or accept or reject the item, then record their comments in the system. After each meeting, reports were forwarded to New Meridian. The reports were generated by the item banking system and summarized feedback from the committee reviewers.

All new development for the summative assessments is being created within the ABBI system, which employs templates to control the consistency of the underlying scoring logic and QTI creation for each item type. The ABBI system incorporates a previewer that allows the reviewers to validate the content of the item and validate the expected scoring of tasks. It supports the full range of review activities, including content review, bias and sensitivity review, expert editorial review, data review, and test construction review. It provides insight into the item edit process through versioning. A series of metadata validations at key points in the development cycle provide support for metadata consistency. The bank can be queried on the full range of metadata values to support bank analysis.

## 9.2 Quality Control of Test Form Development

Test forms were built based upon targets and the established blueprints set. The construction process started with specification and requirement capture to create the test specification document. From there items were pulled into forms based on the criteria approved in the test specifications document. After forms composition, the forms went through a review process that involved subject matter experts from New Meridian Illinois. Quality control steps were conducted on the items and forms evaluating several item characteristics (e.g., content accuracy, completeness, style guide conformity, tools function). Revisions were incorporated into the forms before final review and approval. Section 2.2 provides more details on the form development process.

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 validates 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. (Quality processes for paper forms are described in Section 9.3.)

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 the 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 (OE) responses string that contained a valid range of characters. The test maps also were validated against the form layout and item types for correctness as part of these tests.

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, validated that tools and functions in the test delivery system, TestNav, were accurately applied, and verified that the style and layout met all requirements. In addition, answer keys were validated through a formal key review process. More details on the test development procedures are provided in Section 2.

## 9.3 Quality Control of Test Materials

Pearson provided high quality materials in a timely and efficient manner to meet the test administration needs. Since the majority of printing work was done in-house, it was possible to fully control the production environment, press schedule, and quality process for print materials. Additionally, strict security requirements were employed to protect secure materials production; Section 3 provides details on the secure handling of test materials. 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 (DPP) for digital print methodology, or plateroom for offset print methodology. Both the DPP 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 (DPL) 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:

- 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 (QA) department personnel inspected print output prior to collation and shipment. QA also supported process improvement, work area documentation, audited process adherence, and established training programs for employees.

## 9.4 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 (OMR) 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.

## 9.5 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, 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. These included student n-count checks to make certain of the following:

- Students were placed under the correct header.
- All sheets belonged to the appropriate document.
- Documents were not scanned twice.
- 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.

## 9.6 Quality Control of Answer Document Processing and Scoring

Quality control of answer document processing and scoring involves all aspects of the scoring procedures, including key-based and rule-based machine scoring, and hand-scoring for constructed-response items and performance tasks.

For the 2015 operational administration, Pearson's validation team prepared test plans used throughout the scoring process. Test plan preparation was organized around detailed specifications.

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 non-score; 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/non-updatable fields; negative test for non-existing record or empty file)

The following components were added to the quality control process specifically for the program. These additional steps were introduced to address issues with item-level scoring that were identified in the 2014 field-test administration:

- XML Validation: A combination of automated validation against 100 percent 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 key checks) 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.

In addition to the steps described above, the following quality control process for answer keys and scoring that was implemented for the first operational administration was used:

1. Pearson's psychometrics team conducted empirical analyses based on preliminary data files and flagged items based on statistical criteria.
2. Pearson content team reviewed the flagged items and provided feedback on the accuracy of content, answer keys, and scoring.
3. Items potentially requiring changes were added to the product validation (PV) log for further investigation by other Pearson teams.
4. Staff was notified of items for which keys or scoring changes were recommended.
5. Participating states and agencies approved/rejected scoring changes.
6. All approved scoring changes were implemented and validated prior to the generation of the data files used for psychometric processing.

## 9.7 Quality Control of Psychometric Processes

High quality psychometric work for the operational administrations was necessary to provide accurate and reliable results of student performance. Pearson was responsible for the psychometric analyses of the operational administration and implemented measures to ensure the quality of work. The psychometric analyses were all conducted according to well-defined specifications. 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.

Described below is an overview of the quality control steps performed at different stages of the psychometric analyses. Greater detail is provided in Sections 5 (Classical Item Analysis), 6 (Differential Item Functioning), 7 (IRT Model and Parameters), and 12 (Scale Scores).

### Data Screening

Data screening is an important first step to ensure quality data input for meaningful analysis. The Pearson Customer Data Quality (CDQ) team validated all student data files used in the operational psychometric analyses. The data validation for the student data files (SDF) and item response files (IRF) 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 (UINs), 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 (SIRB) to the item scores.
6. Flagged student records with inconsistencies for further investigation.

## Classical Item Analysis

Classical item analysis (IA) produces item level statistics (e.g., item difficulty and item-total correlations). The IA results were reviewed by Pearson psychometricians. Items flagged for unusual statistical properties were reviewed by the content team. If items were identified as having key issues, scoring issues, or content issues, they were presented to the Priority Alert Task Force, whose task was to make decisions on whether to exclude them from the calculation of reported student scores. Refer to Section 5.4 for classical IA item flagging criteria.

## Conversion Tables

Conversion tables were computed and constructed by New Meridian psychometrics. Conversion tables must be accurate because they are used to generate reported scores for students. Comprehensive records were meticulously 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 psychometric team members 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 12.3 for the procedure to create conversion tables.

# Section 10: Operational Test Forms

Each operational test form is constructed to reflect the blueprints for Illinois Assessment of Readiness. Multiple operational forms are constructed for each grade/subject. The test construction process determined the CCSS that are assessed in more than one evidence statement when selecting the items for the spring 2023 blueprint. The reduction of items attempted to keep the proportion of subclaims close to the original, while still maintaining enough points to report at the subclaim level. The process adhered to the CCSSO criteria for procuring and evaluating high-quality assessments.

Core forms are the operational test forms consisting of only those items that will count toward a student's score. Core forms are constructed to meet the blueprint and psychometric properties outlined in the test construction specifications. New Meridian creates multiple core forms for a given assessment to enhance test security and to support opportunity for item release. The number of core operational forms per grade/subject and mode is provided in Table 10.1.

Table 10.1 Number of Core Operational Forms per Grade/Subject and Mode

| Grade/Subject | ELA/L | | Mathematics | |
|---|---|---|---|---|
| | CBT | PBT | CBT | PBT |
| 3 | 2 | 1 | 2 | 1 |
| 4 | 2 | 1 | 2 | 1 |
| 5 | 2 | 1 | 2 | 1 |
| 6 | 2 | 1 | 2 | 1 |
| 7 | 2 | 1 | 2 | 1 |
| 8 | 2 | 1 | 2 | 1 |

CBT = computer-based test; PBT = paper-based test.

In addition to the operational core forms, appropriate forms were identified as accessibility and accommodated forms. ELA/L assessments have two operational accommodated forms and mathematics assessments have three accommodated forms. The forms are accommodated to support Braille, large print, human reader/human signers, assistive technology, text-to-speech, closed captioning, and Spanish. Human reader/human signers and Spanish are provided for mathematics assessments only. Closed captioning is provided for ELA/L assessments only.

# Section 11: Student Characteristics

## 11.1 Overview of Test Taking Population

Almost 800,000 forms were administered in Illinois during the 2022–2023 school year. Assessments were administered for ELA/L and mathematics in grades 3 through 8 as computer-based tests (CBT) and paper-based tests (PBT). Paper-based tests were offered only as an accommodation. The majority of students tested online with small numbers of paper testers.

## 11.2 Rules for Inclusion of Students in Analyses

Criteria for inclusion of students were implemented prior to all operational analyses. These rules were established by Pearson psychometricians in consultation with Illinois to determine which, if any, student records should be removed from analyses. This data screening process resulted in higher quality, albeit slightly smaller, data sets.

Student response data were included in analyses if

1. Valid form numbers were observed for each unit for online assessments or for the full form for paper assessments.
2. Student records were not flagged as “void” (i.e., do not score).
3. The student attempted at least 25 percent of the items in each unit or form.

Additionally, in cases where students had more than one valid record, the record with the higher raw score was retained and the other record(s) deleted. Records for students with administration issues or anomalies were excluded from analyses.

## 11.3 Students by Grade and Mode

Table 11.1 presents, for each grade of ELA/L, the number and percentage of students who took the test in each mode (CBT or PBT). This information is provided for all participating states combined. Table 11.2 presents the same type of information for all students who took the mathematics assessments. Note that Table 11.2 includes mathematics data for both English-language and Spanish-language test forms combined. Table 11.3 provides this information for students who took the mathematics assessments in Spanish only.

Markedly more students tested online than on paper across all grades for both content areas. This is expected since PBT were offered only as an accommodation. For ELA/L, the percentages of online students by grade level were greater than 99 percent. For all mathematics students, the percentages of students testing online was greater than 99 percent. The percentages of students taking Spanish-language mathematics online forms was greater than or equal to 99 percent, except in grades 7 and 8 where the percentage of students was approximately 98 percent. Overall, fewer students tested at the higher grades for both content areas.

Table 11.1 ELA/L Students by Grade and Mode

| Grade | No. of Valid Cases | CBT | | PBT | |
|---|---|---|---|---|---|
| | | N | % | N | % |
| 3 | 128,356 | 127,546 | 99.4 | 810 | 0.6 |
| 4 | 127,980 | 127,426 | 99.6 | 554 | 0.4 |
| 5 | 129,738 | 129,178 | 99.6 | 560 | 0.4 |
| 6 | 133,179 | 132,658 | 99.6 | 521 | 0.4 |
| 7 | 134,267 | 133,713 | 99.6 | 554 | 0.4 |
| 8 | 138,908 | 138,412 | 99.6 | 496 | 0.4 |
| Grand Total | 792,428 | 788,933 | 99.6 | 3,495 | 0.4 |

Note: Includes students taking accommodated forms of ELA/L. CBT = computer-based test; PBT = paper-based test.

Table 11.2 Mathematics Students by Grade and Mode

| Grade | No. of Valid Cases | CBT | | PBT | |
|---|---|---|---|---|---|
| | | N | % | N | % |
| 3 | 128,109 | 127,432 | 99.5 | 677 | 0.5 |
| 4 | 127,833 | 127,260 | 99.6 | 573 | 0.4 |
| 5 | 129,562 | 128,986 | 99.6 | 576 | 0.4 |
| 6 | 132,858 | 132,338 | 99.6 | 520 | 0.4 |
| 7 | 133,956 | 133,378 | 99.6 | 578 | 0.4 |
| 8 | 138,558 | 138,044 | 99.6 | 514 | 0.4 |
| Grand Total | 790,876 | 787,438 | 99.6 | 3,438 | 0.4 |

Note: Includes students taking mathematics in English, students taking Spanish-language forms for mathematics, and students taking accommodated forms. CBT = computer-based test; PBT = paper-based test.

Table 11.3 Spanish-Language Mathematics Students by Grade and Mode

| Grade | No. of Valid Cases | CBT | | PBT | |
|---|---|---|---|---|---|
| | | N | % | N | % |
| 3 | 3,541 | 3,521 | 99.4 | 20 | 0.6 |
| 4 | 2,588 | 2,572 | 99.4 | 16 | 0.6 |
| 5 | 2,321 | 2,303 | 99.2 | 18 | 0.8 |
| 6 | 1,985 | 1,973 | 99.4 | 12 | 0.6 |
| 7 | 1,485 | 1,451 | 97.7 | 34 | 2.3 |
| 8 | 1,489 | 1,460 | 98.1 | 29 | 1.9 |
| Grand Total | 13,409 | 13,280 | 99.0 | 129 | 1.0 |

Note: CBT = computer-based test; PBT = paper-based test.

## 11.4 Demographics

Appendix 11 presents student demographic information for the following characteristics: gender, race/ethnicity (American Indian/Alaska Native; Asian; Black/African American; Hispanic/Latino; White/Caucasian; Native Hawaiian/Pacific Islander; two or more races reported; race not reported), English learners (EL), economic disadvantage status, and students with disabilities. Student demographic information was provided by Illinois and captured in PearsonAccess$^{next}$ by means of a student data upload. The demographic data was verified by Illinois prior to score reporting. Not all demographics were provided for all students. Students missing information on one or more demographic variables were omitted from the corresponding subgroup analyses.

All tables of demographic information are organized by subject and grade. Percentages are not reported in which fewer than 20 students tested in a given grade for the subject.

# Section 12: Scale Scores

Illinois reports results according to five performance levels that delineate the knowledge, skills, and practices students are able to demonstrate:

Level 5: Exceeded expectations

Level 4: Met expectations

Level 3: Approached expectations

Level 2: Partially met expectations

Level 1: Did not yet meet expectations

The assessments are designed to measure and report results in categories called master claims and subclaims. Master claims (or simply "claims") are at a higher level than subclaims with content representing multiple subclaims contributing to each claim outcome. In addition, four scale scores are reported for the assessments. A summative scale score is reported for each mathematics assessment. A summative scale score and separate claim scores for Reading and Writing are reported for each English language arts/literacy (ELA/L) assessment.

Subclaim outcomes describe student performance for content-specific subsets of the item scores contributing to a particular claim. For example, Written Expression and Knowledge of Conventions subclaim outcomes are reported along with Writing claim scores. Subclaim outcomes are reported as *Below Expectations*, *Nearly Meets Expectations*, or *Meets or Exceeds Expectations*.

## 12.1 Operational Test Content (Claims and Subclaims)

A claim is a statement about student performance based on how students respond to test questions. The tests 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 Common Core State Standards (CCSS).

The number of items associated with each claim and subclaim outcome varies depending on subject and grade. The item types vary in terms of the number of points associated with them, so that both the number of items and the number of points are important in evaluating the quality of a claim or subclaim score.

### 12.1.1 English Language Arts/Literacy

Table 12.1[3] includes the number of items and the number of points by subclaim and claim for ELA/L grade 3. Corresponding information is provided in Appendix 12.1 for all ELA/L grades.

Table 12.1 Form Composition for ELA/L Grade 3

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 4–7 | 8–17 |
| | Reading Informational Text | 4–7 | 11–20 |
| | Vocabulary | 4–5 | 8–10 |
| | **Claim Total** | **12–14** | **30–31** |
| Writing | | | |
| | Written Expression | 1 | 18 |
| | Knowledge of Conventions | 1 | 6 |
| | **Claim Total** | **2** | **24** |
| **Summative Total** | | **14–16** | **54–55** |

Note: Each Prose Constructed-Response (PCR) trait is identified as a separate item in this table for the two writing subclaims and, in some cases, either the Reading Literary Text or the Reading Informational Text subclaim.

[3] Table A.12.1 in Appendix 12.1 is identical to Table 12.1.

Each ELA/L form contains items of varying types. The Prose Constructed-Response (PCR) traits contribute to different claims and the aggregate of the traits contributes to the summative scale score. ELA/L assessments consist of two prose constructed-response tasks. The following details the number of possible points and the associated subclaims for the three PCR tasks:

Literary Analysis Task

Research Simulation Task

Narrative Writing Task

All ELA/L assessments include the 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 two traits: Reading Comprehension and 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; the Written Expression traits are 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. The maximum possible points for ELA/L PCR items are provided in Table 12.2.

Table 12.2 Contribution of Prose Constructed-Response Items to ELA/L for all Grades

| | | Possible Points | | |
|---|---|---|---|---|
| **Grade** | **Score** | **Literary Analysis Task*** | **Research Simulation Task*** | **Narrative Writing Task*** |
| 3 | Reading | 3 | 3 | 0 |
| | Written Expression | 9 | 9 | 9 |
| | Knowledge of Conventions | 3 | 3 | 3 |
| | **Summative Total** | **15** | **15** | **12** |
| 4–5 | Reading | 4 | 4 | 0 |
| | Written Expression | 12 | 12 | 9 |
| | Knowledge of Conventions | 3 | 3 | 3 |
| | **Summative Total** | **19** | **19** | **12** |
| 6–8 | Reading | 4 | 4 | 0 |
| | Written Expression | 12 | 12 | 12 |
| | Knowledge of Conventions | 3 | 3 | 3 |
| | **Summative Total** | **19** | **19** | **15** |

*ELA/L assessments consist of the Research Simulation Task and either the Literary Analysis Task or the Narrative Writing Task.

## 12.1.2 Mathematics

Table 12.3[4] includes the numbers of items and points associated with subclaim scores for mathematics grade 3, as an example of the composition of the mathematics tests. Because there is substantial variation in the composition of the tests corresponding information is provided in the tables in Appendix 12.1 for all mathematics grades/courses.

Table 12.3 Mathematics Form Composition for Grade 3

| | **Subclaims** | **Number of Items** | **Number of Points** |
|---|---|---|---|
| Mathematics | | | |
| | Major Content | 18 | 20 |
| | Additional & Supporting Content | 9 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| **Summative Total** | | **33** | **52** |

[4] Table A.12.9 in Appendix 12.1 is identical to Table 12.3.

## 12.2 Establishing the Reporting Scales

Reporting scales designate student performance into one of five performance levels[5] 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 performance level setting (PLS) 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.

### 12.2.1 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 is 650 and the highest obtainable scale score is 850. The threshold 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 $(\boldsymbol{\theta}_{2015})$ scale. The theta values associated with the Level 2 and Level 4 performance levels were identified using the test characteristic curve associated with the performance level setting form. With Levels 2 and 4 scale scores fixed at 700 and 750, respectively, the relationship between theta $(\boldsymbol{\theta}_{2015})$ and scale scores($ScaleScore_{2015}$) was established as

$$ScaleScore_{2015} = A_{2015} \times \theta_{2015} + B_{2015} \quad (12\text{-}1)$$

where $\boldsymbol{A}_{2015}$ is the slope and $\boldsymbol{B}_{2015}$ is the intercept. The slope and intercept were established as

$$A_{2015} = \frac{750-700}{\theta_{2015_{Level4}} - \theta_{2015_{Level2}}} \quad (12\text{-}2)$$

and

$$B_{2015} = 750 - A_{2015} \times \theta_{2015_{Level4}} \quad (12\text{-}3)$$

As indicated by these formulas, the slope and intercept for the summative scale scores were based on the theta scale, and by default the IRT parameter scale, established in 2015. Since the spring 2016 IRT parameter scale is the base scale for the IRT parameters, the scaling constants $A_{2015}$ and $B_{2015}$ 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 ($ScaleScore_{2015}$):

$$ScaleScore_{2015} = SA_{2016} \times \theta_{2016} + SB_{2016} \quad (12\text{-}4)$$

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 ($\theta_{2015}$) and the 2016 theta scale ($\theta_{2016}$). 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:

[5] Section 8 provides an overview of the performance level setting process, and detailed information can be found in the Performance Level Setting Technical Report.

$$SA_{2016} = \frac{A_{2015}}{slope_{2015\_to\_2016}} \quad (12\text{-}5)$$

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:

$$SB_{2016} = B_{2015} - A_{2016} \times intercept_{2015\_to\_2016} \quad (12\text{-}6)$$

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.

*A* and *B* values resulting from these calculations as well as the theta values associated with the threshold performance levels are included in Appendix 12.2. Also, the 2015–2016 technical report includes raw to scale score conversion tables for the performance level setting forms.

## 12.2.2 ELA/L 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 test characteristic curve associated with the performance level 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. The formulas used for this are provided in Table 12.4.

Table 12.4 Calculating Scaling Constants for Reading and Writing Claim Scores

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

A and B values resulting from these calculations are included in Appendix 12.

## 12.2.3 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 career and college 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; or

Meets or Exceeds Expectations.

The subclaim performance levels are designated through the IRT theta ($\theta$) 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 test characteristic curve. 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*. Other students whose raw subclaim score fell between the Level 3 and 4 thresholds were designated as *Nearly Meets Expectations*.

## 12.3 Creating Conversion Tables

A conversion table relates the number of points earned by a student on the ELA/L summative score, the mathematics summative score, the Reading claim score, or the Writing claim score to the corresponding scale score for the test form administered to that student. An IRT inverse test characteristic curve (TCC) approach is used to develop the relationship between point scores and theta, $\theta_s$, (IRT ability estimates). In carrying out 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:

$$s_i(\theta_j) = \sum_{m=0}^{M_i-1} m p_{im}(\theta_j) \quad (12\text{-}7)$$

$$p_{im}(\theta_j) = \frac{exp[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})]}{\sum_{v=0}^{M_i-1} exp[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{iv})]} \quad (12\text{-}8)$$

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, $\theta_j$; $p_{im}(\theta_j)$ 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 $k$; $v$ is the number of score categories.

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

$$T_j = \sum_{i=1}^{I} w_i \, s_i(\theta_j) \quad (12\text{-}9)$$

where $T_j$ is the expected (weighted) test score on theta, $\theta_j$; $w_i$ 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); $I$ 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:

$$CSEM_j = \sqrt{\frac{1}{\sum_{i=1}^{I} L_i(\theta_j)}} \quad (12\text{-}10)$$

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

$$s_{i2}(\theta_j) = \sum_{m=0}^{M_i-1} m^2 p_{im}(\theta_j) \quad (12\text{-}12)$$

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

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

**Step 5:** Calculate the reported scale score. Using the $A$ and $B$scaling constants in Appendix 12.2, convert each theta value to a scale score and each theta CSEM to a scale score CSEM:

$$ScaleScore = A \times \theta + B \quad (12\text{-}13)$$

$$CSEM = CSEM_{\theta} \times A \quad (12\text{-}14)$$

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.

Figure 12.1 contains TCCs, estimated CSEM curves, and estimated information (INF) curves for ELA/L grade 3. The curves in each figure are for the two core online forms (O1 and O2), one core paper form (P1), and one or more accommodated forms A(O). The curves are reported on the theta scale. Vertical dotted lines indicate the performance level cuts on the theta scale. For ELA/L grade 3, all forms had similar TCCs. CSEM and INF curves were also similar.

Appendix 12 contains TCC, CSEM, and INF curves for all ELA/L grades and all mathematics grades/courses. The curves are based on IRT parameters from prior operational or field-test administrations.

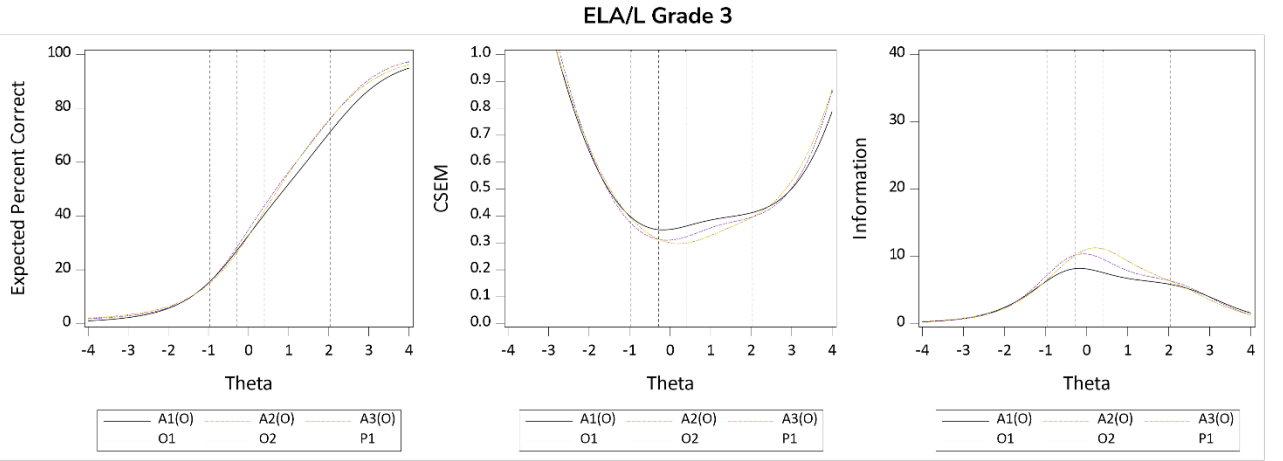



Figure 12.1 TCC, CSEM, and TIC for ELA/L Grade 3

# 12.4 Score Distributions

## 12.4.1 Score Distributions for ELA/L

Figures 12.2 through 12.4 graphically represent the distributions of scale scores for grades 3 through 8 ELA/L summative, Reading, and Writing, respectively. The vertical axis of each graph, labeled "Density," represents the proportion of students earning the scale score point indicated along the horizontal axis. For the summative distributions, the *y*-axis ranges from 0 to 0.02 and the *x*-axis from 650 to 850. For the Reading distributions, the *y*-axis ranges from 0 to 0.05 and the *x*-axis from 10 to 90. For the Writing distributions, the *y*-axis ranges from 0 to 0.10 and the *x*-axis from 10 to 60.

The distributions of the ELA/L summative scale scores were roughly symmetrical and centered around the Level 4 cut score (750) or slightly below. Some distributions, particularly grade 8, exhibited a slight bimodal distribution.

Reading scale scores tended to be centered around or slightly below the Level 4 cut score of 50 and were slightly more irregular than the summative scale scores. Distributions tended to be more evenly distributed.

Writing scale score distributions were noticeably less smooth than 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 less 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 ten in Writing across all ELA/L grades. Students with zero 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, there are relatively few students between 11 and about 20, depending on the grade.[6] As noted in Section 12.2.2, the scale score task force selected ten as the LOSS. This value 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 12.4). For Writing, the lowest theta estimate associated with raw scores ranging from one to two are linearly transformed to scale score values generally between 15 and 20, meaning that there may be multiple scale scores between 11 and 20 that are not assigned to a raw score. Whereas 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 ten and the scale score values around seventeen to nineteen is an artifact of scale score task force selecting the LOSS value of ten.

[6] 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.

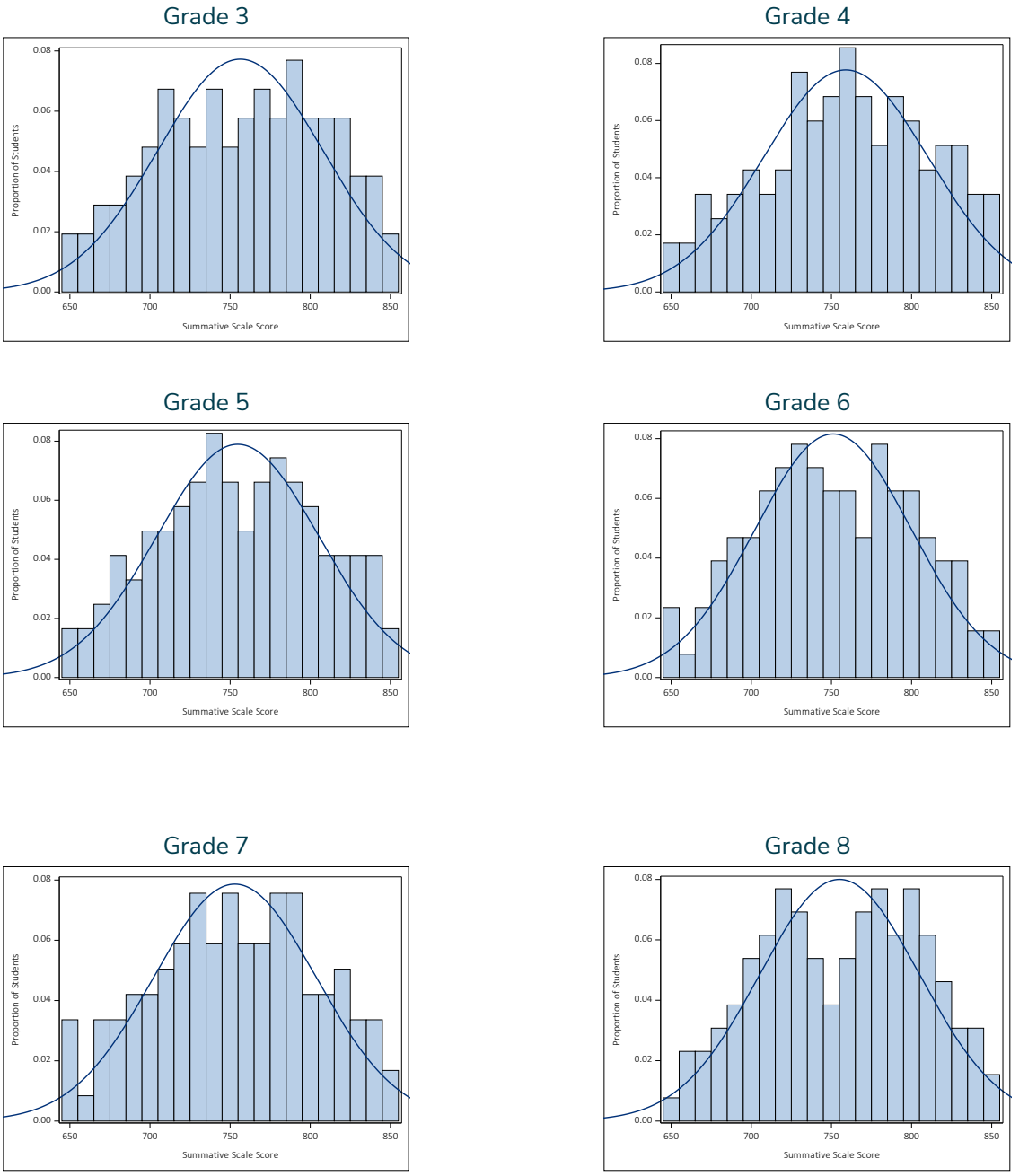



Figure 12.2 Distributions of ELA/L Scale Scores: Grades 3–8

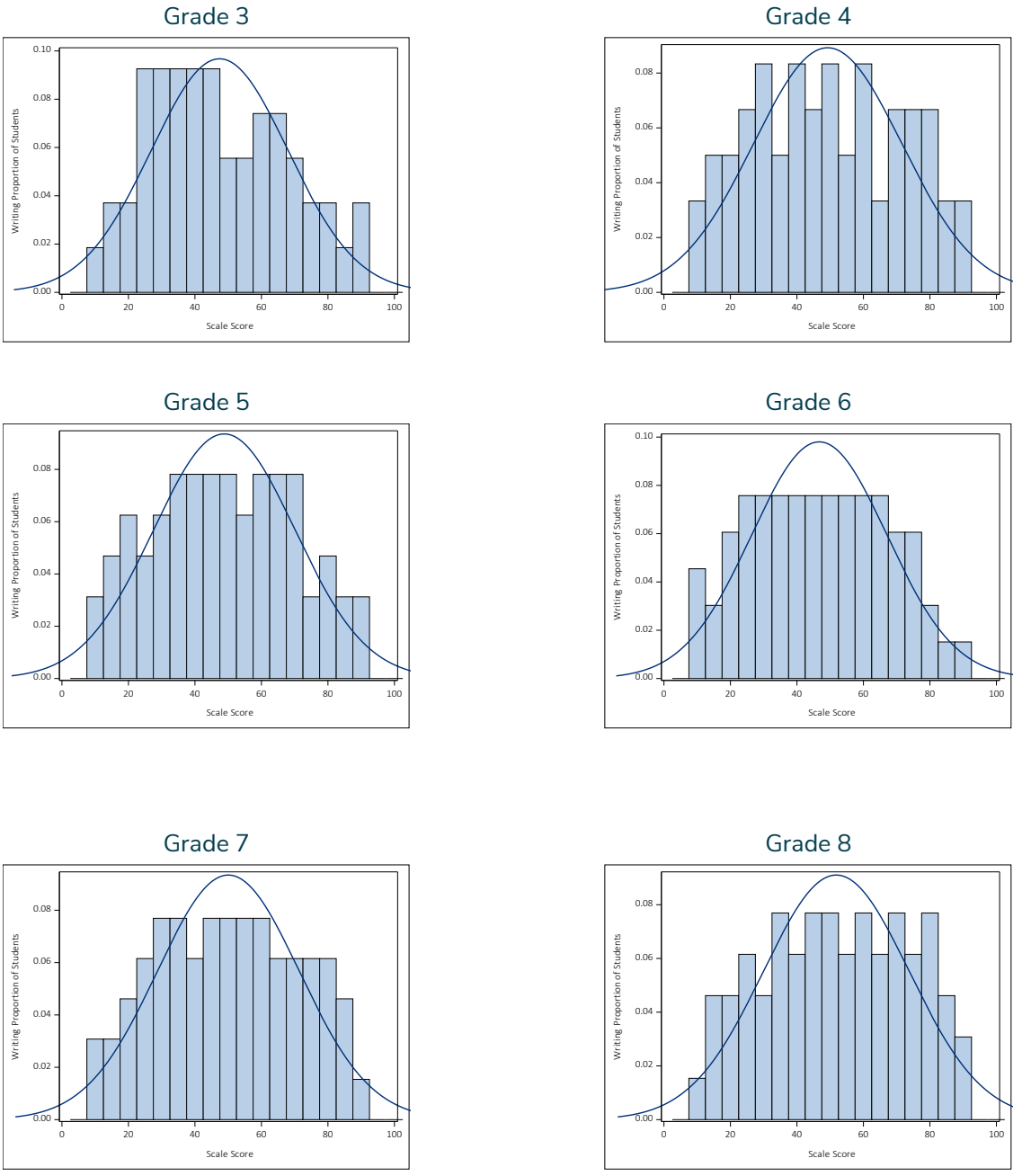



Figure 12.3 Distributions of Reading Scale Scores: Grades 3–8

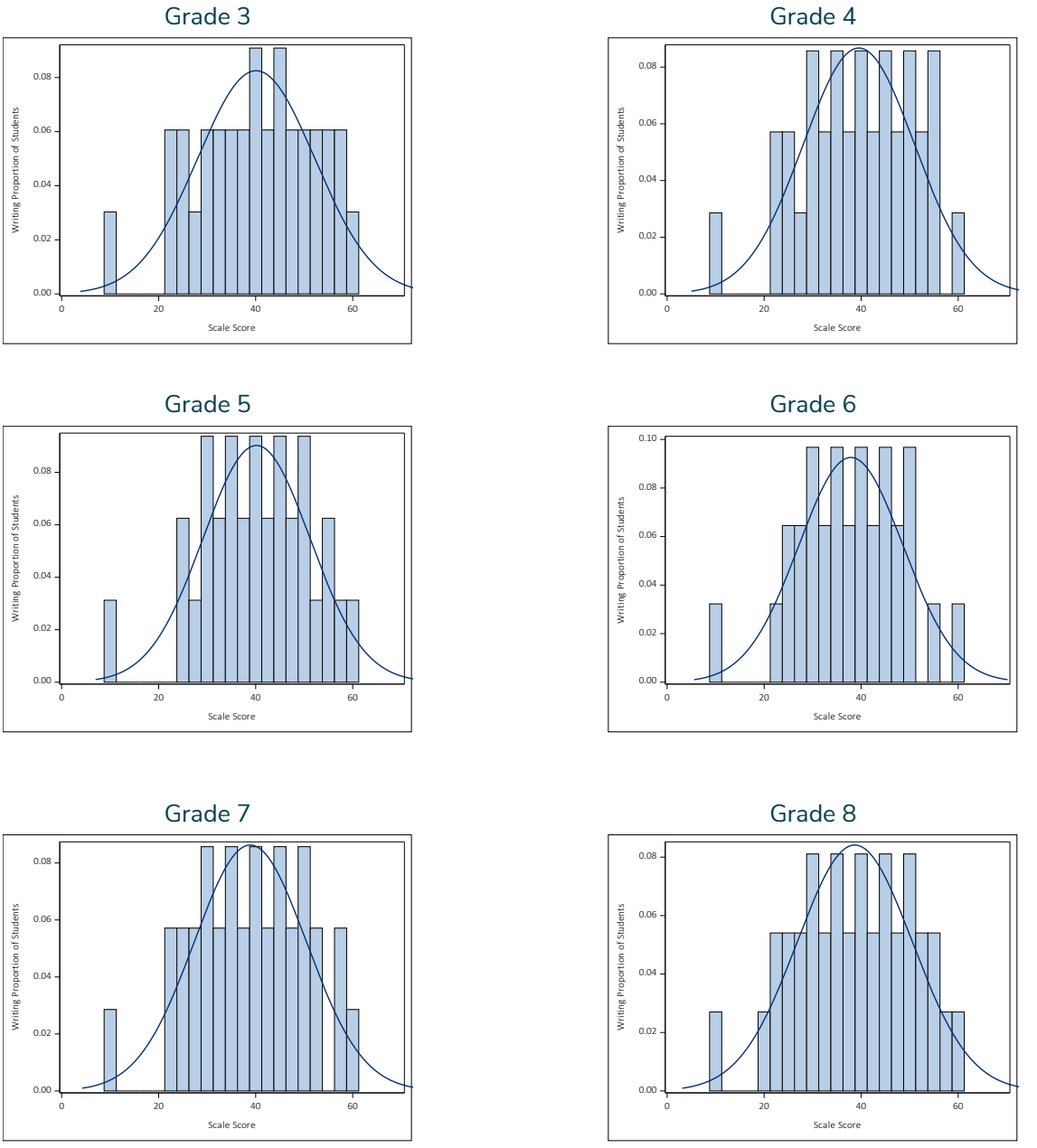



Figure 12.4 Distributions of Writing Scale Scores: Grades 3–8

## 12.4.2 Scale Score Cumulative Frequencies for ELA/L

The cumulative frequency distribution for the summative scale score is presented in Appendix 12 for ELA/L assessments.

## 12.4.3 Summary Scale Score Statistics for ELA/L Groups

Subgroup statistics for ELA/L full summative, Reading, and Writing scale scores are presented in Tables 12.5[7]. The results for all ELA/L grades are provided in Appendix 12.

[7] Due to omitted demographic values, subgroup sample sizes may not sum to the total sample size.

Table 12.5 Subgroup Performance for ELA/L Scale Scores: Grade 3

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **128,356** | **723.35** | **40.77** | **650** | **850** |
| Gender | Female | 62,746 | 727.61 | 41.48 | 650 | 850 |
| | Male | 65,604 | 719.27 | 39.66 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 335 | 708.15 | 38.16 | 650 | 827 |
| | Asian | 7,234 | 746.99 | 41.15 | 650 | 850 |
| | Black/African American | 21,218 | 704.73 | 36.21 | 650 | 850 |
| | Hispanic/Latino | 34,170 | 709.96 | 37.70 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 99 | 731.10 | 37.96 | 650 | 835 |
| | Two or more races | 5,585 | 728.73 | 42.01 | 650 | 850 |
| | White | 58,710 | 734.54 | 38.67 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 738.17 | 39.39 | 650 | 850 |
| | Economically Disadvantaged | 65,055 | 708.93 | 36.72 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,257 | 728.26 | 40.60 | 650 | 850 |
| | English Learner | 25,099 | 703.15 | 34.83 | 650 | 846 |
| Disabilities | Students without Disabilities | 103,803 | 727.92 | 40.11 | 650 | 850 |
| | Student with Disability (SWD) | 24,553 | 704.04 | 37.80 | 650 | 850 |
| **Reading Summative Score** | | **128,356** | **41.22** | **16.83** | **10** | **90** |
| Gender | Female | 62,746 | 42.63 | 17.03 | 10 | 90 |
| | Male | 65,604 | 39.87 | 16.53 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 335 | 35.10 | 15.42 | 10 | 76 |
| | Asian | 7,234 | 50.67 | 17.16 | 10 | 90 |
| | Black/African American | 21,218 | 33.85 | 14.86 | 10 | 90 |
| | Hispanic/Latino | 34,170 | 35.68 | 15.33 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 99 | 44.01 | 15.76 | 10 | 86 |
| | Two or more races | 5,585 | 43.66 | 17.53 | 10 | 90 |
| | White | 58,710 | 45.75 | 16.18 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 47.28 | 16.48 | 10 | 90 |
| | Economically Disadvantaged | 65,055 | 35.31 | 14.97 | 10 | 90 |
| English Learner Status | Non-English Learner | 103,257 | 43.30 | 16.85 | 10 | 90 |
| | English Learner | 25,099 | 32.67 | 13.81 | 10 | 90 |
| Disabilities | Students without Disabilities | 103,803 | 43.04 | 16.58 | 10 | 90 |
| | Student with Disability (SWD) | 24,553 | 33.52 | 15.69 | 10 | 90 |
| **Writing Summative Score** | | **128,356** | **23.71** | **13.04** | **10** | **60** |
| Gender | Female | 62,746 | 25.38 | 13.18 | 10 | 60 |
| | Male | 65,604 | 22.11 | 12.70 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 335 | 19.57 | 12.07 | 10 | 53 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Asian | 7,234 | 30.57 | 12.79 | 10 | 60 |
| | Black/African American | 21,218 | 18.42 | 11.43 | 10 | 60 |
| | Hispanic/Latino | 34,170 | 20.26 | 12.16 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 99 | 26.27 | 12.61 | 10 | 52 |
| | Two or more races | 5,585 | 24.88 | 13.30 | 10 | 60 |
| | White | 58,710 | 26.69 | 12.88 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 27.66 | 12.96 | 10 | 60 |
| | Economically Disadvantaged | 65,055 | 19.86 | 11.92 | 10 | 60 |
| English Learner Status | Non-English Learner | 103,257 | 24.88 | 13.12 | 10 | 60 |
| | English Learner | 25,099 | 18.91 | 11.54 | 10 | 60 |
| Disabilities | Students without Disabilities | 103,803 | 24.94 | 13.06 | 10 | 60 |
| | Student with Disability (SWD) | 24,553 | 18.50 | 11.58 | 10 | 60 |

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

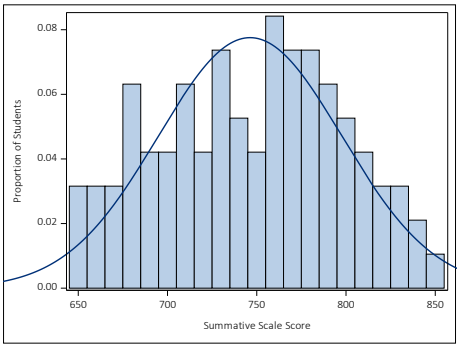
### 12.4.4 Score Distributions for Mathematics

Figure 12.5 graphically represents the distributions of scale scores for grades 3 through 8 mathematics. The *y*-axis for these distributions ranges from 0 to 0.02 and the *x*-axis from 650 to 850. Scale score distributions generally peaked between approximately 700 and the Level 4 performance level cut of 750.

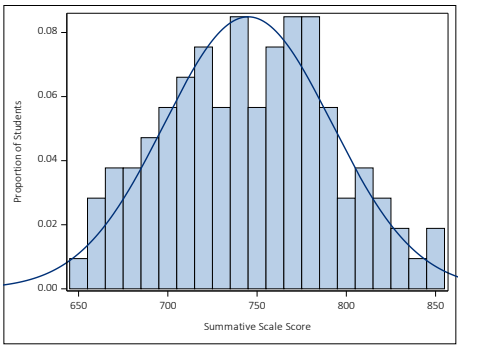
### 12.4.5 Scale Score Cumulative Frequencies for Mathematics

The cumulative frequency distribution for the summative scale score is presented in Appendix 12 for mathematics assessments.

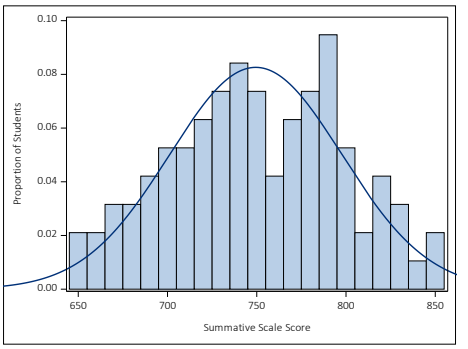
Grade 3



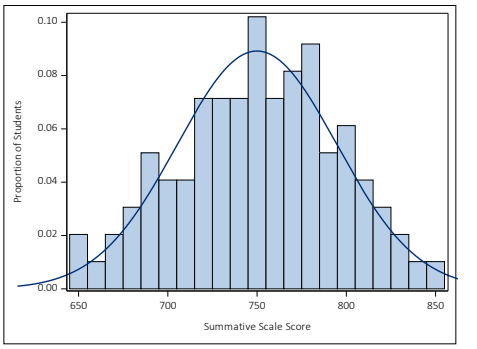

Grade 4



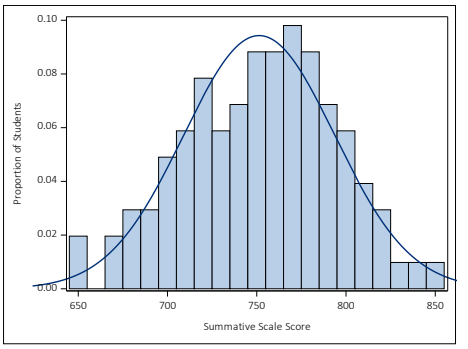

Grade 5




Grade 6




Grade 7




Grade 8

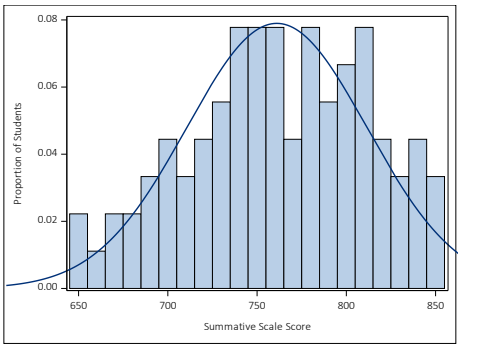



Figure 12.5 Distributions of Mathematics Scale Scores: Grades 3–8

### 12.4.6 Summary Scale Score Statistics for Mathematics Groups

Subgroup statistics for mathematics scale scores are presented in Table 12.6[8] for grade 3. Students taking the Spanish-language form were included in the overall subgroup analyses as well as analyzed separately. Students using the Spanish-language form tended to have lower mean scores. Corresponding tables for all grades/courses are presented in Appendix 12.

Table 12.6 Subgroup Performance for Mathematics Scale Scores: Grade 3

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **128,109** | **731.72** | **37.74** | **650** | **850** |
| Gender | Female | 62,609 | 730.17 | 36.29 | 650 | 850 |
| | Male | 65,494 | 733.21 | 39.03 | 650 | 850 |
| Ethnicity | American Indian or Alaska Native | 336 | 721.81 | 36.82 | 650 | 850 |
| | Asian | 7,235 | 758.43 | 37.67 | 650 | 850 |
| | Black/African American | 21,145 | 709.62 | 32.72 | 650 | 850 |
| | Hispanic/Latino | 34,085 | 719.26 | 33.00 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 99 | 737.60 | 38.22 | 650 | 840 |
| | Two or more races | 5,576 | 735.17 | 39.15 | 650 | 850 |
| | White | 58,627 | 743.49 | 35.29 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,219 | 746.60 | 35.99 | 650 | 850 |
| | Economically Disadvantaged | 64,890 | 717.23 | 33.54 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,065 | 735.34 | 38.10 | 650 | 850 |
| | English Learner | 25,044 | 716.86 | 32.22 | 650 | 850 |
| Disabilities | Students without Disabilities | 103,628 | 735.88 | 36.39 | 650 | 850 |
| | Student with Disability (SWD) | 24,481 | 714.14 | 38.32 | 650 | 850 |
| Language Form | Spanish | 3,541 | 707.45 | 29.67 | 650 | 830 |

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

## 12.5 Interpreting Claim Scores and Subclaim Scores

### 12.5.1 Interpreting Claim Scores

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; therefore, 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 Individual Student Report (ISR) provides the student scale score results and the average scale score results for the school, district, and state.

### 12.5.2 Interpreting Subclaim Scores

Within each reporting category are specific skill sets (subclaims) students demonstrate on the summative assessments. Subclaim categories are not reported using scale scores or performance levels. Subclaim performance for the assessments is reported using

[8] Due to omitted demographic values, subgroup sample sizes in these tables may not sum to total sample size.

graphical representations that indicate how the student performed relative to the Level 3 and Level 4 performance levels for the content area.

Subclaim indicators represent how well students performed in a subclaim category relative to Level 3 and Level 4 thresholds for the items associated with the subclaim category. To determine a student's subclaim performance, the Level 3 and Level 4 thresholds corresponding to the IRT based performance for the items for a given subclaim determined the reference points for *Approached Expectations* and *Did Not Yet Meet Expectations* or *Partially Met Expectations*, respectively.

Student performance for each subclaim is marked with a subclaim performance indicator.

An 'up' arrow for the specified subclaim indicates that the student *Met or Exceeded Expectations*, meaning that the student's subclaim performance reflects a level of proficiency consistent with Performance Level 4 or 5. Students in this subclaim category are likely academically well prepared to engage successfully in further studies in the subclaim content area and may need instructional enrichment.

A 'bidirectional' arrow for the specified subclaim indicates that the student *Approached Expectations*, meaning that the student's subclaim performance reflects a level of proficiency consistent with Performance Level 3. Students in this subclaim category likely need academic support to engage successfully in further studies in the subclaim content area.

A 'down' arrow for the specified subclaim indicates that the student *Did Not Yet Meet or Partially Met Expectations* meaning that the student's subclaim performance reflects a level of proficiency consistent with Performance Level 1 or 2. Students in this subclaim category are likely not academically well prepared to engage successfully in further studies in the subclaim content area. Such students likely need instructional interventions to increase achievement in the subclaim content area.

# Section 13: Reliability

## 13.1 Overview

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, especially which form of the test the student is administered and which persons are assigned to score responses to constructed-response questions. In statistical terms, the variance in the distributions of test scores, essentially 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.

There are several different ways of estimating reliability. The type of raw score reliability estimate reported here is an internal-consistency measure, which is derived from analysis of the consistency of the performance of individuals across items within a test. It is used because it serves as a good estimate of alternate forms reliability, but it does not take into account 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 account for the conversions between raw scores and scale scores.

Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain very 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 in the tables to follow attempt to answer the question, "How consistent would the scores of these students be over replications of the entire testing process?"

Reliability of classification estimates the proportion of students who are accurately classified into proficiency levels. There are two kinds of classification reliability statistics: decision accuracy and decision consistency. Decision accuracy is the agreement between the classifications actually made and the classifications that would be made if the test scores were perfectly reliable. Decision consistency is the agreement between the classifications that would be made on two independent forms of the test.

Another index is inter-rater reliability for the human-scored constructed-response items, which measures the agreement between individual raters (scorers). The inter-rater reliability coefficient answers the question, "How consistent is the scoring such that a set of similarly trained raters would produce similar scores to those obtained?"

Standard error of measurement (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 and SEM estimates were calculated at the full assessment level, and at the claim and subclaim levels. In addition, conditional SEMs were calculated and reported in Appendix 13.

## 13.2 Reliability and SEM Estimation

### 13.2.1 Raw Score Reliability Estimation

Coefficient alpha (Cronbach, 1951), which measures internal consistency reliability, is the most commonly used measure of reliability. Coefficient alpha is estimated by substituting sample estimates for the parameters in the formula below:

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

where $_n$ is the number of items, $\sigma_i^2$ is the variance of scores on the $i$th item, and $\sigma_X^2$ is the variance of the total score (sum of scores on the individual items). Other things being equal, the more items a test includes, the higher the internal consistency reliability.

Since the test forms have mixed item types (dichotomous and polytomous items), it is more appropriate to report stratified alpha (Feldt & Brennan, 1989). Stratified alpha 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:

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

Where $\sigma_{X_h}^2$ is the variance for part $h$ of the test, $\sigma_X^2$ is the variance of the total scores, and $\alpha_h$ 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 standard error of measurement is:

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

Where $\sigma_X$ is the standard deviation of the test raw score and $\rho_{xx'}$ is the reliability estimated by substitution of appropriate statistics for the parameters in equation 13-1 or 13-2.

In this section, reliability estimates are reported for overall summative scores, claim scores, and subclaim scores. Estimates are also reported for subgroups for summative scores. Cronbach's alpha and stratified alpha coefficients 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). A decrease in the number of items may result in a decrease in stratified alpha estimates. The decrease in sample size and the homogeneity of the samples is likely to result in lower stratified alpha estimates. A smaller more homogenous sample will likely result in lower stratified alpha estimates. 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 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).

### 13.2.2 Scale Score Reliability Estimation

Like the stratified alpha coefficients, scale score reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals 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. 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,

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

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

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

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,

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

# 13.3 Reliability Results for Total Group

## 13.3.1 Raw Score Reliability Results

Tables 13.1 and 13.2 summarize test reliability estimates for the total testing group for English language arts/literacy (ELA/L) and mathematics, respectively. The tables provide the average reliability, which is estimated by averaging the internal consistency estimates computed for all the individual forms of the test and the raw score SEMs. In addition, the number of forms, the sample size of the minimum reliability, sample size of the maximum reliability, and the average maximum possible score for each set of tests are provided. Estimates were calculated only for groups of 100 or more students administered a specific test form.

### English Language Arts/Literacy

The average reliability estimates for grades 3 through 8 ELA/L range from a low of 0.83 to a high of 0.86. The average raw score SEM is consistently between about 4 percent and 5 percent of the maximum possible score.

Table 13.1 Summary of ELA/L Test Reliability Estimates for Total Group

| Grade Level | Number of Forms | Avg. Max Possible Score | Avg. Raw Score SEM | Average Reliability | Minimum Reliability | | Maximum Reliability | |
|---|---|---|---|---|---|---|---|---|
| | | | | | N | Alpha | N | Alpha |
| 3 | 5 | 54 | 3.18 | 0.86 | 1,287 | 0.76 | 62,159 | 0.90 |
| 4 | 5 | 70 | 3.86 | 0.83 | 1,489 | 0.74 | 61,639 | 0.90 |
| 5 | 5 | 70 | 3.89 | 0.86 | 1,493 | 0.75 | 62,583 | 0.89 |
| 6 | 5 | 72 | 4.29 | 0.86 | 1,594 | 0.80 | 64,326 | 0.90 |
| 7 | 5 | 72 | 4.52 | 0.86 | 1,511 | 0.77 | 65,225 | 0.89 |
| 8 | 5 | 72 | 4.29 | 0.85 | 1,539 | 0.76 | 67,487 | 0.89 |

### Mathematics

The average reliability estimates for mathematics assessments range from 0.88 to 0.90. The raw score SEM consistently ranges from about 3 to 4 percent of the maximum score.

Table 13.2 Summary of Mathematics Test Reliability Estimates for Total Group

| Grade Level | Number of Forms | Avg. Max Possible Score | Avg. Raw Score SEM | Average Reliability | Minimum Reliability | | Maximum Reliability | |
|---|---|---|---|---|---|---|---|---|
| | | | | | N | Alpha | N | Alpha |
| 3 | 5 | 52 | 3.33 | 0.90 | 41,861 | 0.89 | 21,154 | 0.91 |
| 4 | 5 | 52 | 3.36 | 0.89 | 994 | 0.89 | 20,775 | 0.91 |
| 5 | 5 | 52 | 3.40 | 0.88 | 953 | 0.85 | 21,191 | 0.90 |
| 6 | 5 | 52 | 3.38 | 0.88 | 814 | 0.84 | 20,821 | 0.89 |
| 7 | 5 | 52 | 3.34 | 0.89 | 45,382 | 0.88 | 20,451 | 0.91 |
| 8 | 5 | 52 | 3.09 | 0.88 | 712 | 0.81 | 48,880 | 0.89 |

## 13.3.2 Scale Score Reliability Results

Tables 13.3–13.4 summarize scale score reliability estimates for the total testing group for ELA/L and mathematics for spring 2023. The tables provide average reliabilities by grade, which are estimated by averaging the reliability estimates computed for all forms of the test within the grade level. In addition, the number of forms, the total sample size across all forms, and the average maximum possible score for each set of tests are provided. Scale score reliability requires an ability distribution, which is not reasonable to assume for Integrated Mathematics due to the low sample sizes.

### English Language Arts/Literacy

Reliability estimates for ELA/L are presented in Table 13.3. Average reliabilities range from 0.83 to 0.86. The average SEM ranges from 12.64 to 13.45.

Table 13.3 Summary of ELA/L Test Pre-Equated Scale Score Reliability Estimates for Total Group

| Grade Level | Number of Forms | Avg. Scale Score SEM | Avg. Scale Score Reliability | Min. Scale Score Reliability | Max. Scale Score Reliability |
|---|---|---|---|---|---|
| 3 | 5 | 12.97 | 0.88 | 0.76 | 0.93 |
| 4 | 5 | 13.45 | 0.86 | 0.67 | 0.92 |
| 5 | 5 | 13.30 | 0.85 | 0.69 | 0.91 |
| 6 | 5 | 12.64 | 0.85 | 0.74 | 0.91 |
| 7 | 5 | 13.19 | 0.89 | 0.79 | 0.93 |
| 8 | 5 | 13.07 | 0.89 | 0.79 | 0.94 |

### Mathematics

The scale score reliability estimates for the mathematics assessments are presented in Table 13.4. Average scale score reliability estimates for the grades 3 through 8 mathematics assessments range from 0.88 to 0.95. For grades 3–8, the average scale score SEM ranges from 10.40 to 13.87.

Table 13.4 Summary of Mathematics Test Scale Score Reliability Estimates for Total Group

| Grade Level | Number of Forms | Avg. Scale Score SEM | Avg. Raw Score Reliability | Min. Scale Score Reliability | Max. Scale Score Reliability |
|---|---|---|---|---|---|
| 3 | 5 | 10.50 | 0.95 | 0.94 | 0.95 |
| 4 | 5 | 10.40 | 0.93 | 0.90 | 0.93 |
| 5 | 5 | 11.08 | 0.91 | 0.87 | 0.93 |
| 6 | 5 | 12.33 | 0.88 | 0.83 | 0.90 |
| 7 | 5 | 11.80 | 0.86 | 0.81 | 0.88 |
| 8 | 5 | 13.87 | 0.90 | 0.84 | 0.92 |

## 13.4 Reliability Results for Subgroups of Interest

When the sample size was sufficiently large, raw score reliability and SEM were estimated for the groups identified for DIF analysis. Estimates were calculated only for groups of 100 or more students administered a specific test form.

Tables 13.5 and 13.6 summarize test reliability for groups of interest for ELA/L grade 3 and mathematics grade 3, respectively. Corresponding information is provided in Appendix 13 for all ELA/L and mathematics grades. For each group, the average, minimum, and maximum reliability estimates are listed, as well as the sample sizes of the reported minimum and maximum reliabilities. Note that reliability estimates are dependent on score variance, and subgroups with smaller variance are likely to have lower reliability estimates than the total group.

### 13.4.1 Reliability Results for Gender

#### English Language Arts/Literacy

The average reliability estimates and the average SEMs for males and females reflect the corresponding reliabilities for the total group. For most tests, the reliabilities between males and females are equal or within 0.01. The SEMs for females were slightly higher than for males for ELA/L assessments.

#### Mathematics

As with the ELA/L test components, the average reliability estimates and SEMs for males and females reflect the corresponding reliabilities for the total group. For most tests, the reliabilities between males and females are equal or within 0.03. The SEMs for females are slightly higher than for males for the majority of tests.

### 13.4.2 Reliability Results for Ethnicity

#### English Language Arts/Literacy

The majority of the average reliabilities for the ethnicity groups are equal to 0.04 lower than for the total group. There is not a consistent difference among the average reliabilities for white, black/African American, Asian/Pacific Islander, Hispanic/Latino, and multiple-ethnicity students, with the majority of the reliabilities between 0.80 and 0.90. Average SEMs were generally slightly higher for white and Asian/Pacific Islander students than for black/African American and Hispanic/Latino students.

#### Mathematics

As with the ELA/L reliabilities, the reliabilities for ethnicity groups are marginally lower than for the total group of students. While there is variation across tests, the average reliabilities are often highest for multiple-ethnicity students. The average SEMs reflect the total group SEMs. Average SEMs were generally higher for white, Asian/Pacific Islander, and multiple-ethnicity students than for Hispanic, black/African American, and American Indian/Alaska Native students.

## 13.4.3 Reliability Results for Special Education Needs

### English Language Arts/Literacy

The average reliabilities for five groups of students (economically disadvantaged, not economically disadvantaged, non-English learner, students with disabilities, and students without disabilities) are generally equal to or 0.01 to 0.03 less than the average reliability for the total group of students. Average reliabilities for English learner students are lower, ranging from 0.79 to 0.86. The SEMs are generally higher for the larger student groups (not economically disadvantaged students, non-English learner students, and students without disabilities).

### Mathematics

The average reliabilities for the larger student groups (not economically disadvantaged, non-English learner, and students without disabilities) are generally equal to or 0.01 to 0.04 less than the average reliability for the total group of students. For economically disadvantaged, English learner, and students with disabilities, the average reliabilities are lower than those for the total group, notably in high school. The SEMs are generally higher for the larger student groups (not economically disadvantaged students, non-English learner students, and students without disabilities).

## 13.4.4 Reliability Results for Students Taking Accommodated Forms

### English Language Arts/Literacy

Reliability information for accommodated forms is sparse due to small sample sizes. Reliability for forms with less than 100 students is not reported. Reliability tended to be slightly lower for accommodated forms versus non-accommodated forms.

### Mathematics

Reliability information for accommodated forms is sparse due to small sample sizes. Reliability for forms with less than 100 students is not reported. The text-to-speech and human reader forms had sufficient sample sizes for reliability and SEM estimation across grades/subjects. For almost all tests, text-to-speech reliabilities are similar to the total group reliabilities, with SEMs slightly lower than the total group SEMs.

## 13.4.5 Reliability Results of Students Taking Translated Forms

### Mathematics

There were sufficient numbers of students taking the Spanish-language form for reliability and SEM estimation for grades 3 through 8. The average reliability ranged from 0.67 to 0.87. The SEMs are generally lower for the students administered the Spanish-language forms.

Table 13.5 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 3

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 54 | 3.18 | 0.86 | 1,287 | 0.76 | 62,159 | 0.90 |
| **Gender** | | | | | | | |
| Male | 54 | 3.10 | 0.85 | 824 | 0.76 | 31,493 | 0.90 |
| Female | 54 | 3.26 | 0.86 | 463 | 0.75 | 30,664 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 54 | 2.83 | 0.83 | 292 | 0.72 | 10,174 | 0.89 |
| Asian/Pacific Islander | 54 | 3.84 | 0.89 | 3,628 | 0.89 | 3,628 | 0.89 |
| Hispanic/Latino | 54 | 2.96 | 0.82 | 385 | 0.62 | 16,537 | 0.89 |
| American Indian/Alaska Native | 54 | 3.20 | 0.89 | 153 | 0.89 | 171 | 0.90 |
| Two or more races | 54 | 3.60 | 0.90 | 2,637 | 0.90 | 2,637 | 0.90 |
| White | 54 | 3.35 | 0.85 | 541 | 0.78 | 28,691 | 0.89 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 54 | 2.95 | 0.83 | 868 | 0.72 | 31,252 | 0.89 |
| Not Economically Disadvantaged | 54 | 3.40 | 0.86 | 419 | 0.79 | 30,907 | 0.89 |
| English Learner | 54 | 2.85 | 0.79 | 283 | 0.54 | 12,118 | 0.88 |
| Non-English Learner | 54 | 3.25 | 0.86 | 1,004 | 0.78 | 50,041 | 0.90 |
| Students with Disabilities (SWD) | 54 | 2.92 | 0.86 | 1,287 | 0.76 | 10,399 | 0.90 |
| Students without Disabilities | 54 | 3.57 | 0.89 | 51,956 | 0.89 | 51,760 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 54 | 3.04 | 0.81 | 236 | 0.81 | 236 | 0.81 |
| Non-Screen Reader | 54 | 3.20 | 0.82 | 187 | 0.82 | 187 | 0.82 |
| Screen Reader | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 54 | 2.60 | 0.75 | 2,647 | 0.75 | 2,647 | 0.75 |

Note: n/r = not reported.

Table 13.6 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 3

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.33 | 0.90 | 41,861 | 0.89 | 24,430 | 0.91 |
| **Gender** | | | | | | | |
| Male | 52 | 3.33 | 0.90 | 21,017 | 0.89 | 12,736 | 0.91 |
| Female | 52 | 3.34 | 0.89 | 20,840 | 0.88 | 11,692 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.94 | 0.89 | 171 | 0.88 | 5,514 | 0.89 |
| Asian/Pacific Islander | 52 | 3.61 | 0.89 | 2,709 | 0.86 | 989 | 0.92 |
| Hispanic/Latino | 52 | 3.15 | 0.88 | 10,772 | 0.88 | 243 | 0.89 |
| American Indian/Alaska Native | 52 | 3.06 | 0.89 | 116 | 0.89 | 116 | 0.89 |
| Two or more races | 52 | 3.41 | 0.91 | 2,226 | 0.90 | 617 | 0.93 |
| White | 52 | 3.46 | 0.89 | 23,490 | 0.87 | 6,293 | 0.92 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 3.10 | 0.89 | 17,515 | 0.88 | 574 | 0.89 |
| Not Economically Disadvantaged | 52 | 3.51 | 0.89 | 24,346 | 0.87 | 7,784 | 0.91 |
| English Learner | 52 | 3.08 | 0.88 | 219 | 0.88 | 5,846 | 0.89 |
| Non-English Learner | 52 | 3.38 | 0.90 | 36,850 | 0.89 | 15,273 | 0.92 |
| Students with Disabilities (SWD) | 52 | 3.13 | 0.90 | 6,282 | 0.89 | 5,532 | 0.91 |
| Students without Disabilities | 52 | 3.38 | 0.89 | 36,006 | 0.88 | 18,148 | 0.91 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 3.09 | 0.89 | 460 | 0.89 | 460 | 0.89 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.17 | 0.90 | 42,064 | 0.90 | 42,064 | 0.90 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.77 | 0.87 | 1,733 | 0.86 | 1,787 | 0.87 |

Note: n/r = not reported.

## 13.5 Reliability Results for English Language Arts/Literacy Claims and Subclaims

Participating states and agencies developed subclaims in addition to major claims based on the Common Core State Standards. ELA/L has two major claims relating to Reading and Writing. The major claim for Reading is that students read and comprehend a range of sufficiently complex texts independently. The major claim for Writing is that students write effectively when using and/or analyzing sources. Refer to Table 13.7 for a summary of the ELA/L claims and subclaims.

Table 13.7 Descriptions of ELA/L Claims and Subclaims

| English Language Arts/Literacy | | |
|---|---|---|
| **Major Claim** | **Subclaim** | **Description** |
| Reading | Reading Literature | Students demonstrate comprehension and draw evidence from readings of grade-level, complex literary text. |
| Reading | Reading Information | Students demonstrate comprehension and draw evidence from readings of grade-level, complex informational text. |
| Reading | Reading Vocabulary | Students use context to determine the meaning of words and phrases. |
| Writing | Writing Written Expression | Students produce clear and coherent writing in which the development, organization, and style are appropriate to the task, purpose, and audience. |
| Writing | Writing Knowledge Language and Conventions | Students demonstrate knowledge of conventions and other important elements of language. |

Reliability indices were calculated for each major claim and subclaim. Table 13.8 presents the average reliability estimates for all forms of the test at the specified grade for the ELA/L tests. To assist in understanding the reliability estimates, range of maximum number of points for each major claim and subclaim is also provided.

Average reliabilities for the Reading claim for grades 3 through 8 range from 0.79 to 0.83. They are based on maximum scores of 40–44 points per form, except for grade 3 which is 30–31 points. The Writing claim average reliabilities are based on a lower number of points than those for the Reading claim, and are also slightly lower than Reading, ranging from 0.74 to 0.81. The reliabilities for the Writing claim for grade 3 are based on a maximum raw score of 24 points, the average reliabilities for grades 4 and 5 are based on between 27 and 30 points per form, and the average reliabilities for the grades 6 through 8 Writing claims are based on a maximum score of 30 points.

The average reliabilities of the Reading Literature subclaim scores vary from 0.64 to 0.72. The maximum number of points per form ranges from 11 to 18. The average reliabilities of the Reading Information subclaim scores vary from 0.57 to 0.65, with 11–16 points per form. The average reliabilities of the Reading Vocabulary subclaim scores vary from 0.50 to 0.55. The maximum number of points per form for this subclaim ranges from 6 to 14.

The Writing Written Expression subclaim is based on 18 points for grade 3, 21–24 points for grades 4 and 5, and 24 points for grades 6 through 8. The average reliabilities range from 0.66 to 0.79. The Writing Knowledge of Language and Conventions subclaims are all based on six points. The reliabilities range from 0.72 to 0.80.

Table 13.8 Average ELA/L Reliability Estimates for Subscores

| | Reading: Total | | Reading: Literature | | Reading: Information | | Reading: Vocabulary | | Writing: Total | | Writing: Written Expression | | Writing: Knowledge Language and Conventions | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade Level | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability |
| 3 | 30-31 | 0.83 | 11-14 | 0.72 | 11-11 | 0.58 | 06-08 | 0.53 | 24-24 | 0.75 | 18-18 | 0.66 | 06-06 | 0.76 |
| 4 | 40-44 | 0.79 | 16-18 | 0.64 | 12-14 | 0.56 | 10-12 | 0.51 | 27-30 | 0.74 | 21-24 | 0.67 | 06-06 | 0.72 |
| 5 | 40-44 | 0.82 | 16-18 | 0.71 | 14-16 | 0.58 | 08-12 | 0.53 | 27-30 | 0.78 | 21-24 | 0.73 | 06-06 | 0.77 |
| 6 | 40-44 | 0.82 | 16-18 | 0.69 | 14-16 | 0.57 | 08-14 | 0.55 | 30-30 | 0.80 | 24-24 | 0.76 | 06-06 | 0.79 |
| 7 | 40-44 | 0.82 | 16-18 | 0.68 | 14-16 | 0.65 | 08-10 | 0.50 | 30-30 | 0.80 | 24-24 | 0.77 | 06-06 | 0.80 |
| 8 | 40-44 | 0.81 | 16-18 | 0.65 | 14-16 | 0.61 | 08-10 | 0.51 | 30-30 | 0.81 | 24-24 | 0.79 | 06-06 | 0.80 |

## 13.6 Reliability Results for Mathematics Subclaims

For mathematics, there are four subclaims related to whether students are on track or ready for college and careers:

- Subclaim A: Students solve problems involving the major content for their grade level with connections to the Standards for Mathematical Practice.
- Subclaim B: Students solve problems involving the additional and supporting content for their grade level with connections to the Standards for Mathematical Practice.
- Subclaim C: Students express grade-level appropriate mathematical reasoning by constructing viable mathematical arguments and critiquing the reasoning of others, and/or attending to precision when making mathematical statements.
- Subclaim D: Students solve real-world problems with a degree of difficulty appropriate to the grade by applying knowledge and skills articulated in the standards and by engaging particularly in the modeling practice.

Reliability estimates were calculated for each subclaim for mathematics. Table 13.9 presents the average reliability estimates for mathematics subclaims.

Subclaims with greater numbers of points tend to have greater reliability estimates. The Major Content subclaim has the largest number of points for each assessment and, accordingly, has higher average reliabilities than the other three subclaims. For grades 3 through 8, the median of the average reliabilities for the Major Content ranges from 0.77 to 0.83. The maximum number of points per form for this subclaim ranges from 18 to 22 depending on grade.

The median of the average reliabilities for the Additional and Supporting Content subclaim for grades 3 through 8 ranges from 0.49 to 0.71. The maximum number of points per form for this subclaim ranges from 8 to 10 depending on grade.

The average reliabilities for Mathematics Reasoning ranges from 0.62 to 0.68 for grades 3 through 8. The maximum number of points for this subclaim is 10 for all grades and forms.

For the Modeling Practice subclaim, the average reliabilities for grades 3 through 8 ranges from 0.57 to 0.67. The maximum number of points for this subclaim is 12 for all grades and forms.

Table 13.9 Average Mathematics Reliability Estimates for Subscores

| | Major Content | | Additional & Supporting Content | | Mathematics Reasoning | | Modeling Practice | |
|---|---|---|---|---|---|---|---|---|
| Grade Level | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability | Range of Raw Score | Average Reliability |
| 3 | 20-20 | 0.83 | 10-10 | 0.71 | 10-10 | 0.62 | 12-12 | 0.65 |
| 4 | 21-21 | 0.83 | 09-09 | 0.62 | 10-10 | 0.68 | 12-12 | 0.65 |
| 5 | 20-20 | 0.79 | 10-10 | 0.66 | 10-10 | 0.64 | 12-12 | 0.57 |
| 6 | 20-20 | 0.77 | 10-10 | 0.54 | 10-10 | 0.66 | 12-12 | 0.63 |
| 7 | 20-22 | 0.82 | 08-10 | 0.61 | 10-10 | 0.68 | 12-12 | 0.67 |
| 8 | 18-20 | 0.77 | 10-10 | 0.49 | 10-10 | 0.62 | 12-12 | 0.60 |

## 13.7 Reliability of Classification

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 on the basis of 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 non-overlapping, equally difficult forms of the test.

Decision consistency values are always lower than the corresponding decision accuracy values, because in decision consistency, both of the classifications are subject to measurement error. In decision accuracy, only one of the classifications is based on a score that contains error. 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.

### 13.7.1 English Language Arts/Literacy

Table 13.10 provides information about the accuracy and the consistency of two types of classifications made on the basis of the summative scale scores on the grades 3 through 8 ELA/L assessments. The columns labeled “Exact level” provide the estimates of the indices based on classifications of students into one of five performance levels. The columns labeled “Level 4 or higher vs. 3 or lower” provide 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). Performance Level 4 is considered the College and Career Readiness standard on the summative assessments.

The table shows that for classifying each student into one of the five performance levels, the portion accurately classified ranges from 0.65 to 0.71; the proportion who would be consistently classified on two different test forms ranges from 0.55 to 0.62. For classifying each student as being at Level 4 or higher vs. being at Level 3 or lower, the proportion accurately classified ranges from 0.88 to 0.91; the proportion who would be consistently classified this way on two different test forms ranges from 0.84 to 0.87.

Table 13.10 Reliability of Classification: Summary for ELA/L

| | Decision Accuracy | | Decision Consistency | |
|---|---|---|---|---|
| Level | Exact Level | Level 4 or higher vs. 3 or lower | Exact Level | Level 4 or higher vs. 3 or lower |
| 3 | 0.71 | 0.91 | 0.62 | 0.87 |
| 4 | 0.65 | 0.89 | 0.55 | 0.85 |
| 5 | 0.68 | 0.89 | 0.57 | 0.84 |
| 6 | 0.67 | 0.88 | 0.56 | 0.84 |
| 7 | 0.68 | 0.90 | 0.57 | 0.86 |
| 8 | 0.70 | 0.90 | 0.59 | 0.86 |

Table 13.11 provides more detailed information about the accuracy and the consistency of the classification of students into performance levels for ELA/L grade 3. Each cell in the 5-by-5 table 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 13.10. For “Level 4 and higher vs. 3 and lower” found in Table 13.10, the sum of the shaded values in Table 13.11 is approximately equal to the level of decision accuracy or consistency presented in Table 13.10. Note that the sums based on values in Table 13.11 may not exactly match the values in Table 13.10 due to truncation and rounding error.

Detailed information for all ELA/L spring 2023 results are provided in Appendix 13 Tables A.13.18 through A.13.23. The structure of these tables is the same as that of Table 13.11 and the values in the tables should be interpreted in the same manner.

Table 13.11 Reliability of Classification: Grade 3 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.27 | 0.04 | 0.00 | 0.00 | 0.00 | 0.30 |
| | 700-724 | 0.05 | 0.11 | 0.05 | 0.00 | 0.00 | 0.21 |
| | 725-749 | 0.00 | 0.04 | 0.11 | 0.05 | 0.00 | 0.21 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.22 | 0.02 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Decision Consistency | 650-699 | 0.26 | 0.05 | 0.01 | 0.00 | 0.00 | 0.32 |
| | 700-724 | 0.05 | 0.08 | 0.05 | 0.01 | 0.00 | 0.20 |
| | 725-749 | 0.01 | 0.05 | 0.08 | 0.05 | 0.00 | 0.19 |
| | 750-809 | 0.00 | 0.01 | 0.06 | 0.20 | 0.01 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |

### 13.7.2 Mathematics

Table 13.12 provides information about the accuracy and the consistency of two types of classifications made on the basis of the summative scale scores on the mathematics assessments. For the grades 3 through 8 mathematics tests, the table shows that for classifying each student into one of the five performance levels, the proportion accurately classified ranges from 0.69 to 0.79; the proportion who would be consistently classified on two different test forms ranges from 0.57 to 0.70.

For classifying each student as being at Level 4 or higher vs. being at Level 3 or lower, for the grades 3 through 8 mathematics tests, the proportion accurately classified ranges from 0.90 to 0.94; the proportion who would be consistently classified on two different test forms is 0.86 to 0.91 for grades 3 and 8.

Appendix 13 Tables A.13.24 through A.13.29 provide more detailed information about the accuracy and the consistency of the classification of students into performance levels for mathematics. Each cell in the 5-by-5 table shows the estimated proportion of students who would be classified into a particular combination of performance levels.

Table 13.12 Reliability of Classification: Summary for Mathematics

| | Decision Accuracy | | Decision Consistency | |
|---|---|---|---|---|
| Level | Exact Level | Level 4 or higher vs. 3 or lower | Exact Level | Level 4 or higher vs. 3 or lower |
| 3 | 0.79 | 0.94 | 0.70 | 0.91 |
| 4 | 0.77 | 0.93 | 0.68 | 0.90 |
| 5 | 0.74 | 0.92 | 0.64 | 0.89 |
| 6 | 0.70 | 0.91 | 0.60 | 0.88 |
| 7 | 0.69 | 0.90 | 0.57 | 0.86 |
| 8 | 0.71 | 0.92 | 0.61 | 0.89 |

## 13.8 Inter-rater Agreement

Inter-rater agreement is the agreement between the first and second scores assigned to student responses. Inter-rater agreement measurements include exact, adjacent, and nonadjacent agreement. Pearson scoring staff used these statistics as one factor in determining the needs for continuing training and intervention on both individual and group levels. Table 13.13 displays both the expectations and the actual agreement percentages for perfect agreement and perfect plus adjacent agreement.

Table 13.13 Inter-rater Agreement Expectations and Results

| Subject | Score Point Range | Perfect Agreement Expectation | Perfect Agreement Result | Within One Point Expectation | Within One Point Result |
|---|---|---|---|---|---|
| Mathematics | 0–1 | 90% | 98% | 96% | 100% |
| Mathematics | 0–2 | 80% | 97% | 96% | 100% |
| Mathematics | 0–3 | 70% | 96% | 96% | 99% |
| Mathematics | 0–4 | 65% | 94% | 95% | 99% |
| Mathematics | 0–5 | 65% | 91% | 95% | 98% |
| Mathematics | 0–6 | 65% | 95% | 95% | 98% |
| ELA/L | Multi-trait | 65% | 83% | 96% | 100% |

Note: A 0 or 1 score compared to a blank score will have a disagreement greater than 1 point.

Pearson's ePEN2 scoring 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. Table 13.13 shows that the actual percentages for perfect reader agreement were higher than the inter-rater agreement expectations, and the percentages for within one point were very close. Refer to Section 4 for more information on hand-scoring.

# Section 14: Validity

## 14.1 Overview

The Standards for Educational and Psychological Testing, issued jointly by the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014), reports:

> 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, item development, and psychometric characteristics. The 2022–2023 operational assessments provided an opportunity to gather evidence of validity based on both test content and on the internal structure of the tests.

Pearson applies the principles of universal design, as articulated in materials developed by the National Center for Educational Outcomes (NCEO) at the University of Minnesota (Thompson et al., 2002).

## 14.2 Evidence Based on Test Content

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 Common Core State 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.

Pearson and New Meridian built spreadsheets at the evidence statement level that incorporate the probability statements from the test blueprints and attrition rates at committee review and data review. The basis of our entire item development is driven by the use of these item development target spreadsheets. Before beginning item development, Pearson uses these target spreadsheets to develop an internal item development plan to correlate with the expectations of the test design. These are reviewed and approved by state or agency leads and New Meridian. All parties acknowledge that each assessment has multiple parts and each part specifies the types of tasks and standards eligible for assessment.

In addition to the evidence statements, content is aligned through the articulation of performance in the performance level descriptors. At the policy level, the performance level descriptors 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-level descriptors are the foundation for the subject- and grade-specific performance level descriptors, which, along with the evidence frameworks, guide the development of the items and tasks.

The college- and career-ready determinations (CCRD) in English language arts/literacy (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 percent likelihood of earning a grade of “C” or better in credit-bearing courses without the need for remedial coursework. After reviewing the standards and assessment design, the Governing Board (made up of the K–12 education chiefs in participating states or 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 (NAEP), Trends in International Mathematics and Science Study (TIMSS), Programme of International Student Assessment (PISA), and Progress in International Reading Literacy Study (PIRLS) tests were conducted (McClarty et al., 2015).

Gathering construct validity evidence for the assessments is embedded in the process by which the assessment content is developed and validated. At each step in the assessment development process, participating states or agencies involved hundreds of educators, assessment experts, and bias and sensitivity experts in review of text, items, and tasks for accuracy, appropriateness, and freedom from bias. See Section 2 for an overview of the content development process. In the early stages of development, Pearson conducted research studies to validate the item and task development approach. One such study was a student task interaction study designed to collect data on the student's experience with the assessment tasks and technological functionalities, as well as the amount of time needed for answering each task. Pearson also conducted a rubric choice study that compared the functioning of two rubrics developed to score the Prose Constructed-Response (PCR) tasks in ELA/L. Quantitative and qualitative evidence was collected to support the use of a condensed or expanded trait scoring rubric in scoring student responses.

The items and tasks were field tested prior to their use on an assessment. During the initial field test administration in 2014, 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 future development of additional items and tasks.

All item developers and item writers are provided with an electronic version of the accessibility guidelines and the linguistic complexity rubric. Items and passages are reviewed internally by accessibility and fairness experts trained in the principles of universal design and who become well versed in the accessibility guidelines. Items received internal review for alignment to evidence tables, task generation model, item selection guidelines, and accessibility and fairness reviews.

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. New Meridian convened bias and sensitivity committees to review all items. Additionally, content experts facilitated reviews of all items. All reviewers were trained using the bias and sensitivity guidelines, and the guidelines were used to review items and ELA/L passages. Accommodations were made available based on individual need documented in the student's approved IEP, 504 Plan, or if required by the participating state or agency, an English Learner (EL) Plan. An accessibility specialist worked in consultation with the accessibility specialist to review forms and determine which forms should be used for students with accommodations.

The ELA/L and mathematics operational test forms, as described in Section 2, were carefully constructed to align with the test blueprints and specifications that are based on the Common Core State Standards (CCSS). During the fall of 2016, content experts representing various participating states and agencies, along with other content experts, held a series of meetings to review the operational forms for ELA/L and mathematics. These meetings provided opportunity to evaluate test forms in their entirety and recommend changes. Requested item replacements were accommodated to the extent possible while striving to maintain the integrity of the various linking designs required for the operational test analyses. Psychometricians were available throughout this process to provide guidance with regard to implications of item replacements for the linking and statistical requirements.

Further information regarding the college- and career-ready content standards, performance level descriptors, and accessibility features and accommodations is provided at http://resources.newmeridiancorp.org/.

## 14.3 Evidence Based on Internal Structure

Analyses of the internal structure of a test typically involve studies of the relationships among test items and/or test components (i.e., subclaims) in the interest of establishing the degree to which the items or components appear to reflect the construct on which a test score interpretation is based (AERA, APA, & NCME, 2014, p. 16). The term "construct" is used here to refer to the characteristics that a test is intended to measure; in the case of the operational tests, the characteristics of interest are the knowledge and skills defined by the test blueprint for ELA/L and for mathematics.

The summative assessments provide a full summative test score, the ELA/L Reading and Writing claim scores, and the ELA/L subclaim scores, as well as the mathematics subclaim scores. The goal of reporting at this level is to provide criterion-referenced data to assess the strengths and weaknesses of a student's achievement in specific components of each content area. This information can then be used by teachers to plan for further instruction, to plan for curriculum development, and to report progress

to parents. The results can also be used as one factor in making administrative decisions about program effectiveness, teacher effectiveness, class grouping, and needs assessment.

### 14.3.1 Intercorrelations

The ELA/L full summative tests comprise two claim scores, Reading (RD) and Writing (WR), 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 RD claim score is a composite of RL, RI, and RV. The writing claim score, a composite of WE and WKL, comprises only PCR items, and the same PCR items are in each subclaim. The ELA/L operational test analyses were performed by evaluating the separate trait scores of WE and WKL, and for some PCR items also RL or RI; therefore, the trait scores were used for the intercorrelations.

The mathematics full summative tests have four subclaim scores – Major Content (MC), Mathematical Reasoning (MR), Modeling Practice (MP), and Additional and Supporting Content (ASC).

High total group internal consistencies as well as similar reliabilities across subgroups provide additional evidence of validity. High reliability of test scores implies that the test items within a domain are measuring a single construct, which is a necessary condition for validity when the intention is to measure a single construct. Refer to Section 13 for reliability estimates for the overall population, subgroups of interest, as well as for claims and subclaims for ELA/L and subclaims for mathematics.

Another way to assess the internal structure of a test is through the evaluation of correlations among scores. These analyses were conducted between the ELA/L Reading and Writing claim scores and the ELA/L subclaims (RL, RI, RV, WE, and WKL) and between the mathematics subclaims. If these components within a content area are strongly related to each other, this is evidence of unidimensionality.

A series of tables are provided to summarize the results for the spring 2023 administration. Tables 14.1 through 14.6 present the Pearson correlations observed between the ELA/L Reading and Writing claim scores and subclaim scores for each grade. The tables provide the weighted average intercorrelations by averaging the intercorrelations computed for all the core operational forms of the test within each grade level. The total sample size across all forms is provided in the upper triangle portion of the tables. The WR, WE, and WKL scores tended to be highly correlated; this is expected given that these three intercorrelations are based on the trait scores from the same Writing items. RL, RI, and RV, all subclaims of Reading, are moderately to highly correlated. Additionally, the WR claim and the WE and WKL subclaims are moderately correlated with RD subclaims (of RL, RI, and RV). These moderate to high ELA/L intercorrelations amongst 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 intercorrelations estimates for mathematics are provided in Tables 14.7 through 14.12. The average intercorrelations are provided in the lower portion of the table and the total sample sizes are provided in the upper portion of the table. Please refer to Appendix 12 (Form Composition) for information about the number of items and number of score points in each claim and subclaim.

The mathematics intercorrelations are moderate. The main observable pattern in the mathematics intercorrelations is that the MC subclaim generally has slightly higher correlations with the ASC, MR, and MP subclaims; the intercorrelations amongst 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 Intercorrelations and Reliability between Grade 3 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 128,356 | 128,356 | 128,356 | 128,356 | 128,356 | 128,356 |
| **RL** | 0.89 | 1 | 128,356 | 128,356 | 128,356 | 128,356 | 128,356 |
| **RI** | 0.81 | 0.58 | 1 | 128,356 | 128,356 | 128,356 | 128,356 |
| **RV** | 0.80 | 0.58 | 0.53 | 1 | 128,356 | 128,356 | 128,356 |
| **WR** | 0.59 | 0.52 | 0.56 | 0.42 | 1 | 128,356 | 128,356 |
| **WE** | 0.57 | 0.51 | 0.55 | 0.40 | 0.97 | 1 | 128,356 |
| **WKL** | 0.51 | 0.45 | 0.47 | 0.38 | 0.86 | 0.72 | 1 |

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

Table 14.2 Average Intercorrelations and Reliability between Grade 4 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 127,980 | 127,980 | 127,980 | 127,980 | 127,980 | 127,980 |
| **RL** | 0.86 | 1 | 127,980 | 127,980 | 127,980 | 127,980 | 127,980 |
| **RI** | 0.77 | 0.51 | 1 | 127,980 | 127,980 | 127,980 | 127,980 |
| **RV** | 0.82 | 0.54 | 0.49 | 1 | 127,980 | 127,980 | 127,980 |
| **WR** | 0.59 | 0.50 | 0.56 | 0.43 | 1 | 127,980 | 127,980 |
| **WE** | 0.58 | 0.49 | 0.56 | 0.42 | 0.98 | 1 | 127,980 |
| **WKL** | 0.53 | 0.44 | 0.48 | 0.40 | 0.87 | 0.76 | 1 |

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

Table 14.3 Average Intercorrelations and Reliability between Grade 5 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 129,738 | 129,738 | 129,738 | 129,738 | 129,738 | 129,738 |
| **RL** | 0.88 | 1 | 129,738 | 129,738 | 129,738 | 129,738 | 129,738 |
| **RI** | 0.80 | 0.57 | 1 | 129,738 | 129,738 | 129,738 | 129,738 |
| **RV** | 0.81 | 0.57 | 0.51 | 1 | 129,738 | 129,738 | 129,738 |
| **WR** | 0.60 | 0.55 | 0.57 | 0.41 | 1 | 129,738 | 129,738 |
| **WE** | 0.60 | 0.54 | 0.56 | 0.40 | 0.99 | 1 | 129,738 |
| **WKL** | 0.56 | 0.51 | 0.52 | 0.39 | 0.91 | 0.85 | 1 |

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

Table 14.4 Average Intercorrelations and Reliability between Grade 6 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 133,179 | 133,179 | 133,179 | 133,179 | 133,179 | 133,179 |
| **RL** | 0.89 | 1 | 133,179 | 133,179 | 133,179 | 133,179 | 133,179 |
| **RI** | 0.82 | 0.59 | 1 | 133,179 | 133,179 | 133,179 | 133,179 |
| **RV** | 0.82 | 0.62 | 0.53 | 1 | 133,179 | 133,179 | 133,179 |
| **WR** | 0.64 | 0.58 | 0.62 | 0.45 | 1 | 133,179 | 133,179 |
| **WE** | 0.64 | 0.57 | 0.61 | 0.44 | 0.99 | 1 | 133,179 |
| **WKL** | 0.62 | 0.55 | 0.59 | 0.43 | 0.95 | 0.90 | 1 |

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

Table 14.5 Average Intercorrelations and Reliability between Grade 7 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 134,267 | 134,267 | 134,267 | 134,267 | 134,267 | 134,267 |
| **RL** | 0.89 | 1 | 134,267 | 134,267 | 134,267 | 134,267 | 134,267 |
| **RI** | 0.82 | 0.58 | 1 | 134,267 | 134,267 | 134,267 | 134,267 |
| **RV** | 0.78 | 0.55 | 0.50 | 1 | 134,267 | 134,267 | 134,267 |
| **WR** | 0.66 | 0.56 | 0.65 | 0.44 | 1 | 134,267 | 134,267 |
| **WE** | 0.65 | 0.55 | 0.65 | 0.44 | 1.00 | 1 | 134,267 |
| **WKL** | 0.64 | 0.54 | 0.62 | 0.43 | 0.95 | 0.92 | 1 |

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

Table 14.6 Average Intercorrelations and Reliability between Grade 8 ELA/L Subclaims

| | RD | RL | RI | RV | WR | WE | WKL |
|---|---|---|---|---|---|---|---|
| **RD** | 1 | 138,908 | 138,908 | 138,908 | 138,908 | 138,908 | 138,908 |
| **RL** | 0.87 | 1 | 138,908 | 138,908 | 138,908 | 138,908 | 138,908 |
| **RI** | 0.83 | 0.56 | 1 | 138,908 | 138,908 | 138,908 | 138,908 |
| **RV** | 0.78 | 0.54 | 0.52 | 1 | 138,908 | 138,908 | 138,908 |
| **WR** | 0.62 | 0.52 | 0.61 | 0.42 | 1 | 138,908 | 138,908 |
| **WE** | 0.62 | 0.52 | 0.61 | 0.41 | 1.00 | 1 | 138,908 |
| **WKL** | 0.61 | 0.52 | 0.59 | 0.42 | 0.96 | 0.93 | 1 |

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

Table 14.7 Average Intercorrelations and Reliability between Grade 3 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 128,109 | 128,109 | 128,109 |
| **ASC** | 0.72 | 1 | 128,109 | 128,109 |
| **MR** | 0.71 | 0.66 | 1 | 128,109 |
| **MP** | 0.78 | 0.67 | 0.67 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

Table 14.8 Average Intercorrelations and Reliability between Grade 4 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 127,833 | 127,833 | 127,833 |
| **ASC** | 0.75 | 1 | 127,833 | 127,833 |
| **MR** | 0.72 | 0.67 | 1 | 127,833 |
| **MP** | 0.66 | 0.59 | 0.62 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

Table 14.9 Average Intercorrelations and Reliability between Grade 5 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 129,562 | 129,562 | 129,562 |
| **ASC** | 0.71 | 1 | 129,562 | 129,562 |
| **MR** | 0.71 | 0.65 | 1 | 129,562 |
| **MP** | 0.72 | 0.63 | 0.66 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

Table 14.10 Average Intercorrelations and Reliability between Grade 6 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 132,858 | 132,858 | 132,858 |
| **ASC** | 0.75 | 1 | 132,858 | 132,858 |
| **MR** | 0.71 | 0.67 | 1 | 132,858 |
| **MP** | 0.65 | 0.59 | 0.58 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

Table 14.11 Average Intercorrelations and Reliability between Grade 7 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 133,956 | 133,956 | 133,956 |
| **ASC** | 0.76 | 1 | 133,956 | 133,956 |
| **MR** | 0.74 | 0.71 | 1 | 133,956 |
| **MP** | 0.71 | 0.66 | 0.66 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

Table 14.12 Average Intercorrelations and Reliability between Grade 8 Mathematics Subclaims

| | MC | ASC | MR | MP |
|---|---|---|---|---|
| **MC** | 1 | 138,558 | 138,558 | 138,558 |
| **ASC** | 0.69 | 1 | 138,558 | 138,558 |
| **MR** | 0.69 | 0.67 | 1 | 138,558 |
| **MP** | 0.65 | 0.60 | 0.59 | 1 |

Note: MC = Major Content, ASC = Additional and Supporting Content, MR = Mathematical Reasoning, and MP = Modeling Practice.

## 14.3.2 Reliability

Additionally, the reliability analyses presented in Section 13 of this technical report provide information about the internal consistency of the summative assessments. Internal consistency is typically measured via correlations amongst the items on an assessment and provides an indication of how much the items measure the same general construct. The raw score reliability estimates, computed using coefficient alpha (Cronbach, 1951), are presented in Tables 13.1 and 13.2.[9]. Table 13.5 summarizes test reliability for groups of interest for ELA/L grades 3, and Table 13.6 summarizes test reliability for groups of interest for mathematics grade 3. Along with the subclaim intercorrelations, the reliability estimates indicate that the items within each assessment are measuring the same construct and provide further evidence of unidimensionality.

## 14.3.3 Local Item Dependence

In addition to the intercorrelations for ELA/L and mathematics, local item independence was evaluated. Local independence is one of the primary assumptions of item response theory (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 essentially 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 since 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. Specifically, if evidence suggests these items indeed have local dependence, then it might be preferable to sum the item scores into clusters or testlets as a method of minimizing LID. However, if these items do not appear to have strong local item dependence, then retaining the scores as individual item scores in an IRT calibration is preferred since more information concerning item properties is retained. During the initial operational administration of the summative assessments in spring 2015, a study that included two methods of investigating the presence of LID was conducted. A description of the methods along with study findings are summarized below.

First, analyses of the internal consistency in items and testlets were conducted under classical test theory (Wainer & Thissen, 2001) as a way to evaluate the degree of LID. Two estimates of Cronbach's alpha (Cronbach, 1951) were compared based on individual items in a test and those clustered into testlets. Cronbach's alpha is formulated as:

$$\alpha = \frac{l}{l-1}\frac{\sum_{i \neq i'} \sigma_{ii'}}{\sigma_X^2} \quad (14\text{-}1)$$

where $l$ is the total number of items, $\sigma_{ii'}$ is the covariance of items $i$ and $i'$ $(i \neq i')$, and $\sigma_X^2$ is the variance of total scores. To compute an alpha coefficient, sample standard deviations and variances are substituted for the $\sigma_{ii'}$ and $\sigma_X^2$. The alpha for the total test based on individual items is compared with those that form testlets based on larger subparts. If the item-level configuration has appreciably higher levels of internal consistency compared with the testlets, LID may be present.

For IRT-based methods, local dependence can be evaluated using statistics such as $Q3$ (Yen, 1984). The item residual is the difference between observed and expected performance. The $Q3$ index is the correlation between residuals of each item pair defined as

[9] Section 13 provides information on the computations of the reliability estimates.

$$d_i = (O - \hat{E}), \quad (14\text{-}2)$$

$$Q_3 = r(d_i, d_{i'}) \quad (14\text{-}3)$$

where $O$ is the observed score and $\hat{E}$ is the expected value of $O$ under a proposed IRT model and the index is defined as the correlation between the two item residuals.

LID manifests itself as a residual correlation that is nonzero and large. For $Q3$, LID can be either positive or negative. Positive (negative) LID indicates that performance is higher (lower) than expectation. The residual $Q3$ correlation matrix can be inspected to determine if there are any blocks of locally dependent items (e.g., perhaps blocks of items belonging to the same reading passage). For $Q3$, the null hypothesis is that local independence holds true. The expected value of $Q3$ is -1/n-1 where n is the number of items such that the statistic shows a small negative bias. As a rule of thumb, item pairs with moderate levels of LID for $Q3$ are |.2| or greater. Significant levels of LID are present when the statistic is greater than |.4|. An alternative is to use the Fisher r to z transformation and evaluate the resulting p-values.

For the LID comparisons, the following eight test levels administered in spring 2015 were selected:

- Grade 4 for span 3–5 in ELA/L
- Grade 4 for span 3–5 in mathematics
- Grade 7 for span 6–8 in ELA/L
- Grade 7 for span 6–8 in mathematics
- Grade 10 for span 9–11 in ELA/L
- Integrated Mathematics II for Integrated Mathematics I–III
- Algebra I
- Algebra II

One spring 2015 CBT form for each of the eight tests was selected that was roughly at the median in terms of test difficulty. For ELA/L, reading items were summed according to passage assignment. For mathematics, items were summed according to subclaims. Cronbach's alpha was computed for the entire forms using the two different approaches as described above, one involving calculations at the item level and the second utilizing scores on summed items (i.e., testlets). Further description of the data is given in Table 14.13.

To cross-validate the internal consistency analysis, the Q3 statistic was computed from spring CBT data based on grade 4 ELA/L and Integrated Mathematics II items. All items in the pool at that test level were included. The CBT item pool for grade 4 ELA/L contained 125 items while Integrated Mathematics II had 77 items.

The results for the internal consistency analysis are shown in Figure 14.1. In every instance, the item-level Cronbach's alpha is higher than in the testlet configuration. The greatest difference was for Algebra II, which showed a difference of 0.07. Although this was not unexpected, the magnitude of the differences in the respective alpha coefficients in general do not suggest a concerning level of LID. Table 14.14 shows the summary for the Q3 values. Figures 14.2 and 14.3 show graphs of the distribution of Q3 values. Most of the Q3 values were small and negative, again suggesting that LID is not at a level of concern. For these two test levels, the difference in the alpha coefficients was 0.03 and was consistent with the low values of Q3.

In summary, this investigation did not find evidence for the existence of pervasive LID. The results of both the internal consistency analyses and Q3 methods support a claim of minimal LID. For a multiple-choice-only test containing four reading passages with 5 to 12 items associated with a reading passage, Sireci et al. (1991) reported that testlet alpha was approximately 10 percent lower than the item-level coefficient. In comparison, the tests have complex test structures and exhibited smaller differences in alpha coefficients. In addition, the median Q3 values presented in Table 14.20 centered around the expectation of -1/n-1.

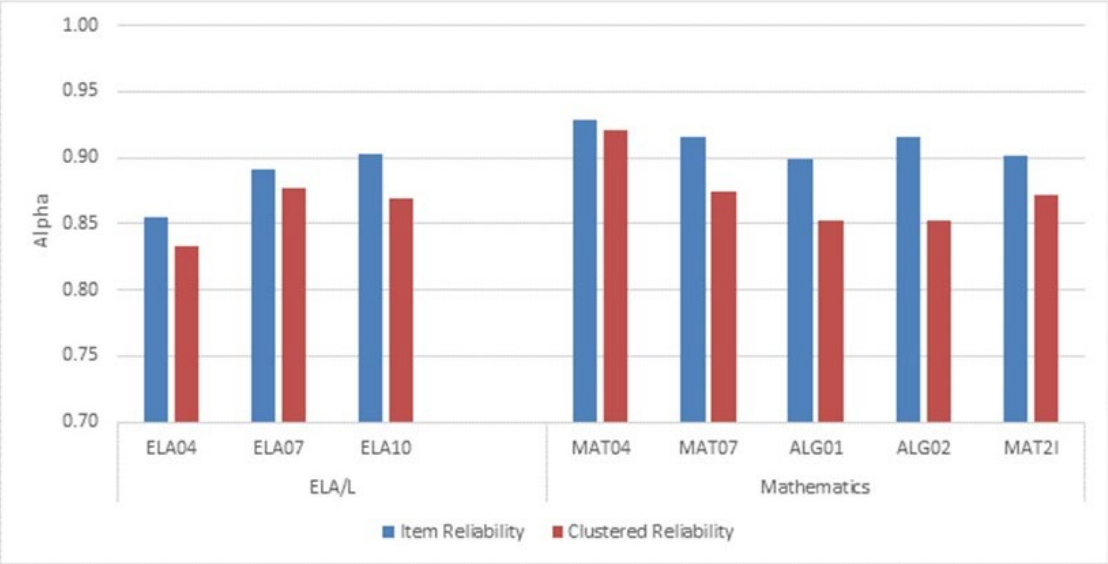



Figure 14.1 Comparison of Internal Consistency by Item and Cluster (Testlet)

Table 14.13 Conditions used in LID Investigation and Results

| Content | Grade/ Course | N Valid | N Complete | Percent Incomplete | No. Items | No. Tasks | Item Rel. | Task Rel. |
|---|---|---|---|---|---|---|---|---|
| | **ELA/L** | | | | | | | |
| ELA/L | 4 | 13,660 | 13,518 | 1.04 | 31 | 5 | 0.86 | 0.83 |
| ELA/L | 7 | 12,757 | 12,685 | 0.56 | 41 | 7 | 0.89 | 0.88 |
| ELA/L | 10 | 3,097 | 3,033 | 2.07 | 41 | 7 | 0.90 | 0.87 |

| Content | Grade/ Course | N Valid | N Complete | Percent Incomplete | No. Items | No. Tasks | Item Rel. | Task Rel. |
|---|---|---|---|---|---|---|---|---|
| | **Mathematics** | | | | | | | |
| Math | 4 | 10,332 | 10,255 | 0.75 | 53 | 4 | 0.93 | 0.92 |
| Math | 7 | 10,295 | 10,188 | 1.04 | 50 | 6 | 0.92 | 0.87 |
| Math | A1 | 5,072 | 4,885 | 3.69 | 52 | 6 | 0.90 | 0.85 |
| Math | A2 | 4,982 | 4,769 | 4.28 | 54 | 6 | 0.92 | 0.85 |
| Math | M2 | 2,708 | 2,645 | 2.33 | 51 | 6 | 0.90 | 0.87 |

Note: A1 = Algebra I, A2 = Algebra II, M2 = Integrated Mathematics II.

Table 14.14 Summary of Q3 Values for ELA/L Grade 4 and Integrated Mathematics II (Spring 2015)

| Min. | Q1 | Median | Mean | Q3 | Max. | SD |
|---|---|---|---|---|---|---|
| | ELA/L Grade 4 | | | | | |
| -0.138 | -0.047 | -0.031 | -0.031 | -0.017 | 0.279 | 0.030 |
| | Integrated Mathematics II | | | | | |
| -0.160 | -0.038 | -0.017 | -0.019 | 0.001 | 0.280 | 0.032 |

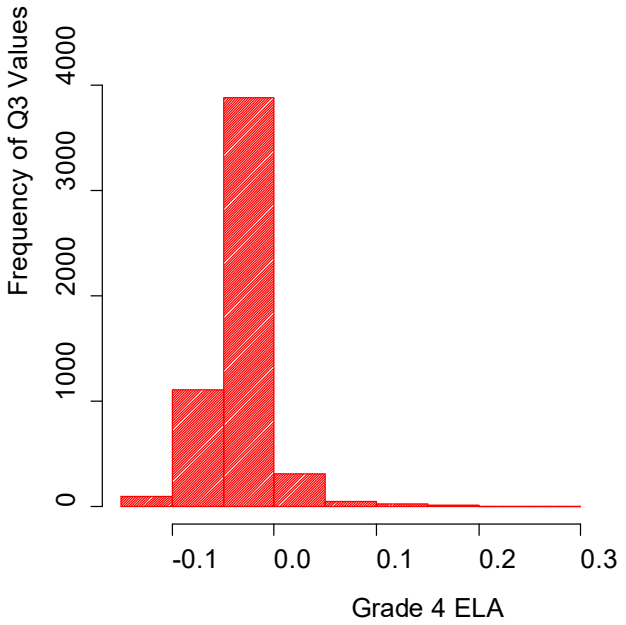



Figure 14.2 Distribution of Q3 Values for Grade 4 ELA/L (Spring 2015)

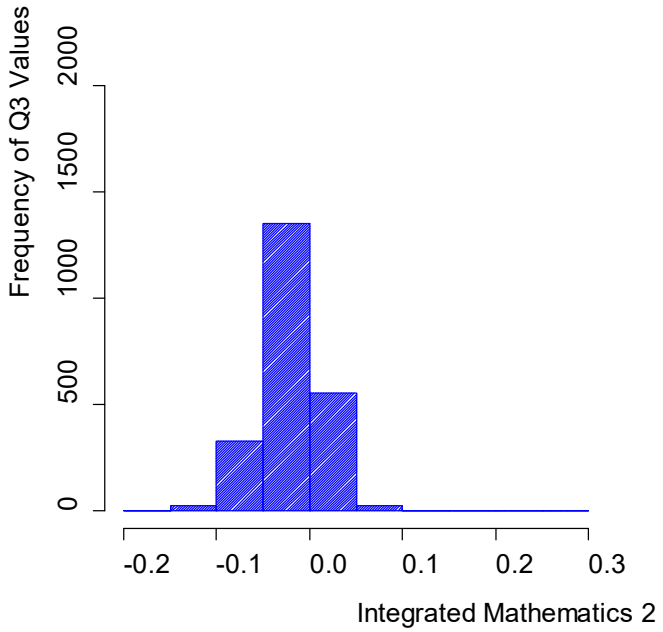



Figure 14.3 Distribution of Q3 Values for Integrated Mathematics II (Spring 2015)

## 14.4 Evidence from Special Studies

Several research studies were conducted to provide additional validity evidence for the participating state and agencies' 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. Some of the special studies conducted include the following:

- Content alignment studies
- Benchmarking study
- Longitudinal study of external validity
- Mode comparability study
- Device comparability study
- Quality Testing Standards study

The following paragraphs briefly describe each of these studies.

### 14.4.1 Content Alignment Studies

In 2016, content of the ELA/L assessments at grades 5, 8, and 11 and the Algebra II and Integrated Mathematics II assessments were evaluated to determine how well the assessments were aligned to the Common Core State Standards (CCSS; Doorey, & Polikoff, 2016; Schultz et al., 2016). These content alignment studies were conducted by the Fordham Institute for grades 5 and 8 and by Human Resources Research Organization (HumRRO) for the high school assessments. Both of these studies used the same methodology by having content experts review the assessment items and answers (for the constructed-response items the rubrics were reviewed). The content experts then 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 English learners and students with disabilities. The authors of both studies noted that 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.

The content studies had the individual content experts review and rate each item; then as a group the content experts came to a consensus on the final ratings for the content alignment, depth of knowledge, and accessibility to all students. In addition to the ratings, the content experts were asked to make comments that provided an explanation of their ratings; these comments were then used by the full group of content experts to provide narrative comments regarding the overall ratings and to provide feedback and recommendation about the assessment programs.

The assessment program was rated as Excellent Match for ELA/L content and depth and Good Match for mathematics content and depth for grades 5 and 8. However, for grade 11 ELA/L content was rated as Excellent Match but depth was rated as Limited/Uneven Match. The high school mathematics assessments were rated at Excellent Match for content and Good Match for depth.

The content studies noted some weaknesses and strengths of the assessments. For ELA/L, it was noted that the assessments include complex texts, a range of cognitive demands, and have a variety of item types. Furthermore, the ELA/L "assessments require close reading, assess writing to sources, research, and inquiry, and emphasize vocabulary and language skills" (Doorey & Polikoff, 2016). The grade 11 ELA/L assessment had a smaller range of depth and included items assessing the higher-demand cognitive level. A weakness of the ELA/L assessments is the lack of a listening and speaking component. It was also suggested that the ELA/L assessments could be enhanced by the inclusion of a research task that requires the use of two or more sources of information.

The strengths of the mathematics assessments include assessments that are aligned to the major work for each grade level. While the grade 5 assessment includes a range of cognitive demand, the grade 8 assessment includes a number of 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. Additionally, the reviewers noted that some of the mathematics items should be carefully reviewed for editorial and mathematical accuracy.

The high school report noted that the assessment program incorporates a number of accessibility features and test accommodations for students with disabilities and for English learners. Furthermore, the assessments included items designed to accommodate the needs of students with disabilities.

In 2017, the Human Resources Research Organization (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). This alignment study followed a similar methodology as the 2016 study. For the study, cognitive complexity was consistent with the current assessments' definition. An item's cognitive complexity is 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 were asked to determine 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 percent of grade 3 and 4 items, 95.5 percent of grade 6 items, and 94.6 percent of grade 7 items were fully aligned. For mathematics, 92.0 percent of grade 3, 91.1 percent of grade 4 items, 83.1 percent of grade 6 items, and 94.0 percent of grade 7 items were fully aligned. The majority of the items that did not fall into fully aligned were considered partially aligned to the standards. 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, for all four grades, at least 93 percent 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 percent), and grade 7 had the highest (96.3 percent). 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, as well as the cognitive complexity levels of these same grades (Schultz et al., 2017). There are five criteria for ELA/L content: close reading, writing, vocabulary and language skills, research and inquiry, and speaking and listening. Reviewers were asked to rate 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. There are four criteria for ELA/L depth: text quality and types, complexity of texts, cognitive demand, and high-quality items and item variety. All grades in this study 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 in the study 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 items on ELA/L and mathematics at grades 3, 4, 6, and 7 (Schultz et al., 2017). Reviewers indicated their agreement with the intended cognitive complexity ratings provided by participating states and agencies 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 math, 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. For grade 7, there was agreement across low, medium, and high in all domains.

### 14.4.2 Benchmarking Study

The purpose of the benchmarking study (McClarty et al., 2015) was to provide information that would inform the performance level setting (PLS) process. An evidence-based standard setting approach (EBSS; McClarty et al., 2013) was used to establish the performance levels for its assessments. 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. The charge of this committee was to suggest a reasonable range for the percentage of students meeting or exceeding the Level 4 threshold score and therefore considered college- and career-ready. Section 8.3.2 of this report provides more information about the pre-policy meeting.

For the benchmarking study, external information was analyzed to provide information about the Level 4 threshold scores for the grade 11 ELA/L, Algebra II, and Integrated Mathematics III assessments, the grade 8 ELA/L and mathematics assessments, and the grade 4 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 (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), National Assessment of Educational Progress (NAEP), 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 participating states and agencies was determined. Across all grades and subjects, the data indicated approximately 25 to 50 percent of students were college- and career-ready or on track to readiness based on the Level 4 expectations.

For details on how the benchmarking study was used during the standard setting process, refer to Section 8 of this technical report.

### 14.4.3 Longitudinal Study of External Validity of Performance Levels (Phase 1)

In 2016–2017, the first phase of a two-part external validity study of claims about the alignment of Level 4 to college readiness was completed (Steedle et al., 2017) using the summative assessment scores from the 2014–2015 and 2015–2016 academic years. Associations between the performance levels and college-readiness benchmarks established by the College Board and ACT were used to study the claim that students who achieve Level 4 have a 0.75 probability of attaining at least a C in entry-level, credit-bearing, postsecondary coursework. Regression estimates measured the relationship between the summative assessment scores and external test scores. The Level 4 benchmark was used to estimate the expected score on an external test, and vice versa. Assessment scores were dichotomized for additional analyses. Cross-tabulation tables provided classification agreement among tests. Logistic regression modeled the relationship between students' summative scores and their probabilities of meeting the external assessment benchmark, and vice versa.

These methods were used to make the following comparisons in mathematics: Algebra I and PSAT10 Math; Geometry and PSAT10 Math; Algebra II and PSAT10 Math; Algebra II and PSAT/NMSQT Math; Algebra II and SAT Math; and Algebra II and ACT Math. The classification agreement (meeting the benchmark on both tests or not meeting the benchmark on both tests) ranged from 62.50 percent to 86.50 percent. The overall trend indicated that students who met the benchmark on a mathematics assessment were likely to meet or exceed the benchmark on an external test (probabilities ranged from 0.509 to 0.886). However, students who met the benchmark on the external test had relatively low probabilities of meeting the mathematics benchmark (0.097 to 0.310).

The following comparisons were made in ELA/L: grade 9 and PSAT10 evidence-based reading and writing (EBRW); grade 10 and PSAT10 EBRW; grade 10 and PSAT/NMSQT EBRW; grade 10 and SAT EBRW; grade 11 and PSAT/NMSQT EBRW; grade 11 and SAT EBRW; grade 11 and ACT English; and grade 11 and ACT reading. In the majority of comparisons, the trend in ELA/L results was similar to mathematics. The classification agreements ranged from 67.30 percent to 79.70 percent. Students meeting the ELA/L benchmark had probabilities between 0.667 and 0.825 of meeting the benchmark on the external assessment. However, a student taking the external test had lower probabilities of meeting the benchmark on the ELA/L assessments (0.326 to 0.513).

Overall, results indicated that a student meeting the benchmark on the summative assessment had a high probability of making the benchmark on the external test, but the converse did not hold for students meeting the benchmark on the external test, for the majority of comparisons. These results suggest that meeting the summative benchmark is an indicator of academic readiness for college. However, it may be that students who meet the summative benchmark have a greater than 0.75 probability of earning a C or higher in first-year college courses.

Phase 1 is a preliminary study using indirect comparisons; therefore, there are limitations to interpretations. Phase 2 of this study was to occur in 2018 and use longitudinal data including academic performance in entry-level college courses for students who took the summative assessments during high school. Currently, this study is on hold due to challenges obtaining student academic data from entry-level college courses and/or matching the data to the student summative scores.

### 14.4.4 Mode and Device Comparability Studies

The summative assessments have been operational since the 2014–2015 school year. In addition to the traditional paper format, the assessments were available for online administration via a variety of electronic devices, including desktop computers, laptop computers, and tablets. The research agenda includes several studies evaluating the interchangeability of scale scores across modes and devices.

This report describes a two-pronged study consisting of a mode comparability analysis and a device comparability analysis. In the mode comparability analysis, scores arising from the paper administration were compared to those arising from any type of online

administration. In the device comparability analysis, online scores arising from tests administered using a tablet are compared with online scores arising 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: 1) to investigate whether assessment items were of similar difficulty across the levels of conditions for each analysis (i.e., paper and online for the mode comparability analysis and tablet and non-tablet for the device comparability analysis); 2) to determine whether the psychometric properties of test scores were similar across the levels of conditions for each analysis; and 3) 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 as well as 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 results of the mode comparability analysis were mixed and found to be consistent with prior research. The item means suggested that items were of similar difficulty on paper and online modes. Only two items were flagged for mode effects, both of which were on the mathematics assessments. C-level differential item functioning (DIF) was present in both analyses. All the items flagged for C-level DIF in the mathematics assessments favored the online students, whereas the majority of items flagged for C-level DIF in the ELA/L assessments favored the paper students. An examination of test reliability displayed comparable reliability values between the two modes; none of the test forms were flagged for mode effects with respect to test reliability. The test-level adjustment analysis as well as the change of the paper 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 paper test form for each assessment except grades 5 and 7 mathematics. However, all adjustments were less than the minimum standard error of Theta except for grade 11 ELA/L, which was the same as 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). Specifically, 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 as well as the change of the students' performance levels after the adjustment constants were applied did not indicate strong evidence of device effects.

The generalizability of the findings from this study may be limited due to the small sample size of both the paper students (for mode comparability) and the tablet students (for device comparability) at the high-school grades; however, it appears that high-quality matching supports the internal validity of this study's findings. For mode and device comparability, there were little to no items 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 paper or tablet condition were within measurement precision.

## 14.5 Evidence Based on Response Processes

As noted in the AERA, APA, and NCME Standards (2014), 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. This type of evidence may be gathered from interacting with students in order to understand what processes underlie their item responses. Evidence may also be derived from feedback provided by test proctors/teachers involved in the administration of the test and raters 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.

New Meridian has undertaken research investigating 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. In addition, the accessibility of the test for students with disabilities and English learners has been examined. This research has included examining students' understanding of the format of the assessments and the use of technology.

One such study involved a series of four component studies that were conducted to evaluate the usability and effect of a drawing tool for online mathematics items. The purpose of these studies was to determine if results could support the use of the drawing

tool, which is a way to expand students' ability to demonstrate their understanding and reasoning, thereby enhancing accessibility and construct validity of the assessment. This goal is in keeping with guidance from the Common Core State Standards (CCSS) and the National Council of Teachers of Mathematics (NCTM) that students should have multiple paths and tools available to express their responses. Additionally, the drawing tool was intended to boost comparability across modes.

The first two studies (Brandt, Bercovitz, McNally, & Zimmerman, 2015; Brandt, Bercovitz, & Zimmerman, 2015) 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 (Steedle & LaSalle, 2016; Minchen et al., 2018) 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 were studied by field testing them 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. Items with access to the drawing tool, however, 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.

Several other research efforts have investigated questions relevant to response processes evidence. Descriptions of the research conducted can be found online.[10]

## 14.6 Interpretations of Test Scores

The summative assessment scores are expressed as scale scores (both total scores and claim scores), along with performance levels to describe how well students met the academic standards for their grade level. Additionally, information on specific skills (the subclaims) is also provided and is reported as *Below Expectations*, *Nearly Meets Expectations*, and *Meets or Exceeds Expectations*. On the basis of a student's total score, an inference is drawn about how much knowledge and skill in the content area the student has acquired. The total score is also used to classify students in terms of their level of knowledge and skill in the content area as students progress in their K–12 education. These levels are called performance levels and are reported as:

Level 5: Exceeded expectations

Level 4: Met expectations

Level 3: Approached expectations

Level 2: Partially met expectations

Level 1: Did not yet meet expectations

Students classified as either Level 4 or Level 5 are meeting or exceeding the grade level expectations. Performance level descriptors (PLDs) assist with the understanding and interpretations of the ELA/L scores (https://resources.newmeridiancorp.org/ela-test-design/) and mathematics scores (https://resources.newmeridiancorp.org/math-test-design/). Additionally, resource information is available online to educators, parents, and students (http://resources.newmeridiancorp.org/). Section 12 of this technical report provides more information on the scale scores and the subclaim scores.

## 14.7 Evidence Based on the Consequences to Testing

The consequence of testing should also be investigated to support the validity evidence for the use of the summative assessments as the standards note that tests are usually administered "with the expectation that some benefit will be realized from the intended use of the scores" (AERA, APA, & NCME, 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 CCSS and the continued administration of the assessments.

Consequences of the tests may vary by state or by school district. For example, some states may require "passing" the assessments as one of several criteria for high school graduation, while other states/districts may not require students to "pass" the assessments for high school graduation. Additionally, some school districts may use the scores along with other information

[10] Various research is described at: http://resources.newmeridiancorp.org/

such as school grades and teacher recommendations for placing students into special programs (e.g., remedial support, gifted and talented program) or for course placement (e.g., Algebra I in grade 8). Because the consequences for the assessments can vary by each state, it is suggested that each member state provide school districts, teachers, parents, and students with information on how to interpret and use the scores. Additionally, the states should monitor how scores are used to ensure that the scores are being used as intended.

## 14.8 Summary

In this section of the technical report, several aspects of validity were included, such as validity evidence based on content, the internal structure of the assessments, relationships across the content assessments, and evidence from special studies.

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). Additionally, items were field tested prior to the initial operational administration, and data and feedback from students, test administrators, and classroom teachers was 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.

The intercorrelations of the subclaims, the reliability analyses, and the local item dependence analyses indicated that the ELA/L and the mathematics assessments are both essentially unidimensional. Furthermore, the correlations between ELA/L and mathematics indicated that the two assessments are measuring different content.

Several studies were conducted as part of the assessment program (e.g., benchmarking study, content evaluation/alignment studies, longitudinal study, and mode and device comparability studies). The benchmarking study was conducted in support of the standard setting meeting. This study indicated students performing at or above Level 4 could be considered to be college- and career-ready or on track to readiness.

The content evaluation/alignment studies performed by the Fordham Institute and HumRRO indicate that the assessments are good to excellent matches to the CCSS in terms of content and depth of knowledge. Thus, the assessments are assessing the college- and career-readiness standards. However, the reports noted that the program could improve by adding a wider range of depth of knowledge to some of the assessments. The reports also suggested enhancing the ELA/L assessments by including a research task that requires the use of two or more sources of information.

In the longitudinal study of external validity, associations between the performance levels and college-readiness benchmarks established by the College Board and ACT were used to study the claim that students who achieve Level 4 have a .75 probability of attaining at least a C in entry-level, credit-bearing, postsecondary coursework. In the first phase of the study, the relationship between the summative assessment and external tests was studied. Overall, results indicated that a student meeting the benchmark on the summative assessment had a high probability of making the benchmark on the external test, but the converse did not hold for students meeting the benchmark on the external test, for the majority of comparisons. These results suggest that meeting the benchmark is an indicator of academic readiness for college. In the next phase of the study, the relationship between scores and performance in first-year college courses will be explored.

The mode comparability study indicated that the comparability across modes was inconsistent across content domains and grade levels. The results of the mode comparability analysis were mixed and found to be consistent with prior research. The results of the device comparability study revealed consistent evidence supporting the comparability between the tablet condition (TC) and the non-tablet condition (NTC). In both the mode and device comparability studies, there were little to no items 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 paper or tablet condition were within measurement precision.

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 5 provides information concerning the test characteristics based on classical test theory.
- Section 6 provides information regarding the differential item functioning (DIF) analyses.
- Section 11 presents information regarding student characteristics for the spring administration of the ELA/L and mathematics administration.

- Section 12 provides detailed information concerning the scores that were reported and the cut scores for ELA/L and mathematics.
- Section 13 provides information on the test reliability (total test score and for subclaims) and includes information on the interrater reliability/agreement.

# Section 15: Student Growth Measures

Student growth percentiles (SGPs) are normative measures of annual progress. Normative measures 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, which 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. SGPs compare a student's performance to that of his or her academic peers both within the state and across the consortium. Academic peers are defined as students in the norm group who took the same assessment as the student in prior years and achieved a similar score.

Some participating states or agencies chose to implement norm groups based on their respective student data. State-specific SGP results are not reported in this Technical Report. As a result, SGPs were only summarized for states using norm groups based on the consortium. The following sections describe the norm groups, the estimation procedure, and the results for SGPs based on consortium norm groups.

The SGP describes a student's location in the distribution of current test scores for all students who performed similarly in the past. SGPs indicate 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, a SGP of 60 on grade 7 ELA/L means that the student scored better than 60 percent of the students in the state or consortium who took grade 7 ELA/L 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..[11] A SGP of 50 represents typical (median) student growth for the state or consortium. Because students are only compared with other students who performed similarly in the past, all students, regardless of starting point, can demonstrate high or low growth.

The 2020–2021 academic year is the seventh year of test administration, including an abbreviated administration to a small number of students in one state in 2020. Data from 2020 was not used in SGP calculations. Students in states that participated in spring 2018 and spring 2019 generally received SGPs based on two prior scores. Students in states that participated in spring 2019 received SGPs based on one prior score. Students who do not have a previous test score, which include any new students and all grade 3 and 4 students, do not receive an SGP.

## 15.1 Norm Groups

The norm groups consisted of students with the same prior scores based on grade or content area progressions (academic peers). SGPs were based on up to two years of prior test scores from spring 2018 and spring 2019 administrations. States administering traditional mathematics assessments in fall 2018 or fall 2019 may also have SGPs based on these prior scores. Tables 15.1–15.8 list the grade or content area progressions required for SGPs based on one prior or two prior test scores for ELA/L grades 3 through 11, mathematics grades 3 through 8, Algebra I, Geometry, Algebra II, Integrated Mathematics I, II, and III, respectively. In general, the progressions of grade levels and content areas are consecutive. The traditional and integrated mathematics courses have progressions that are not consecutive but reflect student progression for high school mathematics courses. SGPs were calculated for all norm groups with at least 1,000 students. Some progressions did not meet the minimum sample size for SGP calculations.

[11] Note: 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 often leads to confusion among non-technical stakeholders who often ask, "How many students are there with exactly the same prior scores?" To avoid explaining regression to non-technical stakeholders, the "similar scores" is often used to finesse the idea of regression without mentioning it.

Table 15.1 ELA/L Grade-Level Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| N/A | N/A | Grade 3* |
| N/A | Grade 3 | Grade 4 |
| Grades 3 and 4 | Grade 4 | Grade 5 |
| Grades 4 and 5 | Grade 5 | Grade 6 |
| Grades 5 and 6 | Grade 6 | Grade 7 |
| Grades 6 and 7 | Grade 7 | Grade 8 |
| Grades 7 and 8 | Grade 8 | Grade 9 |
| Grades 8 and 9 | Grade 9 | Grade 10 |
| Grades 9 and 10 | Grade 10 | Grade 11 |

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

Table 15.2 Mathematics Grade-Level Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| N/A | N/A | Grade 3* |
| N/A | Grade 3 | Grade 4 |
| Grades 3 and 4 | Grade 4 | Grade 5 |
| Grades 4 and 5 | Grade 5 | Grade 6 |
| Grades 5 and 6 | Grade 6 | Grade 7 |
| Grades 6 and 7 | Grade 7 | Grade 8 |

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

Table 15.3 Algebra I Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 5 and 6 | Grade 6 | Algebra I |
| Grades 6 and 7 | Grade 7 | Algebra I |
| Grades 6 or 7 and 8 | Grade 8 | Algebra I |
| Grades 6, 7, or 8 and Geometry | Geometry | Algebra I |
| Grade 8 and Integrated Mathematics I | Integrated Mathematics I | Algebra I |
| Integrated Mathematics I and Integrated Mathematics II | Integrated Mathematics II | Algebra I |

Table 15.4 Geometry Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 5 and 6 | Grade 6 | Geometry |
| Grades 6 and 7 | Grade 7 | Geometry |
| Grades 6 or 7 and 8 | Grade 8 | Geometry |
| Grades 6, 7, or 8 and Algebra I | Algebra I | Geometry |
| Grade 8 and Integrated Mathematics I | Integrated Mathematics I | Geometry |
| Integrated Mathematics I and Integrated Mathematics II | Integrated Mathematics II | Geometry |

Table 15.5 Algebra II Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 6 and 7 | Grade 7 | Algebra II |
| Grades 7 and 8 | Grade 8 | Algebra II |
| Grades 7 or 8 and Algebra I | Algebra I | Algebra II |
| Grade 8 or Algebra I and Geometry | Geometry | Algebra II |
| Grade 8 and Integrated Mathematics I | Integrated Mathematics I | Algebra II |
| Integrated Mathematics I and Integrated Mathematics II | Integrated Mathematics II | Algebra II |

Table 15.6 Integrated Mathematics I Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 5 and 6 | Grade 6 | Integrated Mathematics I |
| Grades 6 and 7 | Grade 7 | Integrated Mathematics I |
| Grades 6 or 7 and 8 | Grade 8 | Integrated Mathematics I |
| Grades 7 or 8 and Algebra I | Algebra I | Integrated Mathematics I |
| Grade 8 or Algebra I and Geometry | Geometry | Integrated Mathematics I |

Table 15.7 Integrated Mathematics II Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 6 and 7 | Grade 7 | Integrated Mathematics II |
| Grades 7 and 8 | Grade 8 | Integrated Mathematics II |
| Grades 7 or 8 and Integrated Mathematics I | Algebra I | Integrated Mathematics II |

Table 15.8 Integrated Mathematics III Grade/Content Area Progressions for One- and Two-year Prior Test Scores

| Two Prior Year Test Scores | One Prior Year Test Score | Current Year Test Score |
|---|---|---|
| Grades 6 and 7 | Grade 7 | Integrated Mathematics III |
| Grades 7 and 8 | Grade 8 | Integrated Mathematics III |
| Grades 7 or 8 and Integrated Mathematics I | Algebra I | Integrated Mathematics III |
| Integrated Mathematics I and Integrated Mathematics II | Integrated Mathematics II | Integrated Mathematics III |

In addition to the above progressions, in 2018 the State Leads approved a state-specific SGP progression for one state. In this state, grade 9 students are not required to take the test. Therefore, grade 10 students were not receiving a SGP. For this state, both mathematics and ELA/L progressions were adjusted (see Table 15.9) such that the grade 10 students would receive growth estimates. Other states were not affected by this change.

Table 15.9 State-specific SGP Progressions

| Two Prior Test Scores | One Prior Test Score | Current Test Score |
|---|---|---|
| ELA/L Grades 7 and 8 | ELA/L Grade 8 | ELA/L Grade 10 |
| Mathematics Grade 7 and 8 | Mathematics Grade 8 | Geometry |
| Mathematics Grade 7 and Algebra I | Algebra I | Geometry |

## 15.2 Student Growth Percentile Estimation

SGPs are calculated using quantile regression, which 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 assessment 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 (NCIEA) performed the analyses using Betebenner's (2009) non-linear quantile-regression based SGP. The analysis was done in the SGP package in R (Betebenner et al., 2017). For details on student growth percentiles, 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 like 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 $t \geqslant 2$), the $\tau$th quantile function for $Y_t$ conditional on prior scores ( $Y_{t-1}, \ldots, Y_1$ ) is

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

where $\phi_{ju}$ ( $j$ =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 $\tau$ 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 a SGP of 60.

SGPs are assumed to be uniformly distributed and uncorrelated with prior achievement. Scale score conditional standard errors of measurement (CSEMs) were incorporated for calculation of SGP standard errors of measurement (SEMs). 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 Student Growth Percentile Results/Model Fit for Total Group

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. A ceiling/floor effects test identifies potential problems at the highest obtainable scale scores (HOSS) and lowest obtainable scale scores (LOSS). Other fit plots compare the observed conditional density of SGP estimates with the theoretical uniform density. If there is perfect model fit, 10 percent of the estimated growth percentiles 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.

Tables 15.10 and 15.11 summarize SGP estimates for the total testing group for ELA/L and mathematics, respectively. SGPs were calculated at the consortium level and, if sample size was sufficient, the state level. Median SGPs were all 50. If the model is a perfect fit, the median is expected to be 50 with norm-referenced data. The minimum SGP is 1 and the maximum SGP is 99. The average standard error for the SGPs is within expectations for these models.

In general, SGPs can be divided into three categories: below 30 indicating that a student is not meeting a year's worth of growth, a SGP of 30–70 indicating that a student did achieve a year's worth of growth, and a 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 12.99–16.10, support these interpretations (Betebenner et al., 2016).

Table 15.9 Summary of ELA/L SGP Estimates for Total Group

| Grade | Sample Size | Average SGP | Average Standard | Median SGP |
|---|---|---|---|---|
| 4 | 96,655 | 49.99 | 14.09 | 50 |
| 5 | 98,944 | 49.99 | 14.13 | 50 |
| 6 | 99,392 | 50.00 | 13.66 | 50 |
| 7 | 99,326 | 50.00 | 13.89 | 50 |
| 8 | 98,201 | 49.98 | 13.89 | 50 |
| 9 | — | — | — | — |
| 10 | — | — | — | — |

Note: "—" indicates insufficient sample for SGP calculation for these tests.

Table 15.10 Summary of Mathematics SGP Estimates for Total Group

| Grade | Sample Size | Average SGP | Average Standard | Median SGP |
|---|---|---|---|---|
| 4 | 94,459 | 50.08 | 13.48 | 50 |
| 5 | 96,780 | 50.01 | 14.15 | 50 |
| 6 | 97,424 | 50.01 | 14.63 | 50 |
| 7 | 94,704 | 49.96 | 15.34 | 50 |
| 8 | 93,869 | 49.90 | 16.75 | 50 |
| A1 | 1,455 | 49.54 | 15.74 | 49 |
| GO | 1,861 | 49.63 | 15.24 | 50 |
| A2 | 1,558 | 49.58 | 15.89 | 49 |

Note: "—" indicates insufficient sample for SGP calculation for these tests. A1 = Algebra I, GO = Geometry, A2 = Algebra II

## 15.4 Student Growth Percentile Results for Subgroups of Interest

Median SGPs are provided for subgroups of interest. With norm-referenced data, the median of all SGPs is expected to be close to 50. Median subgroup growth percentiles below 50 represent growth lower than the median, and median growth percentiles above 50 represent growth higher than the median. Table 15.12 summarizes SGPs for groups of interest for ELA/L grade 5. The ELA/L tables for grades 5–8 and 10 are provided in the Appendix (Tables A.15.1–A.15.6). Table 15.13 summarizes SGPs for groups of interest for mathematics grade 5; the other mathematics subgroup results are provided in the Appendix (Tables A.15.7–A.15.13). Median SGPs for subgroups of interest fell within the band of 30–70, which is considered to be adequate growth. ELA/L grades 11, Algebra I, and Geometry had insufficient sample size for SGP subgroup results to be reported.

### 15.4.1 SGP Results for Gender

#### English Language Arts/Literacy

The median SGPs for females tend to be higher than the median SGPs for males. The median SGP for females ranges from 48 to 54, whereas the median SGP for males ranges from 46 to 50.5. The standard error for males and females is comparable to the total group.

#### Mathematics

There was no consistent pattern between median SGPs for females and males. The median SGP for females ranges from 48 to 51, and the median SGP for males ranges from 49 to 51. The standard errors for both are similar to the total group.

### 15.4.2 SGP Results for Ethnicity

#### English Language Arts/Literacy

The African American group median SGP ranges from 34 to 47, with students in higher grades at the higher range. Asian/Pacific Islanders tend to have the highest median SGPs, over 60 for all tests but grade 10. American Indian/Alaska Native students had median SGPs ranging from 43 to 52 in grades 5–8. The median SGP for Hispanics ranges from 43 to 51. For all ethnicity groups, standard errors are similar to that of the total group.

#### Mathematics

The median SGP for African Americans ranges from 33 to 41, with the highest growth in mathematics grade 8 and Algebra II. Asian/Pacific Islanders tend to have the highest SGPs across all tests, with a minimum of 51 and a maximum of 66. American Indian/Alaska Native had median SGPs ranging from 31 to 46. The median SGP for Hispanics ranges from 42 to 48. For all ethnicities, the standard errors for all groups are under 20 points.

### 15.4.3 SGP Results for Special Instructional Needs

#### English Language Arts/Literacy

Economically disadvantaged and English language learner students tended to have moderate median SGPs. The median SGP ranges from 41 to 48 for economically disadvantaged students and from 40 to 49 for English language learners. Students with disabilities observed median SGP of 40 to 44. The standard errors for special instructional needs subgroups are similar to those observed for the total group.

#### Mathematics

Economically disadvantaged and English language learner students tend to have lower median SGPs than the general population. The median SGP ranges from 39 to 45 for economically disadvantaged students and from 42 to 47 for English language learners. Students with disabilities median SGP ranges from 34.5 to 47, whereas for students without disabilities the median SGP ranges from 51 to 52. The standard errors for special education students are similar to the total group.

Table 15.11 Summary of SGP Estimates for Subgroups: Grade 4 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 49,196 | 49.67 | 14.11 | 50 |
| Female | 47,458 | 50.31 | 14.06 | 50 |
| **Ethnicity** | | | | |
| White | 51,770 | 51.44 | 13.99 | 52 |
| African American | 12,578 | 42.71 | 14.41 | 40 |
| Asian/Pacific Islander | 5,241 | 59.02 | 13.78 | 63 |
| American Indian/Alaska Native | 207 | 45.45 | 14.43 | 44 |
| Hispanic | 21,886 | 48.48 | 14.24 | 48 |
| Multiple | 4,703 | 50.78 | 13.95 | 51 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 40,641 | 45.49 | 14.24 | 44 |
| Not-economically Disadvantaged | 56,014 | 53.25 | 13.98 | 55 |
| English Learner (EL) | 14,992 | 47.00 | 14.37 | 46 |
| Non-English Learner | 81,663 | 50.54 | 14.03 | 51 |
| Students with Disabilities (SWD) | 16,915 | 41.49 | 14.39 | 38 |
| Students without Disabilities | 79,740 | 51.79 | 14.02 | 53 |

## 15.4.4 SGP Results for Students Taking Spanish Forms

### Mathematics

There is a wide range of median growth percentiles for students taking Spanish forms. The sample size is less than 50 for all grade levels. These forms had a slightly higher standard error on average, likely due to lower sample sizes.

Table 15.12 Summary of SGP Estimates for Subgroups: Grade 4 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 48,093 | 50.44 | 13.48 | 51 |
| Female | 46,365 | 49.71 | 13.49 | 50 |
| **Ethnicity** | | | | |
| White | 51,416 | 50.45 | 13.05 | 51 |
| African American | 12,328 | 44.39 | 14.79 | 42 |
| Asian/Pacific Islander | 5,187 | 59.13 | 13.14 | 63 |
| American Indian/Alaska Native | 197 | 49.16 | 13.55 | 48 |
| Hispanic | 20,401 | 50.15 | 13.88 | 50 |
| Multiple | 4,661 | 51.37 | 13.49 | 51 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 39,044 | 46.91 | 14.05 | 46 |
| Not-economically Disadvantaged | 55,415 | 52.32 | 13.08 | 53 |
| English Learner (EL) | 13,590 | 49.62 | 14.13 | 50 |
| Non-English Learner | 80,869 | 50.16 | 13.37 | 50 |
| Students with Disabilities (SWD) | 16,526 | 43.36 | 14.15 | 41 |
| Students without Disabilities | 77,933 | 51.51 | 13.34 | 52 |
| **Spanish Language Form** | 1,268 | 42.63 | 14.57 | 41 |

# References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, 63(4), 602–614.

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. *Bulletin*, Issue 21. Pearson Education, Inc.

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.

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

Brandt, R., Bercovitz, E., McNally, S., & Zimmerman, L. (2015). *Drawing response interaction usability study for PARCC*, July 28–July 30, 2015. Partnership for Assessment of Readiness for College and Careers.

Brandt, R., Bercovitz, E., & Zimmerman, L. (2015). *Drawing response interaction usability study for PARCC. Austin*, TX: Pearson.

Brennan, R. L. (2004). Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy. Version 1 (No. 9). CASMA Research Report.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Chicago, IL: Scientific Software International.

Center for Assessment. (2018). *PARCC comparability review guidelines*. Dover, NH: Center for Assessment.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (Second ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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

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

Doorey, N. & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Washington, DC: 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).* Princeton, NJ: 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). Hillsdale, NJ: Erlbaum.

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

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

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). Hillsdale, NJ: Erlbaum.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15(2), 1–8.

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

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models.* University of Iowa. Version 1.0.

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

Kolen, M. J. (2004). POLYCSEM windows console version [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The 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 Series, No. RR-93-48, Princeton, NJ: ETS.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 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,* 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,* 719–748.

McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson Educational Measurement, Iowa City, IA: 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. (2018). *Alternative blueprinting options 2018 research report*. Austin, TX: Pearson.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16(2),* 159–176.

Muraki, E. & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software International.

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, 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. *Meeting of the National Council on Measurement in Education.* Montreal. Canada.

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

Schultz, S. R., Michaels, H. R., Norman Dvorak, R., & Wiley, C. R. H. (2016). *Evaluating the content and quality of next generation high school assessments*. (HumRRO Report 2016 No. 001). Alexandria, VA: Human Resources Research Organization.

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). Alexandria, VA: Human Resources Research Organization.

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

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

Steedle, J., Quesen, S., & Boyd, A. (2017). *Longitudinal study of external Validity of the PARCC Performance Levels: Phase I Report.* Austin, TX: Pearson.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. Synthesis Report.

Wainer, H., & Thissen, D. (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

Wei, Y., & He, X. (2006). Conditional growth charts. *The 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,* 2–13.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

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*, 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). Hillsdale, NJ: Erlbaum

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items (ETS Research Report RR-97-05).* Princeton, NJ: Educational Testing Service.

# Appendices

## Appendix 6: Summary of Differential Item Function (DIF) Results

Table A.6.1 Pre-Administration Differential Item Functioning for ELA/L Grade 3

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 32 | | | | | 32 | 100 | | | | |
| White vs Black | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Hispanic | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Asian | 32 | | | | | 31 | 97 | 1 | 3 | | |
| White vs American Indian | 32 | | | | | 32 | 100 | | | | |
| White vs Pacific Islander | 32 | | | 1 | 3 | 31 | 97 | | | | |
| White vs Two or more races | 32 | | | | | 32 | 100 | | | | |
| Not vs Economically Disadvantaged | 32 | | | | | 32 | 100 | | | | |
| Non vs English Learners | 32 | 1 | 3 | 1 | 3 | 30 | 94 | | | | |
| Without vs Students with Disabilities | 32 | | | | | 32 | 100 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.2 Pre-Administration Differential Item Functioning for ELA/L Grade 4

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 51 | 1 | 2 | 1 | 2 | 44 | 86 | 4 | 8 | 1 | 2 |
| White vs Black | 51 | 2 | 4 | 2 | 4 | 47 | 92 | | | | |
| White vs Hispanic | 51 | | | 2 | 4 | 49 | 96 | | | | |
| White vs Asian | 51 | | | 1 | 2 | 50 | 98 | | | | |
| White vs American Indian | 51 | | | 1 | 2 | 50 | 98 | | | | |
| White vs Pacific Islander | 51 | | | | | 51 | 100 | | | | |
| White vs Two or more races | 51 | | | | | 51 | 100 | | | | |
| Not vs Economically Disadvantaged | 51 | | | | | 51 | 100 | | | | |
| Non vs English Learners | 51 | 2 | 4 | 4 | 8 | 45 | 88 | | | | |
| Without vs Students with Disabilities | 51 | | | 1 | 2 | 50 | 98 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.3 Pre-Administration Differential Item Functioning for ELA/L Grade 5

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 51 | 1 | 2 | 3 | 6 | 46 | 90 | 1 | 2 | | |
| White vs Black | 51 | | | 1 | 2 | 50 | 98 | | | | |
| White vs Hispanic | 51 | 1 | 2 | 2 | 4 | 48 | 94 | | | | |
| White vs Asian | 51 | | | 1 | 2 | 49 | 96 | 1 | 2 | | |
| White vs American Indian | 51 | | | | | 51 | 100 | | | | |
| White vs Pacific Islander | 51 | | | 1 | 2 | 50 | 98 | | | | |
| White vs Two or more races | 51 | | | | | 51 | 100 | | | | |
| Not vs Economically Disadvantaged | 51 | | | | | 51 | 100 | | | | |
| Non vs English Learners | 51 | 1 | 2 | 7 | 14 | 43 | 84 | | | | |
| Without vs Students with Disabilities | 51 | 1 | 2 | 4 | 8 | 46 | 90 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.4 Pre-Administration Differential Item Functioning for ELA/L Grade 6

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 45 | 1 | 2 | 3 | 7 | 40 | 89 | 1 | 2 | | |
| White vs Black | 45 | 1 | 2 | 1 | 2 | 43 | 96 | | | | |
| White vs Hispanic | 45 | 1 | 2 | 1 | 2 | 43 | 96 | | | | |
| White vs Asian | 45 | | | 1 | 2 | 44 | 98 | | | | |
| White vs American Indian | 45 | | | 3 | 7 | 41 | 91 | 1 | 2 | | |
| White vs Pacific Islander | 45 | | | 1 | 2 | 44 | 98 | | | | |
| White vs Two or more races | 45 | | | | | 45 | 100 | | | | |
| Not vs Economically Disadvantaged | 45 | | | 2 | 4 | 43 | 96 | | | | |
| Non vs English Learners | 45 | 1 | 2 | 6 | 13 | 38 | 84 | | | | |
| Without vs Students with Disabilities | 45 | | | 1 | 2 | 44 | 98 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.5 Pre-Administration Differential Item Functioning for ELA/L Grade 7

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 51 | | | 7 | 14 | 44 | 86 | | | | |
| White vs Black | 51 | | | 1 | 2 | 50 | 98 | | | | |
| White vs Hispanic | 51 | | | 3 | 6 | 48 | 94 | | | | |
| White vs Asian | 51 | | | | | 50 | 98 | 1 | 2 | | |
| White vs American Indian | 51 | | | 3 | 6 | 48 | 94 | | | | |
| White vs Pacific Islander | 51 | 1 | 2 | 2 | 4 | 48 | 94 | | | | |
| White vs Two or more races | 51 | | | | | 51 | 100 | | | | |
| Not vs Economically Disadvantaged | 51 | | | | | 51 | 100 | | | | |
| Non vs English Learners | 51 | 1 | 2 | 6 | 12 | 44 | 86 | | | | |
| Without vs Students with Disabilities | 51 | 1 | 2 | 1 | 2 | 49 | 96 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.6 Pre-Administration Differential Item Functioning for ELA/L Grade 8

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 51 | 1 | 2 | 2 | 4 | 47 | 92 | 1 | 2 | | |
| White vs Black | 51 | | | | | 51 | 100 | | | | |
| White vs Hispanic | 51 | | | 3 | 6 | 48 | 94 | | | | |
| White vs Asian | 51 | | | | | 50 | 98 | | | 1 | 2 |
| White vs American Indian | 51 | | | | | 51 | 100 | | | | |
| White vs Pacific Islander | 51 | | | | | 51 | 100 | | | | |
| White vs Two or more races | 51 | | | | | 51 | 100 | | | | |
| Not vs Economically Disadvantaged | 51 | | | 2 | 4 | 49 | 96 | | | | |
| Non vs English Learners | 51 | 2 | 4 | 6 | 12 | 43 | 84 | | | | |
| Without vs Students with Disabilities | 51 | | | 1 | 2 | 50 | 98 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.7 Pre-Administration Differential Item Functioning for Mathematics Grade 3

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 87 | | | 2 | 2 | 83 | 95 | 2 | 2 | | |
| White vs Black | 87 | 1 | 1 | 7 | 8 | 78 | 90 | 1 | 1 | | |
| White vs Hispanic | 87 | | | 1 | 1 | 85 | 98 | 1 | 1 | | |
| White vs Asian | 87 | | | 1 | 1 | 82 | 94 | 4 | 5 | | |
| White vs American Indian | 87 | 1 | 1 | | | 86 | 99 | | | | |
| White vs Pacific Islander | 87 | | | 2 | 2 | 85 | 98 | | | | |
| White vs Two or more races | 87 | | | 1 | 1 | 85 | 98 | 1 | 1 | | |
| Not vs Economically Disadvantaged | 87 | | | 2 | 2 | 85 | 98 | | | | |
| Non vs English Learners | 87 | | | 2 | 2 | 84 | 97 | 1 | 1 | | |
| Without vs Students with Disabilities | 87 | | | 3 | 3 | 84 | 97 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.8 Pre-Administration Differential Item Functioning for Mathematics Grade 4

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 85 | | | 2 | 2 | 82 | 96 | 1 | 1 | | |
| White vs Black | 85 | | | 6 | 7 | 79 | 93 | | | | |
| White vs Hispanic | 85 | | | | | 85 | 100 | | | | |
| White vs Asian | 85 | | | | | 84 | 99 | 1 | 1 | | |
| White vs American Indian | 85 | | | 3 | 4 | 82 | 96 | | | | |
| White vs Pacific Islander | 85 | | | | | 84 | 99 | 1 | 1 | | |
| White vs Two or more races | 85 | | | 1 | 1 | 83 | 98 | 1 | 1 | | |
| Not vs Economically Disadvantaged | 85 | | | | | 85 | 100 | | | | |
| Non vs English Learners | 85 | | | 1 | 1 | 84 | 99 | | | | |
| Without vs Students with Disabilities | 85 | | | | | 85 | 100 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.9 Pre-Administration Differential Item Functioning for Mathematics Grade 5

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 84 | 1 | 1 | | | 83 | 99 | | | | |
| White vs Black | 84 | | | 3 | 4 | 80 | 95 | 1 | 1 | | |
| White vs Hispanic | 84 | | | | | 84 | 100 | | | | |
| White vs Asian | 84 | | | | | 73 | 87 | 11 | 13 | | |
| White vs American Indian | 84 | | | 3 | 4 | 81 | 96 | | | | |
| White vs Pacific Islander | 84 | | | 1 | 1 | 83 | 99 | | | | |
| White vs Two or more races | 84 | | | | | 84 | 100 | | | | |
| Not vs Economically Disadvantaged | 84 | | | | | 84 | 100 | | | | |
| Non vs English Learners | 84 | | | 2 | 2 | 82 | 98 | | | | |
| Without vs Students with Disabilities | 84 | | | 3 | 4 | 80 | 95 | 1 | 1 | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.10 Pre-Administration Differential Item Functioning for Mathematics Grade 6

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 80 | 1 | 1 | 3 | 4 | 76 | 95 | | | | |
| White vs Black | 80 | | | 6 | 8 | 74 | 93 | | | | |
| White vs Hispanic | 80 | | | 2 | 3 | 78 | 98 | | | | |
| White vs Asian | 80 | | | | | 70 | 88 | 8 | 10 | 2 | 3 |
| White vs American Indian | 80 | | | 1 | 1 | 78 | 98 | 1 | 1 | | |
| White vs Pacific Islander | 80 | | | 1 | 1 | 79 | 99 | | | | |
| White vs Two or more races | 80 | | | | | 80 | 100 | | | | |
| Not vs Economically Disadvantaged | 80 | | | | | 80 | 100 | | | | |
| Non vs English Learners | 80 | | | 4 | 5 | 75 | 94 | 1 | 1 | | |
| Without vs Students with Disabilities | 80 | | | 1 | 1 | 77 | 96 | 1 | 1 | 1 | 1 |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.11 Pre-Administration Differential Item Functioning for Mathematics Grade 7

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 86 | | | 5 | 6 | 79 | 92 | 2 | 2 | | |
| White vs Black | 86 | | | 4 | 5 | 81 | 94 | 1 | 1 | | |
| White vs Hispanic | 86 | | | 4 | 5 | 82 | 95 | | | | |
| White vs Asian | 86 | | | 2 | 2 | 76 | 88 | 4 | 5 | 4 | 5 |
| White vs American Indian | 86 | | | 1 | 1 | 85 | 99 | | | | |
| White vs Pacific Islander | 86 | | | 1 | 1 | 85 | 99 | | | | |
| White vs Two or more races | 86 | | | | | 86 | 100 | | | | |
| Not vs Economically Disadvantaged | 86 | | | | | 86 | 100 | | | | |
| Non vs English Learners | 86 | 2 | 2 | 3 | 3 | 78 | 91 | 3 | 3 | | |
| Without vs Students with Disabilities | 86 | | | 1 | 1 | 85 | 99 | | | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

Table A.6.12 Pre-Administration Differential Item Functioning for Mathematics Grade 8

| DIF Comparison | Total N of Unique Items | C- DIF | | B- DIF | | A DIF | | B+ DIF | | C+ DIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | % of Total | N | % of Total | N | % of Total | N | % of Total | N | % of Total |
| Female vs Male | 82 | | | 1 | 1 | 81 | 99 | | | | |
| White vs Black | 82 | | | 2 | 2 | 79 | 96 | 1 | 1 | | |
| White vs Hispanic | 82 | | | | | 82 | 100 | | | | |
| White vs Asian | 82 | | | 1 | 1 | 70 | 85 | 8 | 10 | 3 | 4 |
| White vs American Indian | 82 | | | 2 | 2 | 80 | 98 | | | | |
| White vs Pacific Islander | 82 | | | | | 82 | 100 | | | | |
| White vs Two or more races | 82 | | | 1 | 1 | 81 | 99 | | | | |
| Not vs Economically Disadvantaged | 82 | | | | | 82 | 100 | | | | |
| Non vs English Learners | 82 | 2 | 2 | 4 | 5 | 76 | 93 | | | | |
| Without vs Students with Disabilities | 82 | | | 1 | 1 | 80 | 98 | 1 | 1 | | |

Note: American Indian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Not = not economically disadvantaged, Non = not an English learner, Without = not student with disability.

# Appendix 7.1: Pre-Equated IRT Results for Spring 2023 English Language Arts/Literacy (ELA/L)

Table A.7.1 Pre-Equated IRT Summary Parameter Estimates for All Items for ELA/L by Grade

| | | | | *b* Estimates Summary | | | | *a* Estimates Summary | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **Item Grouping** | **No. of Score Points** | **No. of Items** | **Mean** | **SD** | **Min** | **Max** | **Mean** | **SD** | **Min** | **Max** |
| 3 | All Items | 80 | 36 | 0.36 | 0.92 | -2.12 | 1.69 | 0.59 | 0.22 | 0.25 | 1.04 |
| | Reading | 56 | 28 | 0.05 | 0.81 | -2.12 | 1.63 | 0.52 | 0.18 | 0.25 | 0.85 |
| | Writing | 24 | 8 | 1.42 | 0.21 | 0.99 | 1.69 | 0.87 | 0.14 | 0.59 | 1.04 |
| 4 | All Items | 125 | 56 | 0.49 | 1.00 | -1.30 | 2.66 | 0.48 | 0.24 | 0.18 | 1.06 |
| | Reading | 92 | 46 | 0.31 | 1.01 | -1.30 | 2.66 | 0.39 | 0.13 | 0.18 | 0.78 |
| | Writing | 33 | 10 | 1.31 | 0.40 | 0.88 | 1.90 | 0.91 | 0.11 | 0.71 | 1.06 |
| 5 | All Items | 126 | 56 | 0.42 | 1.07 | -1.65 | 3.59 | 0.51 | 0.24 | 0.13 | 1.02 |
| | Reading | 92 | 46 | 0.22 | 1.06 | -1.65 | 3.59 | 0.42 | 0.16 | 0.13 | 0.85 |
| | Writing | 34 | 10 | 1.34 | 0.51 | 0.60 | 2.22 | 0.91 | 0.08 | 0.74 | 1.02 |
| 6 | All Items | 109 | 49 | 0.39 | 0.74 | -1.02 | 1.86 | 0.50 | 0.23 | 0.24 | 1.16 |
| | Reading | 82 | 41 | 0.24 | 0.70 | -1.02 | 1.70 | 0.41 | 0.11 | 0.24 | 0.73 |
| | Writing | 27 | 8 | 1.14 | 0.41 | 0.53 | 1.86 | 0.96 | 0.13 | 0.79 | 1.16 |
| 7 | All Items | 125 | 56 | 0.36 | 0.70 | -1.47 | 1.60 | 0.50 | 0.28 | 0.13 | 1.30 |
| | Reading | 92 | 46 | 0.28 | 0.73 | -1.47 | 1.60 | 0.40 | 0.15 | 0.13 | 0.75 |
| | Writing | 33 | 10 | 0.71 | 0.36 | 0.19 | 1.32 | 1.00 | 0.20 | 0.67 | 1.30 |
| 8 | All Items | 125 | 56 | 0.15 | 0.92 | -2.98 | 2.38 | 0.49 | 0.25 | 0.08 | 1.14 |
| | Reading | 92 | 46 | 0.00 | 0.94 | -2.98 | 2.38 | 0.39 | 0.11 | 0.08 | 0.64 |
| | Writing | 33 | 10 | 0.83 | 0.35 | 0.39 | 1.36 | 0.97 | 0.16 | 0.68 | 1.14 |

# Appendix 7.2: Pre-Equated IRT Results for Spring 2023 Mathematics

Table A.7.2 Pre-Equated IRT Summary Parameter Estimates for All Items for Mathematics by Grade

| Grade | Item Grouping | No. of Score Points | No. of Items | b Estimates Summary | | | | a Estimates Summary | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | Mean | SD | Min | Max |
| 3 | All Items | 139 | 87 | -0.24 | 1.24 | -2.56 | 3.68 | 0.77 | 0.28 | 0.19 | 1.42 |
| | SSMC | 26 | 26 | -0.82 | 1.38 | -2.47 | 3.68 | 0.76 | 0.28 | 0.19 | 1.16 |
| | CR | 113 | 61 | 0.00 | 1.10 | -2.56 | 3.04 | 0.77 | 0.28 | 0.25 | 1.42 |
| | Type I | 79 | 71 | -0.45 | 1.26 | -2.56 | 3.68 | 0.82 | 0.27 | 0.19 | 1.42 |
| | Type II | 27 | 8 | 0.56 | 0.59 | -0.29 | 1.49 | 0.49 | 0.09 | 0.38 | 0.63 |
| | Type III | 33 | 8 | 0.85 | 0.27 | 0.47 | 1.28 | 0.56 | 0.08 | 0.46 | 0.71 |
| 4 | All Items | 143 | 85 | -0.05 | 1.09 | -2.65 | 2.36 | 0.71 | 0.23 | 0.31 | 1.46 |
| | SSMC | 21 | 21 | -1.07 | 1.00 | -2.65 | 1.13 | 0.67 | 0.22 | 0.31 | 1.13 |
| | CR | 122 | 64 | 0.28 | 0.90 | -1.93 | 2.36 | 0.73 | 0.23 | 0.42 | 1.46 |
| | Type I | 83 | 69 | -0.25 | 1.10 | -2.65 | 2.36 | 0.73 | 0.24 | 0.31 | 1.46 |
| | Type II | 27 | 8 | 0.78 | 0.33 | 0.20 | 1.15 | 0.63 | 0.10 | 0.42 | 0.75 |
| | Type III | 33 | 8 | 0.79 | 0.52 | 0.20 | 1.65 | 0.64 | 0.18 | 0.42 | 0.92 |
| 5 | All Items | 144 | 84 | 0.01 | 1.15 | -2.34 | 2.13 | 0.69 | 0.25 | 0.18 | 1.50 |
| | SSMC | 24 | 24 | -0.77 | 1.06 | -2.34 | 2.13 | 0.68 | 0.31 | 0.18 | 1.50 |
| | CR | 120 | 60 | 0.32 | 1.04 | -2.16 | 2.09 | 0.70 | 0.23 | 0.31 | 1.24 |
| | Type I | 81 | 67 | -0.23 | 1.15 | -2.34 | 2.13 | 0.73 | 0.27 | 0.18 | 1.50 |
| | Type II | 30 | 9 | 0.97 | 0.37 | 0.28 | 1.42 | 0.56 | 0.12 | 0.47 | 0.81 |
| | Type III | 33 | 8 | 0.92 | 0.68 | -0.32 | 1.56 | 0.56 | 0.11 | 0.46 | 0.75 |
| 6 | All Items | 142 | 80 | 0.45 | 1.12 | -3.57 | 4.46 | 0.71 | 0.30 | 0.16 | 1.54 |
| | SSMC | 24 | 24 | -0.18 | 1.19 | -3.57 | 2.10 | 0.58 | 0.24 | 0.24 | 1.09 |
| | CR | 118 | 56 | 0.73 | 0.98 | -1.66 | 4.46 | 0.77 | 0.31 | 0.16 | 1.54 |
| | Type I | 85 | 64 | 0.32 | 1.17 | -3.57 | 4.46 | 0.73 | 0.33 | 0.16 | 1.54 |
| | Type II | 27 | 8 | 0.65 | 0.70 | -0.37 | 1.73 | 0.62 | 0.16 | 0.44 | 0.89 |
| | Type III | 30 | 8 | 1.35 | 0.53 | 0.49 | 1.93 | 0.64 | 0.18 | 0.49 | 1.03 |
| 7 | All Items | 144 | 86 | 0.61 | 1.08 | -2.23 | 3.42 | 0.73 | 0.33 | 0.19 | 1.72 |
| | SSMC | 31 | 31 | -0.05 | 1.16 | -2.23 | 2.57 | 0.58 | 0.31 | 0.19 | 1.27 |
| | CR | 113 | 55 | 0.99 | 0.83 | -1.18 | 3.42 | 0.82 | 0.32 | 0.34 | 1.72 |
| | Type I | 84 | 70 | 0.52 | 1.14 | -2.23 | 3.42 | 0.76 | 0.36 | 0.19 | 1.72 |
| | Type II | 27 | 8 | 1.04 | 0.76 | -0.45 | 2.07 | 0.64 | 0.08 | 0.56 | 0.80 |
| | Type III | 33 | 8 | 1.02 | 0.46 | 0.39 | 1.85 | 0.55 | 0.11 | 0.35 | 0.69 |
| 8 | All Items | 139 | 82 | 0.95 | 1.00 | -1.70 | 2.70 | 0.60 | 0.28 | 0.10 | 1.44 |
| | SSMC | 26 | 26 | 0.38 | 1.07 | -1.70 | 2.03 | 0.42 | 0.17 | 0.16 | 0.76 |
| | CR | 113 | 56 | 1.21 | 0.85 | -0.71 | 2.70 | 0.68 | 0.29 | 0.10 | 1.44 |
| | Type I | 85 | 67 | 0.77 | 0.98 | -1.70 | 2.53 | 0.60 | 0.30 | 0.10 | 1.44 |
| | Type II | 24 | 7 | 1.83 | 0.63 | 1.08 | 2.70 | 0.63 | 0.18 | 0.46 | 0.91 |
| | Type III | 30 | 8 | 1.70 | 0.68 | 0.60 | 2.63 | 0.56 | 0.18 | 0.35 | 0.87 |

# Appendix 11: Students by Grade/Subject and Mode

Table A.11.1 Number of ELA/L Test Takers, by Grade, Mode, and Gender

| | | | Female | | Male | |
|---|---|---|---|---|---|---|
| **Grade** | **Mode** | **Valid Cases** | **N** | **%** | **N** | **%** |
| 3 | **All** | 128,356 | 62,746 | 48.9 | 65,604 | 51.1 |
| | CBT | 127,540 | 62,423 | 48.6 | 65,117 | 50.7 |
| | PBT | 810 | 323 | 0.3 | 487 | 0.4 |
| 4 | **All** | 127,980 | 62,671 | 49.0 | 65,298 | 51.0 |
| | CBT | 127,415 | 62,412 | 48.8 | 65,003 | 50.8 |
| | PBT | 554 | 259 | 0.2 | 295 | 0.2 |
| 5 | **All** | 129,738 | 63,606 | 49.0 | 66,113 | 51.0 |
| | CBT | 129,159 | 63,346 | 48.8 | 65,813 | 50.7 |
| | PBT | 560 | 260 | 0.2 | 300 | 0.2 |
| 6 | **All** | 133,179 | 65,132 | 48.9 | 68,026 | 51.1 |
| | CBT | 132,637 | 64,895 | 48.7 | 67,742 | 50.9 |
| | PBT | 521 | 237 | 0.2 | 284 | 0.2 |
| 7 | **All** | 134,267 | 65,120 | 48.5 | 69,106 | 51.5 |
| | CBT | 133,672 | 64,873 | 48.3 | 68,799 | 51.3 |
| | PBT | 554 | 247 | 0.2 | 307 | 0.2 |
| 8 | **All** | 138,908 | 67,706 | 48.8 | 71,144 | 51.2 |
| | CBT | 138,354 | 67,479 | 48.6 | 70,875 | 51.0 |
| | PBT | 496 | 227 | 0.2 | 269 | 0.2 |

Note: CBT=computer-based test; PBT=paper-based test; n/a=not applicable. and n/r=not reported due to n<20

Table A.11.2 Number of Mathematics Test Takers, by Grade, Mode, and Gender

| | | | Female | | Male | |
|---|---|---|---|---|---|---|
| **Grade** | **Mode** | **Valid Cases** | **N** | **%** | **N** | **%** |
| | **All** | 128,109 | 62,609 | 48.9 | 65,494 | 51.1 |
| 3 | CBT | 127,426 | 62,321 | 48.6 | 65,105 | 50.8 |
| | PBT | 677 | 288 | 0.2 | 389 | 0.3 |
| | **All** | 127,833 | 62,608 | 49.0 | 65,214 | 51.0 |
| 4 | CBT | 127,249 | 62,339 | 48.8 | 64,910 | 50.8 |
| | PBT | 573 | 269 | 0.2 | 304 | 0.2 |
| | **All** | 129,562 | 63,522 | 49.0 | 66,021 | 51.0 |
| 5 | CBT | 128,967 | 63,250 | 48.8 | 65,717 | 50.7 |
| | PBT | 576 | 272 | 0.2 | 304 | 0.2 |
| | **All** | 132,858 | 64,954 | 48.9 | 67,883 | 51.1 |
| 6 | CBT | 132,317 | 64,712 | 48.7 | 67,605 | 50.9 |
| | PBT | 520 | 242 | 0.2 | 278 | 0.2 |
| | **All** | 133,956 | 64,986 | 48.5 | 68,931 | 51.5 |
| 7 | CBT | 133,339 | 64,725 | 48.3 | 68,614 | 51.2 |
| | PBT | 578 | 261 | 0.2 | 317 | 0.2 |
| | **All** | 138,558 | 67,515 | 48.7 | 70,985 | 51.3 |
| 8 | CBT | 137,986 | 67,273 | 48.6 | 70,713 | 51.1 |
| | PBT | 514 | 242 | 0.2 | 272 | 0.2 |

Note: n/a=not applicable. and n/r=not reported due to n<20.

Table A.11.3 Number of Spanish-Language Mathematics Test Takers, by Grade, Mode, and Gender

| | | | Female | | Male | |
|---|---|---|---|---|---|---|
| Grade | Mode | Valid Cases | N | % | N | % |
| | **All** | 3,541 | 1,792 | 50.6 | 1,748 | 49.4 |
| 3 | CBT | 3,520 | 1,782 | 50.3 | 1,738 | 49.1 |
| | PBT | 20 | n/r | n/r | n/r | n/r |
| | **All** | 2,588 | 1,297 | 50.1 | 1,291 | 49.9 |
| 4 | CBT | 2,572 | 1,288 | 49.8 | 1,284 | 49.6 |
| | PBT | n/r | n/r | n/r | n/r | n/r |
| | **All** | 2,321 | 1,176 | 50.7 | 1,145 | 49.3 |
| 5 | CBT | 2,303 | 1,168 | 50.3 | 1,135 | 48.9 |
| | PBT | n/r | n/r | n/r | n/r | n/r |
| | **All** | 1,985 | 976 | 49.2 | 1,009 | 50.8 |
| 6 | CBT | 1,973 | 966 | 48.7 | 1,007 | 50.7 |
| | PBT | n/r | n/r | n/r | n/r | n/r |
| | **All** | 1,485 | 697 | 46.9 | 788 | 53.1 |
| 7 | CBT | 1,451 | 678 | 45.7 | 773 | 52.1 |
| | PBT | 34 | n/r | n/r | n/r | n/r |
| | **All** | 1,489 | 736 | 49.4 | 753 | 50.6 |
| 8 | CBT | 1,460 | 719 | 48.3 | 741 | 49.8 |
| | PBT | 29 | n/r | n/r | n/r | n/r |

Note: CBT=computer-based test; PBT=paper-based test; n/a=not applicable. and n/r=not reported due to n<20.

Table A.11.4 Percentage of Demographics for ELA/L by Grade

| Demographic | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 |
|---|---|---|---|---|---|---|
| Economically Disadvantaged | 50.7 | 50.4 | 50.2 | 49.9 | 49.7 | 49.8 |
| Students with Disabilities | 19.1 | 19.7 | 20.1 | 19.9 | 19.8 | 19.6 |
| English Learner | 19.6 | 19.1 | 16.5 | 14.6 | 14.1 | 13.2 |
| Male | 51.1 | 51.0 | 51.0 | 51.1 | 51.5 | 51.2 |
| Female | 48.9 | 49.0 | 49.0 | 48.9 | 48.5 | 48.7 |
| American Indian/Alaska Native | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| Asian | 5.6 | 5.8 | 5.7 | 5.5 | 5.5 | 5.4 |
| Black/African American | 16.5 | 16.3 | 16.2 | 16.2 | 16.5 | 16.2 |
| Hispanic/Latino | 26.6 | 26.6 | 27.3 | 27.6 | 27.7 | 28.5 |
| White | 45.7 | 45.9 | 45.5 | 45.6 | 45.5 | 45.1 |
| Native Hawaiian/Pacific Islander | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Two or more races | 4.4 | 4.3 | 4.2 | 4.0 | 3.9 | 3.8 |
| Unknown | 0.8 | 0.8 | 0.7 | 0.6 | 0.6 | 0.6 |

Note: n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.5 Percentage of Demographics for Mathematics by Grade

| Demographic | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 |
|---|---|---|---|---|---|---|
| Economically Disadvantaged | 50.7 | 50.4 | 50.1 | 49.8 | 49.6 | 49.7 |
| Students with Disabilities | 19.1 | 19.7 | 20.1 | 19.8 | 19.8 | 19.5 |
| English Learner | 19.5 | 19.1 | 16.5 | 14.6 | 14.1 | 13.2 |
| Male | 51.1 | 51.0 | 51.0 | 51.1 | 51.5 | 51.2 |
| Female | 48.9 | 49.0 | 49.0 | 48.9 | 48.5 | 48.7 |
| American Indian/Alaska Native | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| Asian | 5.6 | 5.8 | 5.7 | 5.6 | 5.5 | 5.4 |
| Black/African American | 16.5 | 16.3 | 16.2 | 16.2 | 16.4 | 16.2 |
| Hispanic/Latino | 26.6 | 26.6 | 27.3 | 27.6 | 27.7 | 28.5 |
| White | 45.8 | 45.9 | 45.5 | 45.7 | 45.6 | 45.1 |
| Native Hawaiian/Pacific Islander | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Two or more races | 4.4 | 4.3 | 4.2 | 4.0 | 3.9 | 3.8 |
| Unknown | 0.8 | 0.8 | 0.7 | 0.6 | 0.6 | 0.6 |

Note: n/a=not applicable; and n/r=not reported due to n<20.

# Appendix 12.1: Form Composition

Table A.12.1 Form Composition for ELA/L Grade 3

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 4—7 | 8—17 |
| | Reading Informational Text | 4—7 | 11—20 |
| | Vocabulary | 4—5 | 8—10 |
| | Claim Total | 12—14 | 30—31 |
| Writing | | | |
| | Written Expression | 1 | 18 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 24 |
| Summative Total | | 14—16 | 54—55 |

Note: This table is identical to Table 12.1 in Section 12.

Table A.12.2 Form Composition for ELA/L Grade 4

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 5—8 | 14—20 |
| | Reading Informational Text | 5—9 | 18—22 |
| | Vocabulary | 4—7 | 8—14 |
| | Claim Total | 18 | 40—44 |
| Writing | | | |
| | Written Expression | 1 | 21—24 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 27—30 |
| Summative Total | | 20 | 67—74 |

Table A.12.3 Form Composition for ELA/L Grade 5

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 5—8 | 14—20 |
| | Reading Informational Text | 5—9 | 14—22 |
| | Vocabulary | 4—7 | 8—14 |
| | Claim Total | 18 | 40—44 |
| Writing | | | |
| | Written Expression | 1 | 21—24 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 27—30 |
| Summative Total | | 20 | 67—74 |

Table A.12.4 Form Composition for ELA/L Grade 6

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 5—9 | 14—22 |
| | Reading Informational Text | 5—11 | 14—26 |
| | Vocabulary | 4—7 | 8—14 |
| | Claim Total | 18 | 40—44 |
| Writing | | | |
| | Written Expression | 1 | 24 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 30 |
| Summative Total | | 20 | 70—74 |

Table A.12.5 Form Composition for ELA/L Grade 7

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 5—9 | 14—22 |
| | Reading Informational Text | 5—11 | 14—26 |
| | Vocabulary | 4—7 | 8—14 |
| | Claim Total | 18 | 40—44 |
| Writing | | | |
| | Written Expression | 1 | 24 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 30 |
| Summative Total | | 20 | 70—74 |

Table A.12.6 Form Composition for ELA/L Grade 8

| Claims | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| Reading | | | |
| | Reading Literary Text | 5—9 | 14—22 |
| | Reading Informational Text | 5—11 | 14—26 |
| | Vocabulary | 4—7 | 8—14 |
| | Claim Total | 18 | 40—44 |
| Writing | | | |
| | Written Expression | 1 | 24 |
| | Knowledge of Conventions | 1 | 6 |
| | Claim Total | 2 | 30 |
| Summative Total | | 20 | 70—74 |

Table A.12.9 Form Composition for Mathematics Grade 3

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 18 | 20 |
| | Additional & Supporting Content | 9 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 33 | 52 |

Note: This table is identical to Table 12.3 in Section 12.

Table A.12.10 Form Composition for Mathematics Grade 4

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 17 | 21 |
| | Additional & Supporting Content | 8 | 9 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 31 | 52 |

Table A.12.11 Form Composition for Mathematics Grade 5

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 17 | 20 |
| | Additional & Supporting Content | 8 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 31 | 52 |

Table A.12.12 Form Composition for Mathematics Grade 6

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 15 | 20 |
| | Additional & Supporting Content | 8 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 29 | 52 |

Table A.12.13 Form Composition for Mathematics Grade 7

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 18 | 20 |
| | Additional & Supporting Content | 7 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 31 | 52 |

Table A.12.14 Form Composition for Mathematics Grade 8

| | Subclaims | Number of Items | Number of Points |
|---|---|---|---|
| | Major Content | 18 | 20 |
| | Additional & Supporting Content | 6 | 10 |
| | Expressing Mathematical Reasoning | 3 | 10 |
| | Modeling & Applications | 3 | 12 |
| Summative Total | | 30 | 52 |

# Appendix 12.2: Threshold Scores and Scaling Constants

Table A.12.18 Threshold Scores and Scaling Constants for ELA/L Grades 3 to 8

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

Table A.12.19 Threshold Scores and Scaling Constants for Mathematics Grades 3 to 8

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

Table A.12.22 Scaling Constants for Reading and Writing Grades 3 to 10

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

# Appendix 12.3: IRT Test Characteristic Curves, Information Curves, and CSEM Curves

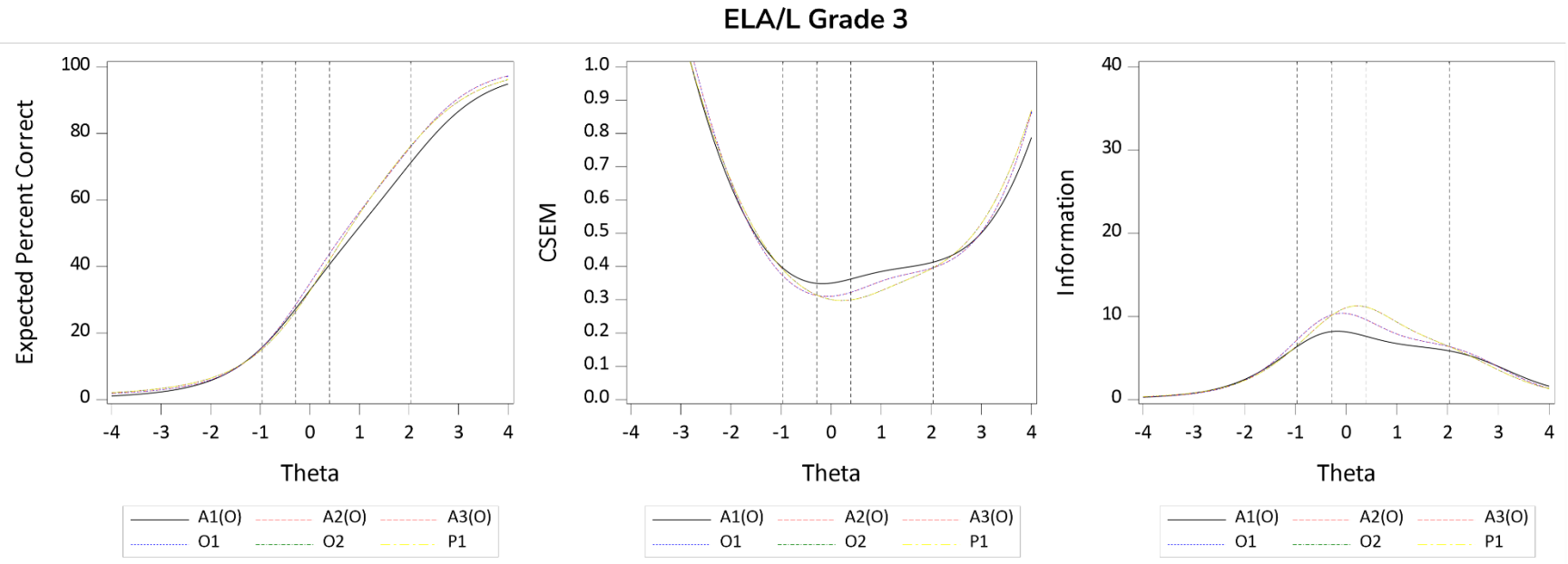



Figure A.12.1 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 3

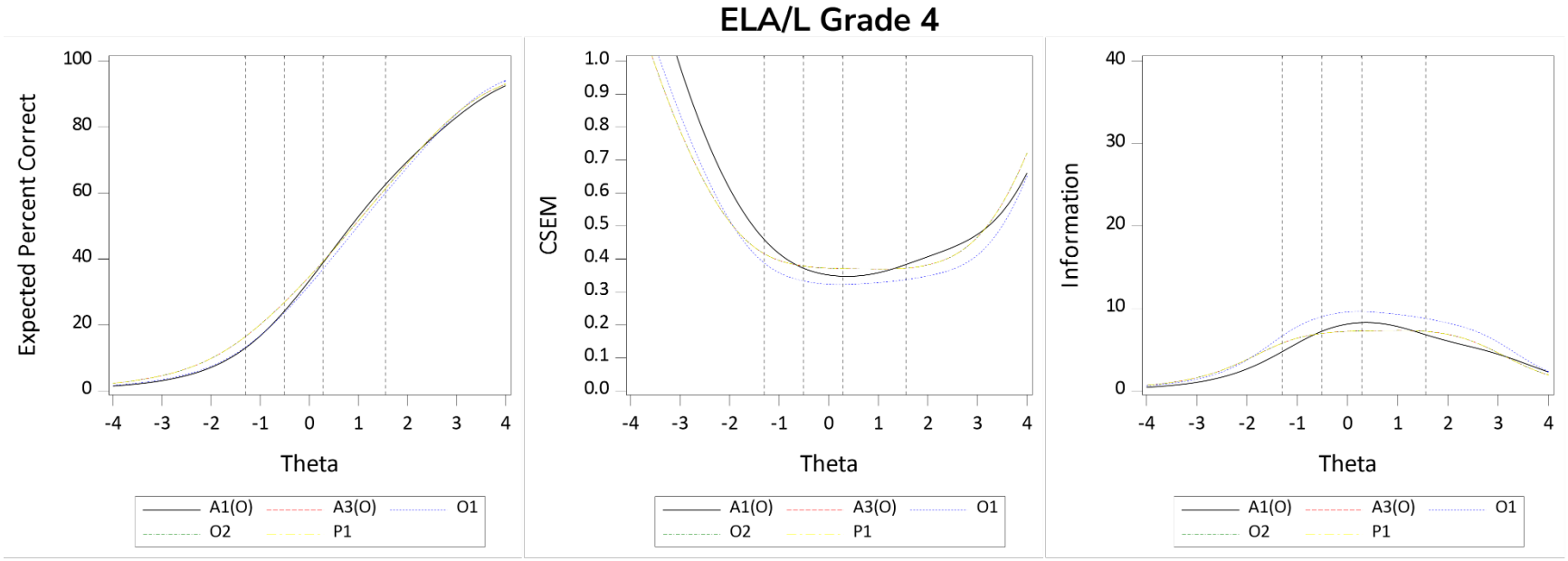



Figure A.12.2 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 4

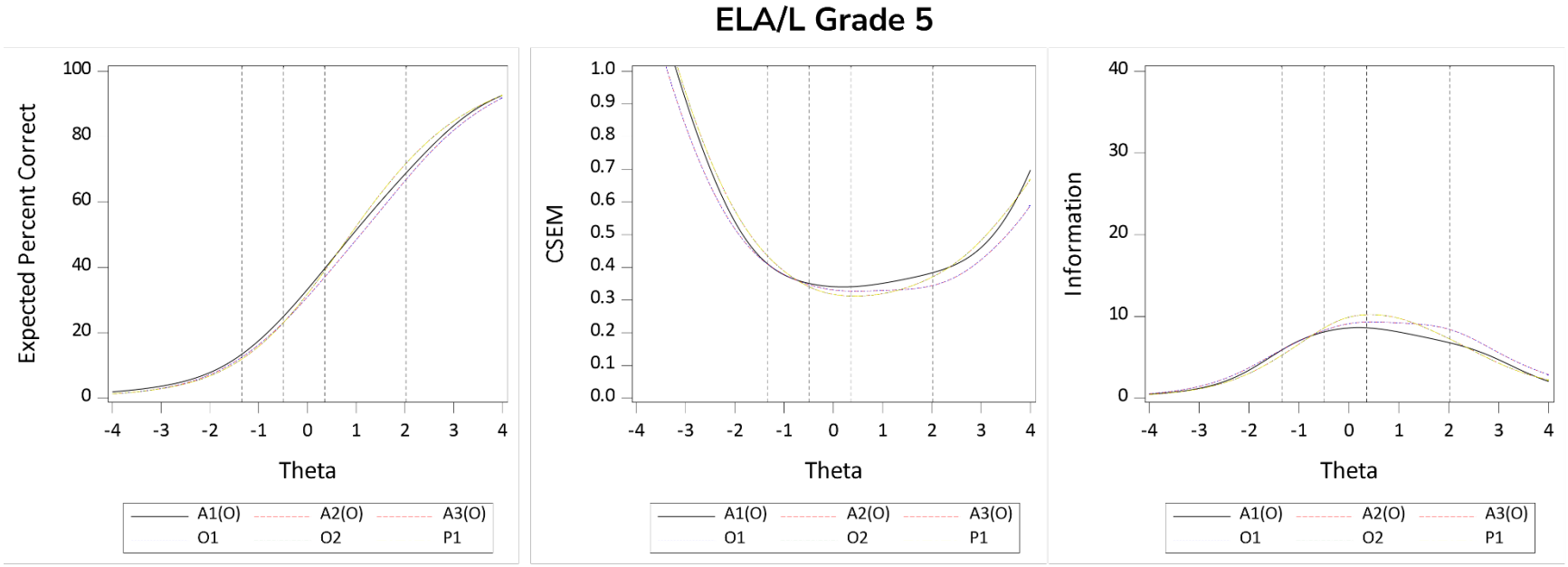



Figure A.12.3 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 5

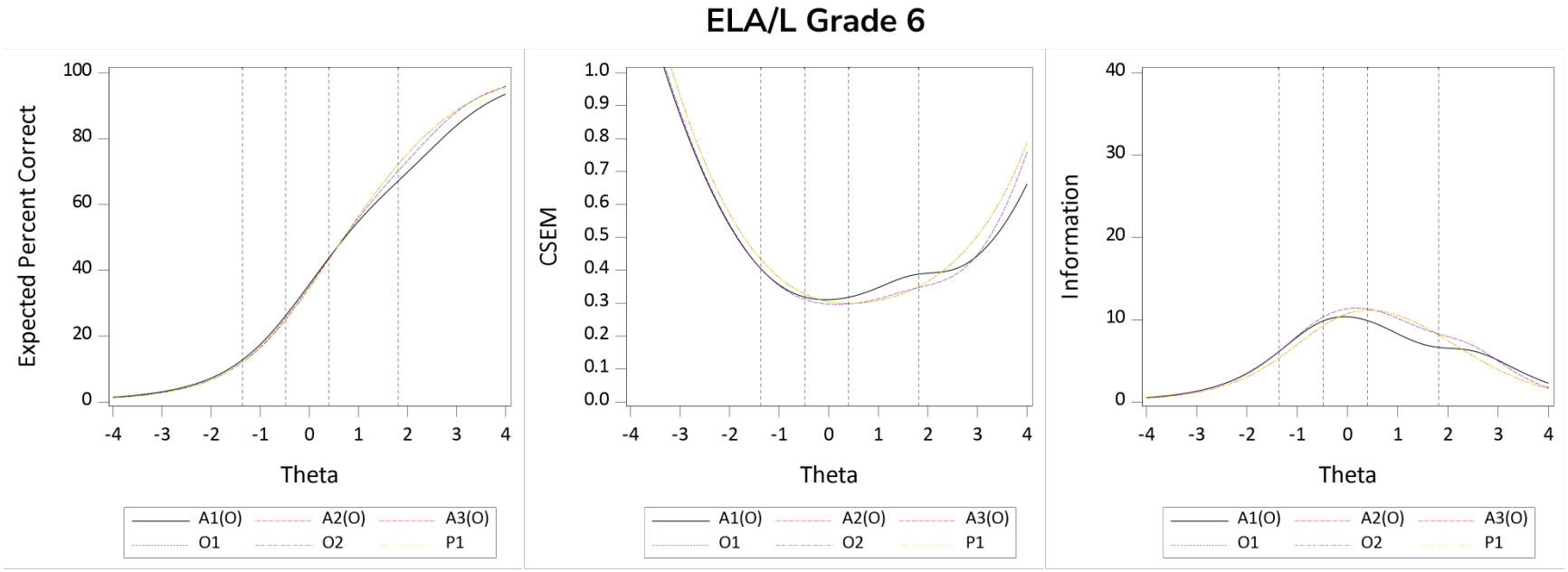



Figure A.12.4 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 6

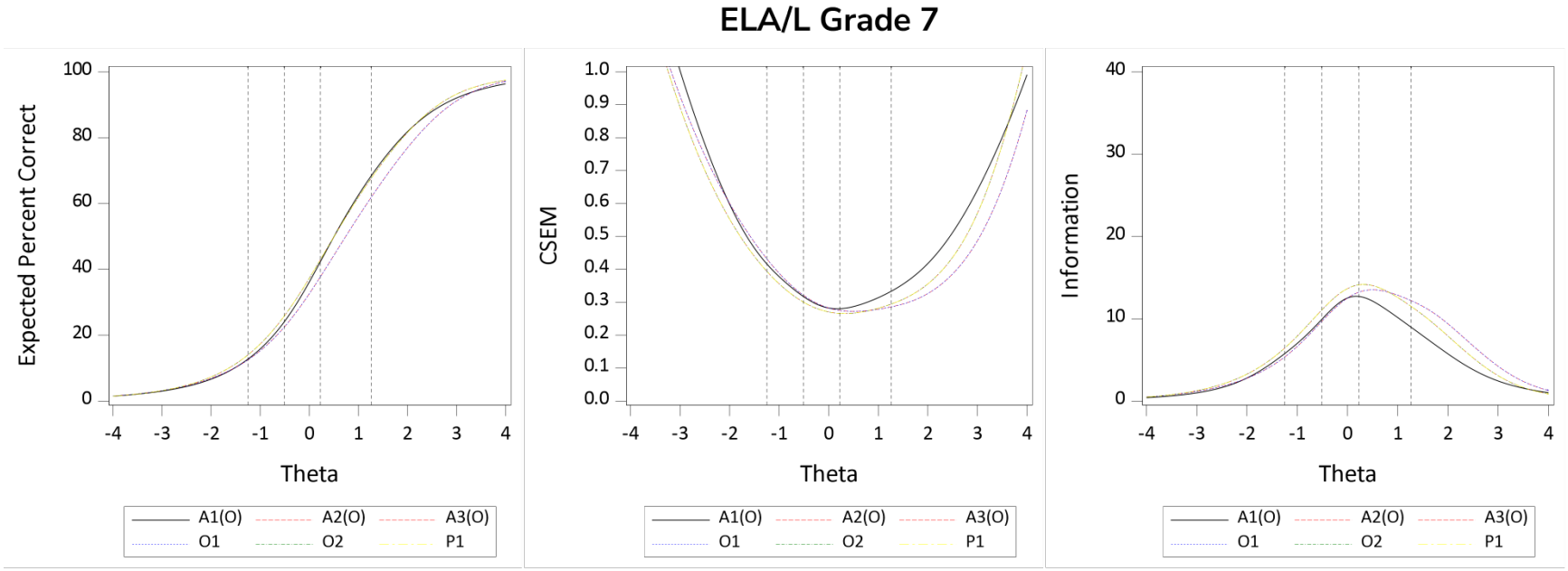



Figure A.12.5 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 7

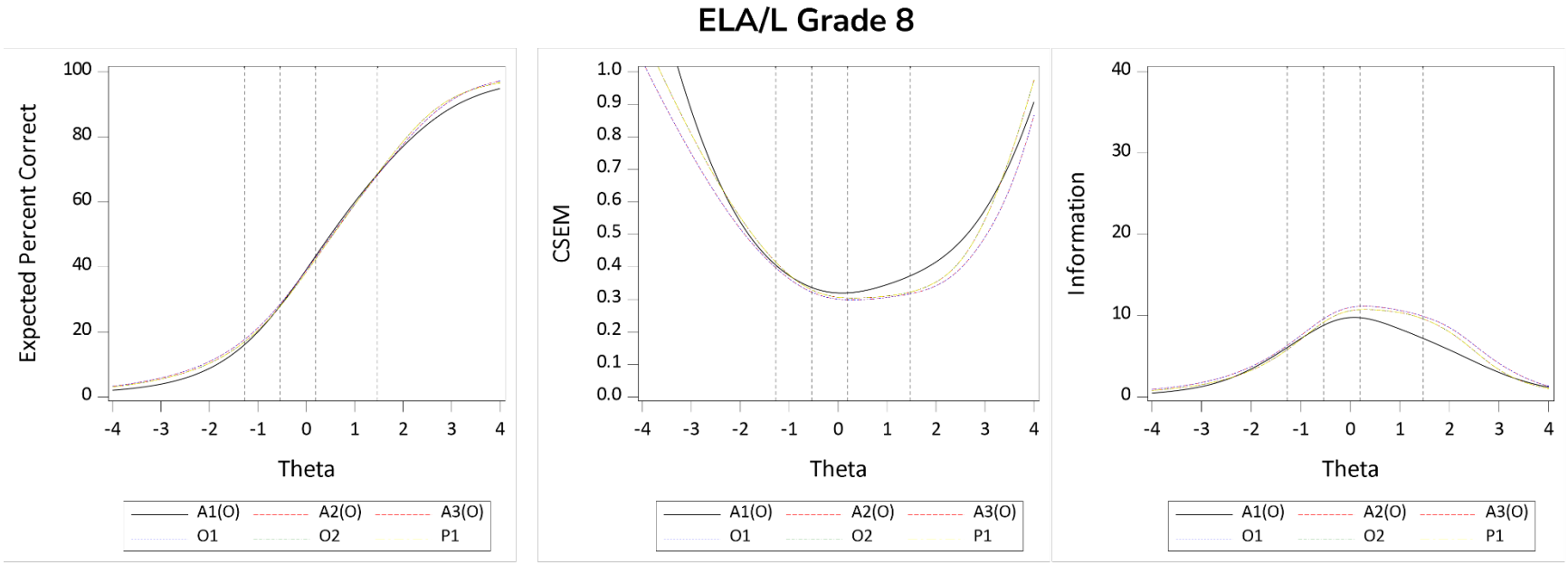



Figure A.12.6 Pre-Equated TCC, CSEM, and TIC for ELA/L Grade 8

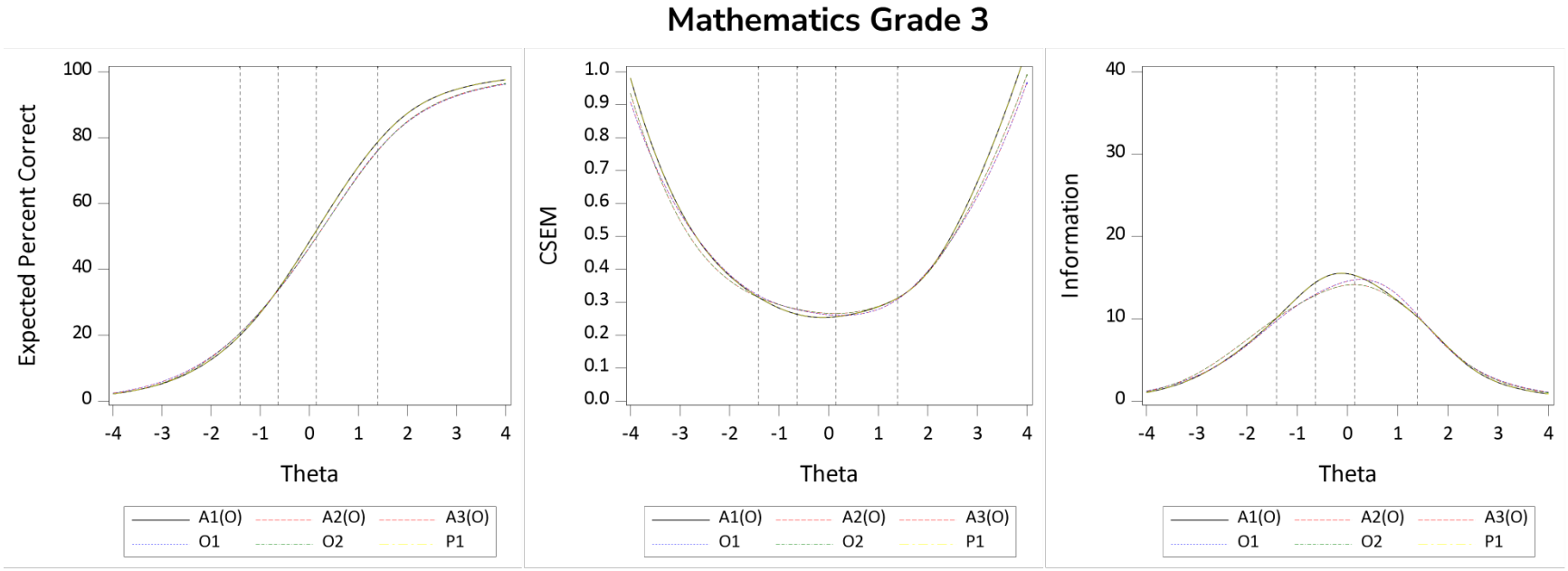



Figure A.12.7 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 3

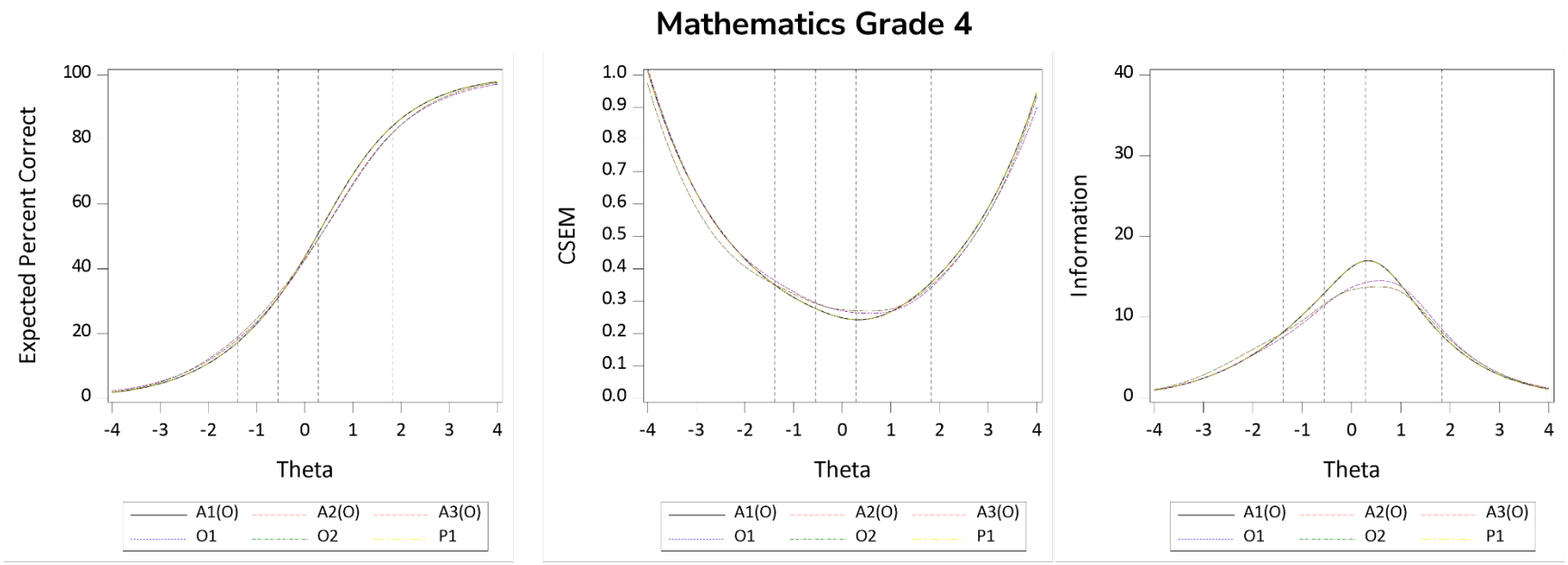



Figure A.12.8 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 4

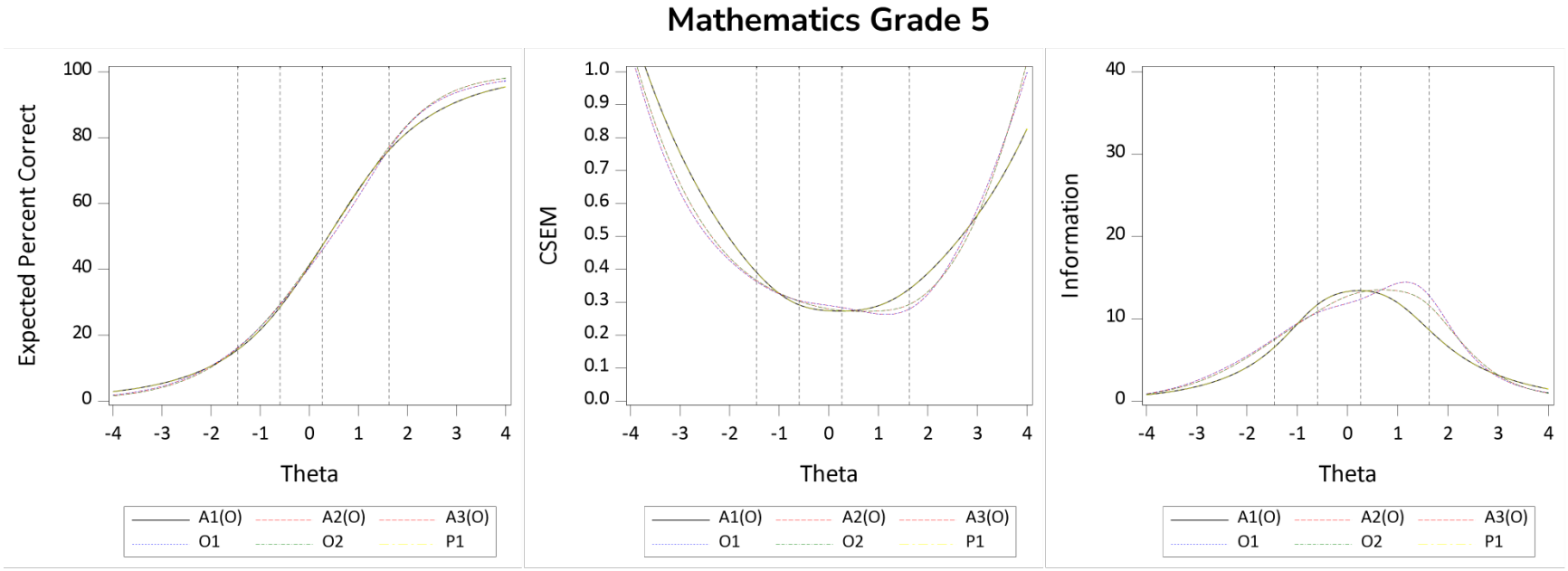



Figure A.12.9 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 5

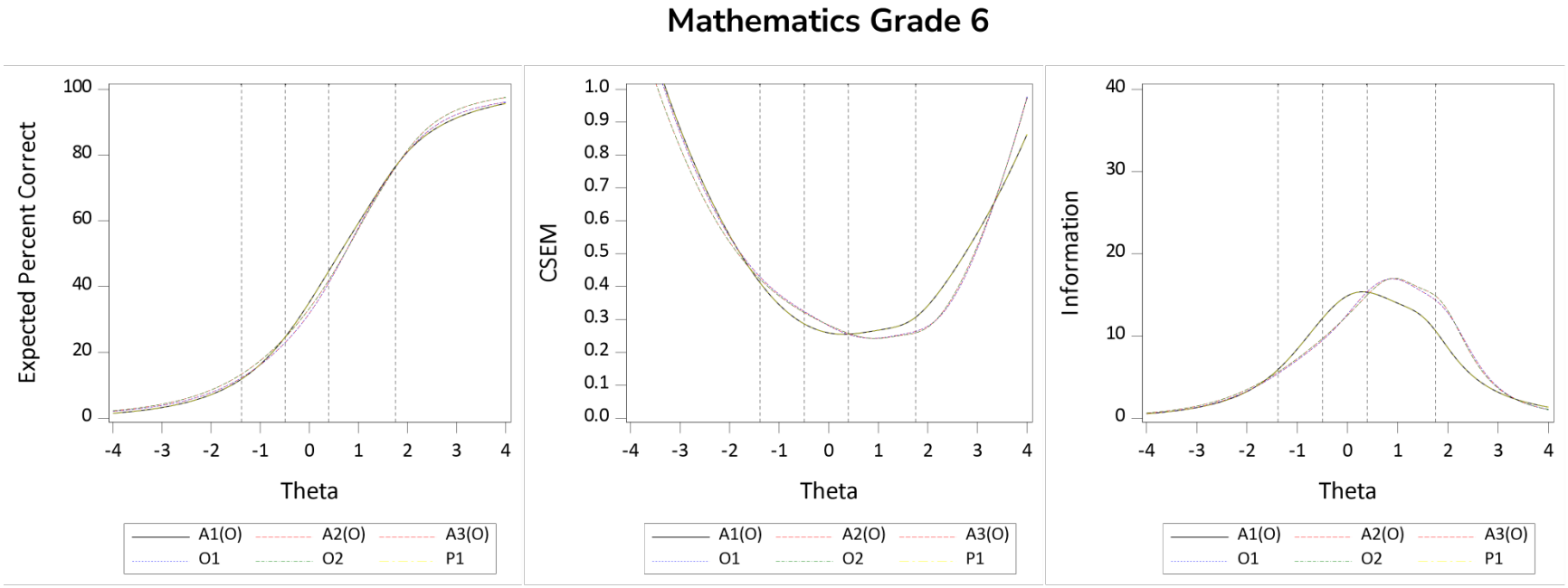



Figure A.12.10 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 6

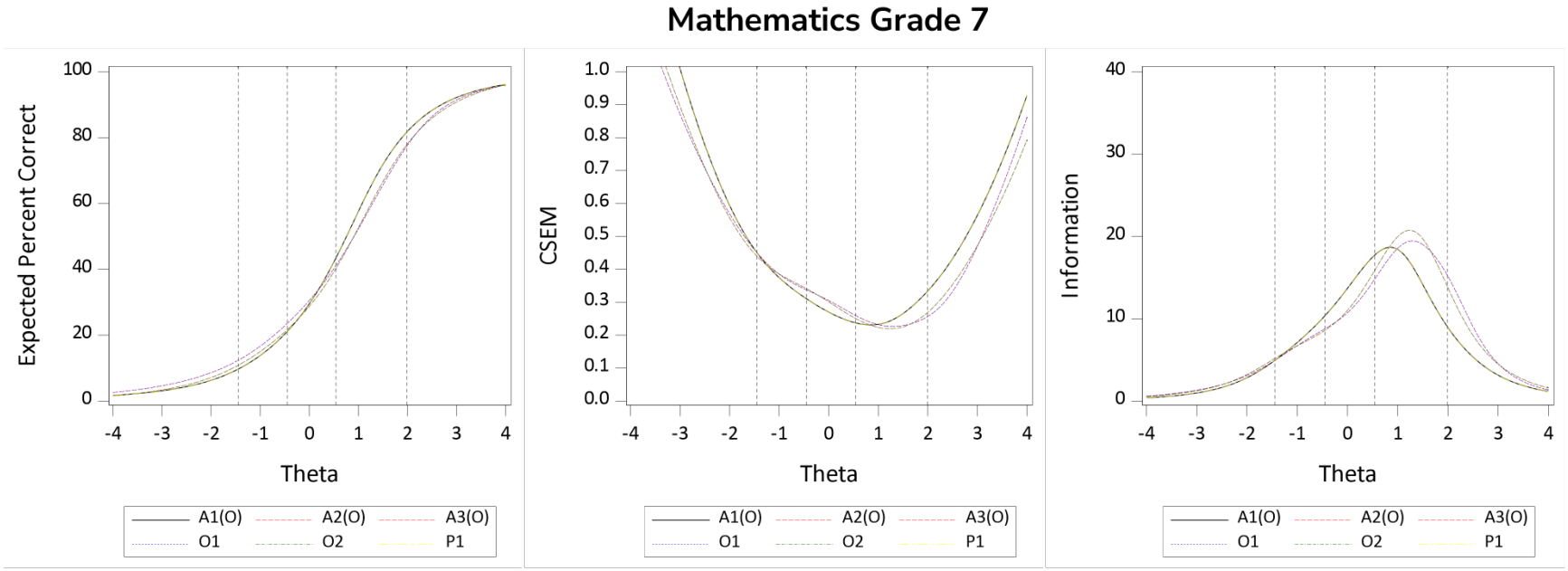



Figure A.12.11 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 7

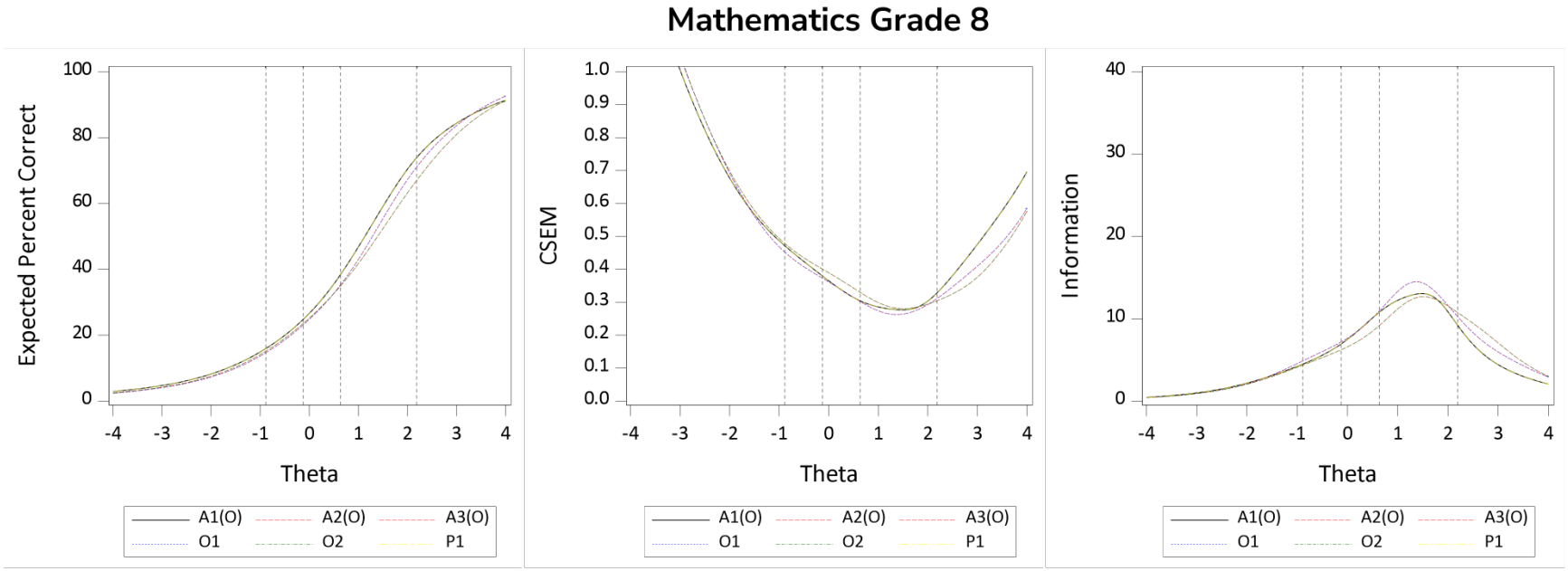



Figure A.12.12 Pre-Equated TCC, CSEM, and TIC for Mathematics Grade 8

# Appendix 12.4: Scale Score Cumulative Frequencies

Table A.12.23 Scale Score Cumulative Frequencies: ELA/L Grade 3

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 6,934 | 5.40 | 6,934 | 5.40 |
| 655—659 | 2,565 | 2.00 | 9,499 | 7.40 |
| 660—664 | 79 | 0.06 | 9,578 | 7.46 |
| 665—669 | 3,354 | 2.61 | 12,932 | 10.07 |
| 670—674 | 3,361 | 2.62 | 16,293 | 12.69 |
| 675—679 | 6,880 | 5.36 | 23,173 | 18.05 |
| 680—684 | 106 | 0.08 | 23,279 | 18.13 |
| 685—689 | 6,480 | 5.05 | 29,759 | 23.18 |
| 690—694 | 5,793 | 4.51 | 35,552 | 27.69 |
| 695—699 | 5,090 | 3.97 | 40,642 | 31.66 |
| 700—704 | 4,462 | 3.48 | 45,104 | 35.14 |
| 705—709 | 5,866 | 4.57 | 50,970 | 39.71 |
| 710—714 | 3,736 | 2.91 | 54,706 | 42.62 |
| 715—719 | 3,835 | 2.99 | 58,541 | 45.61 |
| 720—724 | 6,962 | 5.42 | 65,503 | 51.03 |
| 725—729 | 5,143 | 4.01 | 70,646 | 55.04 |
| 730—734 | 5,016 | 3.91 | 75,662 | 58.95 |
| 735—739 | 6,462 | 5.03 | 82,124 | 63.98 |
| 740—744 | 6,463 | 5.04 | 88,587 | 69.02 |
| 745—749 | 3,120 | 2.43 | 91,707 | 71.45 |
| 750—754 | 6,045 | 4.71 | 97,752 | 76.16 |
| 755—759 | 4,205 | 3.28 | 101,957 | 79.44 |
| 760—764 | 5,270 | 4.11 | 107,227 | 83.55 |
| 765—769 | 2,572 | 2.00 | 109,799 | 85.55 |
| 770—774 | 4,564 | 3.56 | 114,363 | 89.11 |
| 775—779 | 3,007 | 2.34 | 117,370 | 91.45 |
| 780—784 | 1,782 | 1.39 | 119,152 | 92.84 |
| 785—789 | 2,775 | 2.16 | 121,927 | 95.00 |
| 790—794 | 1,066 | 0.83 | 122,993 | 95.83 |
| 795—799 | 1,807 | 1.41 | 124,800 | 97.24 |
| 800—804 | 757 | 0.59 | 125,557 | 97.83 |
| 805—809 | 606 | 0.47 | 126,163 | 98.30 |
| 810—814 | 844 | 0.66 | 127,007 | 98.96 |
| 815—819 | 368 | 0.29 | 127,375 | 99.25 |
| 820—824 | 307 | 0.24 | 127,682 | 99.49 |
| 825—829 | 178 | 0.14 | 127,860 | 99.63 |
| 830—834 | 136 | 0.11 | 127,996 | 99.74 |
| 835—839 | 114 | 0.09 | 128,110 | 99.83 |
| 840—844 | 105 | 0.08 | 128,215 | 99.91 |
| 845—849 | 31 | 0.02 | 128,246 | 99.93 |
| 850 | 6,934 | 5.40 | 6,934 | 5.40 |

Table A.12.24 Scale Score Cumulative Frequencies: ELA/L Grade 4

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 2,075 | 1.62 | 2,075 | 1.62 |
| 655—659 | 1,063 | 0.83 | 3,138 | 2.45 |
| 660—664 | 42 | 0.03 | 3,180 | 2.48 |
| 665—669 | 2,504 | 1.96 | 5,684 | 4.44 |
| 670—674 | 3,185 | 2.49 | 8,869 | 6.93 |
| 675—679 | 50 | 0.04 | 8,919 | 6.97 |
| 680—684 | 3,616 | 2.83 | 12,535 | 9.80 |
| 685—689 | 3,862 | 3.02 | 16,397 | 12.82 |
| 690—694 | 3,972 | 3.10 | 20,369 | 15.92 |
| 695—699 | 3,823 | 2.99 | 24,192 | 18.91 |
| 700—704 | 5,563 | 4.35 | 29,755 | 23.26 |
| 705—709 | 5,469 | 4.27 | 35,224 | 27.53 |
| 710—714 | 3,681 | 2.88 | 38,905 | 30.41 |
| 715—719 | 7,242 | 5.66 | 46,147 | 36.07 |
| 720—724 | 5,237 | 4.09 | 51,384 | 40.16 |
| 725—729 | 7,047 | 5.51 | 58,431 | 45.67 |
| 730—734 | 6,799 | 5.31 | 65,230 | 50.98 |
| 735—739 | 4,920 | 3.84 | 70,150 | 54.82 |
| 740—744 | 6,581 | 5.14 | 76,731 | 59.96 |
| 745—749 | 6,510 | 5.09 | 83,241 | 65.05 |
| 750—754 | 4,518 | 3.53 | 87,759 | 68.58 |
| 755—759 | 5,894 | 4.61 | 93,653 | 73.19 |
| 760—764 | 5,504 | 4.30 | 99,157 | 77.49 |
| 765—769 | 5,017 | 3.92 | 104,174 | 81.41 |
| 770—774 | 4,570 | 3.57 | 108,744 | 84.98 |
| 775—779 | 4,800 | 3.75 | 113,544 | 88.73 |
| 780—784 | 3,220 | 2.52 | 116,764 | 91.25 |
| 785—789 | 2,032 | 1.59 | 118,796 | 92.84 |
| 790—794 | 2,269 | 1.77 | 121,065 | 94.61 |
| 795—799 | 1,910 | 1.49 | 122,975 | 96.10 |
| 800—804 | 1,494 | 1.17 | 124,469 | 97.27 |
| 805—809 | 1,077 | 0.84 | 125,546 | 98.11 |
| 810—814 | 627 | 0.49 | 126,173 | 98.60 |
| 815—819 | 641 | 0.50 | 126,814 | 99.10 |
| 820—824 | 473 | 0.37 | 127,287 | 99.47 |
| 825—829 | 226 | 0.18 | 127,513 | 99.65 |
| 830—834 | 176 | 0.14 | 127,689 | 99.79 |
| 835—839 | 107 | 0.08 | 127,796 | 99.87 |
| 840—844 | 49 | 0.04 | 127,845 | 99.91 |
| 845—849 | 43 | 0.03 | 127,888 | 99.94 |
| 850 | 92 | 0.07 | 127,980 | 100.00 |

Table A.12.25 Scale Score Cumulative Frequencies: ELA/L Grade 5

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 1,797 | 1.39 | 1,797 | 1.39 |
| 655—659 | 37 | 0.03 | 1,834 | 1.42 |
| 660—664 | 1,003 | 0.77 | 2,837 | 2.19 |
| 665—669 | 1,216 | 0.94 | 4,053 | 3.13 |
| 670—674 | 1,400 | 1.08 | 5,453 | 4.21 |
| 675—679 | 3,273 | 2.52 | 8,726 | 6.73 |
| 680—684 | 1,967 | 1.52 | 10,693 | 8.25 |
| 685—689 | 1,899 | 1.46 | 12,592 | 9.71 |
| 690—694 | 3,895 | 3.00 | 16,487 | 12.71 |
| 695—699 | 5,867 | 4.52 | 22,354 | 17.23 |
| 700—704 | 5,950 | 4.59 | 28,304 | 21.82 |
| 705—709 | 3,692 | 2.85 | 31,996 | 24.67 |
| 710—714 | 5,473 | 4.22 | 37,469 | 28.89 |
| 715—719 | 7,338 | 5.66 | 44,807 | 34.55 |
| 720—724 | 5,247 | 4.04 | 50,054 | 38.59 |
| 725—729 | 7,201 | 5.55 | 57,255 | 44.14 |
| 730—734 | 6,768 | 5.22 | 64,023 | 49.36 |
| 735—739 | 6,703 | 5.17 | 70,726 | 54.53 |
| 740—744 | 7,758 | 5.98 | 78,484 | 60.51 |
| 745—749 | 6,101 | 4.70 | 84,585 | 65.21 |
| 750—754 | 7,260 | 5.60 | 91,845 | 70.81 |
| 755—759 | 5,369 | 4.14 | 97,214 | 74.95 |
| 760—764 | 6,230 | 4.80 | 103,444 | 79.75 |
| 765—769 | 5,564 | 4.29 | 109,008 | 84.04 |
| 770—774 | 3,907 | 3.01 | 112,915 | 87.05 |
| 775—779 | 4,176 | 3.22 | 117,091 | 90.27 |
| 780—784 | 2,887 | 2.23 | 119,978 | 92.50 |
| 785—789 | 2,848 | 2.20 | 122,826 | 94.70 |
| 790—794 | 1,795 | 1.38 | 124,621 | 96.08 |
| 795—799 | 1,475 | 1.14 | 126,096 | 97.22 |
| 800—804 | 1,148 | 0.88 | 127,244 | 98.10 |
| 805—809 | 620 | 0.48 | 127,864 | 98.58 |
| 810—814 | 646 | 0.50 | 128,510 | 99.08 |
| 815—819 | 327 | 0.25 | 128,837 | 99.33 |
| 820—824 | 256 | 0.20 | 129,093 | 99.53 |
| 825—829 | 253 | 0.20 | 129,346 | 99.73 |
| 830—834 | 81 | 0.06 | 129,427 | 99.79 |
| 835—839 | 90 | 0.07 | 129,517 | 99.86 |
| 840—844 | 84 | 0.06 | 129,601 | 99.92 |
| 845—849 | 18 | 0.01 | 129,619 | 99.93 |
| 850 | 119 | 0.09 | 129,738 | 100.00 |

Table A.12.26 Scale Score Cumulative Frequencies: ELA/L Grade 6

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 2,055 | 1.54 | 2,055 | 1.54 |
| 655—659 | — | — | — | — |
| 660—664 | 32 | 0.02 | 2,087 | 1.56 |
| 665—669 | 2,341 | 1.76 | 4,428 | 3.32 |
| 670—674 | 47 | 0.04 | 4,475 | 3.36 |
| 675—679 | 3,056 | 2.29 | 7,531 | 5.65 |
| 680—684 | 3,642 | 2.73 | 11,173 | 8.38 |
| 685—689 | 3,819 | 2.87 | 14,992 | 11.25 |
| 690—694 | 3,669 | 2.75 | 18,661 | 14.00 |
| 695—699 | 3,479 | 2.61 | 22,140 | 16.61 |
| 700—704 | 4,929 | 3.70 | 27,069 | 20.31 |
| 705—709 | 6,053 | 4.55 | 33,122 | 24.86 |
| 710—714 | 3,098 | 2.33 | 36,220 | 27.19 |
| 715—719 | 7,187 | 5.40 | 43,407 | 32.59 |
| 720—724 | 7,800 | 5.86 | 51,207 | 38.45 |
| 725—729 | 7,382 | 5.54 | 58,589 | 43.99 |
| 730—734 | 6,081 | 4.57 | 64,670 | 48.56 |
| 735—739 | 9,077 | 6.82 | 73,747 | 55.38 |
| 740—744 | 8,699 | 6.53 | 82,446 | 61.91 |
| 745—749 | 5,615 | 4.22 | 88,061 | 66.13 |
| 750—754 | 8,163 | 6.13 | 96,224 | 72.26 |
| 755—759 | 7,657 | 5.75 | 103,881 | 78.01 |
| 760—764 | 6,816 | 5.12 | 110,697 | 83.13 |
| 765—769 | 4,136 | 3.11 | 114,833 | 86.24 |
| 770—774 | 4,362 | 3.28 | 119,195 | 89.52 |
| 775—779 | 3,874 | 2.91 | 123,069 | 92.43 |
| 780—784 | 2,998 | 2.25 | 126,067 | 94.68 |
| 785—789 | 2,000 | 1.50 | 128,067 | 96.18 |
| 790—794 | 1,517 | 1.14 | 129,584 | 97.32 |
| 795—799 | 1,145 | 0.86 | 130,729 | 98.18 |
| 800—804 | 964 | 0.72 | 131,693 | 98.90 |
| 805—809 | 427 | 0.32 | 132,120 | 99.22 |
| 810—814 | 399 | 0.30 | 132,519 | 99.52 |
| 815—819 | 159 | 0.12 | 132,678 | 99.64 |
| 820—824 | 219 | 0.16 | 132,897 | 99.80 |
| 825—829 | 75 | 0.06 | 132,972 | 99.86 |
| 830—834 | 97 | 0.07 | 133,069 | 99.93 |
| 835—839 | 51 | 0.04 | 133,120 | 99.97 |
| 840—844 | — | — | — | — |
| 845—849 | 24 | 0.02 | 133,144 | 99.99 |
| 850 | 35 | 0.03 | 133,179 | 100.00 |

Table A.12.27 Scale Score Cumulative Frequencies: ELA/L Grade 7

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 3,535 | 2.63 | 3,535 | 2.63 |
| 655—659 | — | — | — | — |
| 660—664 | 1,513 | 1.13 | 5,048 | 3.76 |
| 665—669 | 1,323 | 0.99 | 6,371 | 4.75 |
| 670—674 | 3,386 | 2.52 | 9,757 | 7.27 |
| 675—679 | 58 | 0.04 | 9,815 | 7.31 |
| 680—684 | 3,831 | 2.85 | 13,646 | 10.16 |
| 685—689 | 3,976 | 2.96 | 17,622 | 13.12 |
| 690—694 | 3,858 | 2.87 | 21,480 | 15.99 |
| 695—699 | 3,853 | 2.87 | 25,333 | 18.86 |
| 700—704 | 5,285 | 3.94 | 30,618 | 22.80 |
| 705—709 | 3,610 | 2.69 | 34,228 | 25.49 |
| 710—714 | 6,799 | 5.06 | 41,027 | 30.55 |
| 715—719 | 3,396 | 2.53 | 44,423 | 33.08 |
| 720—724 | 6,406 | 4.77 | 50,829 | 37.85 |
| 725—729 | 6,282 | 4.68 | 57,111 | 42.53 |
| 730—734 | 7,305 | 5.44 | 64,416 | 47.97 |
| 735—739 | 5,763 | 4.29 | 70,179 | 52.26 |
| 740—744 | 8,214 | 6.12 | 78,393 | 58.38 |
| 745—749 | 5,350 | 3.98 | 83,743 | 62.36 |
| 750—754 | 7,634 | 5.69 | 91,377 | 68.05 |
| 755—759 | 4,870 | 3.63 | 96,247 | 71.68 |
| 760—764 | 6,938 | 5.17 | 103,185 | 76.85 |
| 765—769 | 6,420 | 4.78 | 109,605 | 81.63 |
| 770—774 | 3,878 | 2.89 | 113,483 | 84.52 |
| 775—779 | 4,404 | 3.28 | 117,887 | 87.80 |
| 780—784 | 3,908 | 2.91 | 121,795 | 90.71 |
| 785—789 | 2,645 | 1.97 | 124,440 | 92.68 |
| 790—794 | 2,770 | 2.06 | 127,210 | 94.74 |
| 795—799 | 1,758 | 1.31 | 128,968 | 96.05 |
| 800—804 | 1,380 | 1.03 | 130,348 | 97.08 |
| 805—809 | 882 | 0.66 | 131,230 | 97.74 |
| 810—814 | 971 | 0.72 | 132,201 | 98.46 |
| 815—819 | 525 | 0.39 | 132,726 | 98.85 |
| 820—824 | 578 | 0.43 | 133,304 | 99.28 |
| 825—829 | 211 | 0.16 | 133,515 | 99.44 |
| 830—834 | 261 | 0.19 | 133,776 | 99.63 |
| 835—839 | 141 | 0.11 | 133,917 | 99.74 |
| 840—844 | 111 | 0.08 | 134,028 | 99.82 |
| 845—849 | 45 | 0.03 | 134,073 | 99.85 |
| 850 | 194 | 0.14 | 134,267 | 100.00 |

Table A.12.28 Scale Score Cumulative Frequencies: ELA/L Grade 8

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 1,947 | 1.40 | 1,947 | 1.40 |
| 655—659 | 1,691 | 1.22 | 3,638 | 2.62 |
| 660—664 | — | — | — | — |
| 665—669 | 2,290 | 1.65 | 5,928 | 4.27 |
| 670—674 | 1,594 | 1.15 | 7,522 | 5.42 |
| 675—679 | 1,290 | 0.93 | 8,812 | 6.35 |
| 680—684 | 3,101 | 2.23 | 11,913 | 8.58 |
| 685—689 | 3,175 | 2.29 | 15,088 | 10.87 |
| 690—694 | 3,314 | 2.39 | 18,402 | 13.26 |
| 695—699 | 3,437 | 2.47 | 21,839 | 15.73 |
| 700—704 | 6,445 | 4.64 | 28,284 | 20.37 |
| 705—709 | 3,302 | 2.38 | 31,586 | 22.75 |
| 710—714 | 4,965 | 3.57 | 36,551 | 26.32 |
| 715—719 | 6,496 | 4.68 | 43,047 | 31.00 |
| 720—724 | 5,189 | 3.74 | 48,236 | 34.74 |
| 725—729 | 6,626 | 4.77 | 54,862 | 39.51 |
| 730—734 | 8,282 | 5.96 | 63,144 | 45.47 |
| 735—739 | 6,558 | 4.72 | 69,702 | 50.19 |
| 740—744 | 6,677 | 4.81 | 76,379 | 55.00 |
| 745—749 | 6,759 | 4.87 | 83,138 | 59.87 |
| 750—754 | 6,613 | 4.76 | 89,751 | 64.63 |
| 755—759 | 6,394 | 4.60 | 96,145 | 69.23 |
| 760—764 | 7,786 | 5.61 | 103,931 | 74.84 |
| 765—769 | 6,839 | 4.92 | 110,770 | 79.76 |
| 770—774 | 5,078 | 3.66 | 115,848 | 83.42 |
| 775—779 | 4,255 | 3.06 | 120,103 | 86.48 |
| 780—784 | 3,742 | 2.69 | 123,845 | 89.17 |
| 785—789 | 3,207 | 2.31 | 127,052 | 91.48 |
| 790—794 | 2,776 | 2.00 | 129,828 | 93.48 |
| 795—799 | 2,345 | 1.69 | 132,173 | 95.17 |
| 800—804 | 1,882 | 1.35 | 134,055 | 96.52 |
| 805—809 | 1,320 | 0.95 | 135,375 | 97.47 |
| 810—814 | 1,063 | 0.77 | 136,438 | 98.24 |
| 815—819 | 706 | 0.51 | 137,144 | 98.75 |
| 820—824 | 450 | 0.32 | 137,594 | 99.07 |
| 825—829 | 424 | 0.31 | 138,018 | 99.38 |
| 830—834 | 248 | 0.18 | 138,266 | 99.56 |
| 835—839 | 313 | 0.23 | 138,579 | 99.79 |
| 840—844 | 99 | 0.07 | 138,678 | 99.86 |
| 845—849 | 57 | 0.04 | 138,735 | 99.90 |
| 850 | 173 | 0.12 | 138,908 | 100.00 |

Table A.12.29 Scale Score Cumulative Frequencies: Mathematics Grade 3

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 2,564 | 2.00 | 2,564 | 2.00 |
| 655—659 | 1,235 | 0.96 | 3,799 | 2.96 |
| 660—664 | 932 | 0.73 | 4,731 | 3.69 |
| 665—669 | 2,607 | 2.03 | 7,338 | 5.72 |
| 670—674 | 50 | 0.04 | 7,388 | 5.76 |
| 675—679 | 3,177 | 2.48 | 10,565 | 8.24 |
| 680—684 | 3,422 | 2.67 | 13,987 | 10.91 |
| 685—689 | 3,733 | 2.91 | 17,720 | 13.82 |
| 690—694 | 3,958 | 3.09 | 21,678 | 16.91 |
| 695—699 | 4,100 | 3.20 | 25,778 | 20.11 |
| 700—704 | 4,236 | 3.31 | 30,014 | 23.42 |
| 705—709 | 8,543 | 6.67 | 38,557 | 30.09 |
| 710—714 | 4,373 | 3.41 | 42,930 | 33.50 |
| 715—719 | 4,384 | 3.42 | 47,314 | 36.92 |
| 720—724 | 8,681 | 6.78 | 55,995 | 43.70 |
| 725—729 | 4,172 | 3.26 | 60,167 | 46.96 |
| 730—734 | 8,122 | 6.34 | 68,289 | 53.30 |
| 735—739 | 7,427 | 5.80 | 75,716 | 59.10 |
| 740—744 | 3,682 | 2.87 | 79,398 | 61.97 |
| 745—749 | 6,797 | 5.31 | 86,195 | 67.28 |
| 750—754 | 6,388 | 4.99 | 92,583 | 72.27 |
| 755—759 | 5,841 | 4.56 | 98,424 | 76.83 |
| 760—764 | 5,361 | 4.18 | 103,785 | 81.01 |
| 765—769 | 2,560 | 2.00 | 106,345 | 83.01 |
| 770—774 | 4,720 | 3.68 | 111,065 | 86.69 |
| 775—779 | 4,142 | 3.23 | 115,207 | 89.92 |
| 780—784 | 1,836 | 1.43 | 117,043 | 91.35 |
| 785—789 | 3,222 | 2.52 | 120,265 | 93.87 |
| 790—794 | 1,532 | 1.20 | 121,797 | 95.07 |
| 795—799 | 1,240 | 0.97 | 123,037 | 96.04 |
| 800—804 | 1,992 | 1.55 | 125,029 | 97.59 |
| 805—809 | 851 | 0.66 | 125,880 | 98.25 |
| 810—814 | 372 | 0.29 | 126,252 | 98.54 |
| 815—819 | 299 | 0.23 | 126,551 | 98.77 |
| 820—824 | 538 | 0.42 | 127,089 | 99.19 |
| 825—829 | 263 | 0.21 | 127,352 | 99.40 |
| 830—834 | 152 | 0.12 | 127,504 | 99.52 |
| 835—839 | 164 | 0.13 | 127,668 | 99.65 |
| 840—844 | 91 | 0.07 | 127,759 | 99.72 |
| 845—849 | — | — | — | — |
| 850 | 350 | 0.27 | 128,109 | 100.00 |

Table A.12.30 Scale Score Cumulative Frequencies: Mathematics Grade 4

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 1,075 | 0.84 | 1,075 | 0.84 |
| 655—659 | 1,524 | 1.19 | 2,599 | 2.03 |
| 660—664 | 30 | 0.02 | 2,629 | 2.05 |
| 665—669 | 2,559 | 2.00 | 5,188 | 4.05 |
| 670—674 | 3,405 | 2.66 | 8,593 | 6.71 |
| 675—679 | 51 | 0.04 | 8,644 | 6.75 |
| 680—684 | 3,955 | 3.09 | 12,599 | 9.84 |
| 685—689 | 4,133 | 3.23 | 16,732 | 13.07 |
| 690—694 | 4,527 | 3.54 | 21,259 | 16.61 |
| 695—699 | 4,630 | 3.62 | 25,889 | 20.23 |
| 700—704 | 4,933 | 3.86 | 30,822 | 24.09 |
| 705—709 | 7,023 | 5.49 | 37,845 | 29.58 |
| 710—714 | 7,202 | 5.63 | 45,047 | 35.21 |
| 715—719 | 6,935 | 5.43 | 51,982 | 40.64 |
| 720—724 | 6,758 | 5.29 | 58,740 | 45.93 |
| 725—729 | 6,343 | 4.96 | 65,083 | 50.89 |
| 730—734 | 6,257 | 4.89 | 71,340 | 55.78 |
| 735—739 | 7,831 | 6.13 | 79,171 | 61.91 |
| 740—744 | 7,092 | 5.55 | 86,263 | 67.46 |
| 745—749 | 6,448 | 5.04 | 92,711 | 72.50 |
| 750—754 | 5,705 | 4.46 | 98,416 | 76.96 |
| 755—759 | 4,964 | 3.88 | 103,380 | 80.84 |
| 760—764 | 4,566 | 3.57 | 107,946 | 84.41 |
| 765—769 | 4,005 | 3.13 | 111,951 | 87.54 |
| 770—774 | 3,439 | 2.69 | 115,390 | 90.23 |
| 775—779 | 2,993 | 2.34 | 118,383 | 92.57 |
| 780—784 | 2,571 | 2.01 | 120,954 | 94.58 |
| 785—789 | 1,163 | 0.91 | 122,117 | 95.49 |
| 790—794 | 2,028 | 1.59 | 124,145 | 97.08 |
| 795—799 | 789 | 0.62 | 124,934 | 97.70 |
| 800—804 | 724 | 0.57 | 125,658 | 98.27 |
| 805—809 | 632 | 0.49 | 126,290 | 98.76 |
| 810—814 | 524 | 0.41 | 126,814 | 99.17 |
| 815—819 | 409 | 0.32 | 127,223 | 99.49 |
| 820—824 | 123 | 0.10 | 127,346 | 99.59 |
| 825—829 | 150 | 0.12 | 127,496 | 99.71 |
| 830—834 | 75 | 0.06 | 127,571 | 99.77 |
| 835—839 | 90 | 0.07 | 127,661 | 99.84 |
| 840—844 | — | — | — | — |
| 845—849 | 49 | 0.04 | 127,710 | 99.88 |
| 850 | 123 | 0.10 | 127,833 | 100.00 |

Table A.12.31 Scale Score Cumulative Frequencies: Mathematics Grade 5

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 1,657 | 1.28 | 1,657 | 1.28 |
| 655—659 | 15 | 0.01 | 1,672 | 1.29 |
| 660—664 | 1,144 | 0.88 | 2,816 | 2.17 |
| 665—669 | 967 | 0.75 | 3,783 | 2.92 |
| 670—674 | 1,915 | 1.48 | 5,698 | 4.40 |
| 675—679 | 1,452 | 1.12 | 7,150 | 5.52 |
| 680—684 | 4,214 | 3.25 | 11,364 | 8.77 |
| 685—689 | 4,923 | 3.80 | 16,287 | 12.57 |
| 690—694 | 5,254 | 4.06 | 21,541 | 16.63 |
| 695—699 | 5,514 | 4.26 | 27,055 | 20.89 |
| 700—704 | 5,554 | 4.29 | 32,609 | 25.18 |
| 705—709 | 5,621 | 4.34 | 38,230 | 29.52 |
| 710—714 | 11,010 | 8.50 | 49,240 | 38.02 |
| 715—719 | 5,308 | 4.10 | 54,548 | 42.12 |
| 720—724 | 9,869 | 7.62 | 64,417 | 49.74 |
| 725—729 | 4,595 | 3.55 | 69,012 | 53.29 |
| 730—734 | 8,686 | 6.70 | 77,698 | 59.99 |
| 735—739 | 7,880 | 6.08 | 85,578 | 66.07 |
| 740—744 | 5,045 | 3.89 | 90,623 | 69.96 |
| 745—749 | 6,122 | 4.73 | 96,745 | 74.69 |
| 750—754 | 5,330 | 4.11 | 102,075 | 78.80 |
| 755—759 | 4,740 | 3.66 | 106,815 | 82.46 |
| 760—764 | 4,102 | 3.17 | 110,917 | 85.63 |
| 765—769 | 3,641 | 2.81 | 114,558 | 88.44 |
| 770—774 | 3,051 | 2.35 | 117,609 | 90.79 |
| 775—779 | 2,725 | 2.10 | 120,334 | 92.89 |
| 780—784 | 2,247 | 1.73 | 122,581 | 94.62 |
| 785—789 | 1,999 | 1.54 | 124,580 | 96.16 |
| 790—794 | 1,256 | 0.97 | 125,836 | 97.13 |
| 795—799 | 1,374 | 1.06 | 127,210 | 98.19 |
| 800—804 | 602 | 0.46 | 127,812 | 98.65 |
| 805—809 | 462 | 0.36 | 128,274 | 99.01 |
| 810—814 | 398 | 0.31 | 128,672 | 99.32 |
| 815—819 | 317 | 0.24 | 128,989 | 99.56 |
| 820—824 | 218 | 0.17 | 129,207 | 99.73 |
| 825—829 | 75 | 0.06 | 129,282 | 99.79 |
| 830—834 | 79 | 0.06 | 129,361 | 99.85 |
| 835—839 | 57 | 0.04 | 129,418 | 99.89 |
| 840—844 | — | — | — | — |
| 845—849 | 57 | 0.04 | 129,475 | 99.93 |
| 850 | 87 | 0.07 | 129,562 | 100.00 |

Table A.12.32 Scale Score Cumulative Frequencies: Mathematics Grade 6

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 2,720 | 2.05 | 2,720 | 2.05 |
| 655—659 | — | — | — | — |
| 660—664 | 31 | 0.02 | 2,751 | 2.07 |
| 665—669 | 1,939 | 1.46 | 4,690 | 3.53 |
| 670—674 | 2,265 | 1.70 | 6,955 | 5.23 |
| 675—679 | 2,982 | 2.24 | 9,937 | 7.47 |
| 680—684 | 3,348 | 2.52 | 13,285 | 9.99 |
| 685—689 | 3,732 | 2.81 | 17,017 | 12.80 |
| 690—694 | 8,012 | 6.03 | 25,029 | 18.83 |
| 695—699 | 4,413 | 3.32 | 29,442 | 22.15 |
| 700—704 | 8,158 | 6.14 | 37,600 | 28.29 |
| 705—709 | 7,650 | 5.76 | 45,250 | 34.05 |
| 710—714 | 6,894 | 5.19 | 52,144 | 39.24 |
| 715—719 | 9,041 | 6.81 | 61,185 | 46.05 |
| 720—724 | 5,630 | 4.24 | 66,815 | 50.29 |
| 725—729 | 9,882 | 7.44 | 76,697 | 57.73 |
| 730—734 | 4,382 | 3.30 | 81,079 | 61.03 |
| 735—739 | 7,588 | 5.71 | 88,667 | 66.74 |
| 740—744 | 6,427 | 4.84 | 95,094 | 71.58 |
| 745—749 | 6,840 | 5.15 | 101,934 | 76.73 |
| 750—754 | 4,684 | 3.53 | 106,618 | 80.26 |
| 755—759 | 6,042 | 4.55 | 112,660 | 84.81 |
| 760—764 | 3,351 | 2.52 | 116,011 | 87.33 |
| 765—769 | 4,494 | 3.38 | 120,505 | 90.71 |
| 770—774 | 2,534 | 1.91 | 123,039 | 92.62 |
| 775—779 | 2,841 | 2.14 | 125,880 | 94.76 |
| 780—784 | 2,302 | 1.73 | 128,182 | 96.49 |
| 785—789 | 1,534 | 1.15 | 129,716 | 97.64 |
| 790—794 | 890 | 0.67 | 130,606 | 98.31 |
| 795—799 | 777 | 0.58 | 131,383 | 98.89 |
| 800—804 | 558 | 0.42 | 131,941 | 99.31 |
| 805—809 | 277 | 0.21 | 132,218 | 99.52 |
| 810—814 | 237 | 0.18 | 132,455 | 99.70 |
| 815—819 | 167 | 0.13 | 132,622 | 99.83 |
| 820—824 | 65 | 0.05 | 132,687 | 99.88 |
| 825—829 | 39 | 0.03 | 132,726 | 99.91 |
| 830—834 | 53 | 0.04 | 132,779 | 99.95 |
| 835—839 | 28 | 0.02 | 132,807 | 99.97 |
| 840—844 | — | — | — | — |
| 845—849 | — | — | — | — |
| 850 | 51 | 0.04 | 132,858 | 100.00 |

Table A.12.33 Scale Score Cumulative Frequencies: Mathematics Grade 7

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 1,482 | 1.11 | 1,482 | 1.11 |
| 655—659 | — | — | — | — |
| 660—664 | — | — | — | — |
| 665—669 | 1,330 | 0.99 | 2,812 | 2.10 |
| 670—674 | 1,191 | 0.89 | 4,003 | 2.99 |
| 675—679 | 2,005 | 1.50 | 6,008 | 4.49 |
| 680—684 | 1,915 | 1.43 | 7,923 | 5.92 |
| 685—689 | 2,478 | 1.85 | 10,401 | 7.77 |
| 690—694 | 2,755 | 2.06 | 13,156 | 9.83 |
| 695—699 | 5,973 | 4.46 | 19,129 | 14.29 |
| 700—704 | 6,408 | 4.78 | 25,537 | 19.07 |
| 705—709 | 6,712 | 5.01 | 32,249 | 24.08 |
| 710—714 | 6,561 | 4.90 | 38,810 | 28.98 |
| 715—719 | 6,653 | 4.97 | 45,463 | 33.95 |
| 720—724 | 12,819 | 9.57 | 58,282 | 43.52 |
| 725—729 | 6,053 | 4.52 | 64,335 | 48.04 |
| 730—734 | 10,778 | 8.05 | 75,113 | 56.09 |
| 735—739 | 8,990 | 6.71 | 84,103 | 62.80 |
| 740—744 | 7,461 | 5.57 | 91,564 | 68.37 |
| 745—749 | 6,551 | 4.89 | 98,115 | 73.26 |
| 750—754 | 6,707 | 5.01 | 104,822 | 78.27 |
| 755—759 | 6,529 | 4.87 | 111,351 | 83.14 |
| 760—764 | 4,532 | 3.38 | 115,883 | 86.52 |
| 765—769 | 4,642 | 3.47 | 120,525 | 89.99 |
| 770—774 | 3,142 | 2.35 | 123,667 | 92.34 |
| 775—779 | 3,135 | 2.34 | 126,802 | 94.68 |
| 780—784 | 1,849 | 1.38 | 128,651 | 96.06 |
| 785—789 | 1,515 | 1.13 | 130,166 | 97.19 |
| 790—794 | 1,343 | 1.00 | 131,509 | 98.19 |
| 795—799 | 755 | 0.56 | 132,264 | 98.75 |
| 800—804 | 419 | 0.31 | 132,683 | 99.06 |
| 805—809 | 390 | 0.29 | 133,073 | 99.35 |
| 810—814 | 188 | 0.14 | 133,261 | 99.49 |
| 815—819 | 268 | 0.20 | 133,529 | 99.69 |
| 820—824 | 108 | 0.08 | 133,637 | 99.77 |
| 825—829 | 88 | 0.07 | 133,725 | 99.84 |
| 830—834 | — | — | — | — |
| 835—839 | 125 | 0.09 | 133,850 | 99.93 |
| 840—844 | — | — | — | — |
| 845—849 | — | — | — | — |
| 850 | 106 | 0.08 | 133,956 | 100.00 |

Table A.12.34 Scale Score Cumulative Frequencies: Mathematics Grade 8

| Score Band | Count | Percent | Cumulative Count | Cumulative Percent |
|---|---|---|---|---|
| 650—654 | 6,862 | 4.95 | 6,862 | 4.95 |
| 655—659 | 46 | 0.03 | 6,908 | 4.98 |
| 660—664 | — | — | — | — |
| 665—669 | 6,353 | 4.59 | 13,261 | 9.57 |
| 670—674 | 65 | 0.05 | 13,326 | 9.62 |
| 675—679 | 8,428 | 6.08 | 21,754 | 15.70 |
| 680—684 | 79 | 0.06 | 21,833 | 15.76 |
| 685—689 | 9,444 | 6.82 | 31,277 | 22.58 |
| 690—694 | 71 | 0.05 | 31,348 | 22.63 |
| 695—699 | 9,738 | 7.03 | 41,086 | 29.66 |
| 700—704 | 9,103 | 6.57 | 50,189 | 36.23 |
| 705—709 | 8,506 | 6.14 | 58,695 | 42.37 |
| 710—714 | 39 | 0.03 | 58,734 | 42.40 |
| 715—719 | 7,457 | 5.38 | 66,191 | 47.78 |
| 720—724 | 7,078 | 5.11 | 73,269 | 52.89 |
| 725—729 | 6,233 | 4.50 | 79,502 | 57.39 |
| 730—734 | 7,949 | 5.74 | 87,451 | 63.13 |
| 735—739 | 4,533 | 3.27 | 91,984 | 66.40 |
| 740—744 | 6,033 | 4.35 | 98,017 | 70.75 |
| 745—749 | 5,109 | 3.69 | 103,126 | 74.44 |
| 750—754 | 3,052 | 2.20 | 106,178 | 76.64 |
| 755—759 | 5,359 | 3.87 | 111,537 | 80.51 |
| 760—764 | 3,370 | 2.43 | 114,907 | 82.94 |
| 765—769 | 4,053 | 2.93 | 118,960 | 85.87 |
| 770—774 | 2,601 | 1.88 | 121,561 | 87.75 |
| 775—779 | 3,296 | 2.38 | 124,857 | 90.13 |
| 780—784 | 2,752 | 1.99 | 127,609 | 92.12 |
| 785—789 | 1,211 | 0.87 | 128,820 | 92.99 |
| 790—794 | 2,146 | 1.55 | 130,966 | 94.54 |
| 795—799 | 1,349 | 0.97 | 132,315 | 95.51 |
| 800—804 | 1,237 | 0.89 | 133,552 | 96.40 |
| 805—809 | 1,145 | 0.83 | 134,697 | 97.23 |
| 810—814 | 996 | 0.72 | 135,693 | 97.95 |
| 815—819 | 573 | 0.41 | 136,266 | 98.36 |
| 820—824 | 509 | 0.37 | 136,775 | 98.73 |
| 825—829 | 395 | 0.29 | 137,170 | 99.02 |
| 830—834 | 361 | 0.26 | 137,531 | 99.28 |
| 835—839 | 303 | 0.22 | 137,834 | 99.50 |
| 840—844 | 220 | 0.16 | 138,054 | 99.66 |
| 845—849 | 110 | 0.08 | 138,164 | 99.74 |
| 850 | 394 | 0.28 | 138,558 | 100.00 |

# Appendix 12.5: Subgroup Scale Score Performance

Table A.12.35 Subgroup Performance for ELA/L Scale Scores: Grade 3

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **128,356** | **723.35** | **40.77** | **650** | **850** |
| Gender | Female | 62,746 | 727.61 | 41.48 | 650 | 850 |
| | Male | 65,604 | 719.27 | 39.66 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 335 | 708.15 | 38.16 | 650 | 827 |
| | Asian | 7,234 | 746.99 | 41.15 | 650 | 850 |
| | Black/African American | 21,218 | 704.73 | 36.21 | 650 | 850 |
| | Hispanic/Latino | 34,170 | 709.96 | 37.70 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 99 | 731.10 | 37.96 | 650 | 835 |
| | Two or more races | 5,585 | 728.73 | 42.01 | 650 | 850 |
| | White | 58,710 | 734.54 | 38.67 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 738.17 | 39.39 | 650 | 850 |
| | Economically Disadvantaged | 65,055 | 708.93 | 36.72 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,257 | 728.26 | 40.60 | 650 | 850 |
| | English Learner | 25,099 | 703.15 | 34.83 | 650 | 846 |
| Disabilities | Students without Disabilities | 103,803 | 727.92 | 40.11 | 650 | 850 |
| | Student with Disability (SWD) | 24,553 | 704.04 | 37.80 | 650 | 850 |
| **Reading Summative Score** | | **128,356** | **41.22** | **16.83** | **10** | **90** |
| Gender | Female | 62,746 | 42.63 | 17.03 | 10 | 90 |
| | Male | 65,604 | 39.87 | 16.53 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 335 | 35.10 | 15.42 | 10 | 76 |
| | Asian | 7,234 | 50.67 | 17.16 | 10 | 90 |
| | Black/African American | 21,218 | 33.85 | 14.86 | 10 | 90 |
| | Hispanic/Latino | 34,170 | 35.68 | 15.33 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 99 | 44.01 | 15.76 | 10 | 86 |
| | Two or more races | 5,585 | 43.66 | 17.53 | 10 | 90 |
| | White | 58,710 | 45.75 | 16.18 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 47.28 | 16.48 | 10 | 90 |
| | Economically Disadvantaged | 65,055 | 35.31 | 14.97 | 10 | 90 |
| English Learner Status | Non-English Learner | 103,257 | 43.30 | 16.85 | 10 | 90 |
| | English Learner | 25,099 | 32.67 | 13.81 | 10 | 90 |
| Disabilities | Students without Disabilities | 103,803 | 43.04 | 16.58 | 10 | 90 |
| | Student with Disability (SWD) | 24,553 | 33.52 | 15.69 | 10 | 90 |
| **Writing Summative Score** | | **128,356** | **23.71** | **13.04** | **10** | **60** |
| Gender | Female | 62,746 | 25.38 | 13.18 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Male | 65,604 | 22.11 | 12.70 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 335 | 19.57 | 12.07 | 10 | 53 |
| | Asian | 7,234 | 30.57 | 12.79 | 10 | 60 |
| | Black/African American | 21,218 | 18.42 | 11.43 | 10 | 60 |
| | Hispanic/Latino | 34,170 | 20.26 | 12.16 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 99 | 26.27 | 12.61 | 10 | 52 |
| | Two or more races | 5,585 | 24.88 | 13.30 | 10 | 60 |
| | White | 58,710 | 26.69 | 12.88 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 63,301 | 27.66 | 12.96 | 10 | 60 |
| | Economically Disadvantaged | 65,055 | 19.86 | 11.92 | 10 | 60 |
| English Learner Status | Non-English Learner | 103,257 | 24.88 | 13.12 | 10 | 60 |
| | English Learner | 25,099 | 18.91 | 11.54 | 10 | 60 |
| Disabilities | Students without Disabilities | 103,803 | 24.94 | 13.06 | 10 | 60 |
| | Student with Disability (SWD) | 24,553 | 18.50 | 11.58 | 10 | 60 |

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

Table A.12.36 Subgroup Performance for ELA/L Scale Scores: Grade 4

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **127,980** | **734.11** | **37.78** | **650** | **850** |
| Gender | Female | 62,671 | 737.97 | 38.03 | 650 | 850 |
| | Male | 65,298 | 730.39 | 37.16 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 318 | 729.54 | 38.92 | 650 | 826 |
| | Asian | 7,380 | 757.09 | 37.59 | 650 | 850 |
| | Black/African American | 20,905 | 715.55 | 33.73 | 650 | 850 |
| | Hispanic/Latino | 34,018 | 722.47 | 35.43 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 108 | 743.19 | 40.73 | 665 | 850 |
| | Two or more races | 5,530 | 738.65 | 38.50 | 650 | 850 |
| | White | 58,744 | 744.26 | 35.43 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,483 | 747.85 | 35.97 | 650 | 850 |
| | Economically Disadvantaged | 64,497 | 720.58 | 34.49 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,511 | 738.82 | 37.28 | 650 | 850 |
| | English Learner | 24,469 | 714.17 | 33.07 | 650 | 850 |
| Disabilities | Students without Disabilities | 102,775 | 739.46 | 36.12 | 650 | 850 |
| | Student with Disability (SWD) | 25,205 | 712.28 | 36.52 | 650 | 850 |
| **Reading Summative Score** | | **127,980** | **44.55** | **15.47** | **10** | **90** |
| Gender | Female | 62,671 | 45.64 | 15.41 | 10 | 90 |
| | Male | 65,298 | 43.50 | 15.45 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 318 | 42.62 | 15.82 | 10 | 83 |
| | Asian | 7,380 | 53.53 | 15.51 | 10 | 90 |
| | Black/African American | 20,905 | 37.37 | 13.88 | 10 | 90 |
| | Hispanic/Latino | 34,018 | 39.82 | 14.36 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 108 | 48.53 | 16.95 | 17 | 90 |
| | Two or more races | 5,530 | 46.59 | 15.92 | 10 | 90 |
| | White | 58,744 | 48.56 | 14.68 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 63,483 | 50.05 | 14.91 | 10 | 90 |
| | Economically Disadvantaged | 64,497 | 39.13 | 14.03 | 10 | 90 |
| English Learner Status | Non-English Learner | 103,511 | 46.51 | 15.32 | 10 | 90 |
| | English Learner | 24,469 | 36.24 | 13.14 | 10 | 90 |
| Disabilities | Students without Disabilities | 102,775 | 46.63 | 14.82 | 10 | 90 |
| | Student with Disability (SWD) | 25,205 | 36.06 | 15.14 | 10 | 90 |
| **Writing Summative Score** | | **127,980** | **26.73** | **13.26** | **10** | **60** |
| Gender | Female | 62,671 | 28.53 | 13.13 | 10 | 60 |
| | Male | 65,298 | 25.00 | 13.16 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 318 | 25.46 | 13.56 | 10 | 55 |
| | Asian | 7,380 | 33.85 | 12.07 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Black/African American | 20,905 | 20.67 | 12.23 | 10 | 60 |
| | Hispanic/Latino | 34,018 | 23.38 | 12.84 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 108 | 28.88 | 13.84 | 10 | 60 |
| | Two or more races | 5,530 | 27.67 | 13.41 | 10 | 60 |
| | White | 58,744 | 29.89 | 12.62 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 63,483 | 30.86 | 12.57 | 10 | 60 |
| | Economically Disadvantaged | 64,497 | 22.66 | 12.65 | 10 | 60 |
| English Learner Status | Non-English Learner | 103,511 | 27.99 | 13.16 | 10 | 60 |
| | English Learner | 24,469 | 21.40 | 12.32 | 10 | 55 |
| Disabilities | Students without Disabilities | 102,775 | 28.38 | 12.95 | 10 | 60 |
| | Student with Disability (SWD) | 25,205 | 19.98 | 12.32 | 10 | 60 |

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

Table A.12.37 Subgroup Performance for ELA/L Scale Scores: Grade 5

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **129,738** | **734.37** | **35.29** | **650** | **850** |
| Gender | Female | 63,606 | 737.64 | 35.25 | 650 | 850 |
| | Male | 66,113 | 731.22 | 35.04 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 302 | 726.40 | 33.42 | 650 | 850 |
| | Asian | 7,432 | 757.52 | 35.11 | 650 | 850 |
| | Black/African American | 21,007 | 716.72 | 31.21 | 650 | 850 |
| | Hispanic/Latino | 35,407 | 723.31 | 32.74 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 131 | 743.81 | 32.10 | 650 | 810 |
| | Two or more races | 5,428 | 737.94 | 36.16 | 650 | 850 |
| | White | 59,067 | 744.17 | 33.11 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 64,669 | 747.55 | 33.42 | 650 | 850 |
| | Economically Disadvantaged | 65,069 | 721.28 | 32.07 | 650 | 850 |
| English Learner Status | Non-English Learner | 108,377 | 739.29 | 34.49 | 650 | 850 |
| | English Learner | 21,361 | 709.44 | 28.02 | 650 | 821 |
| Disabilities | Students without Disabilities | 103,644 | 739.66 | 33.43 | 650 | 850 |
| | Student with Disability (SWD) | 26,094 | 713.37 | 34.66 | 650 | 850 |
| **Reading Summative Score** | | **129,738** | **44.59** | **14.47** | **10** | **90** |
| Gender | Female | 63,606 | 45.22 | 14.03 | 10 | 90 |
| | Male | 66,113 | 43.98 | 14.85 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 302 | 41.53 | 13.73 | 10 | 90 |
| | Asian | 7,432 | 53.64 | 14.60 | 10 | 90 |
| | Black/African American | 21,007 | 37.71 | 12.89 | 10 | 90 |
| | Hispanic/Latino | 35,407 | 40.06 | 13.33 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 131 | 48.06 | 13.11 | 10 | 72 |
| | Two or more races | 5,428 | 46.11 | 14.85 | 10 | 90 |
| | White | 59,067 | 48.51 | 13.70 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 64,669 | 49.89 | 13.84 | 10 | 90 |
| | Economically Disadvantaged | 65,069 | 39.32 | 13.09 | 10 | 90 |
| English Learner Status | Non-English Learner | 108,377 | 46.61 | 14.17 | 10 | 90 |
| | English Learner | 21,361 | 34.33 | 11.23 | 10 | 82 |
| Disabilities | Students without Disabilities | 103,644 | 46.63 | 13.73 | 10 | 90 |
| | Student with Disability (SWD) | 26,094 | 36.47 | 14.46 | 10 | 90 |
| **Writing Summative Score** | | **129,738** | **25.55** | **13.70** | **10** | **60** |
| Gender | Female | 63,606 | 27.56 | 13.68 | 10 | 60 |
| | Male | 66,113 | 23.62 | 13.43 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 302 | 22.81 | 13.15 | 10 | 60 |
| | Asian | 7,432 | 33.64 | 12.39 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Black/African American | 21,007 | 19.34 | 12.20 | 10 | 60 |
| | Hispanic/Latino | 35,407 | 22.19 | 13.02 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 131 | 29.63 | 12.71 | 10 | 52 |
| | Two or more races | 5,428 | 26.44 | 13.90 | 10 | 60 |
| | White | 59,067 | 28.72 | 13.33 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 64,669 | 29.83 | 13.23 | 10 | 60 |
| | Economically Disadvantaged | 65,069 | 21.30 | 12.80 | 10 | 60 |
| English Learner Status | Non-English Learner | 108,377 | 27.01 | 13.64 | 10 | 60 |
| | English Learner | 21,361 | 18.19 | 11.41 | 10 | 55 |
| Disabilities | Students without Disabilities | 103,644 | 27.30 | 13.49 | 10 | 60 |
| | Student with Disability (SWD) | 26,094 | 18.60 | 12.21 | 10 | 60 |

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

Table A.12.38 Subgroup Performance for ELA/L Scale Scores: Grade 6

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **133,179** | **733.46** | **33.18** | **650** | **850** |
| Gender | Female | 65,132 | 738.32 | 33.18 | 650 | 850 |
| | Male | 68,026 | 728.80 | 32.51 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 294 | 723.71 | 37.37 | 650 | 833 |
| | Asian | 7,383 | 755.26 | 31.95 | 650 | 850 |
| | Black/African American | 21,617 | 717.51 | 30.35 | 650 | 839 |
| | Hispanic/Latino | 36,751 | 723.52 | 31.17 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 138 | 737.67 | 34.14 | 650 | 838 |
| | Two or more races | 5,365 | 736.66 | 33.50 | 650 | 850 |
| | White | 60,780 | 742.42 | 30.99 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 66,758 | 745.24 | 31.20 | 650 | 850 |
| | Economically Disadvantaged | 66,421 | 721.61 | 30.82 | 650 | 850 |
| English Learner Status | Non-English Learner | 113,730 | 737.92 | 32.19 | 650 | 850 |
| | English Learner | 19,449 | 707.36 | 26.13 | 650 | 813 |
| Disabilities | Students without Disabilities | 106,701 | 738.68 | 31.23 | 650 | 850 |
| | Student with Disability (SWD) | 26,478 | 712.42 | 32.51 | 650 | 845 |
| **Reading Summative Score** | | **133,179** | **43.90** | **13.22** | **10** | **90** |
| Gender | Female | 65,132 | 45.21 | 13.11 | 10 | 90 |
| | Male | 68,026 | 42.66 | 13.21 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 294 | 39.97 | 14.77 | 10 | 77 |
| | Asian | 7,383 | 52.08 | 12.90 | 10 | 90 |
| | Black/African American | 21,617 | 38.07 | 12.35 | 10 | 85 |
| | Hispanic/Latino | 36,751 | 40.00 | 12.41 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 138 | 45.62 | 13.38 | 10 | 81 |
| | Two or more races | 5,365 | 45.39 | 13.47 | 10 | 90 |
| | White | 60,780 | 47.29 | 12.43 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 66,758 | 48.51 | 12.50 | 10 | 90 |
| | Economically Disadvantaged | 66,421 | 39.28 | 12.28 | 10 | 90 |
| English Learner Status | Non-English Learner | 113,730 | 45.69 | 12.84 | 10 | 90 |
| | English Learner | 19,449 | 33.45 | 10.26 | 10 | 77 |
| Disabilities | Students without Disabilities | 106,701 | 45.86 | 12.46 | 10 | 90 |
| | Student with Disability (SWD) | 26,478 | 36.03 | 13.29 | 10 | 90 |
| **Writing Summative Score** | | **133,179** | **27.47** | **12.51** | **10** | **60** |
| Gender | Female | 65,132 | 29.83 | 12.07 | 10 | 60 |
| | Male | 68,026 | 25.22 | 12.51 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 294 | 24.78 | 12.98 | 10 | 60 |
| | Asian | 7,383 | 34.79 | 10.39 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Black/African American | 21,617 | 21.76 | 12.00 | 10 | 60 |
| | Hispanic/Latino | 36,751 | 24.39 | 12.31 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 138 | 28.90 | 12.31 | 10 | 60 |
| | Two or more races | 5,365 | 28.21 | 12.47 | 10 | 60 |
| | White | 60,780 | 30.50 | 11.69 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 66,758 | 31.26 | 11.55 | 10 | 60 |
| | Economically Disadvantaged | 66,421 | 23.66 | 12.28 | 10 | 60 |
| English Learner Status | Non-English Learner | 113,730 | 28.84 | 12.21 | 10 | 60 |
| | English Learner | 19,449 | 19.46 | 11.17 | 10 | 51 |
| Disabilities | Students without Disabilities | 106,701 | 29.29 | 11.95 | 10 | 60 |
| | Student with Disability (SWD) | 26,478 | 20.15 | 12.02 | 10 | 60 |

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

Table A.12.39 Subgroup Performance for ELA/L Scale Scores: Grade 7

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **134,267** | **735.41** | **38.12** | **650** | **850** |
| Gender | Female | 65,120 | 740.76 | 37.98 | 650 | 850 |
| | Male | 69,106 | 730.36 | 37.56 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 290 | 728.62 | 37.47 | 650 | 850 |
| | Asian | 7,439 | 760.65 | 37.34 | 650 | 850 |
| | Black/African American | 22,091 | 717.99 | 34.31 | 650 | 847 |
| | Hispanic/Latino | 37,144 | 724.57 | 35.58 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 117 | 738.91 | 38.37 | 652 | 827 |
| | Two or more races | 5,230 | 738.45 | 38.38 | 650 | 850 |
| | White | 61,126 | 745.12 | 36.34 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 67,550 | 748.46 | 36.31 | 650 | 850 |
| | Economically Disadvantaged | 66,717 | 722.20 | 35.25 | 650 | 850 |
| English Learner Status | Non-English Learner | 115,292 | 740.20 | 37.14 | 650 | 850 |
| | English Learner | 18,975 | 706.31 | 30.22 | 650 | 847 |
| Disabilities | Students without Disabilities | 107,687 | 741.41 | 35.99 | 650 | 850 |
| | Student with Disability (SWD) | 26,580 | 711.10 | 36.84 | 650 | 850 |
| **Reading Summative Score** | | **134,267** | **44.83** | **15.47** | **10** | **90** |
| Gender | Female | 65,120 | 46.18 | 15.26 | 10 | 90 |
| | Male | 69,106 | 43.56 | 15.55 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 290 | 42.30 | 14.94 | 10 | 87 |
| | Asian | 7,439 | 54.57 | 15.58 | 10 | 90 |
| | Black/African American | 22,091 | 38.20 | 13.82 | 10 | 90 |
| | Hispanic/Latino | 37,144 | 40.57 | 14.34 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 117 | 47.12 | 16.36 | 11 | 83 |
| | Two or more races | 5,230 | 46.37 | 15.75 | 10 | 90 |
| | White | 61,126 | 48.55 | 14.95 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 67,550 | 50.03 | 14.96 | 10 | 90 |
| | Economically Disadvantaged | 66,717 | 39.57 | 14.14 | 10 | 90 |
| English Learner Status | Non-English Learner | 115,292 | 46.80 | 15.11 | 10 | 90 |
| | English Learner | 18,975 | 32.87 | 11.75 | 10 | 82 |
| Disabilities | Students without Disabilities | 107,687 | 47.13 | 14.64 | 10 | 90 |
| | Student with Disability (SWD) | 26,580 | 35.50 | 15.20 | 10 | 90 |
| **Writing Summative Score** | | **134,267** | **27.89** | **13.18** | **10** | **60** |
| Gender | Female | 65,120 | 30.32 | 12.86 | 10 | 60 |
| | Male | 69,106 | 25.60 | 13.07 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 290 | 25.39 | 13.40 | 10 | 60 |
| | Asian | 7,439 | 35.67 | 11.36 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Black/African American | 22,091 | 22.31 | 12.51 | 10 | 60 |
| | Hispanic/Latino | 37,144 | 24.72 | 12.78 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 117 | 28.52 | 12.77 | 10 | 49 |
| | Two or more races | 5,230 | 28.35 | 13.27 | 10 | 60 |
| | White | 61,126 | 30.90 | 12.57 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 67,550 | 31.76 | 12.39 | 10 | 60 |
| | Economically Disadvantaged | 66,717 | 23.96 | 12.79 | 10 | 60 |
| English Learner Status | Non-English Learner | 115,292 | 29.20 | 12.97 | 10 | 60 |
| | English Learner | 18,975 | 19.93 | 11.58 | 10 | 60 |
| Disabilities | Students without Disabilities | 107,687 | 29.72 | 12.72 | 10 | 60 |
| | Student with Disability (SWD) | 26,580 | 20.46 | 12.39 | 10 | 60 |

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

Table A.12.40 Subgroup Performance for ELA/L Scale Scores: Grade 8

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **138,908** | **738.26** | **37.97** | **650** | **850** |
| Gender | Female | 67,706 | 744.55 | 37.92 | 650 | 850 |
| | Male | 71,144 | 732.25 | 37.03 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 326 | 736.25 | 36.59 | 650 | 825 |
| | Asian | 7,527 | 762.72 | 36.35 | 650 | 850 |
| | Black/African American | 22,566 | 721.50 | 34.48 | 650 | 850 |
| | Hispanic/Latino | 39,623 | 727.39 | 35.92 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 156 | 744.26 | 37.41 | 650 | 828 |
| | Two or more races | 5,269 | 740.56 | 38.38 | 650 | 850 |
| | White | 62,635 | 748.23 | 35.94 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 69,744 | 750.80 | 35.98 | 650 | 850 |
| | Economically Disadvantaged | 69,164 | 725.61 | 35.66 | 650 | 850 |
| English Learner Status | Non-English Learner | 120,531 | 743.18 | 36.64 | 650 | 850 |
| | English Learner | 18,377 | 705.97 | 29.87 | 650 | 839 |
| Disabilities | Students without Disabilities | 111,690 | 744.04 | 35.78 | 650 | 850 |
| | Student with Disability (SWD) | 27,218 | 714.52 | 37.46 | 650 | 850 |
| **Reading Summative Score** | | **138,908** | **47.11** | **15.79** | **10** | **90** |
| Gender | Female | 67,706 | 49.08 | 15.71 | 10 | 90 |
| | Male | 71,144 | 45.22 | 15.63 | 10 | 90 |
| Ethnicity | American Indian/Alaska Native | 326 | 46.32 | 15.06 | 10 | 80 |
| | Asian | 7,527 | 56.88 | 15.31 | 10 | 90 |
| | Black/African American | 22,566 | 40.63 | 14.61 | 10 | 90 |
| | Hispanic/Latino | 39,623 | 42.60 | 14.99 | 10 | 90 |
| | Native Hawaiian/Pacific Islander | 156 | 49.62 | 16.07 | 10 | 87 |
| | Two or more races | 5,269 | 48.56 | 16.13 | 10 | 90 |
| | White | 62,635 | 51.07 | 14.94 | 10 | 90 |
| Economic Status* | Not Economically Disadvantaged | 69,744 | 52.29 | 14.98 | 10 | 90 |
| | Economically Disadvantaged | 69,164 | 41.89 | 14.84 | 10 | 90 |
| English Learner Status | Non-English Learner | 120,531 | 49.21 | 15.23 | 10 | 90 |
| | English Learner | 18,377 | 33.33 | 12.01 | 10 | 83 |
| Disabilities | Students without Disabilities | 111,690 | 49.44 | 14.88 | 10 | 90 |
| | Student with Disability (SWD) | 27,218 | 37.54 | 15.83 | 10 | 90 |
| **Writing Summative Score** | | **138,908** | **27.67** | **13.13** | **10** | **60** |
| Gender | Female | 67,706 | 30.25 | 12.71 | 10 | 60 |
| | Male | 71,144 | 25.22 | 13.06 | 10 | 60 |
| Ethnicity | American Indian/Alaska Native | 326 | 27.05 | 12.96 | 10 | 51 |
| | Asian | 7,527 | 35.24 | 11.20 | 10 | 60 |

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| | Black/African American | 22,566 | 22.20 | 12.45 | 10 | 60 |
| | Hispanic/Latino | 39,623 | 24.73 | 12.74 | 10 | 60 |
| | Native Hawaiian/Pacific Islander | 156 | 29.99 | 12.21 | 10 | 51 |
| | Two or more races | 5,269 | 27.63 | 13.33 | 10 | 60 |
| | White | 62,635 | 30.65 | 12.59 | 10 | 60 |
| Economic Status* | Not Economically Disadvantaged | 69,744 | 31.34 | 12.41 | 10 | 60 |
| | Economically Disadvantaged | 69,164 | 23.98 | 12.81 | 10 | 60 |
| English Learner Status | Non-English Learner | 120,531 | 28.97 | 12.91 | 10 | 60 |
| | English Learner | 18,377 | 19.18 | 11.30 | 10 | 60 |
| Disabilities | Students without Disabilities | 111,690 | 29.39 | 12.70 | 10 | 60 |
| | Student with Disability (SWD) | 27,218 | 20.62 | 12.53 | 10 | 60 |

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

Table A.12.41 Subgroup Performance for Mathematics Scale Scores: Grade 3

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **128,109** | **731.72** | **37.74** | **650** | **850** |
| Gender | Female | 62,609 | 730.17 | 36.29 | 650 | 850 |
| | Male | 65,494 | 733.21 | 39.03 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 336 | 721.81 | 36.82 | 650 | 850 |
| | Asian | 7,235 | 758.43 | 37.67 | 650 | 850 |
| | Black/African American | 21,145 | 709.62 | 32.72 | 650 | 850 |
| | Hispanic/Latino | 34,085 | 719.26 | 33.00 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 99 | 737.60 | 38.22 | 650 | 840 |
| | Two or more races | 5,576 | 735.17 | 39.15 | 650 | 850 |
| | White | 58,627 | 743.49 | 35.29 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,219 | 746.60 | 35.99 | 650 | 850 |
| | Economically Disadvantaged | 64,890 | 717.23 | 33.54 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,065 | 735.34 | 38.10 | 650 | 850 |
| | English Learner | 25,044 | 716.86 | 32.22 | 650 | 850 |
| Disabilities | Students without Disabilities | 103,628 | 735.88 | 36.39 | 650 | 850 |
| | Student with Disability (SWD) | 24,481 | 714.14 | 38.32 | 650 | 850 |
| Language Form | Spanish | 3,541 | 707.45 | 29.67 | 650 | 830 |

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

Table A.12.42 Subgroup Performance for Mathematics Scale Scores: Grade 4

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **127,833** | **728.98** | **34.50** | **650** | **850** |
| Gender | Female | 62,608 | 727.39 | 33.02 | 650 | 850 |
| | Male | 65,214 | 730.50 | 35.79 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 319 | 725.79 | 36.71 | 650 | 850 |
| | Asian | 7,375 | 756.35 | 34.92 | 650 | 850 |
| | Black/African American | 20,862 | 707.17 | 29.04 | 650 | 850 |
| | Hispanic/Latino | 33,965 | 717.70 | 29.99 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 108 | 738.05 | 36.71 | 650 | 836 |
| | Two or more races | 5,522 | 732.38 | 35.73 | 650 | 850 |
| | White | 58,698 | 739.67 | 31.91 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 63,451 | 742.81 | 33.11 | 650 | 850 |
| | Economically Disadvantaged | 64,382 | 715.35 | 30.14 | 650 | 850 |
| English Learner Status | Non-English Learner | 103,387 | 732.54 | 34.78 | 650 | 850 |
| | English Learner | 24,446 | 713.91 | 28.75 | 650 | 850 |
| Disabilities | Students without Disabilities | 102,690 | 733.22 | 33.11 | 650 | 850 |
| | Student with Disability (SWD) | 25,143 | 711.64 | 34.61 | 650 | 850 |
| Language Form | Spanish | 2,588 | 703.17 | 26.05 | 650 | 793 |

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

Table A.12.43 Subgroup Performance for Mathematics Scale Scores: Grade 5

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **129,562** | **727.88** | **33.19** | **650** | **850** |
| Gender | Female | 63,522 | 726.90 | 31.47 | 650 | 850 |
| | Male | 66,021 | 728.83 | 34.73 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 301 | 721.97 | 31.07 | 650 | 815 |
| | Asian | 7,429 | 756.44 | 34.95 | 650 | 850 |
| | Black/African American | 20,951 | 707.20 | 26.89 | 650 | 828 |
| | Hispanic/Latino | 35,354 | 717.46 | 28.30 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 130 | 735.32 | 28.37 | 673 | 792 |
| | Two or more races | 5,423 | 730.35 | 35.24 | 650 | 850 |
| | White | 59,008 | 737.79 | 31.20 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 64,614 | 741.30 | 32.27 | 650 | 850 |
| | Economically Disadvantaged | 64,948 | 714.53 | 28.35 | 650 | 850 |
| English Learner Status | Non-English Learner | 108,236 | 731.56 | 33.39 | 650 | 850 |
| | English Learner | 21,326 | 709.23 | 24.85 | 650 | 832 |
| Disabilities | Students without Disabilities | 103,531 | 732.09 | 32.01 | 650 | 850 |
| | Student with Disability (SWD) | 26,031 | 711.14 | 32.48 | 650 | 850 |
| Language Form | Spanish | 2,321 | 703.94 | 25.44 | 650 | 799 |

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

Table A.12.44 Subgroup Performance for Mathematics Scale Scores: Grade 6

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **132,858** | **725.16** | **32.56** | **650** | **850** |
| Gender | Female | 64,954 | 724.54 | 31.52 | 650 | 850 |
| | Male | 67,883 | 725.75 | 33.52 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 294 | 716.72 | 33.33 | 650 | 834 |
| | Asian | 7,384 | 754.51 | 33.76 | 650 | 850 |
| | Black/African American | 21,532 | 705.47 | 26.66 | 650 | 838 |
| | Hispanic/Latino | 36,634 | 714.74 | 28.19 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 138 | 726.15 | 32.67 | 650 | 807 |
| | Two or more races | 5,346 | 727.55 | 33.82 | 650 | 850 |
| | White | 60,688 | 734.83 | 30.45 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 66,656 | 737.95 | 31.58 | 650 | 850 |
| | Economically Disadvantaged | 66,202 | 712.28 | 28.16 | 650 | 850 |
| English Learner Status | Non-English Learner | 113,483 | 728.85 | 32.47 | 650 | 850 |
| | English Learner | 19,375 | 703.55 | 23.44 | 650 | 816 |
| Disabilities | Students without Disabilities | 106,495 | 729.19 | 31.78 | 650 | 850 |
| | Student with Disability (SWD) | 26,363 | 708.89 | 30.54 | 650 | 850 |
| Language Form | Spanish | 1,985 | 703.38 | 24.23 | 650 | 792 |

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

Table A.12.45 Subgroup Performance for Mathematics Scale Scores: Grade 7

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **133,956** | **731.18** | **30.21** | **650** | **850** |
| Gender | Female | 64,986 | 730.40 | 29.42 | 650 | 850 |
| | Male | 68,931 | 731.91 | 30.93 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 287 | 725.57 | 30.33 | 650 | 825 |
| | Asian | 7,429 | 757.52 | 32.18 | 650 | 850 |
| | Black/African American | 21,990 | 713.36 | 25.35 | 650 | 838 |
| | Hispanic/Latino | 37,044 | 722.28 | 26.62 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 117 | 734.21 | 33.82 | 665 | 838 |
| | Two or more races | 5,221 | 733.33 | 31.16 | 650 | 850 |
| | White | 61,037 | 739.74 | 28.19 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 67,455 | 742.52 | 29.17 | 650 | 850 |
| | Economically Disadvantaged | 66,501 | 719.68 | 26.70 | 650 | 850 |
| English Learner Status | Non-English Learner | 115,033 | 734.51 | 29.92 | 650 | 850 |
| | English Learner | 18,923 | 710.91 | 23.29 | 650 | 817 |
| Disabilities | Students without Disabilities | 107,492 | 735.57 | 28.64 | 650 | 850 |
| | Student with Disability (SWD) | 26,464 | 713.36 | 29.88 | 650 | 850 |
| Language Form | Spanish | 1,485 | 705.54 | 21.19 | 650 | 767 |

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

Table A.12.46 Subgroup Performance for Mathematics Scale Scores: Grade 8

| Group Type | Subgroup | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| **Full Summative Score** | | **138,558** | **723.86** | **41.04** | **650** | **850** |
| Gender | Female | 67,515 | 724.53 | 40.07 | 650 | 850 |
| | Male | 70,985 | 723.21 | 41.93 | 650 | 850 |
| Ethnicity | American Indian/Alaska Native | 324 | 721.63 | 39.11 | 650 | 831 |
| | Asian | 7,526 | 762.47 | 44.82 | 650 | 850 |
| | Black/African American | 22,478 | 701.11 | 32.32 | 650 | 850 |
| | Hispanic/Latino | 39,512 | 711.95 | 35.50 | 650 | 850 |
| | Native Hawaiian/Pacific Islander | 156 | 733.37 | 42.67 | 650 | 850 |
| | Two or more races | 5,253 | 725.71 | 42.77 | 650 | 850 |
| | White | 62,510 | 734.96 | 39.41 | 650 | 850 |
| Economic Status* | Not Economically Disadvantaged | 69,641 | 738.67 | 41.07 | 650 | 850 |
| | Economically Disadvantaged | 68,917 | 708.90 | 35.16 | 650 | 850 |
| English Learner Status | Non-English Learner | 120,229 | 728.17 | 41.02 | 650 | 850 |
| | English Learner | 18,329 | 695.64 | 27.90 | 650 | 843 |
| Disabilities | Students without Disabilities | 111,472 | 729.27 | 40.11 | 650 | 850 |
| | Student with Disability (SWD) | 27,086 | 701.60 | 37.16 | 650 | 850 |
| Language Form | Spanish | 1,489 | 689.72 | 24.88 | 650 | 812 |

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

# Appendix 13.1: Reliability by Content and Grade/Subject

Table A.13.1 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 3

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 54 | 3.18 | 0.86 | 1,287 | 0.76 | 62,159 | 0.90 |
| **Gender** | | | | | | | |
| Male | 54 | 3.10 | 0.85 | 824 | 0.76 | 31,493 | 0.90 |
| Female | 54 | 3.26 | 0.86 | 463 | 0.75 | 30,664 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 54 | 2.83 | 0.83 | 292 | 0.72 | 10,174 | 0.89 |
| Asian/Pacific Islander | 54 | 3.84 | 0.89 | 3,628 | 0.89 | 3,628 | 0.89 |
| Hispanic/Latino | 54 | 2.96 | 0.82 | 385 | 0.62 | 16,537 | 0.89 |
| American Indian/Alaska Native | 54 | 3.20 | 0.89 | 153 | 0.89 | 171 | 0.90 |
| Multiple | 54 | 3.60 | 0.90 | 2,637 | 0.90 | 2,637 | 0.90 |
| White | 54 | 3.35 | 0.85 | 541 | 0.78 | 28,691 | 0.89 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 54 | 2.95 | 0.83 | 868 | 0.72 | 31,252 | 0.89 |
| Not Economically Disadvantaged | 54 | 3.40 | 0.86 | 419 | 0.79 | 30,907 | 0.89 |
| English Learner | 54 | 2.85 | 0.79 | 283 | 0.54 | 12,118 | 0.88 |
| Non-English Learner | 54 | 3.25 | 0.86 | 1,004 | 0.78 | 50,041 | 0.90 |
| Students with Disabilities | 54 | 2.92 | 0.86 | 1,287 | 0.76 | 10,399 | 0.90 |
| Students without Disabilities | 54 | 3.57 | 0.89 | 51,956 | 0.89 | 51,760 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 54 | 3.04 | 0.81 | 236 | 0.81 | 236 | 0.81 |
| Non-Screen Reader | 54 | 3.20 | 0.82 | 187 | 0.82 | 187 | 0.82 |
| Screen Reader | 54 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 54 | 2.60 | 0.75 | 2,647 | 0.75 | 2,647 | 0.75 |

Note: n/r = not reported.

Table A.13.2 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 4

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 70 | 3.86 | 0.83 | 1,489 | 0.74 | 61,639 | 0.90 |
| **Gender** | | | | | | | |
| Male | 70 | 3.75 | 0.83 | 958 | 0.71 | 31,187 | 0.89 |
| Female | 70 | 3.96 | 0.84 | 637 | 0.76 | 30,447 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 70 | 3.44 | 0.80 | 314 | 0.68 | 10,011 | 0.88 |
| Asian/Pacific Islander | 70 | 4.74 | 0.87 | 3,779 | 0.85 | 3,641 | 0.89 |
| Hispanic/Latino | 70 | 3.62 | 0.79 | 513 | 0.66 | 16,247 | 0.88 |
| American Indian/Alaska Native | 70 | 4.25 | 0.88 | 157 | 0.87 | 152 | 0.90 |
| Multiple | 70 | 4.40 | 0.88 | 2,601 | 0.87 | 2,755 | 0.90 |
| White | 70 | 4.00 | 0.83 | 602 | 0.77 | 28,511 | 0.88 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 70 | 3.59 | 0.80 | 1,036 | 0.68 | 30,712 | 0.88 |
| Not Economically Disadvantaged | 70 | 4.08 | 0.84 | 453 | 0.79 | 30,927 | 0.88 |
| English Learner | 70 | 3.50 | 0.78 | 364 | 0.59 | 11,615 | 0.86 |
| Non-English Learner | 70 | 3.93 | 0.84 | 1,125 | 0.76 | 50,024 | 0.89 |
| Students with Disabilities | 70 | 3.56 | 0.84 | 1,489 | 0.74 | 10,545 | 0.90 |
| Students without Disabilities | 70 | 4.37 | 0.87 | 51,601 | 0.86 | 51,094 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 67 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 67 | 4.22 | 0.87 | 121 | 0.87 | 121 | 0.87 |
| Human Reader | 67 | 3.61 | 0.85 | 196 | 0.85 | 196 | 0.85 |
| Non-Screen Reader | 67 | 3.52 | 0.79 | 264 | 0.79 | 264 | 0.79 |
| Screen Reader | 67 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 74 | 2.81 | 0.74 | 3,018 | 0.74 | 3,018 | 0.74 |

Note: n/r = not reported.

Table A.13.3 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 5

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 70 | 3.89 | 0.86 | 1,493 | 0.75 | 62,583 | 0.89 |
| **Gender** | | | | | | | |
| Male | 70 | 3.73 | 0.86 | 965 | 0.74 | 31,701 | 0.90 |
| Female | 70 | 4.05 | 0.86 | 528 | 0.77 | 30,874 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 70 | 3.39 | 0.81 | 317 | 0.65 | 10,097 | 0.87 |
| Asian/Pacific Islander | 70 | 4.84 | 0.88 | 3,659 | 0.87 | 3,820 | 0.89 |
| Hispanic/Latino | 70 | 3.61 | 0.83 | 468 | 0.72 | 17,004 | 0.88 |
| American Indian/Alaska Native | 70 | 4.10 | 0.88 | 139 | 0.87 | 151 | 0.89 |
| Multiple | 70 | 4.48 | 0.89 | 2,582 | 0.88 | 2,674 | 0.90 |
| White | 70 | 4.08 | 0.85 | 607 | 0.78 | 28,528 | 0.88 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 70 | 3.58 | 0.83 | 1,048 | 0.72 | 31,100 | 0.88 |
| Not Economically Disadvantaged | 70 | 4.17 | 0.86 | 445 | 0.78 | 607 | 0.88 |
| English Learner | 70 | 3.30 | 0.79 | 380 | 0.67 | 179 | 0.85 |
| Non-English Learner | 70 | 3.99 | 0.86 | 1,113 | 0.76 | 52,509 | 0.89 |
| Students with Disabilities | 70 | 3.53 | 0.86 | 1,493 | 0.75 | 10,999 | 0.90 |
| Students without Disabilities | 70 | 4.41 | 0.87 | 138 | 0.81 | 51,584 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 74 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 74 | 3.92 | 0.91 | 195 | 0.91 | 195 | 0.91 |
| Human Reader | 74 | 3.59 | 0.81 | 274 | 0.81 | 274 | 0.81 |
| Non-Screen Reader | 74 | 3.59 | 0.82 | 335 | 0.82 | 335 | 0.82 |
| Screen Reader | 74 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 74 | 2.85 | 0.78 | 3,068 | 0.78 | 3,068 | 0.78 |

Note: n/r = not reported.

Table A.13.4 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 6

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 72 | 4.29 | 0.86 | 1,594 | 0.80 | 64,326 | 0.90 |
| **Gender** | | | | | | | |
| Male | 72 | 4.09 | 0.86 | 1,030 | 0.79 | 32,575 | 0.90 |
| Female | 72 | 4.46 | 0.86 | 564 | 0.81 | 31,742 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 72 | 3.78 | 0.82 | 352 | 0.64 | 10,299 | 0.89 |
| Asian/Pacific Islander | 72 | 5.14 | 0.87 | 3,781 | 0.86 | 3,679 | 0.89 |
| Hispanic/Latino | 72 | 3.96 | 0.84 | 511 | 0.74 | 17,771 | 0.89 |
| American Indian/Alaska Native | 72 | 4.74 | 0.90 | 148 | 0.88 | 131 | 0.92 |
| Multiple | 72 | 4.93 | 0.88 | 2,607 | 0.87 | 2,605 | 0.90 |
| White | 72 | 4.48 | 0.86 | 624 | 0.83 | 289 | 0.89 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 72 | 3.96 | 0.83 | 1,062 | 0.74 | 31,774 | 0.89 |
| Not Economically Disadvantaged | 72 | 4.54 | 0.87 | 532 | 0.84 | 32,552 | 0.89 |
| English Learner | 72 | 3.53 | 0.79 | 392 | 0.66 | 121 | 0.84 |
| Non-English Learner | 72 | 4.39 | 0.86 | 1,202 | 0.81 | 55,062 | 0.89 |
| Students with Disabilities | 72 | 3.86 | 0.87 | 1,594 | 0.80 | 11,116 | 0.91 |
| Students without Disabilities | 72 | 4.91 | 0.87 | 128 | 0.84 | 53,210 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 74 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 74 | 4.31 | 0.90 | 136 | 0.90 | 136 | 0.90 |
| Human Reader | 74 | 3.72 | 0.86 | 238 | 0.86 | 238 | 0.86 |
| Non-Screen Reader | 74 | 3.46 | 0.88 | 236 | 0.88 | 236 | 0.88 |
| Screen Reader | 74 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 74 | 3.22 | 0.79 | 3,157 | 0.79 | 3,157 | 0.79 |

Note: n/r = not reported.

Table A.13.5 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 7

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 72 | 4.52 | 0.86 | 1,511 | 0.77 | 65,225 | 0.89 |
| **Gender** | | | | | | | |
| Male | 72 | 4.28 | 0.86 | 977 | 0.77 | 33,111 | 0.90 |
| Female | 72 | 4.72 | 0.86 | 534 | 0.76 | 32,090 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 72 | 3.94 | 0.82 | 361 | 0.69 | 10,599 | 0.87 |
| Asian/Pacific Islander | 72 | 5.36 | 0.88 | 3,789 | 0.87 | 3,713 | 0.89 |
| Hispanic/Latino | 72 | 4.13 | 0.82 | 469 | 0.63 | 18,033 | 0.88 |
| American Indian/Alaska Native | 72 | 4.95 | 0.88 | 146 | 0.87 | 131 | 0.89 |
| Multiple | 72 | 5.11 | 0.89 | 2,542 | 0.88 | 2,550 | 0.90 |
| White | 72 | 4.71 | 0.86 | 603 | 0.81 | 680 | 0.88 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 72 | 4.14 | 0.83 | 1,025 | 0.71 | 32,248 | 0.88 |
| Not Economically Disadvantaged | 72 | 4.81 | 0.86 | 486 | 0.81 | 32,977 | 0.89 |
| English Learner | 72 | 3.71 | 0.77 | 372 | 0.63 | 9,055 | 0.82 |
| Non-English Learner | 72 | 4.61 | 0.86 | 1,139 | 0.79 | 56,170 | 0.89 |
| Students with Disabilities | 72 | 4.07 | 0.86 | 1,511 | 0.77 | 11,285 | 0.90 |
| Students without Disabilities | 72 | 5.16 | 0.87 | 128 | 0.76 | 53,940 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 70 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 74 | 4.47 | 0.91 | 199 | 0.91 | 199 | 0.91 |
| Human Reader | 70 | 4.16 | 0.81 | 239 | 0.81 | 239 | 0.81 |
| Non-Screen Reader | 70 | 4.05 | 0.81 | 191 | 0.81 | 191 | 0.81 |
| Screen Reader | 70 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 74 | 3.30 | 0.78 | 3,027 | 0.78 | 3,027 | 0.78 |

Note: n/r = not reported.

Table A.13.6 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 8

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 72 | 4.52 | 0.86 | 1,511 | 0.77 | 65,225 | 0.89 |
| **Gender** | | | | | | | |
| Male | 72 | 4.28 | 0.86 | 977 | 0.77 | 33,111 | 0.90 |
| Female | 72 | 4.72 | 0.86 | 534 | 0.76 | 32,090 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 72 | 3.94 | 0.82 | 361 | 0.69 | 10,599 | 0.87 |
| Asian/Pacific Islander | 72 | 5.36 | 0.88 | 3,789 | 0.87 | 3,713 | 0.89 |
| Hispanic/Latino | 72 | 4.13 | 0.82 | 469 | 0.63 | 18,033 | 0.88 |
| American Indian/Alaska Native | 72 | 4.95 | 0.88 | 146 | 0.87 | 131 | 0.89 |
| Multiple | 72 | 5.11 | 0.89 | 2,542 | 0.88 | 2,550 | 0.90 |
| White | 72 | 4.71 | 0.86 | 603 | 0.81 | 680 | 0.88 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 72 | 4.14 | 0.83 | 1,025 | 0.71 | 32,248 | 0.88 |
| Not Economically Disadvantaged | 72 | 4.81 | 0.86 | 486 | 0.81 | 32,977 | 0.89 |
| English Learner | 72 | 3.71 | 0.77 | 372 | 0.63 | 9,055 | 0.82 |
| Non-English Learner | 72 | 4.61 | 0.86 | 1,139 | 0.79 | 56,170 | 0.89 |
| Students with Disabilities | 72 | 4.07 | 0.86 | 1,511 | 0.77 | 11,285 | 0.90 |
| Students without Disabilities | 72 | 5.16 | 0.87 | 128 | 0.76 | 53,940 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 70 | n/r | n/r | n/r | n/r | n/r | n/r |
| Closed Caption | 74 | 4.47 | 0.91 | 199 | 0.91 | 199 | 0.91 |
| Human Reader | 70 | 4.16 | 0.81 | 239 | 0.81 | 239 | 0.81 |
| Non-Screen Reader | 70 | 4.05 | 0.81 | 191 | 0.81 | 191 | 0.81 |
| Screen Reader | 70 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 74 | 3.30 | 0.78 | 3,027 | 0.78 | 3,027 | 0.78 |

Note: n/r = not reported.

Table A.13.7 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 3

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.33 | 0.90 | 41,861 | 0.89 | 24,430 | 0.91 |
| **Gender** | | | | | | | |
| Male | 52 | 3.33 | 0.90 | 21,017 | 0.89 | 12,736 | 0.91 |
| Female | 52 | 3.34 | 0.89 | 20,840 | 0.88 | 11,692 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.94 | 0.89 | 171 | 0.88 | 5,514 | 0.89 |
| Asian/Pacific Islander | 52 | 3.61 | 0.89 | 2,709 | 0.86 | 989 | 0.92 |
| Hispanic/Latino | 52 | 3.15 | 0.88 | 10,772 | 0.88 | 243 | 0.89 |
| American Indian/Alaska Native | 52 | 3.06 | 0.89 | 116 | 0.89 | 116 | 0.89 |
| Multiple | 52 | 3.41 | 0.91 | 2,226 | 0.90 | 617 | 0.93 |
| White | 52 | 3.46 | 0.89 | 23,490 | 0.87 | 6,293 | 0.92 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 3.10 | 0.89 | 17,515 | 0.88 | 574 | 0.89 |
| Not Economically Disadvantaged | 52 | 3.51 | 0.89 | 24,346 | 0.87 | 7,784 | 0.91 |
| English Learner | 52 | 3.08 | 0.88 | 219 | 0.88 | 5,846 | 0.89 |
| Non-English Learner | 52 | 3.38 | 0.90 | 36,850 | 0.89 | 15,273 | 0.92 |
| Students with Disabilities | 52 | 3.13 | 0.90 | 6,282 | 0.89 | 5,532 | 0.91 |
| Students without Disabilities | 52 | 3.38 | 0.89 | 36,006 | 0.88 | 18,148 | 0.91 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 3.09 | 0.89 | 460 | 0.89 | 460 | 0.89 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.17 | 0.90 | 42,064 | 0.90 | 42,064 | 0.90 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.77 | 0.87 | 1,733 | 0.86 | 1,787 | 0.87 |

Note: n/r = not reported.

Table A.13.8 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 4

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.36 | 0.89 | 994 | 0.89 | 20,775 | 0.91 |
| **Gender** | | | | | | | |
| Male | 52 | 3.38 | 0.90 | 562 | 0.89 | 10,890 | 0.91 |
| Female | 52 | 3.33 | 0.89 | 20,633 | 0.88 | 9,884 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.88 | 0.87 | 115 | 0.85 | 5,273 | 0.87 |
| Asian/Pacific Islander | 52 | 3.71 | 0.89 | 2,816 | 0.86 | 962 | 0.92 |
| Hispanic/Latino | 52 | 3.16 | 0.87 | 214 | 0.85 | 7,710 | 0.88 |
| American Indian/Alaska Native | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Multiple | 52 | 3.44 | 0.91 | 2,071 | 0.90 | 652 | 0.92 |
| White | 52 | 3.49 | 0.89 | 23,310 | 0.86 | 5,991 | 0.92 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 3.09 | 0.87 | 472 | 0.87 | 13,659 | 0.88 |
| Not Economically Disadvantaged | 52 | 3.54 | 0.89 | 24,228 | 0.87 | 7,693 | 0.91 |
| English Learner | 52 | 3.08 | 0.86 | 188 | 0.86 | 5,888 | 0.87 |
| Non-English Learner | 52 | 3.40 | 0.89 | 36,420 | 0.88 | 15,364 | 0.91 |
| Students with Disabilities | 52 | 3.12 | 0.89 | 6,815 | 0.86 | 5,592 | 0.90 |
| Students without Disabilities | 52 | 3.42 | 0.89 | 138 | 0.88 | 17,097 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 3.07 | 0.90 | 375 | 0.90 | 375 | 0.90 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.18 | 0.90 | 42,121 | 0.90 | 42,121 | 0.90 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.74 | 0.83 | 1,147 | 0.82 | 1,419 | 0.84 |

Note: n/r = not reported.

Table A.13.9 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 5

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.40 | 0.88 | 953 | 0.85 | 21,191 | 0.90 |
| **Gender** | | | | | | | |
| Male | 52 | 3.41 | 0.89 | 550 | 0.87 | 11,137 | 0.91 |
| Female | 52 | 3.39 | 0.87 | 403 | 0.82 | 10,052 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.93 | 0.84 | 122 | 0.81 | 5,463 | 0.85 |
| Asian/Pacific Islander | 52 | 3.75 | 0.90 | 2,850 | 0.88 | 921 | 0.92 |
| Hispanic/Latino | 52 | 3.20 | 0.85 | 192 | 0.81 | 8,191 | 0.86 |
| American Indian/Alaska Native | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Multiple | 52 | 3.45 | 0.90 | 2,082 | 0.89 | 625 | 0.92 |
| White | 52 | 3.54 | 0.88 | 261 | 0.86 | 5,972 | 0.91 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 3.13 | 0.85 | 443 | 0.84 | 14,257 | 0.86 |
| Not Economically Disadvantaged | 52 | 3.58 | 0.88 | 510 | 0.86 | 6,934 | 0.91 |
| English Learner | 52 | 2.96 | 0.81 | 4,094 | 0.80 | 162 | 0.82 |
| Non-English Learner | 52 | 3.46 | 0.89 | 791 | 0.85 | 15,666 | 0.91 |
| Students with Disabilities | 52 | 3.15 | 0.87 | 7,075 | 0.85 | 5,763 | 0.89 |
| Students without Disabilities | 52 | 3.46 | 0.88 | 115 | 0.79 | 14,516 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 3.05 | 0.85 | 326 | 0.85 | 326 | 0.85 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.14 | 0.89 | 42,368 | 0.89 | 42,368 | 0.89 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.71 | 0.82 | 950 | 0.81 | 1,350 | 0.83 |

Note: n/r = not reported.

Table A.13.10 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 6

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.38 | 0.88 | 814 | 0.84 | 20,821 | 0.89 |
| **Gender** | | | | | | | |
| Male | 52 | 3.42 | 0.88 | 464 | 0.85 | 10,957 | 0.90 |
| Female | 52 | 3.33 | 0.88 | 350 | 0.83 | 9,862 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.72 | 0.83 | 5,579 | 0.81 | 5,565 | 0.84 |
| Asian/Pacific Islander | 52 | 4.00 | 0.89 | 2,942 | 0.88 | 836 | 0.92 |
| Hispanic/Latino | 52 | 3.06 | 0.85 | 10,049 | 0.84 | 151 | 0.86 |
| American Indian/Alaska Native | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Multiple | 52 | 3.48 | 0.90 | 2,082 | 0.88 | 591 | 0.91 |
| White | 52 | 3.59 | 0.88 | 237 | 0.85 | 5,596 | 0.91 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 2.97 | 0.85 | 359 | 0.80 | 13,993 | 0.86 |
| Not Economically Disadvantaged | 52 | 3.68 | 0.88 | 455 | 0.85 | 7,496 | 0.90 |
| English Learner | 52 | 2.60 | 0.77 | 6,572 | 0.74 | 4,865 | 0.78 |
| Non-English Learner | 52 | 3.48 | 0.88 | 708 | 0.85 | 15,956 | 0.90 |
| Students with Disabilities | 52 | 3.03 | 0.86 | 7,168 | 0.81 | 5,805 | 0.89 |
| Students without Disabilities | 52 | 3.48 | 0.88 | 15,847 | 0.88 | 14,018 | 0.89 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 2.78 | 0.82 | 254 | 0.82 | 254 | 0.82 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.06 | 0.88 | 41,864 | 0.88 | 41,864 | 0.88 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.51 | 0.79 | 720 | 0.72 | 1,252 | 0.82 |

Note: n/r = not reported.

Table A.13.11 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 7

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.34 | 0.89 | 45,382 | 0.88 | 20,451 | 0.91 |
| **Gender** | | | | | | | |
| Male | 52 | 3.32 | 0.90 | 23,188 | 0.89 | 10,908 | 0.91 |
| Female | 52 | 3.35 | 0.88 | 22,565 | 0.87 | 9,543 | 0.90 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.73 | 0.84 | 5,355 | 0.81 | 5,535 | 0.86 |
| Asian/Pacific Islander | 52 | 4.06 | 0.89 | 2,953 | 0.87 | 858 | 0.93 |
| Hispanic/Latino | 52 | 3.05 | 0.86 | 9,636 | 0.85 | 8,039 | 0.88 |
| American Indian/Alaska Native | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Multiple | 52 | 3.41 | 0.90 | 2,040 | 0.89 | 488 | 0.93 |
| White | 52 | 3.51 | 0.89 | 25,081 | 0.87 | 5,406 | 0.92 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 2.96 | 0.86 | 18,746 | 0.85 | 13,755 | 0.88 |
| Not Economically Disadvantaged | 52 | 3.62 | 0.89 | 26,718 | 0.87 | 6,696 | 0.92 |
| English Learner | 52 | 2.64 | 0.79 | 6,097 | 0.77 | 126 | 0.84 |
| Non-English Learner | 52 | 3.42 | 0.89 | 41,361 | 0.88 | 15,742 | 0.91 |
| Students with Disabilities | 52 | 2.93 | 0.88 | 6,817 | 0.83 | 6,024 | 0.90 |
| Students without Disabilities | 52 | 3.45 | 0.89 | 39,178 | 0.88 | 13,673 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 2.57 | 0.87 | 136 | 0.87 | 136 | 0.87 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 3.08 | 0.89 | 41,016 | 0.89 | 41,016 | 0.89 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.41 | 0.70 | 737 | 0.68 | 713 | 0.74 |

Note: n/r = not reported.

Table A.13.12 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 8

| | Avg. Max. Raw Score | Avg. SEM | Avg. Reliability | Minimum Reliability N | Minimum Reliability Alpha | Maximum Reliability N | Maximum Reliability Alpha |
|---|---|---|---|---|---|---|---|
| **Total Group** | 52 | 3.09 | 0.88 | 712 | 0.81 | 22,308 | 0.89 |
| **Gender** | | | | | | | |
| Male | 52 | 3.06 | 0.89 | 397 | 0.83 | 11,720 | 0.90 |
| Female | 52 | 3.11 | 0.87 | 315 | 0.77 | 10,584 | 0.89 |
| **Ethnicity** | | | | | | | |
| Black/African American | 52 | 2.59 | 0.81 | 5,319 | 0.79 | 5,575 | 0.83 |
| Asian/Pacific Islander | 52 | 3.72 | 0.90 | 2,894 | 0.89 | 803 | 0.92 |
| Hispanic/Latino | 52 | 2.84 | 0.85 | 146 | 0.80 | 9,888 | 0.86 |
| American Indian/Alaska Native | 52 | 3.10 | 0.86 | 108 | 0.86 | 108 | 0.86 |
| Multiple | 52 | 3.13 | 0.90 | 2,061 | 0.88 | 534 | 0.92 |
| White | 52 | 3.25 | 0.87 | 151 | 0.71 | 5,352 | 0.91 |
| **Special Instruction Needs** | | | | | | | |
| Economically Disadvantaged | 52 | 2.78 | 0.83 | 330 | 0.67 | 15,247 | 0.86 |
| Not Economically Disadvantaged | 52 | 3.33 | 0.89 | 382 | 0.85 | 7,061 | 0.91 |
| English Learner | 52 | 2.42 | 0.71 | 111 | 0.43 | 3,773 | 0.76 |
| Non-English Learner | 52 | 3.16 | 0.88 | 601 | 0.82 | 16,273 | 0.90 |
| Students with Disabilities | 52 | 2.71 | 0.85 | 6,737 | 0.80 | 6,526 | 0.89 |
| Students without Disabilities | 52 | 3.19 | 0.89 | 39,220 | 0.87 | 15,488 | 0.90 |
| **Students Taking Accommodated Forms** | | | | | | | |
| ASL | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Human Reader | 52 | 2.66 | 0.69 | 140 | 0.69 | 140 | 0.69 |
| Non-Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Screen Reader | 52 | n/r | n/r | n/r | n/r | n/r | n/r |
| Text-to-Speech | 52 | 2.74 | 0.90 | 41,937 | 0.90 | 41,937 | 0.90 |
| **Students Taking Translated Forms** | | | | | | | |
| Spanish Language | 52 | 2.24 | 0.65 | 810 | 0.63 | 649 | 0.69 |

Note: n/r = not reported.

# Appendix 13.2: Reliability of Classification by Grade/Subject

Table A.13.18 Reliability of Classification: Grade 3 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.27 | 0.04 | 0.00 | 0.00 | 0.00 | 0.30 |
| | 700-724 | 0.05 | 0.11 | 0.05 | 0.00 | 0.00 | 0.21 |
| | 725-749 | 0.00 | 0.04 | 0.11 | 0.05 | 0.00 | 0.21 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.22 | 0.02 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Decision Consistency | 650-699 | 0.26 | 0.05 | 0.01 | 0.00 | 0.00 | 0.32 |
| | 700-724 | 0.05 | 0.08 | 0.05 | 0.01 | 0.00 | 0.20 |
| | 725-749 | 0.01 | 0.05 | 0.08 | 0.05 | 0.00 | 0.19 |
| | 750-809 | 0.00 | 0.01 | 0.06 | 0.20 | 0.01 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |

Table A.13.19 Reliability of Classification: Grade 4 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.14 | 0.03 | 0.00 | 0.00 | 0.00 | 0.18 |
| | 700-724 | 0.04 | 0.12 | 0.05 | 0.00 | 0.00 | 0.22 |
| | 725-749 | 0.00 | 0.05 | 0.14 | 0.05 | 0.00 | 0.25 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.21 | 0.04 | 0.30 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.04 | 0.05 |
| Decision Consistency | 650-699 | 0.14 | 0.05 | 0.01 | 0.00 | 0.00 | 0.20 |
| | 700-724 | 0.04 | 0.10 | 0.06 | 0.01 | 0.00 | 0.21 |
| | 725-749 | 0.01 | 0.06 | 0.11 | 0.06 | 0.00 | 0.24 |
| | 750-809 | 0.00 | 0.01 | 0.07 | 0.17 | 0.03 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.03 | 0.04 | 0.07 |

Table A.13.20 Reliability of Classification: Grade 5 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.13 | 0.03 | 0.00 | 0.00 | 0.00 | 0.16 |
| | 700-724 | 0.04 | 0.13 | 0.05 | 0.00 | 0.00 | 0.23 |
| | 725-749 | 0.00 | 0.06 | 0.16 | 0.06 | 0.00 | 0.27 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.25 | 0.02 | 0.34 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 |
| Decision Consistency | 650-699 | 0.12 | 0.05 | 0.01 | 0.00 | 0.00 | 0.18 |
| | 700-724 | 0.04 | 0.10 | 0.06 | 0.01 | 0.00 | 0.22 |
| | 725-749 | 0.01 | 0.06 | 0.12 | 0.07 | 0.00 | 0.25 |
| | 750-809 | 0.00 | 0.01 | 0.07 | 0.22 | 0.02 | 0.33 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.01 | 0.03 |

Table A.13.21 Reliability of Classification: Grade 6 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.12 | 0.03 | 0.00 | 0.00 | 0.00 | 0.15 |
| | 700-724 | 0.04 | 0.13 | 0.05 | 0.00 | 0.00 | 0.23 |
| | 725-749 | 0.00 | 0.06 | 0.17 | 0.06 | 0.00 | 0.29 |
| | 750-809 | 0.00 | 0.00 | 0.06 | 0.24 | 0.03 | 0.33 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 |
| Decision Consistency | 650-699 | 0.12 | 0.04 | 0.01 | 0.00 | 0.00 | 0.17 |
| | 700-724 | 0.04 | 0.11 | 0.06 | 0.01 | 0.00 | 0.22 |
| | 725-749 | 0.01 | 0.06 | 0.13 | 0.07 | 0.00 | 0.27 |
| | 750-809 | 0.00 | 0.01 | 0.07 | 0.20 | 0.03 | 0.31 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.01 | 0.04 |

Table A.13.22 Reliability of Classification: Grade 7 ELA/L

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.15 | 0.03 | 0.00 | 0.00 | 0.00 | 0.18 |
| | 700-724 | 0.04 | 0.12 | 0.05 | 0.00 | 0.00 | 0.20 |
| | 725-749 | 0.00 | 0.05 | 0.15 | 0.05 | 0.00 | 0.25 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.21 | 0.03 | 0.29 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.08 |
| Decision Consistency | 650-699 | 0.14 | 0.04 | 0.01 | 0.00 | 0.00 | 0.19 |
| | 700-724 | 0.04 | 0.09 | 0.06 | 0.01 | 0.00 | 0.20 |
| | 725-749 | 0.01 | 0.05 | 0.11 | 0.06 | 0.00 | 0.23 |
| | 750-809 | 0.00 | 0.01 | 0.06 | 0.17 | 0.04 | 0.28 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.10 |

Table A.13.23 Reliability of Classification: Grade 8 ELA/L

| | Full Summative Scale Score | 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 | 0.15 |
| | 700-724 | 0.03 | 0.12 | 0.04 | 0.00 | 0.00 | 0.20 |
| | 725-749 | 0.00 | 0.05 | 0.15 | 0.05 | 0.00 | 0.25 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.27 | 0.03 | 0.36 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.05 |
| Decision Consistency | 650-699 | 0.12 | 0.04 | 0.01 | 0.00 | 0.00 | 0.16 |
| | 700-724 | 0.03 | 0.09 | 0.06 | 0.01 | 0.00 | 0.19 |
| | 725-749 | 0.00 | 0.05 | 0.12 | 0.06 | 0.00 | 0.23 |
| | 750-809 | 0.00 | 0.01 | 0.07 | 0.23 | 0.03 | 0.34 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.07 |

Table A.13.24 Reliability of Classification: Grade 3 Mathematics

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.18 | 0.02 | 0.00 | 0.00 | 0.00 | 0.20 |
| | 700-724 | 0.03 | 0.17 | 0.03 | 0.00 | 0.00 | 0.23 |
| | 725-749 | 0.00 | 0.04 | 0.17 | 0.03 | 0.00 | 0.24 |
| | 750-809 | 0.00 | 0.00 | 0.03 | 0.22 | 0.02 | 0.27 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.05 | 0.06 |
| Decision Consistency | 650-699 | 0.17 | 0.04 | 0.00 | 0.00 | 0.00 | 0.21 |
| | 700-724 | 0.03 | 0.15 | 0.05 | 0.00 | 0.00 | 0.23 |
| | 725-749 | 0.00 | 0.05 | 0.14 | 0.04 | 0.00 | 0.24 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.20 | 0.02 | 0.26 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.04 | 0.06 |

Table A.13.25 Reliability of Classification: Grade 4 Mathematics

| | Full Summative Scale Score | 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 | 0.03 | 0.19 | 0.04 | 0.00 | 0.00 | 0.26 |
| | 725-749 | 0.00 | 0.04 | 0.19 | 0.04 | 0.00 | 0.27 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.21 | 0.01 | 0.25 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 |
| Decision Consistency | 650-699 | 0.16 | 0.04 | 0.00 | 0.00 | 0.00 | 0.21 |
| | 700-724 | 0.04 | 0.16 | 0.06 | 0.00 | 0.00 | 0.25 |
| | 725-749 | 0.00 | 0.06 | 0.15 | 0.05 | 0.00 | 0.26 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.19 | 0.01 | 0.25 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.03 |

Table A.13.26 Reliability of Classification: Grade 5 Mathematics

| | Full Summative Scale Score | 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 | 0.21 | 0.04 | 0.00 | 0.00 | 0.29 |
| | 725-749 | 0.00 | 0.05 | 0.17 | 0.04 | 0.00 | 0.26 |
| | 750-809 | 0.00 | 0.00 | 0.03 | 0.17 | 0.01 | 0.21 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.03 |
| Decision Consistency | 650-699 | 0.16 | 0.05 | 0.00 | 0.00 | 0.00 | 0.22 |
| | 700-724 | 0.05 | 0.17 | 0.06 | 0.00 | 0.00 | 0.28 |
| | 725-749 | 0.00 | 0.07 | 0.14 | 0.05 | 0.00 | 0.25 |
| | 750-809 | 0.00 | 0.00 | 0.05 | 0.15 | 0.01 | 0.22 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.04 |

Table A.13.27 Reliability of Classification: Grade 6 Mathematics

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.18 | 0.04 | 0.00 | 0.00 | 0.00 | 0.21 |
| | 700-724 | 0.04 | 0.19 | 0.05 | 0.00 | 0.00 | 0.29 |
| | 725-749 | 0.00 | 0.06 | 0.17 | 0.05 | 0.00 | 0.28 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.15 | 0.02 | 0.21 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |
| Decision Consistency | 650-699 | 0.17 | 0.06 | 0.00 | 0.00 | 0.00 | 0.23 |
| | 700-724 | 0.05 | 0.15 | 0.07 | 0.01 | 0.00 | 0.27 |
| | 725-749 | 0.00 | 0.07 | 0.13 | 0.05 | 0.00 | 0.26 |
| | 750-809 | 0.00 | 0.01 | 0.06 | 0.13 | 0.02 | 0.21 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.03 |

Table A.13.28 Reliability of Classification: Grade 7 Mathematics

| | Full Summative Scale Score | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Category Total |
|---|---|---|---|---|---|---|---|
| Decision Accuracy | 650-699 | 0.10 | 0.03 | 0.00 | 0.00 | 0.00 | 0.13 |
| | 700-724 | 0.04 | 0.20 | 0.05 | 0.00 | 0.00 | 0.29 |
| | 725-749 | 0.00 | 0.07 | 0.20 | 0.05 | 0.00 | 0.32 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.17 | 0.02 | 0.24 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.02 |
| Decision Consistency | 650-699 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.15 |
| | 700-724 | 0.04 | 0.16 | 0.07 | 0.01 | 0.00 | 0.28 |
| | 725-749 | 0.00 | 0.08 | 0.16 | 0.06 | 0.00 | 0.30 |
| | 750-809 | 0.00 | 0.01 | 0.07 | 0.15 | 0.02 | 0.24 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.03 |

Table A.13.29 Reliability of Classification: Grade 8 Mathematics

| | Full Summative Scale Score | 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 | 0.29 |
| | 700-724 | 0.05 | 0.14 | 0.05 | 0.00 | 0.00 | 0.24 |
| | 725-749 | 0.00 | 0.05 | 0.12 | 0.04 | 0.00 | 0.21 |
| | 750-809 | 0.00 | 0.00 | 0.04 | 0.16 | 0.02 | 0.22 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.03 |
| Decision Consistency | 650-699 | 0.24 | 0.06 | 0.01 | 0.00 | 0.00 | 0.31 |
| | 700-724 | 0.05 | 0.11 | 0.06 | 0.01 | 0.00 | 0.23 |
| | 725-749 | 0.01 | 0.06 | 0.09 | 0.04 | 0.00 | 0.20 |
| | 750-809 | 0.00 | 0.01 | 0.05 | 0.15 | 0.02 | 0.22 |
| | 810-850 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.04 |

# Appendix 15: Growth

Appendix 15 provides the summary growth results for subgroups for grade 4–8 ELA/L and mathematics 4–8.

Table A.15.1 Summary of SGP Estimates for Subgroups: Grade 4 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 49,196 | 49.67 | 14.11 | 50 |
| Female | 47,458 | 50.31 | 14.06 | 50 |
| **Ethnicity** | | | | |
| White | 51,770 | 51.44 | 13.99 | 52 |
| African American | 12,578 | 42.71 | 14.41 | 40 |
| Asian/Pacific Islander | 5,241 | 59.02 | 13.78 | 63 |
| American Indian/Alaska Native | 207 | 45.45 | 14.43 | 44 |
| Hispanic | 21,886 | 48.48 | 14.24 | 48 |
| Multiple | 4,703 | 50.78 | 13.95 | 51 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 40,641 | 45.49 | 14.24 | 44 |
| Not-economically Disadvantaged | 56,014 | 53.25 | 13.98 | 55 |
| English Learner (EL) | 14,992 | 47.00 | 14.37 | 46 |
| Non-English Learner | 81,663 | 50.54 | 14.03 | 51 |
| Students with Disabilities (SWD) | 16,915 | 41.49 | 14.39 | 38 |
| Students without Disabilities | 79,740 | 51.79 | 14.02 | 53 |

Table A.15.2 Summary of SGP Estimates for Subgroups: Grade 5 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 50,517 | 48.42 | 14.33 | 48 |
| Female | 48,422 | 51.64 | 13.91 | 52 |
| **Ethnicity** | | | | |
| White | 53,370 | 50.53 | 13.85 | 51 |
| African American | 12,790 | 45.64 | 14.89 | 44 |
| Asian/Pacific Islander | 5,293 | 60.02 | 13.40 | 64 |
| American Indian/Alaska Native | 186 | 49.89 | 14.28 | 50.5 |
| Hispanic | 22,422 | 48.79 | 14.55 | 48 |
| Multiple | 4,596 | 50.66 | 14.12 | 51 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 41,252 | 46.51 | 14.59 | 45 |
| Not-economically Disadvantaged | 57,692 | 52.49 | 13.80 | 53 |
| English Learner (EL) | 12,853 | 46.21 | 15.07 | 44 |
| Non-English Learner | 86,091 | 50.56 | 13.99 | 51 |
| Students with Disabilities (SWD) | 17,328 | 41.53 | 15.22 | 38 |
| Students without Disabilities | 81,616 | 51.79 | 13.90 | 53 |

Table A.15.3 Summary of SGP Estimates for Subgroups: Grade 6 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 51,180 | 47.85 | 13.79 | 47 |
| Female | 48,204 | 52.28 | 13.52 | 53 |
| **Ethnicity** | | | | |
| White | 53,631 | 51.49 | 13.43 | 52 |
| African American | 12,982 | 45.27 | 14.30 | 43 |
| Asian/Pacific Islander | 5,245 | 56.51 | 13.44 | 59 |
| American Indian/Alaska Native | 173 | 47.55 | 13.67 | 46 |
| Hispanic | 22,641 | 47.79 | 13.91 | 47 |
| Multiple | 4,473 | 50.00 | 13.61 | 50 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 41,534 | 46.56 | 14.05 | 45 |
| Not-economically Disadvantaged | 57,858 | 52.47 | 13.38 | 54 |
| English Learner (EL) | 10,517 | 43.85 | 14.59 | 41 |
| Non-English Learner | 88,875 | 50.73 | 13.55 | 51 |
| Students with Disabilities (SWD) | 17,314 | 41.29 | 14.66 | 38 |
| Students without Disabilities | 82,078 | 51.84 | 13.45 | 53 |

Table A.15.4 Summary of SGP Estimates for Subgroups: Grade 7 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 50,999 | 47.61 | 13.98 | 47 |
| Female | 48,315 | 52.51 | 13.80 | 54 |
| **Ethnicity** | | | | |
| White | 53,630 | 51.74 | 13.84 | 52 |
| African American | 12,925 | 44.79 | 13.95 | 43 |
| Asian/Pacific Islander | 5,264 | 58.51 | 13.88 | 62 |
| American Indian/Alaska Native | 213 | 51.16 | 13.55 | 49 |
| Hispanic | 22,702 | 46.93 | 13.97 | 46 |
| Multiple | 4,322 | 49.70 | 13.96 | 50 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 40,845 | 46.27 | 13.91 | 45 |
| Not-economically Disadvantaged | 58,481 | 52.60 | 13.88 | 54 |
| English Learner (EL) | 9,526 | 42.89 | 14.16 | 40 |
| Non-English Learner | 89,800 | 50.75 | 13.86 | 51 |
| Students with Disabilities (SWD) | 17,135 | 42.45 | 14.14 | 40 |
| Students without Disabilities | 82,191 | 51.57 | 13.84 | 52 |

Table A.15.5 Summary of SGP Estimates for Subgroups: Grade 8 ELA/L

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 50,489 | 46.95 | 13.94 | 46 |
| Female | 47,703 | 53.18 | 13.84 | 55 |
| **Ethnicity** | | | | |
| White | 53,427 | 50.84 | 13.84 | 51 |
| African American | 13,197 | 46.73 | 14.10 | 45 |
| Asian/Pacific Islander | 5,096 | 57.57 | 13.93 | 61 |
| American Indian/Alaska Native | 218 | 48.89 | 13.58 | 50 |
| Hispanic | 21,949 | 48.16 | 13.88 | 47 |
| Multiple | 4,089 | 49.28 | 13.97 | 49 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 40,330 | 47.40 | 13.91 | 46 |
| Not-economically Disadvantaged | 57,871 | 51.77 | 13.88 | 52 |
| English Learner (EL) | 8,635 | 44.34 | 14.22 | 42 |
| Non-English Learner | 89,566 | 50.52 | 13.86 | 51 |
| Students with Disabilities (SWD) | 16,560 | 43.94 | 14.37 | 41 |
| Students without Disabilities | 81,641 | 51.20 | 13.79 | 52 |

Table A.15.6 Summary of SGP Estimates for Subgroups: Grade 4 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 48,093 | 50.44 | 13.48 | 51 |
| Female | 46,365 | 49.71 | 13.49 | 50 |
| **Ethnicity** | | | | |
| White | 51,416 | 50.45 | 13.05 | 51 |
| African American | 12,328 | 44.39 | 14.79 | 42 |
| Asian/Pacific Islander | 5,187 | 59.13 | 13.14 | 63 |
| American Indian/Alaska Native | 197 | 49.16 | 13.55 | 48 |
| Hispanic | 20,401 | 50.15 | 13.88 | 50 |
| Multiple | 4,661 | 51.37 | 13.49 | 51 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 39,044 | 46.91 | 14.05 | 46 |
| Not-economically Disadvantaged | 55,415 | 52.32 | 13.08 | 53 |
| English Learner (EL) | 13,590 | 49.62 | 14.13 | 50 |
| Non-English Learner | 80,869 | 50.16 | 13.37 | 50 |
| Students with Disabilities (SWD) | 16,526 | 43.36 | 14.15 | 41 |
| Students without Disabilities | 77,933 | 51.51 | 13.34 | 52 |
| **Spanish Language Form** | 1,268 | 42.63 | 14.57 | 41 |

Table A.15.7 Summary of SGP Estimates for Subgroups: Grade 5 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 49,440 | 49.34 | 14.11 | 49 |
| Female | 47,335 | 50.72 | 14.18 | 51 |
| **Ethnicity** | | | | |
| White | 53,050 | 50.56 | 13.71 | 51 |
| African American | 12,483 | 44.07 | 15.47 | 41 |
| Asian/Pacific Islander | 5,260 | 59.32 | 13.52 | 63 |
| American Indian/Alaska Native | 175 | 53.42 | 13.85 | 54 |
| Hispanic | 20,963 | 49.89 | 14.64 | 50 |
| Multiple | 4,557 | 49.74 | 14.06 | 50 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 39,735 | 46.91 | 14.83 | 46 |
| Not-economically Disadvantaged | 57,045 | 52.17 | 13.67 | 53 |
| English Learner (EL) | 11,639 | 48.20 | 15.20 | 47 |
| Non-English Learner | 85,141 | 50.26 | 14.00 | 50 |
| Students with Disabilities (SWD) | 16,962 | 41.68 | 14.93 | 38 |
| Students without Disabilities | 79,818 | 51.78 | 13.98 | 53 |
| **Spanish Language Form** | 1,212 | 43.79 | 15.07 | 42 |

Table A.15.8 Summary of SGP Estimates for Subgroups: Grade 6 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 50,175 | 49.59 | 14.66 | 49 |
| Female | 47,242 | 50.45 | 14.61 | 51 |
| **Ethnicity** | | | | |
| White | 53,275 | 51.66 | 14.29 | 52 |
| African American | 12,619 | 42.78 | 15.75 | 40 |
| Asian/Pacific Islander | 5,205 | 58.76 | 13.93 | 63 |
| American Indian/Alaska Native | 165 | 49.52 | 14.29 | 46 |
| Hispanic | 21,490 | 48.28 | 15.02 | 47 |
| Multiple | 4,428 | 48.98 | 14.56 | 49 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 40,200 | 46.47 | 15.26 | 45 |
| Not-economically Disadvantaged | 57,224 | 52.49 | 14.19 | 53 |
| English Learner (EL) | 9,684 | 42.27 | 15.93 | 39 |
| Non-English Learner | 87,740 | 50.86 | 14.49 | 51 |
| Students with Disabilities (SWD) | 16,941 | 40.97 | 15.40 | 37 |
| Students without Disabilities | 80,483 | 51.91 | 14.47 | 53 |
| **Spanish Language Form** | 858 | 44.62 | 15.83 | 42 |

Table A.15.9 Summary of SGP Estimates for Subgroups: Grade 7 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 48,644 | 50.81 | 15.24 | 51 |
| Female | 46,050 | 49.07 | 15.45 | 49 |
| **Ethnicity** | | | | |
| White | 52,045 | 50.48 | 15.14 | 51 |
| African American | 12,299 | 45.45 | 15.96 | 44 |
| Asian/Pacific Islander | 4,958 | 56.11 | 15.30 | 58 |
| American Indian/Alaska Native | 196 | 52.79 | 15.55 | 52.5 |
| Hispanic | 21,179 | 50.26 | 15.47 | 50 |
| Multiple | 3,836 | 48.44 | 15.42 | 48 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 39,749 | 47.63 | 15.60 | 47 |
| Not-economically Disadvantaged | 54,955 | 51.65 | 15.15 | 52 |
| English Learner (EL) | 8,686 | 45.73 | 16.11 | 44 |
| Non-English Learner | 86,018 | 50.39 | 15.26 | 51 |
| Students with Disabilities (SWD) | 16,357 | 41.86 | 15.85 | 38 |
| Students without Disabilities | 78,347 | 51.65 | 15.23 | 52 |
| **Spanish Language Form** | 444 | 44.72 | 16.10 | 42 |

Table A.15.10 Summary of SGP Estimates for Subgroups: Grade 8 Mathematics

| | Total Sample Size | Average SGP | Average Standard Error | Median SGP |
|---|---|---|---|---|
| **Gender** | | | | |
| Male | 48,284 | 48.89 | 16.65 | 48 |
| Female | 45,576 | 50.97 | 16.85 | 51 |
| **Ethnicity** | | | | |
| White | 51,836 | 51.18 | 16.10 | 52 |
| African American | 12,607 | 44.40 | 18.68 | 42 |
| Asian/Pacific Islander | 4,805 | 56.59 | 14.72 | 59 |
| American Indian/Alaska Native | 197 | 48.42 | 16.98 | 47 |
| Hispanic | 20,646 | 48.70 | 17.64 | 48 |
| Multiple | 3,614 | 49.08 | 16.94 | 48 |
| **Special Instruction Needs** | | | | |
| Economically Disadvantaged | 39,280 | 47.12 | 17.92 | 46 |
| Not-economically Disadvantaged | 54,589 | 51.90 | 15.90 | 53 |
| English Learner (EL) | 7,980 | 46.14 | 19.03 | 44 |
| Non-English Learner | 85,889 | 50.25 | 16.53 | 50 |
| Students with Disabilities (SWD) | 15,864 | 44.92 | 18.51 | 43 |
| Students without Disabilities | 78,005 | 50.91 | 16.39 | 51 |
| **Spanish Language Form** | 273 | 47.37 | 19.97 | 44 |