New Meridian

Technical Report 2018–2019
Alternate Blueprint
Illinois Assessment of Readiness (IAR)

February 28, 2020

Table of Contents

List of Tables

List of Figures

Executive Summary

The purpose of this report is to describe the technical qualities of the 2018–2019 operational administration of the English language arts/literacy (ELA/L) and mathematics assessments in grades 3 through 8 and high school. Committees of educators, state education agency staff, and national experts led the work in the development of the summative assessments that are aligned to the Common Core State Standards (CCSS) and are intended to measure more complex skills like critical thinking, persuasive writing, and problem-solving. New Meridian assumes the responsibility for management of the summative assessments, as well as item development and forms construction. 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 was available in spring 2019. For the academic year 2018-2019, participating states and agencies included the Bureau of Indian Education, Illinois, New Jersey, and New Mexico.

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 access 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; and
- 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).

In Illinois, the first administration of the Illinois Assessment of Readiness (IAR) occurred in spring 2019. The IAR was administered to students in grades 3 through 8 in the content areas of English language arts/literacy (ELA/L) and mathematics.

<This page intentionally left blank>

Section 1: Introduction

1.1 Background

States 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 including professional development and communications tools.

The partnership develops and administers next-generation assessments that, compared to traditional K–12 assessments, more accurately measure student progress toward college and career readiness. The assessments are aligned to the Common Core State Standards (CCSS) and include 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 are 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, 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 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 also contracted with New Meridian to manage the IP 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.

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 was available in spring 2019 in addition to the original blueprint. New Meridian's state-centric solution to educational assessment allowed states the flexibility of selecting the assessment solution that best fit their specific needs. For the academic year 2018–2019, participating states and agencies included the Bureau of Indian Education, Illinois, New Jersey, and New Mexico.

The purpose of this technical report is to describe the operational administration of the summative assessments in the 2018–2019 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 summative assessments are 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 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.

1.4 Intended Population

The tests are intended for students taking ELA/L in grades 3 through 11, and/or mathematics in grades 3 through 8, as well as students taking high school mathematics (i.e., Algebra I, Geometry, Algebra II, and Integrated Mathematics I–III). 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.

1.5 Groups and Organizations Involved with the Summative Assessments

New Meridian is a nonprofit organization that assumes the responsibility for management of the assessments, as well as item development and forms construction of the assessments.

Committees of educators, state education agency staff, and national experts lead the work of the assessments. These committees include:

- the Governing Board that makes major policy and operational decisions;
- the Technical Advisory Committee that helps ensure all assessments will provide reliable results to inform valid instructional and accountability decisions;
- the State Lead Council that coordinates all aspects of development of the summative assessment system and serves as the conduit to the Technical Advisory Committee and the Governing Board; and
- ELA/L, Mathematics, and Accessibility and Accommodation Features operational working groups.

Pearson serves as the primary contractor for the operational administration and is responsible for producing all testing materials, packaging and distribution, receiving and scanning of materials, and scoring, as well as program management and customer service. In addition, test and item development activities are conducted by Pearson under the guidance and oversight of New Meridian.

Pearson Psychometrics is responsible for all psychometric analyses of the operational test data. This includes classical item analyses, differential item functioning (DIF) analyses, item calibrations based on item response theory (IRT), scaling, and development of all conversion tables.

Human Resources Research Organization (HumRRO) serves as a subcontractor and is responsible for replicating item calibrations based on item response theory (IRT), scaling, and development of all conversion tables.

Pearson Psychometrics is also responsible for reviewing and comparing the results obtained independently from Pearson and from HumRRO including IRT calibrations, conversion tables, summative and claim scale scores, performance level classifications, and subclaim performance level classifications.

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 procedure, 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
This section describes the test design and the procedures followed during the development of operational test forms.

Section 3 – Test Administration
This section presents the operational administration schedule, information regarding test security and confidentiality, accessibility features and accommodations, and testing irregularities and security breaches.

Section 4 – Item Scoring
The key-based and rule-based processes for machine-scored items, as well as the training and monitoring processes for human-scored items, are provided in this section.

Section 5 – Classical Item Analysis
The 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 are presented in this section.

Section 6 – Differential Item Functioning
In this section, the methods for conducting differential item functioning analyses as well as corresponding flagging criteria are described. This is followed by definitions of the comparison groups and subsequent results for the comparison groups.

Section 7 – IRT Calibration and Scaling
This section presents the information related to the calibration and scaling of item response data including: data preparation, the calibration process, model fit evaluation, and items excluded from score reporting. In addition, the scaling process is described and evaluated.

Section 8 – Performance Level Setting
Performance levels and policy definitions, as well as the processes followed to establish performance level thresholds, are described in this section.

Section 9 – Quality Control Procedures
All aspects of quality control are presented in this section. These activities range from quality assurance of item banking, test form construction, and all testing materials to 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
This section describes the operational test forms including high level blueprints for the assessments.

Section 11 – Student Characteristics
This section describes the 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

This section provides an overview of the claims and subclaims, describes the development of the 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

The results of 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, is provided in this section. 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 handscored 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

This section provides 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, and so on.

Addendum

The addendum presents the results of analyses for the fall operational administration. These results are reported separately from the spring results because fall testing involved a nonrepresentative subset of students testing only ELA/L grades 9, 10, and 11, as well as Algebra I, Geometry, and Algebra II.

To organize the addendum, tables are numbered sequentially according to the section represented by the tables. For example, the first addendum table for Section 11 is numbered ADD.11.1, the second addendum table for Section 11 is numbered ADD.11.2, and so on.

1.7 Glossary of Abbreviations

Table 1.1 Glossary of Abbreviations and Acronyms

Abbreviation/Acronym	Definition
1PL/PC	One-parameter/Partial Credit Model
2PL/GPC	Two-parameter Logistic/Generalized Partial Credit Model
3PL/GPC	Three-parameter Logistic/Generalized Partial Credit Model
A1	Algebra I
A2	Algebra II
AAF	Accessibility, Accommodations, and Fairness
ABBI	Assessment Banking for Building and Interoperability
AERA	American Educational Research Association
AIS	Average Item Score
AIQ	Assessment and Information Quality
AmerIndian	American Indian/Alaska Native
APA	American Psychological Association
ASC	Additional and Supporting Content (Mathematics)
ASL	American Sign Language
ATA	Automatic Test Assembler
CBT	Computer-Based Test
CCSS	Common Core State Standards
CDQ	Customer Data Quality
CSEM	Conditional Standard Error of Measurement
DIF	Differential Item Functioning
DPL	Digital Production Line
DPP	Digital Pre-press
EcnDis	Economically disadvantaged
EBSS	Evidence-based Standard Setting
ELA/L	English Language Arts/Literacy
EL	English Learners
ELN	Not an English learner
ELY	English Learners
EOC	End-of-Course
EOY	End-of-Year
ePEN2	Electronic Performance Evaluation Network second generation
ESEA	Elementary and Secondary Education Act
FRL	Free or Reduced-price Lunch
FS	Full Summative
FT	Field Test
GO	Geometry
HOSS	Highest Obtainable Scale Score
IA	Item Analysis
ICC	Item Characteristic Curve
IDEA	Individuals with Disabilities Education Act
IEP	Individualized Education Program
INF	Information Curve
IP	Intellectual Property

Abbreviation/Acronym	Definition
IRA	Inter-rater Agreement
IRF	Item Response File
IRT	Item Response Theory
ISR	Individual Student Report
K–12	Kindergarten to Grade 12
LEA	Local Education Agency
LID	Local Item Dependence
LOSS	Lowest Obtainable Scale Score
LP	Large Print
M1	Integrated Mathematics I
M2	Integrated Mathematics II
M3	Integrated Mathematics III
MAD	Mean Absolute Difference
MC	Major Content (Mathematics)
MH	Mantel-Haenszel
MP	Modeling Practice (Mathematics)
MR	Mathematical Reasoning
Multiracial	Multiple Races Selected
NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind
NCME	National Council on Measurement in Education
NoEcnDis	Not economically disadvantaged,
NSLP	National School Lunch Program
OE responses	Open-ended responses
OMR	Optical Mark Reading
OWG	Operational Working Group
Pacific Islander	Native Hawaiian or Pacific Islander
PARCC	Partnership for Assessment of Readiness for College and Careers
PBA	Performance-Based Assessment
PBT	Paper-Based Test
PCR	Prose Constructed Response (ELA/L)
PEJ	Postsecondary Educators' Judgment
PLD	Performance Level Descriptor
PLS	Performance Level Setting
PV	Product Validation
QA	Quality Assurance
RD	Reading (ELA/L)
RI	Reading Information (ELA/L)
RL	Reading Literature (ELA/L)
RMSD	Root Mean Square Difference
RV	Reading Vocabulary (ELA/L)
RST	Raw-score-to-theta
SD	Standard Deviation
SDF	Student Data File
SE	Standard Error
SEJ	Standard Error of Judgment
SEM	Standard Error of Measurement
SIRB	Scored Item Response Block

Abbreviation/Acronym	Definition
SMD	Standardized Mean Difference
SSMC	Single Select Multiple Choice
SWD	Students with Disabilities
SWDN	Not student with disability
SWDY	Students with Disabilities
TCC	Test Characteristic Curve
TTS	Text to Speech
UIN	Unique Item Number
WE	Writing Written Expression (ELA/L)
WKL	Writing Knowledge Language and Conventions (ELA/L)
WLS	Weighted Least Squares
WR	Writing (ELA/L)
WRMSD	Weighted Root Mean Square Difference

Section 2: Test Development

2.1 Overview of the Summative Assessments, Claims, and Design

Aligned to the Common Core State Standards (CCSS) as articulated in the Model Content Frameworks, the summative assessments are designed to determine whether students are college- and career-ready or on track, 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 is an ongoing process involving educators, researchers, psychometricians, subject matter professionals, and assessment experts who participate in the development of the test design and its underlying foundational documents; develop and review passages and items used to build the summative assessments; monitor the program for quality, accessibility, and fairness for all students; and construct, review, and score the assessments.

The summative assessments include both English language arts/literacy (ELA/L) and mathematics assessments in grades 3 through 8 and high school. The high school mathematics tests include traditional mathematics and integrated mathematics course pathways. 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 performance-based 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 text(s). The summative assessment also contains literary and informational reading passages with comprehension and vocabulary questions.

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 evidences outlined in the evidence tables for reading and writing (refer to the test specifications documents). The claims and evidences 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. From these, more than 2,000 educators, researchers, and psychometricians have developed the test specifications documents that guide the development of test items and the composition of the tests. 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 website for further information about these documents.

2.2.1 Item Development Process

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, 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 is 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. Prior to the 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, Pearson recruited, trained, and managed the contracted subject matter experts 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 allow for a range of standards/evidences to be demonstrated to meet the assessment claims. ELA/L tests are based on authentic texts, including multi-media stimulus. 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. Pearson 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

Guided by foundational documents, Pearson recruited and trained the item writers and managed the item writing to develop the number of items specified in the annual asset development plan. Prior to further committee reviews,

the assessment teams at Pearson 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 to Pearson and participating states and agencies 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 is a review and approval by the Text Review Committee of the texts eligible for item development. Participants reviewed and provided feedback to Pearson and participating states and agencies 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 make up the committee that reviews items and tasks to confirm that there are 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 consists 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 the field test, educator and bias committee members met to evaluate test items and associated performance data with regard to 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, Pearson constructed the operational forms to adhere to the test blueprints and the assessment goals outlined in the form creation specifications. These goals were:

- 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; and
- 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 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 and to be constructed as "parallel" as possible from a content and test-taking experience. Evaluation criteria for parallelism included adherence to blueprint; sequencing of content across the forms; statistical averages and distributions for difficulty (e.g., p-value) and discrimination (e.g., polyserial correlation); item type and cognitive complexity; and passage characteristics for ELA/L including genre, topics, word count, and text complexity.

Additionally, appropriate forms were identified as accessibility and 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.

Test Construction Activities

After the data review meetings and prior to the test construction meetings, Pearson assessment specialists constructed initial versions of all the core forms. Content specialists constructed the initial core forms based on the support documents and specific processes to achieve fair parallel forms. The following steps were used to construct the operational core forms taken to the Test Construction Committee 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 Pearson 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.

Pearson assessment specialists identified forms for each grade/subject suitable for use as the accommodated forms. Pearson 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 Pearson 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 Construction 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

- 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

- closed captioning

- text-to-speech first form
- Spanish online
- Spanish text-to-speech

Accommodated Base 3 (mathematics only)

- 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 transadaption of the items. Transadaption 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 transadaption 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. Across-administration linking, or year-to-year linking, consisted of common items included in two different administrations. 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 forms to collect data for psychometric analysis necessary to support the assessment system for future administrations. Field-test administration entailed paper and computer administration modes, with computer administration as the dominant mode. The ELA/L unit of field-test items were administered to a sample of students.

Field-test sets were constructed to balance the expected cognitive load and difficulty across forms, reflected in the number of points, distribution of task types, and balance of passages for ELA/L. Forms for each content area were spiraled at the student level. The data collection design entailed three conditions. Condition 1, which comprised the mathematics assessment, was an embedded census field-test model in which all students taking the summative assessment participated in the field test.

Under Condition 2, which comprised the ELA/L assessment, approximately one-third of the schools were sampled across some of the participating states. Students in the sampled schools or districts took forms containing ELA/L embedded field-test tasks. Schools or districts were selected so that the sample for each ELA/L assessment was representative of the general testing populations in terms of achievement (i.e., average scale score and percentage of students at Level 4 and Level 5 in the previous year) and demographics (i.e., ethnicity composition, percentage of economically disadvantaged, English learners, and students with disabilities). The sampling plan was created such that if a given school was part of the ELA/L field test one year (e.g., spring 2017), it would not be required to participate in the field test for the subsequent two years (e.g., spring 2018 and spring 2019).

For Condition 3, states or agencies may select to field-test two ELA/L grade levels rather than all grade levels. The grade levels selected participate in a census field-test where all students are administered the embedded field-test items. The remaining grade levels do not participate in field-testing. The selected grade levels are rotated across years.

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). Above 97 percent of the summative assessments administered during the 2018–2019 administration were online assessments, and less than 3 percent were paper-based assessments (see Tables 11.1 – 11.3).

3.1.1 Secure vs. NonSecure Materials

Participating states and agencies define secure materials 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. Nonsecure 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 (includes 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 that are to be provided by Test Administrators to students include test booklets (grade 3) or answer documents (grades 4 through high school). Nonscorable secure materials that are 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. All eligible students will participate in the ELA/L and mathematics assessments. 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 Individualized Education Program (IEP);
2. students with a Section 504 plan who have a physical or mental impairment 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; and
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, participating states and agencies 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 what they know and can do. However, the accessibility features available to students should minimize the need for accommodations during testing and ensure the inclusive, accessible, and fair testing of the diverse students being assessed.

3.2.3 What are Accessibility Features?

On the 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 determine which are appropriate for use on the assessment. Consideration should be given to the supports a student finds helpful and consistently uses during instruction. 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 considered to be 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; and
- 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 each state's policies and procedures for notifying the state assessment office:

- 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

An 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. An 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 must be 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 plan, 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:

Electronic Devices

- Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner) while secure test materials are still distributed, while students are testing, after a student turns in 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 still distributed or while students are testing
- Leaving students unattended for any period of time while secure test materials are still distributed or while students are testing
- Deviating from testing time procedures
- Allowing cheating of any kind
- Providing unauthorized persons with access to secure materials
- 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 is 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.

[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.

3.4.3 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). A Weighted Least Squares (WLS) regression methodology was evaluated and recommended by the Technical Advisory Committee (TAC) for implementation starting 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.

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 Handscored Items

Constructed-response items were scored by human scorers in a process referred to as handscoring. 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.
- Project managers documented the procedures, identified risks, and managed day-to-day administrative matters.
- A program 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. Qualification sets 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 11). In mathematics, a prototype training set was built for a grouping of similar items for a total of approximately three to five 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 task 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. 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. 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. 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. 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. 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 define and document 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). 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). The subsequent ELA/L abbreviated items do not include qualification sets.

4.2.2 Scorer Qualification

In order 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 handscoring specifications document that detailed the handscoring 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 of the first and second scores, the content specialist or scoring director would apply an adjudication score (fourth score).

Table 4.3 Scoring Hierarchy Rules

If a response was scored more than once, the following rules were applied to determine the final score:

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 in order to confirm that the scores were correctly assigned and to give feedback and remediation to individual scorers. Pearson backread approximately 5 percent of the handscored 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 zero or 1 score compared to a blank 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%	100%	100%
Mathematics	0–2	80%	97%	100%	100%
Mathematics	0–3	70%	95%	100%	99%
Mathematics	0–4	65%	94%	99%	99%
Mathematics	0–5	65%	93%	99%	98%
Mathematics	0–6	65%	95%	99%	98%
ELA/L	Multi-trait	65%	80%	100%	99%

*A zero 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.

The perfect agreement rate for mathematics responses scored by two scorers ranged from 93 to 98 percent and the within one point rate ranged from 98 to 100 percent. For all ELA/L responses scored by two scorers, the perfect agreement rate was 80 percent and the within one point rate was 99 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 at the time IEA was trained. 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 two human scores 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 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 2019 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); and

- 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 2019, 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 following ten comparison groups:

Table 4.6 Comparison Groups

Group Type	Comparison Groups
Sex	Female
	Male
Ethnicity	American Indian/Alaska Native
	Asian
	Black/African American
	Hispanic/Latino
	Native Hawaiian or Other Pacific Islander
	White
Special Instructional Needs	English Language Learners (ELL)
	Students with Disabilities (SWD)

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 task items are scored on two traits: (1) Reading Comprehension and Written Expression and (2) Knowledge of Language and Conventions.

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 those states choosing the human-scoring option, the second reliability score was assigned by IEA. 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 (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 .74 to .90.

During operational scoring, the PCR agreement rates are monitored for quality and items not meeting the criteria are shared with the handscoring group. After the operational administration, the performance of all the PCRs is provided to the content team as feedback for re-using PCRs and in order to inform development of future PCRs. This provides evidence for continuous improvement of the testing program.

Table 4.7 PCR Average Agreement Indices by Test

			Written Expression				Conventions			
Test	Number of PCRs	Count	Exact	Kappa	QW Kappa	*r*	Exact	Kappa	QW Kappa	*r*
ELA03	5	38,626	71.88	0.54	0.74	0.74	73.56	0.58	0.77	0.77
ELA04	5	41,309	69.22	0.56	0.81	0.82	71.60	0.59	0.83	0.83
ELA05	5	77,241	70.74	0.58	0.84	0.84	71.02	0.59	0.83	0.83
ELA06	5	47,325	74.30	0.64	0.86	0.86	74.66	0.64	0.85	0.85
ELA07	5	61,267	73.36	0.63	0.88	0.88	74.52	0.65	0.87	0.87
ELA08	5	51,067	76.18	0.68	0.90	0.90	77.02	0.69	0.89	0.89
ELA09	5	15,051	71.70	0.61	0.87	0.87	71.50	0.60	0.84	0.84
ELA10	5	24,432	73.20	0.64	0.90	0.90	76.80	0.68	0.89	0.89
ELA11	5	5,991	76.56	0.66	0.85	0.86	77.98	0.67	0.86	0.86

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. In addition, ELA/L assessments were post-equated for some states or agencies for score reporting (see Section 7). For pre-equated tests, the item statistics provided in this section were from prior operational administrations and reflect the statistics that were used at test construction and for score reporting for some states and agencies. For the post-equated tests, the statistics from the spring administration were also provided in this section. 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 handscored 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 by form and by administration mode. 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. The range for p-values is from .00 to 1.00. Items with high p-values are easy items and those with low p-values are difficult items. Dichotomously scored items were flagged for review if the p-value was above .95 (i.e., too easy) or below .25 (i.e., too difficult).

For polytomously scored items, difficulty is indicated by the average item score (AIS). The AIS can range from .00 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. Polytomously scored items were flagged for review if the p-value was above .95 or below .25.

The percentage of students choosing each response option

Selected-response items on the summative assessments refer primarily to single-select multiple-choice 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 omit 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 the incorrect option than the correct response. Such a result could indicate that the item has multiple correct answers or is miskeyed.

Item-total correlation

This statistic describes the relationship between students' performance on a specific item and their performance on the total test. The item-total correlation is usually referred to as the item discrimination index. For operational item analysis, the total score on the assessment was used as the total test score. The polyserial 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 -1.00 to 1.00. Desired values are positive and larger than .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 .15 were flagged for review. Items with extremely low or negative values were considered for exclusion from IRT calibrations or linking (refer to Section 7 for details on item inclusion and exclusion criteria for IRT analyses).

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 (refer to #3 analysis above) for the distractors. Items with distractor-total correlations above .00 were flagged for review as these items may have multiple correct answers, be miskeyed, 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. 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. Items with such response patterns may pose problems during the IRT calibrations and therefore may need to be excluded (refer to Section 7 for more information).

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 .95 for dichotomous items or polytomous items
2. p-value below .25 for dichotomous items or polytomous items
3. item-total correlation below .15
4. any distractor-total correlation above .00
5. greater number of high-performing students (top 20 percent) choosing a distractor rather than the keyed response
6. high percentage of omits: 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

Pearson's psychometric staff carefully reviewed the flagged items and brought items 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. The mathematics assessments were pre-equated, meaning that the scoring was based on item parameters estimated using data from earlier administrations. For the pre-equated grades/subjects, item analysis results in this section are the item statistics from prior administrations that were used to make decisions during test construction and for scoring. The ELA/L assessments were both pre-equated and post-equated. Therefore, the item analysis results from both prior administrations and from the spring operational administration are presented in this section.

- Tables 5.1 and 5.2 present pre-administration and post-administration p-value information by grade for the ELA/L operational items.
- Table 5.3 presents pre-administration p-value information by grade/course for the mathematics operational items.
- Tables 5.4 and 5.5 present pre-administration and post-administration item-total correlations by grade for the ELA/L operational items.
- Table 5.6 presents pre-administration item-total correlations by grade/course 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.

The fall 2018 forms were based on the spring 2018 operational forms; therefore, the item analyses for these forms were reported in the 2017–2018 Technical Report. Some forms on the spring 2019 administration were based on spring 2017 and 2018 administrations; therefore, the item analyses for these forms were reported in the 2016–2017 and the 2017–2018 Technical Reports.

Table 5.1 Summary of Pre-Administration p-Values for ELA/L Operational Items by Grade

Grade	N of Unique Items	Mean p-Value	SD p-Value	Min p-Value	Max p-Value	Median p-Value
3	58	0.47	0.17	0.16	0.82	0.47
4	74	0.47	0.16	0.18	0.86	0.46
5	66	0.47	0.14	0.09	0.83	0.45
6	77	0.48	0.15	0.15	0.92	0.48
7	62	0.48	0.14	0.22	0.83	0.47
8	72	0.48	0.13	0.20	0.85	0.48
9	88	0.43	0.13	0.09	0.78	0.41
10	63	0.42	0.11	0.14	0.64	0.42
11	62	0.36	0.10	0.16	0.65	0.35

Table 5.2 Summary of Post-Administration p-Values for ELA/L Operational Items by Grade

Grade	N of Unique Items	Mean p-Value	SD p-Value	Min p-Value	Max p-Value	Median p-Value
3	58	0.47	0.18	0.18	0.84	0.45
4	74	0.47	0.16	0.21	0.82	0.46
5	66	0.48	0.14	0.10	0.84	0.46
6	77	0.49	0.15	0.18	0.91	0.49
7	62	0.50	0.14	0.24	0.80	0.49
8	72	0.49	0.13	0.22	0.82	0.49
9	88	0.44	0.14	0.09	0.77	0.42
10	63	0.47	0.12	0.15	0.73	0.48
11	62	0.37	0.10	0.21	0.65	0.36

Table 5.3 Summary of p-Values for Mathematics Operational Items by Grade/Course

Grade/ Course	N of Unique Items	Mean p-Value	SD p-Value	Min p-Value	Max p-Value	Median p-Value
3	77	0.57	0.20	0.19	0.91	0.57
4	72	0.52	0.20	0.06	0.91	0.51
5	71	0.50	0.18	0.13	0.84	0.48
6	69	0.41	0.18	0.11	0.78	0.40
7	67	0.39	0.17	0.08	0.75	0.37
8	64	0.33	0.18	0.08	0.68	0.29
A1	111	0.31	0.16	0.05	0.73	0.30
GO	118	0.29	0.17	0.05	0.79	0.28
A2	109	0.27	0.15	0.05	0.82	0.26
M1	42	0.36	0.16	0.08	0.65	0.39
M2	41	0.29	0.20	0.05	0.69	0.25
M3	40	0.27	0.15	0.05	0.61	0.24

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Table 5.4 Summary of Pre-Administration Item-Total Correlations for ELA/L Operational Items by Grade

Grade	N of Unique Items	Mean Polyserial	SD Polyserial	Min Polyserial	Max Polyserial	Median Polyserial
3	58	0.54	0.12	0.23	0.78	0.52
4	74	0.47	0.14	0.23	0.81	0.45
5	66	0.48	0.14	0.19	0.83	0.47
6	77	0.50	0.14	0.20	0.83	0.48
7	62	0.50	0.15	0.26	0.83	0.48
8	72	0.49	0.15	0.26	0.83	0.47
9	88	0.50	0.17	0.25	0.88	0.46
10	63	0.50	0.17	0.18	0.86	0.47
11	62	0.47	0.16	0.17	0.85	0.43

Table 5.5 Summary of Post-Administration Item-Total Correlations for ELA/L Operational Items by Grade

Grade	N of Unique Items	Mean Polyserial	SD Polyserial	Min Polyserial	Max Polyserial	Median Polyserial
3	58	0.56	0.13	0.30	0.81	0.54
4	74	0.49	0.15	0.15	0.82	0.47
5	66	0.52	0.15	0.20	0.86	0.51
6	77	0.53	0.15	0.30	0.87	0.51
7	62	0.53	0.15	0.26	0.86	0.49
8	72	0.52	0.16	0.27	0.88	0.49
9	88	0.51	0.18	0.25	0.88	0.46
10	63	0.52	0.17	0.19	0.88	0.48
11	62	0.48	0.18	0.15	0.86	0.44

Table 5.6 Summary of Item-Total Correlations for Mathematics Operational Items by Grade/Course

Grade/ Course	N of Unique Items	Mean Polyserial	SD Polyserial	Min Polyserial	Max Polyserial	Median Polyserial
3	77	0.52	0.13	0.28	0.81	0.52
4	72	0.52	0.12	0.26	0.76	0.51
5	71	0.52	0.11	0.20	0.77	0.52
6	69	0.54	0.14	0.17	0.92	0.54
7	67	0.51	0.16	0.17	0.82	0.53
8	64	0.49	0.12	0.24	0.73	0.50
A1	111	0.45	0.15	0.15	0.75	0.45
GO	118	0.49	0.16	0.17	0.95	0.48
A2	109	0.49	0.14	0.19	0.84	0.50
M1	42	0.50	0.13	0.24	0.75	0.50
M2	41	0.47	0.15	0.21	0.83	0.46
M3	40	0.46	0.13	0.18	0.69	0.46

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Section 6: Differential Item Functioning

6.1 Overview

Differential item functioning (DIF) analyses were conducted using the data obtained from the operational items. If an item performs differentially across identifiable subgroups (e.g., gender or ethnicity) 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.

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

6.2 DIF Procedures

Dichotomous Items

The Mantel-Haenszel (MH) DIF statistic was calculated for selected-response items and for dichotomously scored constructed-response items. In this method, students are classified to relevant subgroups of interest (e.g., gender or ethnicity). Using the raw score total as the criteria, students in a certain total score category in the focal group (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):

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

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 standardization 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 standardization 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:

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

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 DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values. 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 2. Absolute value of the MH D-DIF is significantly different from one, but is less than 1.5. Positive values are classified as "B+" and negative values as "B-".
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one, and is at least 1.5. Positive values are classified as "C+" and negative values as "C-".

Table 6.2 DIF Categories for Polytomous Constructed-Response Items

DIF Category	Criteria
A (negligible)	Mantel Chi-square p-value > 0.05 or $\|STD\text{-}EISDIF/SD\| \leq 0.17$
B (slight to moderate)	Mantel Chi-square p-value < 0.05 and $\|STD\text{-}EISDIF/SD\| > 0.17$
C (moderate to large)	Mantel Chi-square p-value < 0.05 and $\|STD\text{-}EISDIF/SD\| > 0.25$

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

6.3 Operational Analysis DIF Comparison Groups

DIF analyses were conducted on each test form 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 the states and district and captured in PearsonAccess^{next} by means of a student data upload. The demographic data was verified by the states and district 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 (AmerIndian)	White
	Asian	White
	Black or African American	White
	Hispanic/Latino	White
	Native Hawaiian or Pacific Islander	White
	Multiple Race Selected	White
Economic Status*	Economically Disadvantaged (EcnDis)	Not Economically Disadvantaged (NoEcnDis)
Special Instructional Needs	English Learner (ELY)	Non English Learner (ELN)
	Students with Disabilities (SWDY)	Students without Disabilities (SWDN)

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:

- the smaller group, reference or focal, had at least 100 students, and
- 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). The fall 2018 forms were based on spring 2018 operational forms. The DIF analyses for these forms are reported in the 2017–2018 Technical Report.

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. In addition, "0" indicates that the DIF analysis did not classify any items in the particular DIF category, while "n/a" indicates that the DIF analysis was not performed due to insufficient sample sizes.

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
Male vs Female	52	1	2	.	.	50	96	1	2		
White vs Black	52	.	.	1	2	51	98	.	.		
White vs Hispanic	52	.	.	3	6	49	94	.	.		
White vs Asian	52	1	2	.	.	51	98	.	.		
White vs AmerIndian	52	.	.	.	.	52	100	.	.		
White vs Pacific Islander	52	.	.	1	2	50	96	1	2		
White vs Multiracial	52	.	.	.	.	52	100	.	.		
NoEcnDis vs EcnDis	52	.	.	.	.	52	100	.	.		
ELN vs ELY	52	.	.	2	4	50	96	.	.		
SWDN vs SWDY	52	.	.	.	.	52	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian or Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table 6.5 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
Male vs Female	77			1	1	75	97	1	1	.	.
White vs Black	77			6	8	68	88	3	4	.	.
White vs Hispanic	77			3	4	74	96	.	.	.	.
White vs Asian	77			.	.	67	87	9	12	1	1
White vs AmerIndian	77			2	3	75	97	.	.	.	.
White vs Pacific Islander	77			1	1	75	97	1	1	.	.
White vs Multiracial	77			1	1	75	97	1	1	.	.
NoEcnDis vs EcnDis	77			.	.	77	100	.	.	.	.
ELN vs ELY	77			1	1	76	99	.	.	.	.
SWDN vs SWDY	77			2	3	75	97	.	.	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian or Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Section 7: IRT Calibration and Scaling

7.1 Overview

Multiple operational core forms were administered for each grade in English language arts/literacy (ELA/L) and mathematics assessments. The purpose of the item response theory (IRT) calibration and scaling was to place all operational items for a single grade/subject onto a common scale. For the ELA/L computer-based tests (CBTs), the IRT parameters were post-equated. This section describes procedures used to calibrate and scale the post-equated operational assessments. Because ELA/L paper-based tests (PBTs) and all mathematics tests were pre-equated, much of the discussion in this section will not apply; however, the parameters used to construct the conversion tables for these tests are presented in this section.

In this section of the technical report, the following topics related to IRT calibration and scaling are discussed:

Calibration:

7.2 IRT Data Preparation
7.3 Description of the Calibration Process
7.4 Model Fit Evaluation Criteria
7.5 Items Excluded from Score Reporting

Scaling:

7.6 Scaling Parameter Estimates
7.7 Items Excluded from Linking Sets
7.8 Correlations and Plots of Scaling Item Parameter Estimates
7.9 Scaling Constants
7.10 Summary Statistics and Distributions from IRT Analyses

7.2 IRT Data Preparation

7.2.1 Overview

The post-equating was based on the majority of students testing in the spring administration. All student response data in the samples for operational items were used to create the IRT sparse data matrices for the concurrent calibration. IRT sparse data matrices combine student data across forms within administration mode. Items on the non-accommodated forms are included in the post-equating analysis. Table 7.1 lists the number of items and equating sample size for the post-equated assessments.

Table 7.1 Counts and Number of Items in the ELA/L IRT Calibration Files

Grade	Count	Items
3	287,704	46
4	32,2383	62
5	331,254	64
6	332,892	62
7	325,874	60
8	322,373	62
9	120,185	34
10	183,715	62
11	33,274	44

7.2.2 Student Inclusion/Exclusion Rules

The following are the IRT valid case criteria. These criteria are the same as the student inclusion/exclusion rules used to evaluate and filter data prior to conducting the operational item analysis (IA) and differential item functioning (DIF) analyses (steps 1–5).

1. All records with an invalid form number were excluded.
2. All records that were flagged as "void" were excluded.
3. Records in which the student attempted fewer than 25 percent of the items in any unit were excluded.
4. For students with more than one valid record, the record with the higher raw score was chosen. If the raw scores were the same, the record with the higher attempted rate across all operational units was chosen.
5. Records for students with administration issues or anomalies were excluded.

7.2.3 Items Excluded from IRT Sparse Matrices

Pearson conducted an initial scoring and key check. Items identified by Pearson as "spoiled" (also referred to as "do not use (DNU)") were listed and excluded from the analyses. When the IRT sparse data matrices were created, all items were included in the files unless they were marked as "spoiled" by Pearson.

7.2.4 Omitted, Not Reached, and Not Presented Items

In the student data files, some items were identified as omitted, not reached, or not presented items depending on the student response data. Item response scores for omits were recoded as "0" in the IRT sparse matrix files *unless* the omitted item were at the end of the test or unit. These items were treated as not reached—items that the student probably did not reach or try to answer. Not reached items were counted as missing or no response, and therefore did not contribute to the item statistics.

7.2.5 Quality Control of the IRT Sparse Matrix Data Files

The IRT sparse data matrices were created by the primary analysts and replicators from Pearson and HumRRO. The matrices were checked for quality and accuracy by comparing the number of students (counts), item category frequencies, and item statistics (e.g., average item score values) between Pearson and HumRRO. Since the same inclusion rules for students were used, all counts, category frequencies, and statistics for all items matched. All

discrepancies in counts were resolved. The programs used to create the IRT statistics were independent, so the QC procedure involved parallel computing. Table 7.1 shows the counts and number of items in the CBT IRT sparse data matrices for each grade in ELA/L.

7.3 Description of the Calibration Process

The IRT calibrations were performed only on the ELA/L CBT tests. The form-to-form linking is established through internal and external common items selected during test construction to represent the blueprint.

7.3.1 Two-Parameter Logistic/Generalized Partial Credit Model

The operational IRT analyses were conducted by both Pearson and HumRRO. The operational items in the IRT sparse data matrix were concurrently calibrated with the two-parameter logistic/generalized partial credit model (2PL/GPC: Muraki, 1992). The 2PL/GPC is denoted

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0; p_{im}(\theta_j)$ is the probability of a student with θ_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 $M_i - 1$; v sequences through each response category through $M_i - 1$.

7.3.2 Treatment of Prose Constructed-Response (PCR) Tasks

The prose constructed-response (PCR) tasks were calibrated at the trait score level (and not as aggregated scores). To address the issue of local independence related to PCR items, a single-calibration "model" approach was used. When sample sizes were large (i.e., greater than 10,000 students), the data were manipulated using random assignment, by selecting one of the two traits for each PCR item for each student. Then one calibration was run so that all trait parameters were independently estimated. When sample sizes were smaller (i.e., field-test samples), a multiple-calibration "model" approach was used. In this alternative approach, the same data set was calibrated two times, each trait represented in one of the two data sets for all students. Then the PCR traits were scaled onto the base scale using non-PCR items as anchor items. These two trait calibration approaches addressed the issue of local dependence while allowing for the accurate calculation of claim scores and the proper weighting of traits in the summative scale scores.

7.3.3 IRT Item Exclusion Rules (Before Calibration)

In addition to checking IRT data for accuracy, Pearson conducted item analyses (IA) to identify items that were not performing as expected and should be considered for removal from calibration and score reporting. The following are the criteria Pearson used to flag extremely problematic items to be dropped from calibration. All "non-spoiled"

items were included in the IRT data matrices; however, the IRTPRO calibration software (Cai et al., 2011) control files were used to exclude from calibration items flagged for the following reasons:

1. A weighted polyserial correlation less than 0.0
2. An average item score of 0.0
3. 100 percent of the students having the same item score, such as:
    - 100 percent omitted the item
    - 100 percent received the same score
    - 100 percent of the responses were at the same score after collapsing score categories due to low frequencies, or
    - 100 percent of the responses were not presented or not reached
4. Insufficient sample sizes for the selected IRT model combinations (i.e., 300 for the 2PL/GPC)
5. High omit rates (i.e., greater than 50 percent) on one or more forms (usually an indication that an item may not be functioning correctly on all forms)

A master list of all problematic items before and after calibration was maintained and all flagged and potentially flawed items were brought to the Priority Alert Task Force (consisting of New Meridian and participating State Leads for member states or agencies) for content and statistical reviews. Ultimately, the decisions about whether to keep or exclude an item from score reporting was made by the Priority Alert Task Force.

7.3.4 IRTPRO Calibration Procedures and Convergence Criteria

The data were calibrated concurrently across forms using the 2PL/GPC model combination. The primary goal was to place the operational item data within each content area and grade/subject on a common difficulty scale. The following are the steps used to calibrate the operational item response data:

1. Using the IRT sparse data matrices, concurrent calibrations were conducted using commercially available IRTPRO for Windows (version 4.2) on CBT data within each grade/subject.
2. IRTPRO Calibration Settings: The logistic partial credit model was specified using the scale constant of 1.0. The prior distributions for latent traits were set to a mean of zero and a standard deviation of one. The number of quadrature points used in the estimation was set to 49. And the slope starting value was set or updated before each run.
3. Each IRTPRO run was inspected for convergence and for any unexpected item-parameter estimates. The PRIORS command in IRTPRO provided a prior on IRT parameters to constrain the calibration so that convergence was more likely. Specifically, option “Guessing[0]” indicated that the prior is placed on the lower asymptote for the 3-PL model, and a normal distribution for the priors with mean of -1.4 and standard deviation 1. For these items, an inspection of item-level statistics and modal-data fit plots were sufficient to ensure that item parameters were acceptable if convergence was reached. Item information functions from the IRTPRO output may also be reviewed. Pearson verified that the maximum number of EM (expectation-maximization) cycles was not reached (which indicated the program did not converge).
4. To convert IRTPRO item parameters to the commonly used logistic parameter presentation (called new item parameters), the following formula was used since IRTPRO uses 1.0 for a scaling constant. There was no need to transfer b- and c-parameters from IRTPRO output. Please note that all unscaled and scaled item parameters were kept on the theta scale. For 2PL models:

New *a*-parameter: $a_{new} = \frac{a_{irtpro}}{1.7}$ (7-2)

5. Pearson reported any need for item-calibration decisions, including convergence issues and extreme parameter estimates, along with proposed resolutions, to the Priority Alert Task Force. Anticipated resolutions included fixing the slope parameters to a minimum .10 value, fixing the guessing parameter to a rational value (1 divided by number of options), and fixing the difficulty parameters at an upper or lower bound, depending on the nature of the problem. If extreme *b*-parameter values were observed (e.g., > 100) and the *a*-parameter values for these items were low (i.e., < 0.10), it was recommended that the prior for the *a*-parameter be set to 0.5.
6. Dropping an item from further processing or dropping an item and rerunning IRTPRO was performed only if it was needed after communication with HumRRO and the Priority Alert Task Force.
7. Inspection of model-data fit plots was helpful in deciding parameter constraints and acceptability of parameter fit. Documentation of each step, after resolution of any issues, was provided by Pearson to New Meridian and HumRRO.

7.3.5 Calibration Quality Control

To ensure IRT calibrations and conversion tables were produced accurately, HumRRO replicated the IRT calibrations and the generation of the score conversion tables. Both Pearson and HumRRO used the same calibration software, IRTPRO. Meetings were held, as needed, so that Pearson and HumRRO could provide status reports and discuss issues related to the IRT work. Pearson performed quality control comparisons between the Pearson and HumRRO item parameter estimates to identify any differences.

Specifically, the following quality control analyses/comparisons were completed:

1. Verified all items were treated the same way (i.e., similar score distributions)
2. Compared IRT item parameter estimates by Pearson and HumRRO (i.e., IRT *a*-, *b*-, and *d*-parameter estimates)
3. Compared the scaling constants for the common item linking sets
4. Compared scaled CBT parameter estimates generated by Pearson and HumRRO
5. Compared all conversion tables produced by Pearson and HumRRO

Exact matches were found between all Pearson and HumRRO conversion tables before scores were reported.

7.4 Model Fit Evaluation Criteria

The usefulness of IRT models is dependent on the extent to which they effectively reflect the data. As discussed by Hambleton et al. (1991), “The advantages of item response models can be obtained only when the fit between the model and the test data of interest is satisfactory. A poorly fitting IRT model will not yield invariant item and ability parameters” (p. 53).

After convergence was achieved for each IRT data set, the IRT model fit was evaluated by doing the following:

1. Calculating the Q_1 statistic and comparing it to a criterion score
2. Calculating the G_2 statistic and comparing it to a criterion score

3. Reviewing graphical output for <u>all</u> items

The Q_1 statistic (Yen, 1981) was used as an index of correspondence between observed and expected performance. To compute Q_1, first the estimated item parameters and student response data (along with observed item scores) were used to estimate student ability $(\hat{\theta})$. Next, expected performance was computed for each item using students' ability estimates in combination with estimated item parameters. Differences between expected item performance and observed item performance were then compared at 10 intervals across the range of student achievement (with approximately the same number of students per interval). Q_1 was computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-squared (χ^2) statistic, which can be compared to a critical chi-squared value to make a statistical inference about whether the data (observed item performance) were consistent with what might be observed if the IRT model was true (expected item performance). Q_1 is not directly comparable across different item types because items with different numbers of IRT parameters have different degrees of freedom (*df*). For that reason, a linear transformation (to a *Z*-score, ZQ_1) was applied to Q_1. This transformation also made item fit results easier to interpret and addressed the sensitivity of Q_1 to sample size.

To evaluate item fit, Yen's Q_1 statistic was calculated for all items. Q_1 is a fit statistic that compares observed and expected item performance. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. For dichotomous items, Q_1 was computed as

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_{ij}(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} \tag{7-3}$$

where N_{ij} was the number of students in interval (or group) j for item i, O_{ij} was the observed proportion of the students for the same cell, and E_{ij} was the expected proportions of the students for the same interval. The expected proportion was computed as

$$E_{ij} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_i(\hat{\theta}_a) \tag{7-4}$$

where $P_i(\hat{\theta}_a)$ was the item characteristic function for item i and students a. The summation is taken over students in interval j.

The generalization of Q_1 for items with multiple response categories is

$$\text{Gen } Q_{1i} = \sum_{j=1}^{10} \sum_{k=1}^{m_i} \frac{N_{ij}(O_{ikj} - E_{ikj})^2}{E_{ikj}} \tag{7-5}$$

where

$$E_{ikj} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_{ik}(\hat{\theta}_a) \tag{7-6}$$

Both Q_1 and generalized Q_1 results were transformed to ZQ_1 and were compared to a criterion $ZQ_{1,crit}$ to determine acceptable fit. The conversion formula was

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}} \tag{7-7}$$

and

$$ZQ_{1,crit} = \frac{N}{1500} \times 4 \tag{7-8}$$

where df is the degrees of freedom. The degrees of freedom is equal to the number of independent cells less the number of independent item parameters. For example, the degrees of freedom for polytomous items equals [10 × (number of score categories–1) – number of independent item parameters]. For the GPCM, the number of independent item parameters equals 1 (for the *a* parameter) plus the number of step values (e.g., for an item scored 0, 1, 2, 3: there are 3 independent step values—the *b* parameter is simply the mean of the step values and is not, therefore, independent).

If Q_1 is found to be excessively sensitive (i.e., a large number of items are flagged for poor fit, even if their item fit plots look reasonable), a likelihood-ratio chi-squared statistic may be computed for each item (Muraki & Bock, 1997):

$$G_i^2 = 2 \sum_{j=1}^{J_i} \sum_{k=1}^{m_i} r_{jik} \times \ln\left(\frac{r_{jik}}{N_{ji} P_{ik}(\bar{\theta}_j)} \right) \tag{7-9}$$

where r_{jik} is the observed frequency of the *k*th categorical response to item i in interval j, N_{ji} is the number of students in interval *j* for item i, $P_{ik}(\bar{\theta}_j)$ is the expected probability of observing the *k*th categorical response to item i for the mean θ in interval j, and J_i is the number of intervals remaining after neighboring intervals are merged, if necessary, to avoid expected values, $N_{ji} P_{ik}(\bar{\theta}_j)$, less than 5. To conduct a standard hypothesis test, the number of degrees of freedom is equal to the number of intervals, J_i, multiplied by $m_i - 1$.

As an alternative to a traditional hypothesis test, the "contingency coefficient" (effect size; Barton & Huynh, 2003) was computed:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{7-10}$$

In this formula, G_i^2 was substituted for χ^2, and N is the sample size on which the IRT parameters were estimated. According to Cohen (1988, pp. 224-225), values of *C* below .10 are considered insignificant, .10+ small, .287+ medium, and .447+ large. A threshold of .35 is recommended (i.e., flag items for which $C \geq .35$).

An item fit-plot was created for each item. Item-fit plots show observed and expected average scores for each interval. Figure 7.1 is an example of ELA/L five-category item calibrated with the 2 PL/GPC model. This item had an n-count of 44,658, Q1=1266.64, ZQ1=147.21 and a criterion $ZQ_{1,crit}$ = 237.02.

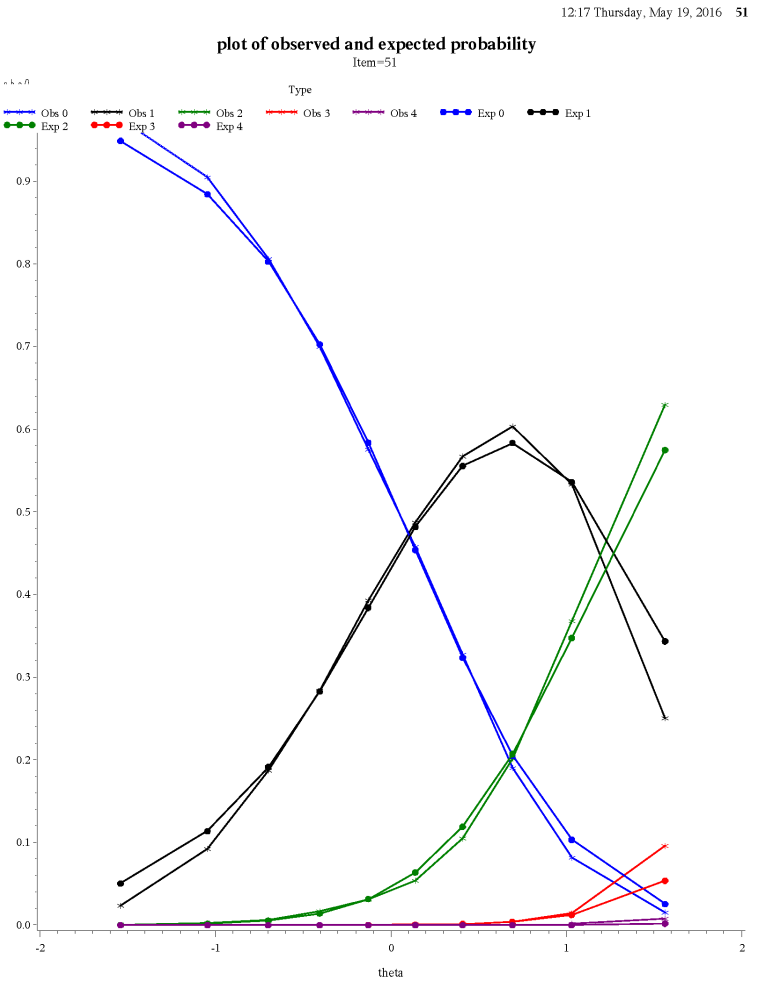

Figure 7.1 ELA/L Item Fit Plot: Observed and Expected Probability

7.5 Items Excluded from Score Reporting

As mentioned previously, after calibration and model fit evaluation were completed, a master list of all problematic items, if warranted, were brought to the Priority Alert Task Force. The Task Force reviewed each item, its content, and the statistical properties, and made decisions about whether to include the item in the operational scores. Sometimes, an item was rejected because it appeared to have content issues, and sometimes an item was excluded because it had unreasonable IRT parameters or showed extremely poor IRT model fit. Ultimately the decision about whether to keep or exclude each flagged item was made by the Task Force.

7.5.1 Item Review Process

The following are the types of problematic items that were brought to the Priority Alert Task Force for evaluation and an "include or exclude" determination was made:

- Extremely difficult items (e.g., an item with a p-value less than 0.02)
- Items with low *a*-parameter estimates (e.g., slope less than 0.10)
- Items flagged for subgroup DIF

The primary goal was to minimize the number of items dropped from the operational test forms. An equally important goal was to not advantage or disadvantage any students.

7.5.2 Count and Percentage of Items Excluded from Score Reporting

All items were calibrated except for 30 items from grade 9 ELA/L and 18 items from grade 11 ELA/L were excluded from IRT calibration because these items were unique to some forms that were administered to small groups of students. For these items, the prior administration item statistics were more stable and more accurate estimates for the item parameters. No items were removed after the IRT calibration. Table 7.2 presents the count and percentage of CBT items excluded from IRT calibration along with the reasons the items were excluded.

Table 7.2 Number and Percentage of ELA/L Items Excluded from IRT Calibration

Grade	Total *n* of CBT Items	*n* of CBT Items Excluded	Percent Excluded	Reason Excluded			
				Small Sample Size	Poor IA Stats	Did Not Calibrate	Other
3	46	0	0%				
4	62	0	0%				
5	64	0	0%				
6	62	0	0%				
7	60	0	0%				
8	62	0	0%				
9	64	30	47%	Yes			
10	62	0	0%				
11	62	18	29%	Yes			

7.6 Scaling Parameter Estimates

Year-to-year linking was performed on all ELA/L CBTs to transform IRT parameters to the base IRT scale. The linking analyses included common-item sets. The linking methodology was based on the Stocking and Lord (1983) test characteristic curve scale transformation method. Year-to-year linking transforms IRT parameters from different years (or administrations) onto the same underlying IRT scale.

HumRRO also used STUIRT (Kim & Kolen, 2004) software to transform their IRTPRO item parameter estimates onto the IRTPRO scales for each grade/subject. HumRRO's scaling constants were compared to those generated by Pearson and found to exactly match.

7.7 Items Excluded from Linking Sets

Robust *Z* (Huynh & Meyer, 2010) and Weighted Root Mean Square Difference (WRMSD) were used to identify outlier items in the linking sets. The following rules were used to identify items for possible exclusion from the linking sets:

1. Exclude an item from the common-item set if different amounts of collapsing resulted in a different number of response categories.
2. Flag and potentially exclude an item from the common-item set if the weighted polyserial correlation, based on the item analysis, was less than 0.10.
3. Exclude items dropped by the Priority Alert Task Force (i.e., due to content or parameter estimation issues).
4. Exclude an item if the scoring rules changed.

After removing items, if necessary, the following steps were performed:

1. Implement the Robust *Z* approach to see if any common items are flagged.
2. Run the initial Stocking and Lord procedure using the STUIRT software.
3. Calculate WRMSD and check to see if any common items exceed the threshold.

4. Re-run STUIRT after removing the items flagged by Robust *Z* and WRMSD.
5. Compare the slopes and intercepts from steps 2 and 4.

Table 7.3 lists the flagging criteria for the WRMSD.

Table 7.3 WRMSD Flagging Criteria for Inspection and Possible Removal of Linking Items

Categories	Points	WRMSD/ Points	WRMSD
2	1	0.100	0.100
3	2	0.075	0.150
4	3	0.075	0.225
5	4	0.075	0.300
6	5	0.075	0.375
7	6	0.075	0.450
>=8	>=7	0.090	0.999

When inspecting items flagged for exclusion from the linking sets, content representation was also considered to avoid removing large numbers of items from the same subclaim. Table 7.4 presents the total number of common items, items excluded from the year-to-year linking sets, and items kept in the linking sets for each grade for ELA/L. The final number of linking items ranged from 8 (in grade 11) to 28 (in grade 8). Grades 3, 4, and 5 had the largest number of items removed from the linking sets due to Robust *Z* for the *a*-parameter and *b*-parameter, some of which were also flagged for high WRMSD.

Table 7.4 Number of ELA/L Items Excluded from the Year-to-Year Linking Sets

Grade	Total *n* of Common Items	Number Excluded	Final Number in Linking Set	Number of Excluded Items by Reason for Exclusion			
				Low Polyserial	Robust *Z* IRT_a	Robust *Z* IRT_b	High WRMSD
3	24	5	19	0	3	2	0
4	27	5	22	0	3	2	0
5	20	5	15	0	1	4	1
6	28	2	26	0	2	0	0
7	24	3	21	0	1	2	1
8	29	1	28	0	1	0	0
9	14	1	13	0	1	0	0
10	31	4	27	0	1	3	0
11	10	2	8	0	1	1	0

Note: WRMSD did not flag any additional items for removal from the common item sets.

7.8 Correlations and Plots of Scaling Item Parameter Estimates

Once the final group of items for each linking set was determined, the *a*- and *b*-parameter estimates were plotted and the correlation between the *a*-parameter estimates and the *b*-parameter estimates were calculated. Table 7.5 presents the number of linking items, total score points of the linking items, and the correlation of the *a*- and *b*-parameter estimates across years.

Table 7.5 Number of Items, Number of Points, and Correlations for ELA/L Year-to-Year Linking Items

	Number		Parameter Correlations	
Grade	**Items**	**Points**	***a*-**	***b*-**
3	19	42	0.9776	0.9960
4	22	49	0.9922	0.9961
5	15	36	0.9759	0.9981
6	26	58	0.9932	0.9887
7	21	48	0.9849	0.9887
8	28	62	0.9838	0.9929
9	13	29	0.9894	0.9927
10	27	60	0.9920	0.9950
11	8	19	0.9721	0.9602

Figures 7.2 and 7.3 are a selection of plots of the *a*- and *b*-parameter estimates for linking items for the year-to-year linking for ELA/L grade 8. For each plot, the *x*-axis is the original (reference) parameter and the *y*-axis is the new parameter after applying the scaling constants.

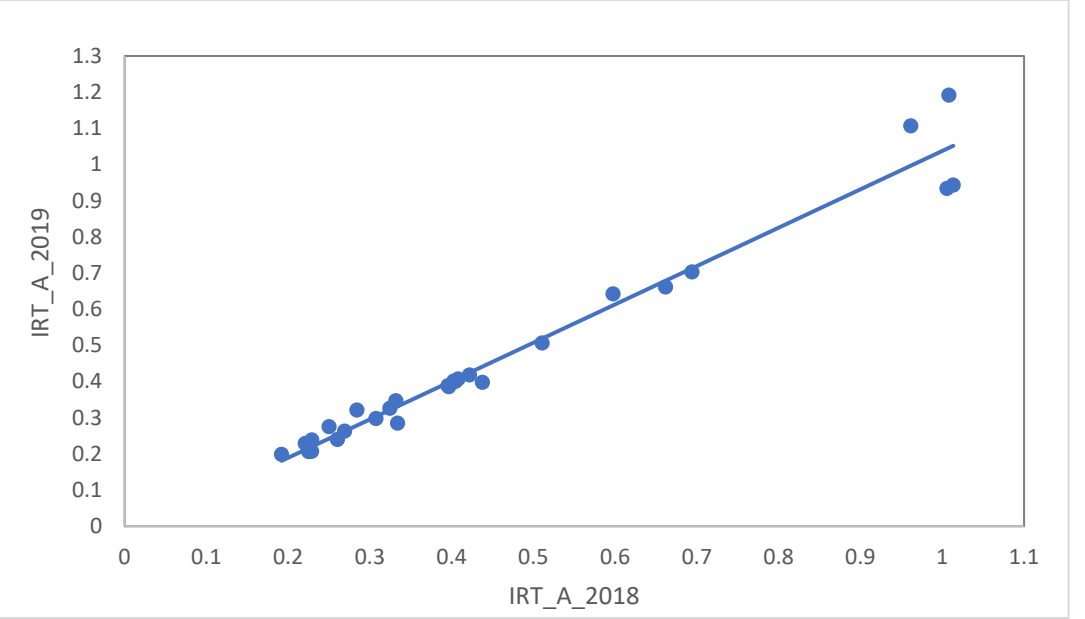

Figure 7.2 ELA/L Grade 8 Transformed New *a*- vs. Reference *a*-Parameter Estimates for Year-to-Year Linking

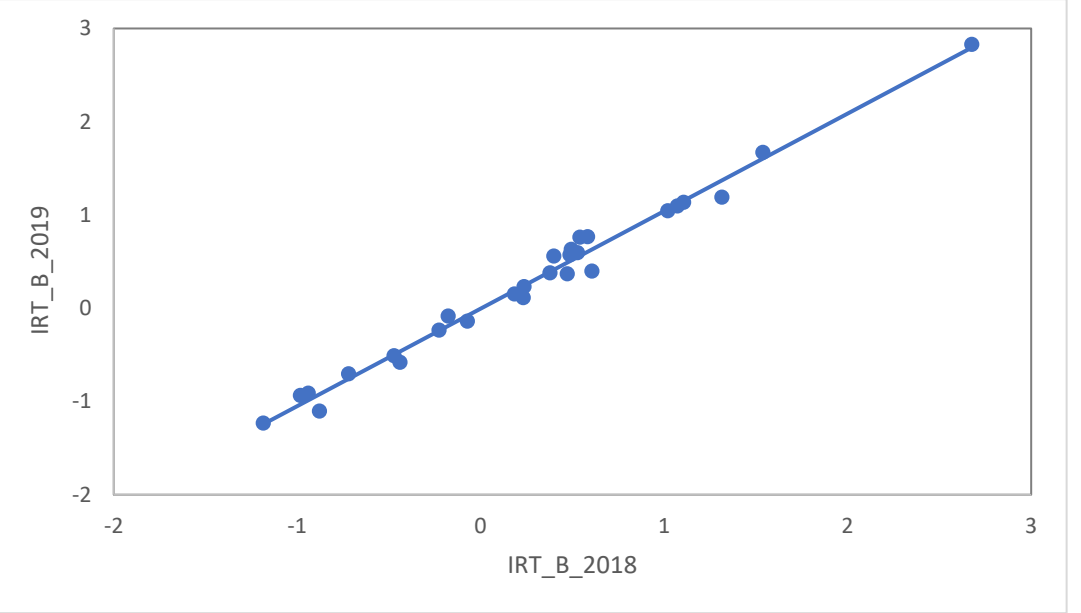

Figure 7.3 ELA/L Grade 8 Transformed New *b*- vs. Reference *b*-Parameter Estimates for Year-to-Year Linking

7.9 Scaling Constants

Table 7.6 presents the slope and intercept scaling constants for ELA/L for the year-to-year linking, derived from STUIRT (Kim & Kolen, 2004) using the Stocking and Lord (1983) test characteristic curve procedure. The slopes and intercepts are similar. The slopes range from 0.9835 to 1.1344, and the intercepts range from 0.0446 to 0.3155.

Table 7.6 Scaling Constants Spring 2018 to Spring 2019 for ELA/L

	Spring 2018 to Spring 2019	
Grade/Subject	**Slope**	**Intercept**
3	1.0292	0.1130
4	1.0759	0.1072
5	1.1013	0.1635
6	1.1049	0.1744
7	1.0993	0.1279
8	1.1344	0.1262
9	1.0849	0.3002
10	1.0806	0.3155
11	0.9835	0.0446

7.10 Summary Statistics and Distributions from IRT Analyses

Tables 7.7 through 7.13 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 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. For ELA/L tests, separate tables were created to display the summary of pre-equated IRT parameter estimates, and the summary of post-equated IRT parameter estimates to reflect the IRT parameters of the items being post-equated. The summary statistics for standard errors of the parameter estimates and the IRT model fit values are only provided for the post-equated ELA/L items.

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.

7.10.1 IRT Summary Statistics for English Language Arts/Literacy

Table 7.7 shows the pre-equated *b*- and *a*-parameter estimates for all ELA/L assessments. Table 7.8 shows the source year for the item statistics for each of the ELA/L assessments that were pre-equated. Table 7.9 summarizes the *b*- and *a*-parameter estimates for the post-equated ELA/L assessments which include post-equated items in spring 2019 and pre-equated items. The number of items in Table 7.9 is consistent with Table 7.7. For forms with too few student responses or special populations, the item parameters were not post-equated. Table 7.10 presents the standard errors (SE) of the post-equated parameters, and Table 7.11 provides model fit information. Only items included in the post-equated calibrations are reported in Tables 7.10 and 7.11. IRT summary statistics are provided in Appendix 7 for ELA/L for all items, reading-only, and writing-only.

Table 7.7 Pre-Equated IRT Summary Parameter Estimates for All Items for ELA/L by Grade

			Summary of *b* Estimates				Summary of *a* Estimates			
Grade	No. of Items	No. of Score Points	Mean	SD	Min	Max	Mean	SD	Min	Max
3	58	128	0.37	0.97	-1.40	3.13	0.59	0.21	0.16	1.01
4	74	164	0.24	1.29	-6.48	2.29	0.45	0.22	0.17	1.02
5	66	145	0.28	1.15	-6.27	2.69	0.49	0.23	0.19	1.06
6	77	172	0.29	0.92	-1.97	4.45	0.51	0.23	0.20	1.13
7	62	139	0.22	0.70	-1.33	1.86	0.49	0.24	0.17	1.18
8	72	159	0.13	0.78	-2.03	2.68	0.47	0.23	0.19	1.12
9	88	197	0.63	0.79	-1.29	2.95	0.52	0.30	0.17	1.44
10	63	141	0.62	0.75	-0.54	2.81	0.50	0.28	0.13	1.24
11	62	139	0.88	0.68	-0.67	2.80	0.46	0.23	0.14	1.10

Table 7.8 Pre-Equated IRT Parameter Distribution by Year for All Items for ELA/L by Grade

Grade	ALL	2014	2015	2016	2017	2018
3	58	0	0	0	21	37
4	74	0	0	0	13	61
5	66	0	0	0	29	37
6	77	0	0	0	27	50
7	62	0	0	10	24	28
8	72	0	0	5	27	40
9	88	0	8	14	9	57
10	63	0	6	4	26	27
11	62	2	2	0	20	38

Table 7.9 Post-Equated IRT Summary Parameter Estimates for All Items for ELA/L by Grade

Grade	No. of Items	No. of Score Points	Summary of *b* Estimates				Summary of *a* Estimates			
			Mean	SD	Min	Max	Mean	SD	Min	Max
3	58	128	0.32	0.91	-1.66	2.05	0.60	0.24	0.22	1.24
4	74	164	0.16	1.55	-9.56	2.35	0.45	0.23	0.12	0.99
5	66	145	0.25	1.06	-5.38	2.63	0.49	0.21	0.10	0.96
6	77	172	0.27	0.87	-1.93	2.95	0.50	0.23	0.18	1.16
7	62	139	0.16	0.73	-1.34	2.37	0.48	0.25	0.17	1.13
8	72	159	0.13	0.82	-1.88	2.83	0.47	0.25	0.18	1.19
9	88	197	0.59	0.80	-1.36	2.95	0.51	0.29	0.14	1.23
10	63	141	0.59	0.79	-0.93	2.85	0.49	0.27	0.14	1.12
11	62	139	0.92	0.83	-0.67	4.55	0.46	0.24	0.08	1.10

Table 7.10 Post-Equated IRT Standard Errors of Parameter Estimates for All Items for ELA/L by Grade

Grade	No. of Items	No. of Score Points	SE of *b* Estimates				SE of *a* Estimates			
			Mean	SD	Min	Max	Mean	SD	Min	Max
3	46	102	0.006	0.003	0.003	0.017	0.006	0.003	0.003	0.017
4	62	137	0.004	0.003	0.002	0.016	0.004	0.003	0.002	0.016
5	64	141	0.005	0.003	0.002	0.014	0.005	0.003	0.002	0.014
6	62	139	0.004	0.003	0.002	0.017	0.004	0.003	0.002	0.017
7	60	135	0.005	0.004	0.001	0.019	0.005	0.004	0.001	0.019
8	62	139	0.005	0.003	0.002	0.019	0.005	0.003	0.002	0.019
9	34	77	0.006	0.004	0.002	0.021	0.006	0.004	0.002	0.021
10	62	139	0.005	0.003	0.002	0.017	0.005	0.003	0.002	0.017
11	44	100	0.012	0.006	0.006	0.029	0.012	0.006	0.006	0.029

Table 7.11 Post-Equated IRT Model Fit for All Items for ELA/L by Grade

			G^2				Q_1			
Grade	No. of Items	No. of Score Points	Mean	SD	Min	Max	Mean	SD	Min	Max
3	46	102	2732.6	2016.8	385.7	10703.0	2574.6	2037.9	360.1	11583.4
4	62	137	3584.0	3089.3	163.8	14358.4	3473.5	3003.8	159.2	14072.8
5	64	141	2920.3	3540.3	151.3	18025.6	2806.2	3505.6	142.6	17306.2
6	62	139	3284.2	2606.4	289.7	13658.8	3055.4	2407.7	291.5	11996.2
7	60	135	3436.0	4207.6	148.1	24499.4	3263.3	4170.4	140.0	26003.2
8	62	139	3502.7	3075.6	125.0	14717.3	3296.8	2871.1	123.0	12427.9
9	34	77	2394.4	2548.5	252.1	13398.9	2225.1	2452.2	226.2	12715.7
10	62	139	2325.6	1874.8	188.5	8318.2	2220.8	1887.5	183.2	8269.9
11	44	100	565.9	320.9	105.9	1718.9	514.5	294.2	104.4	1666.5

7.10.2 IRT Summary Statistics for Mathematics

Table 7.12 shows the *b*- and *a*-parameter estimates for the mathematics assessments. Table 7.13 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 7.12 IRT Summary Parameter Estimates for All Items for Mathematics by Grade/Course

			Summary of *b* Estimates				Summary of *a* Estimates			
Grade	No. of Items	No. of Score Points	Mean	SD	Min	Max	Mean	SD	Min	Max
3	77	110	-0.28	0.98	-2.40	1.90	0.79	0.24	0.32	1.33
4	72	112	-0.15	0.95	-2.61	2.54	0.74	0.20	0.38	1.32
5	71	116	0.02	0.91	-2.21	1.77	0.73	0.27	0.19	1.57
6	69	121	0.36	0.89	-3.02	1.98	0.72	0.24	0.20	1.30
7	67	112	0.75	0.95	-1.03	3.36	0.69	0.29	0.19	1.38
8	64	115	0.91	0.98	-1.12	2.55	0.61	0.21	0.22	1.29
A1	111	209	1.27	1.03	-0.96	3.62	0.58	0.27	0.16	1.41
GO	118	223	1.16	0.94	-1.25	3.83	0.71	0.31	0.19	1.54
A2	109	218	1.41	0.92	-1.53	3.67	0.65	0.29	0.18	1.34
M1	42	81	1.02	0.88	-0.64	2.78	0.62	0.23	0.25	1.39
M2	41	80	1.58	1.30	-0.67	4.68	0.67	0.31	0.17	1.30
M3	40	81	1.39	0.94	-0.35	3.32	0.57	0.27	0.17	1.27

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Table 7.13 IRT Parameter Distribution by Year for All Items for Mathematics by Grade/Course

Grade	ALL	2014	2015	2016	2017	2018
3	77	0	20	10	25	22
4	72	1	20	9	18	24
5	71	0	15	9	16	31
6	69	0	12	7	23	27
7	67	0	15	14	6	32
8	64	0	12	12	13	27
A1	111	0	9	37	27	38
GO	118	0	23	25	33	37
A2	109	0	13	20	36	40
M1	42	0	6	2	21	13
M2	41	0	10	13	10	8
M3	40	0	11	10	6	13

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

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

More detailed descriptions of each performance level, known as policy definitions, are:

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 3–10: 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.

Algebra II, Integrated Mathematics III, and ELA/L Grade 11: Students performing at this level **exceed academic expectations** for the knowledge, skills, and practices contained in the mathematics and ELA/L standards assessed at grade 11. They are very likely to engage successfully in entry-level, credit-bearing courses in mathematics and ELA/L, as well as technical courses requiring an equivalent command of the content area. Students performing at this level are exempt from having to take and pass placement tests in two- and four-year public institutions of higher education designed to determine whether they are academically prepared for such courses without need for remediation.

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 3–10: 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.

Algebra II, Integrated Mathematics III, and ELA/L Grade 11: Students performing at this level **meet academic expectations** for the knowledge, skills, and practices contained in mathematics and ELA/L at grade 11. They are very likely to engage successfully in entry-level, credit-bearing courses in mathematics and ELA/L, as well as technical courses requiring an equivalent command of the content area. Students performing at this level are exempt from having to take and pass placement tests in two- and four-year public institutions of higher education designed to determine whether they are academically prepared for such courses without need for remediation.

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 3–10: 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.

Algebra II, Integrated Mathematics III, and ELA/L Grade 11: Students performing at this level **approach academic expectations** for the knowledge, skills, and practices contained in the ELA/L and mathematics standards assessed at grade 11. They are likely to engage successfully in entry-level, credit-bearing courses in mathematics and ELA/L, as well as technical courses requiring an equivalent command of the content area. **Students performing at Level 3 are strongly encouraged to continue to take challenging high school coursework in English and mathematics through graduation.** Postsecondary institutions are encouraged to use additional information about students performing at Level 3, such as course completion, course grades, and scores on other assessments to determine whether to place them directly into entry-level courses.

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 3–10: 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.

Algebra II, Integrated Mathematics III, and ELA/L Grade 11: Students performing at this level **partially meet academic** expectations for the knowledge, skills, and practices contained in the ELA/L and mathematics standards assessed at grade 11. They will likely need academic support to engage successfully in entry-level, credit-bearing courses, and technical courses requiring an equivalent command of the content area. Students performing at this level are not exempt from having to take and pass placement tests designed to determine whether they are academically prepared for such courses without the need for remediation in two- and four-year public institutions of higher education.

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 3–10: 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.

Algebra II, Integrated Mathematics III, and ELA/L Grade 11: Students performing at this level **do not yet meet academic expectations** for the knowledge, skills, and practices contained in the ELA/L and mathematics standards assessed at grade 11. They will need academic support to engage successfully in entry-level, credit-bearing courses in college algebra, introductory college statistics, and technical courses requiring an equivalent level of mathematics. Students performing at this level are not exempt from having to take and pass placement tests in two- and four-year public institutions of higher education designed to determine whether they are academically prepared for such courses without need for remediation.

8.3 Performance Level Setting Process for the Assessment System

One of the main objectives of the assessment system is to provide information to students, parents, educators, and administrators as to whether students are on track in their learning for success after high school, defined as college- and career-readiness. To set performance levels associated with this objective, 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 in order to establish performance standards for the summative assessments are:

- Step 1: Define outcomes of interest and policy goals
- Step 2: Develop research, data collection, and analysis plans
- Step 3: Synthesize the research results
- Step 4: Conduct pre-policy meeting
- Step 5: Conduct performance level setting (PLS) meetings with panels
- Step 6: Conduct reasonableness review with post-policy panel
- Step 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 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

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

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 full report in the *Performance Level Setting Dry-Run Meeting Report*.

The PLS meetings for the summative assessments were conducted during three one-week sessions. The dates of the twelve PLS committee meetings that were conducted 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 committees' recommendations, this group also voted to conduct a shift in the performance levels to better meet the intended inferences about student performance. Holding the college- and career-ready (or on-track) expectations (i.e., the 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 items on the assessments were developed by Pearson and West Ed and put in the item bank once created.

The 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, accept or reject the item, and 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 groups from New Meridian, Pearson and participating states. 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 QC 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 produce 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:

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

Finally, accuracy checks were performed by checking random documents against scanned data to verify the accuracy of the scanning process.

Once all corrections were made, the scan program was tested a second time to verify all data were valid. When the resulting output showed that no fields were flagged as suspect, the file was considered clean and scoring began. Once all scanning was completed, the right/wrong response data were securely handed off.

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 handscoring 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 keychecks) and categorization of identified issues to help inform investigation by other groups.
- Content Validation: An examination, by subject matter experts, of all items using a combination of automated tools to generate response and scoring data.

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; and
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 and HumRRO implemented quality control procedures to ensure the quality of the work including:

1. Well-defined psychometric specifications
2. Consistently applied data cleaning rules
3. Clear and frequent communication
4. Test run analyses
5. Quality checks of the analyses
6. Checklists for statistical procedures

9.7.1 Pearson Psychometric Quality Control Process

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 Calibration and Scaling), 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.

Pearson Psychometrics and HumRRO established predefined valid case criteria, which were implemented consistently throughout the process. Refer to Section 5.2 for rules for inclusion of students in analyses and Section 7.2 for IRT calibration data preparation criteria and procedures.

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. A subset of items identified as having key issues, scoring issues, or content issues was presented to the Priority Alert Task Force, which made decisions on whether to exclude them from the IRT calibration process and, consequently, the calculation of reported student scores. Refer to Section 5.4 for classical IA item flagging criteria.

Calibrations

Creation of item response theory (IRT) sparse data matrices is an important step before the calibrations can begin. Using the same scored item response data, Pearson and HumRRO teams filtered the data and generated their own sparse data matrices independently. Processing of all data was done in parallel by two psychometricians and compared for number of students. This verification of the data preparation was important to ensure that student exclusion rules were applied consistently across the analyses.

During the calibration process, checks were made to ensure that the correct options for the analyses were selected. Checks were also made on the number of items, number of students with valid scores, IRT item difficulties, standard

errors for the item difficulties, and the consistency between selected IRT statistics to the corresponding statistics obtained during item analyses. Psychometricians also performed detailed reviews of statistics to investigate the extent to which the assumptions of the model fit the observed data. Refer to Section 7.4 for IRT model fit evaluation criteria.

Scaling

During the scaling process, checks were made on the number of linking items, the number of items that were excluded from linking during the stability check of the scaling process, and the scaling constants. Linking items that did not meet the anchor criteria were excluded as linking items. Additionally, items with large weighted root mean square difference (WRMSD) values in Round 1 of scaling were excluded as linking items in Round 2. Finally, reviewers computed the linking constants and then checked that the linking constants were correctly applied. Refer to Section 7.6 for a description of the scaling process.

Conversion Tables

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. All conversion tables were processed in parallel by Pearson and HumRRO 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.

Delivering Item Statistics

Item statistics based on classical item analyses and IRT analyses were obtained during the psychometric analysis process. The statistics were compiled by two data analysts independently to ensure that the correct statistics were delivered for the item bank.

9.7.2 HumRRO Psychometric Quality Control Process

HumRRO served as the psychometric replicator for the operational administration. HumRRO replicated the IRT analyses, scaling analyses, and the conversion file creations. The following steps outline the replication process:

1. Calibrated online data.
2. Sent the item parameter estimates and scaling constants to Pearson for comparison.
3. Reconciled differences, if any, in results with Pearson.
4. Sent data files to Pearson for comparison and reconciled differences, if any.
5. Generated the performance levels, summative, claim, and subclaim conversion tables.
6. Sent conversion tables to Pearson for comparison and reconciled differences, if any.

Section 10: Operational Test Forms

Each operational test form is constructed to reflect the alternate New Meridian blueprint. 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 2019 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 for ELA/L and Mathematics

Grade/Subject	ELA/L		Mathematics	
	CBT	PBT	CBT	PBT
Grade 3	2	1	2	1
Grade 4	2	1	2	1
Grade 5	2	1	2	1
Grade 6	2	1	2	1
Grade 7	2	1	2	1
Grade 8	2	1	2	1
Grade 9	2	1		
Grade 10	2	1		
Grade 11	2	1		
Algebra I			2	1
Geometry			2	1
Algebra II			2	1
Integrated Mathematics I			1	1
Integrated Mathematics II			1	1
Integrated Mathematics III			1	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. Grades 3–11 ELA/L and Integrated Mathematics I, II, and III have two operational accommodated forms and mathematics grades 3–8 and the high school traditional 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.

The summative assessments were administered in either a computer-based test (CBT) or a paper-based test (PBT) format. ELA/L assessments focused on writing effectively when analyzing text. Mathematics assessments focused on applying skills and concepts, and featured multi-step problems that require abstract reasoning and modeling of real-world problems. In both content areas, students also demonstrated their acquired skills and knowledge by answering selected response items and fill-in-the-blank questions. Each assessment was comprised of multiple units; one of the mathematics units was split into calculator and non-calculator sections.

Section 11: Student Characteristics

11.1 Overview of Test Taking Population

Approximately two million students from the Bureau of Indian Education, Illinois, New Jersey, and New Mexico participated in the operational administration of the summative assessments during the 2018–2019 school year. Not all participating states and agencies had students testing in all grades. Assessments were administered for English language arts/literacy (ELA/L) in grades 3 through 11; mathematics assessments were administered in grades 3 through 8, as well as for traditional high school mathematics (Algebra I, Geometry, and Algebra II) and integrated high school mathematics (Integrated Mathematics I, II, and III). A small subset of students tested in ELA/L grades 9, 10, and 11, and Algebra I, Geometry, and Algebra II during fall of 2018. Student characteristics for this group are presented in an addendum. The majority of students tested during the spring administration when all grades and content areas were administered online and on paper.

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 participating states and agencies 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), and
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 chosen. Records for students with administration issues or anomalies were excluded from analyses.

11.3 Students by Grade/Course, Mode, and Gender

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, and Table 11.3 provides this information for students who took the mathematics assessments in Spanish.

Markedly more students tested online than on paper across all grades for both content areas. For ELA/L, the percentages of online students by grade level, for all states combined, ranged from 87.6 percent to 99.5 percent, while the percentages of paper test students ranged from .5 percent to 12.4 percent. For all mathematics students, the percentages of students testing online ranged from 87.7 percent to 100 percent, whereas the percentages of students testing on paper ranged from 0 percent to 12.3 percent. The percentages of students taking Spanish-language mathematics online forms ranged from 84.4 percent to 99.8 percent and the percentages of students taking Spanish-language mathematics paper forms ranged from 0 percent to 15.6 percent. Generally, the percentage of students who tested online increased steadily from the lower grades to the higher grades. For example, about 88 percent of the ELA/L grade 3 students tested online, while about 98 to 100 percent of the high school students tested online. Overall, fewer students tested at the higher grades for both content areas.

Table 11.1 ELA/L Students by Grade and Mode: All States Combined

	No. of Valid	CBT		PBT	
Grade	Cases	N	%	N	%
3	256,870	224,957	87.6	31,913	12.4
4	265,169	259,642	97.9	5,527	2.1
5	271,778	267,807	98.5	3,971	1.5
6	275,277	271,346	98.6	3,931	1.4
7	269,386	265,686	98.6	3,700	1.4
8	266,251	263,370	98.9	2,881	1.1
9	121,619	121,061	99.5	558	0.5
10	118,322	117,751	99.5	571	0.5
11	34,610	34,035	98.3	575	1.7
Grand Total	1,879,282	1,825,655	97.1	53,627	2.9

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

Table 11.2 Mathematics Students by Grade/Course and Mode: All States Combined

Grade/Course	No. of Valid Cases	CBT		PBT	
		N	%	N	%
3	258,807	226,933	87.7	31,874	12.3
4	266,629	261,092	97.9	5,537	2.1
5	272,714	268,724	98.5	3,990	1.5
6	275,732	271,798	98.6	3,934	1.4
7	264,960	261,252	98.6	3,708	1.4
8	225,726	222,869	98.7	2,857	1.3
A1	134,107	133,482	99.5	625	0.5
GO	105,010	104,444	99.5	566	0.5
A2	66,789	66,317	99.3	472	0.7
M1	673	672	99.9	n/r	n/r
M2	541	540	99.8	n/r	n/r
M3	201	201	100.0	n/r	n/r
Grand Total	1,871,889	1,818,324	97.1	53,565	2.9

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; A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III; n/r = not reported due to n<20.

Table 11.3 Spanish-Language Mathematics Students by Grade/Course and Mode: All States Combined

Grade/Course	No. of Valid Cases	CBT		PBT	
		N	%	N	%
3	4,812	4,063	84.4	749	15.6
4	3,691	3,668	99.4	23	0.6
5	3,270	3,247	99.3	23	0.7
6	2,642	2,623	99.3	n/r	n/r
7	2,255	2,243	99.5	n/r	n/r
8	2,118	2,108	99.5	n/r	n/r
A1	2,426	2,400	98.9	26	1.1
GO	1,728	1,721	99.6	n/r	n/r
A2	558	557	99.8	n/r	n/r
M1	n/r	n/r	n/r	n/r	n/r
M2	n/r	n/r	n/r	n/r	n/r
M3	n/r	n/r	n/r	n/r	n/r
Grand Total	23,534	22,664	96.3	870	3.7

Note: CBT = computer-based test; PBT = paper-based test; A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III; n/r = not reported due to n<20.

Tables A.11.1, A.11.2, and A.11.3 in Appendix 11 show the number and percentage of students with valid test scores in each content area (including Spanish-language mathematics), grade/course, and mode of assessment for all states and agencies combined and for each state or agency separately. Tables A.11.4, A.11.5, and A.11.6 present the distribution by content area, grade/course, mode, and gender, for all states combined.

11.4 Demographics

Also presented in Appendix 11 is student demographic information for the following characteristics: economically disadvantaged, students with disabilities, English learners (EL), gender, and race/ethnicity (American Indian/Alaska

Native; Asian; Black/African American; Hispanic/Latino; White/Caucasian; Native Hawaiian or Other Pacific Islander; two or more races reported; race not reported). Student demographic information was provided by the states and districts and captured in PearsonAccess^{next} by means of a student data upload. The demographic data was verified by the states and districts prior to score reporting.

Tables A.11.7 through A.11.15 provide demographic information for students with valid ELA/L scores, and Tables A.11.16 through A.11.26 present demographics for students with valid mathematics scores. All tables of demographic information are organized by grade/course; the results are first aggregated across all participating states and agencies and then presented for each state or agency. Percentages are not reported in which fewer than 20 students tested in a grade/course area.

Section 12: Scale Scores

Participating states and agencies report 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.[3] 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[4] 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.

[3] Addendum 12 presents a summary of results on scale scores for the fall 2018 administration.
[4] Table A.12.1 in Appendix 12.1 is identical to Table 12.1.

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.

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

Grade	Score	Possible Points		
		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
	Total	15	15	12
4–5	Reading	4	4	0
	Written Expression	12	12	9
	Knowledge of Conventions	3	3	3
	Total	19	19	12
6–11	Reading	4	4	0
	Written Expression	12	12	12
	Knowledge of Conventions	3	3	3
	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 [5] 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.

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 and Applications	3	12
TOTAL		33	52

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.

12.2 Establishing the Reporting Scales

Reporting scales designate student performance into one of five performance levels [6] 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.

[5] Table A.12.10 in Appendix 12.1 is identical to Table 12.3.
[6] Section 8 provides an overview of the performance level setting process, and detailed information can be found in the Performance Level Setting Technical Report.

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 is the Level 2 and Level 4. The cuts are the anchors for establishing the linear transformation between the theta scale and the reported scale score. A scale score of 700 is associated with minimum Level 2 performance, and a scale score of 750 is associated with minimum Level 4 performance. Not all possible scale scores may be realized in a scoring table.

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

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

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

$$A_{2015} = \frac{750 - 700}{\theta_{2015_{Level\,4}} - \theta_{2015_{Level\,2}}} \quad (12\text{-}2)$$

and

$$B_{2015} = 750 - A_{2015} \times \theta_{2015_{Level\,4}} \quad (12\text{-}3)$$

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 θ_{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 (θ_{2015}) and the 2016 theta scale (θ_{2016}). These values were included in the scale score formula, and the formulas were used to solve for the slope (SA_{2016}) and (SB_{2016}) intercept for 2016.

The slope (A_{2016}) was updated using the following formula:

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

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 12.2.

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 (θ) scale for the items associated with a particular subclaim. The theta values and corresponding raw scores associated with the Level 3 and Level 4 performance levels were identified using the 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, θ_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:

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

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

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, θ_j; $p_{im}(\theta_j)$ is the probability of a student, j, with θ_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:

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

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

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

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

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

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

Step 4: Match every raw score with a theta. θ_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.[7] The curves in each figure are for the two core online forms (O1 and O2), one core paper form (P1), and two or three 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 very similar TCCs. CSEM and INF curves were also similar.

Appendix 12.3 contains TCC, CSEM, and INF curves for all ELA/L grades and all mathematics grades/courses. Both pre-equated and post-equated curves are provided for ELA/L. The pre-equated curves are based on IRT parameters from a prior operational or field-test administration. The post-equated curves are based on IRT parameters estimated using the spring 2019 post-equating sample. Pre-equated curves are provided for the mathematics assessments.

[7] Grade 3 TCC, CSEM, and INF curves are also included in Appendix Figure A.12.1.

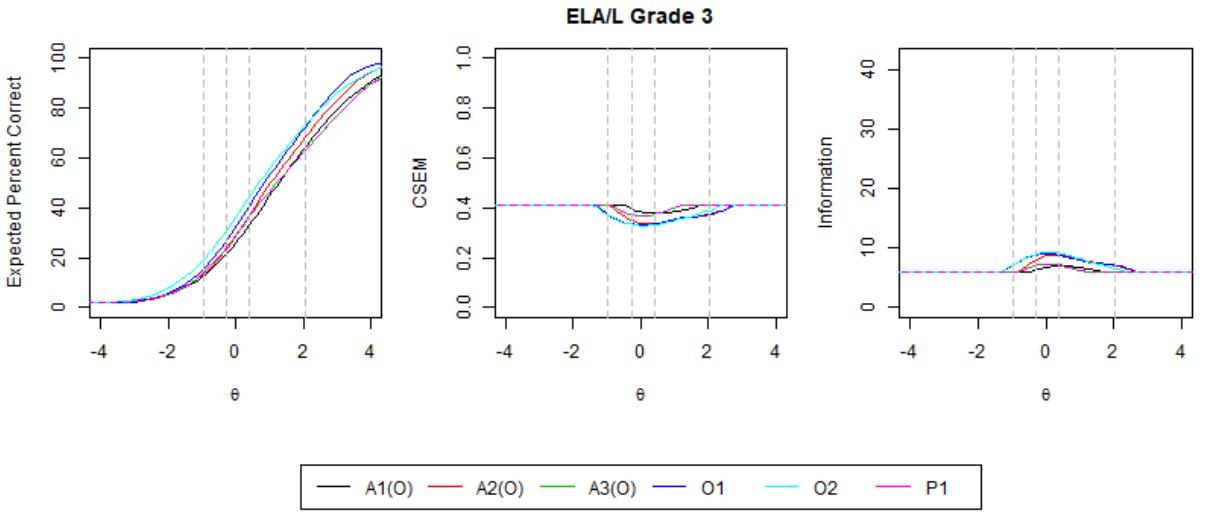

Figure 12.1 Test Characteristic Curves, Conditional Standard Error of Measurement Curves, and Information Curves for ELA/L Grade 3 (Post-Equated)

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 11 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 .02 and the *x*-axis from 650 to 850. For the Reading distributions, the *y*-axis ranges from 0 to .05 and the *x*-axis from 10 to 90. For the Writing distributions, the *y*-axis ranges from 0 to .10 and the *x*-axis from 10 to 60.

The distributions of the ELA/L summative scale scores were fairly symmetrical and centered around the Level 4 cut score (750).

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 fairly symmetric.

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 not likely to be obtained resulting in multiple peaks across the range of the Writing scale score. A noticeable proportion of students earned the LOSS of 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, zero students obtained scale scores in the range of eleven to seventeen.[8] 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 in the range of seventeen to nineteen. Whereas, the Reading lowest theta estimates associated with raw scores ranging from one to two are linearly transformed to scale score values in the range of ten to eleven. 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.

[8] 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.

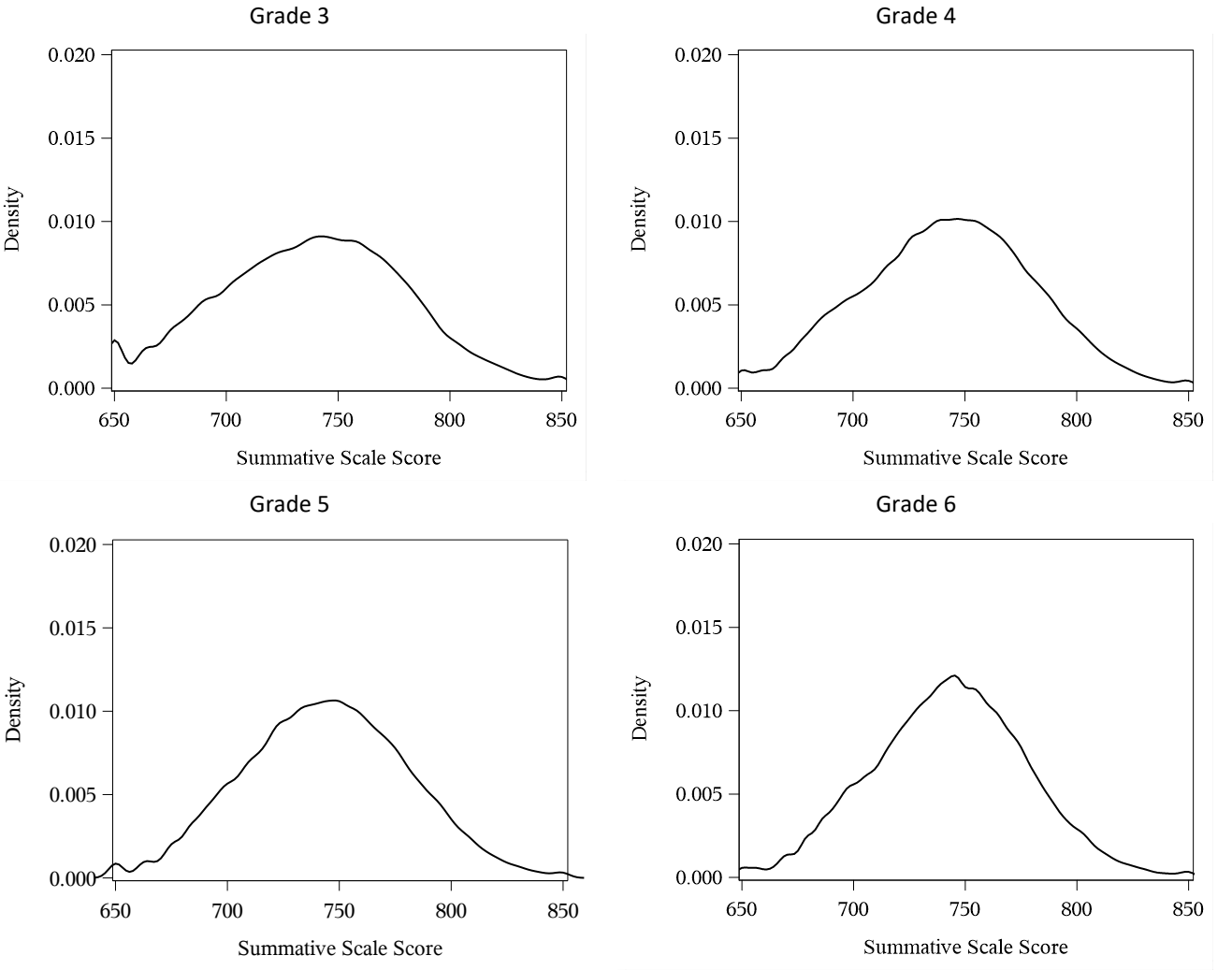

Figure 12.2 Distributions of ELA/L Scale Scores: Grades 3–11

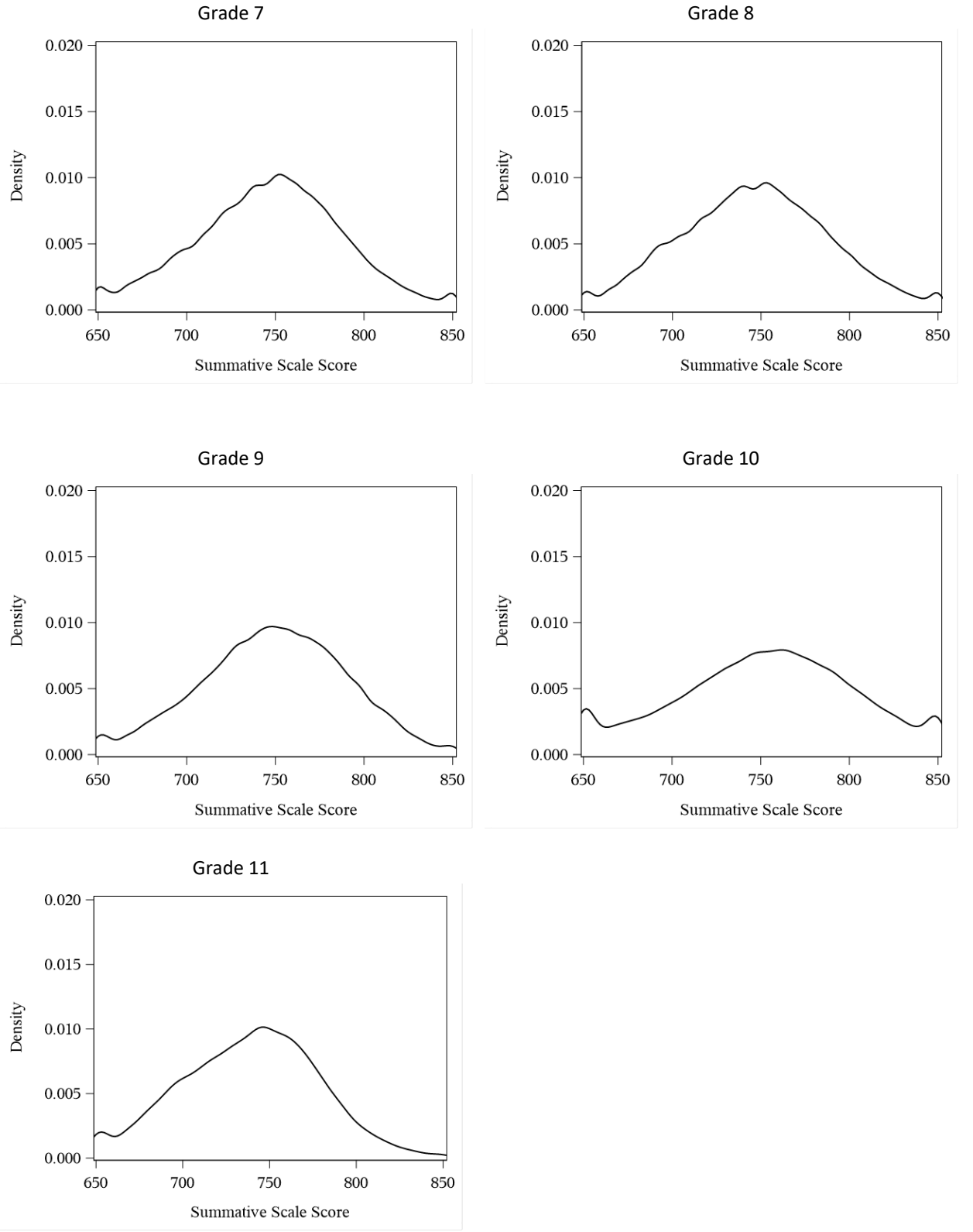

Figure 12.2 (continued) Distributions of ELA/L Scale Scores: Grades 3–11

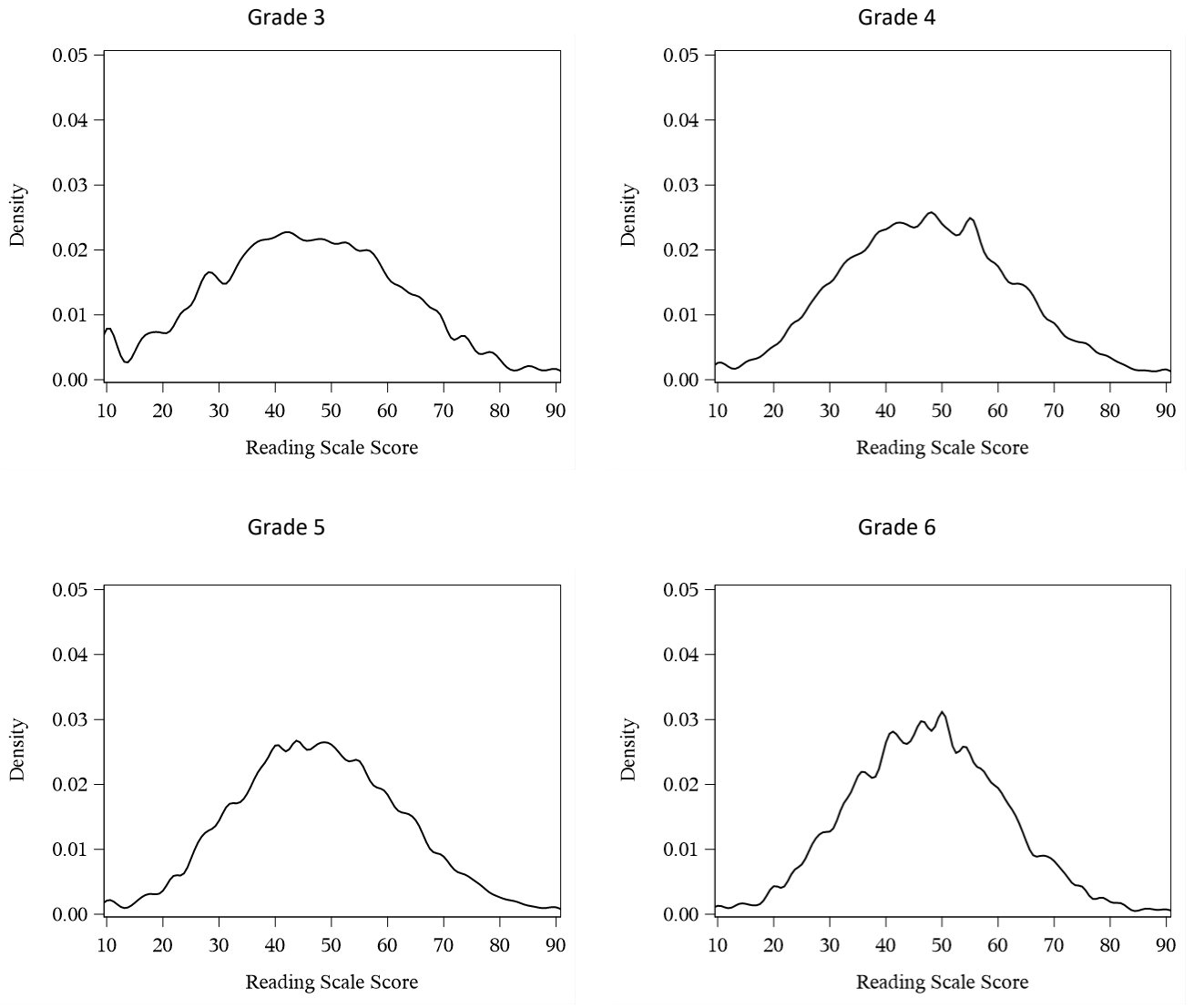

Figure 12.3 Distributions of Reading Scale Scores: Grades 3–11

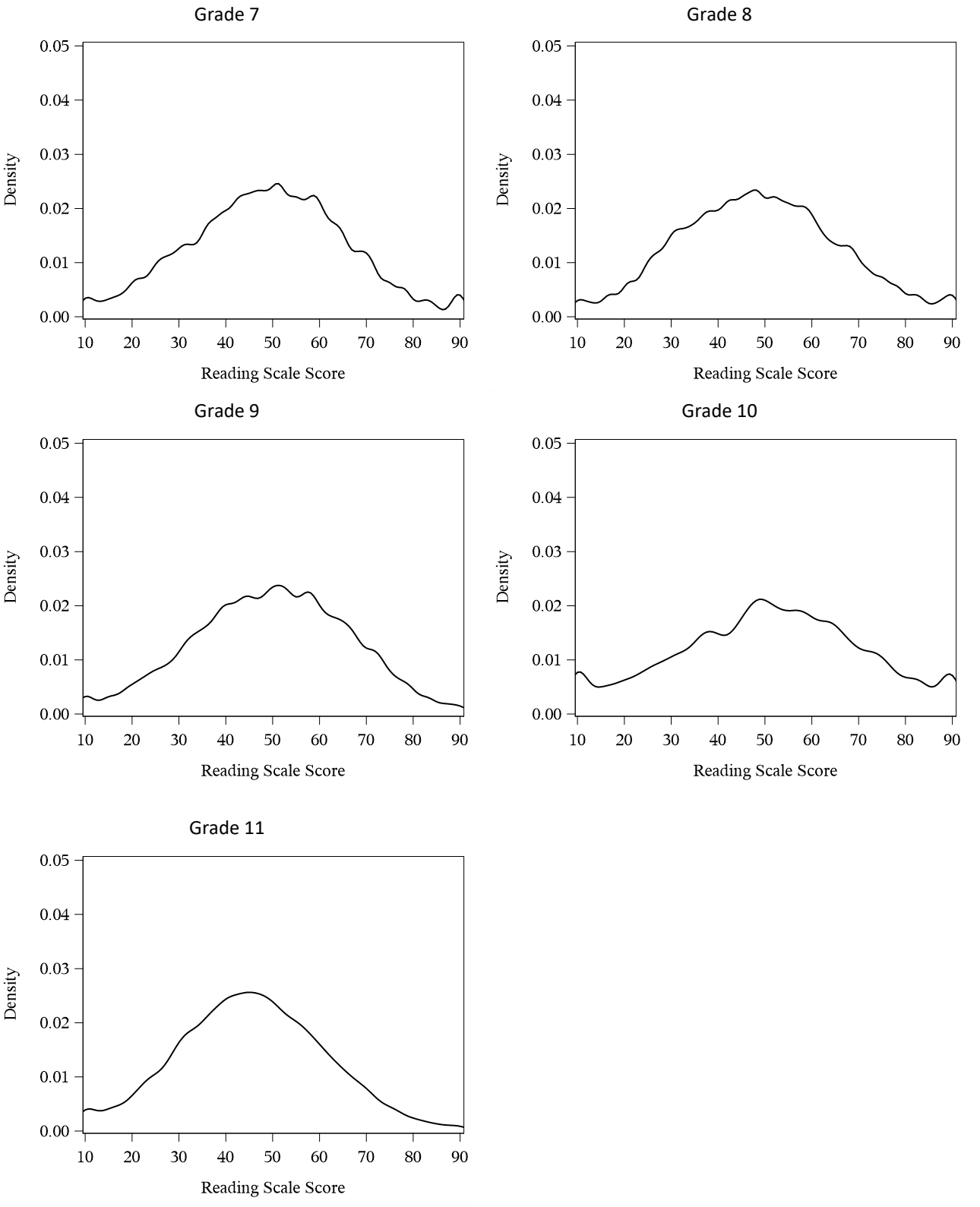

Figure 12.3 (continued) Distributions of Reading Scale Scores: Grades 3–11

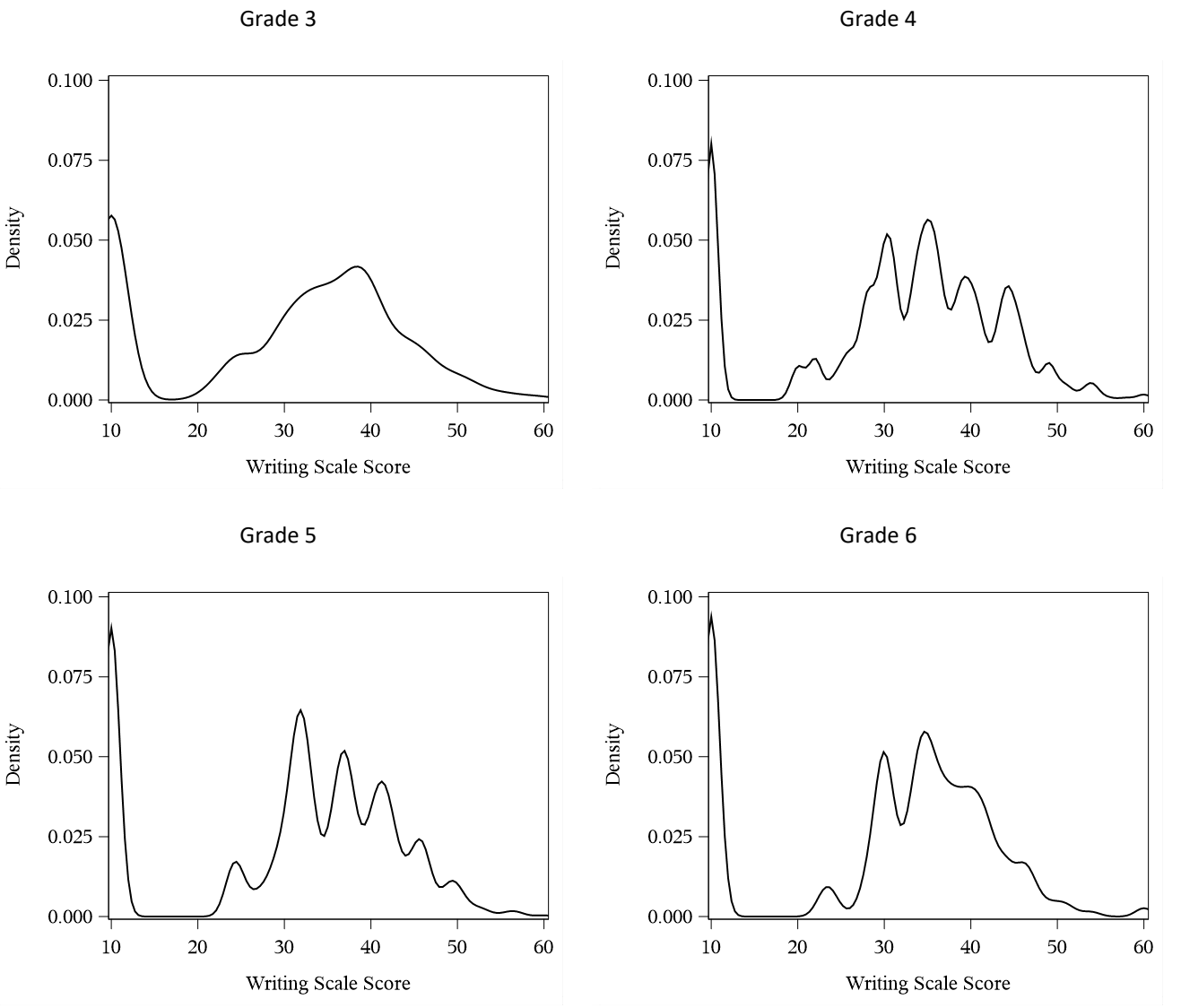

Figure 12.4 Distributions of Writing Scale Scores: Grades 3–11

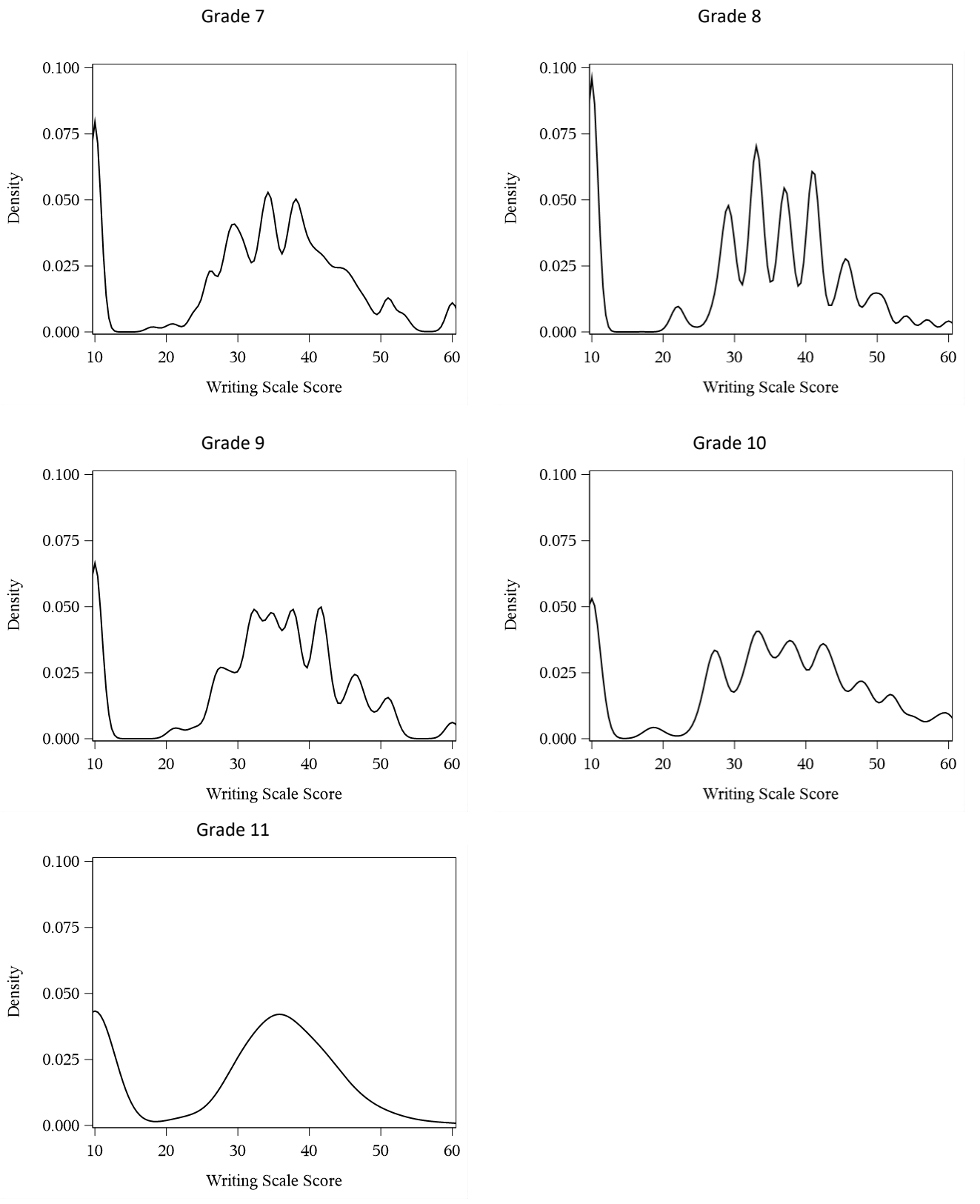

Figure 12.4 (continued) Distributions of Writing Scale Scores: Grades 3–11

12.4.2 Scale Score Cumulative Frequencies for ELA/L

The cumulative frequency distribution for the summative scale score is presented in Appendix 12.4 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 and 12.6 [9] for ELA/L grades 3 and 9, respectively. The results for all ELA/L grades are provided in Appendix 12.5. Grade 3 ELA/L subgroup statistics are presented in Table 12.5. [10] Mean scores were higher for female students relative to male students. Mean scores were highest for Asian students and were lowest for American Indian/Alaska native students. Economically disadvantaged students performed less well than students who are not economically disadvantaged. English learners (EL) performed less well than non-EL students. Students with disabilities performed less well than students without disabilities. Patterns of mean scale scores were similar in grades 4 through 8, although the ordering of ethnicity subgroups varied slightly; corresponding tables for all grades are presented in Appendix 12.5.

[9] Due to omitted demographic values, subgroup sample sizes may not sum to the total sample size.
[10] Table A.12.48 in Appendix 12.5 is identical to Table 12.5.

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

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**256,870**	**738.54**	**42.05**	**650**	**850**
Gender	Female	125,311	743.09	42.23	650	850
	Male	131,559	734.21	41.43	650	850
Ethnicity	American Indian/Alaska Native	4,217	715.63	36.47	650	850
	Asian	17,994	767.23	40.36	650	850
	Black/African American	38,832	722.71	41.08	650	850
	Hispanic/Latino	77,952	727.08	40.23	650	850
	Native Hawaiian or Pacific Islander	357	751.86	39.65	650	850
	Two or more races	8,340	743.03	42.71	650	850
	White	109,159	748.14	38.83	650	850
Economic Status*	Not Economically Disadvantaged	129,667	753.06	39.30	650	850
	Economically Disadvantaged	126,919	723.77	39.53	650	850
English Learner Status	Non-English Learner	218,282	743.05	41.39	650	850
	English Learner	38,373	713.08	36.30	650	850
Disabilities	Students without Disabilities	212,957	744.14	40.21	650	850
	Students with Disabilities	43,174	711.14	40.16	650	850
Reading Summative Score		**256,870**	**45.51**	**16.86**	**10**	**90**
Gender	Female	125,311	46.64	16.74	10	90
	Male	131,559	44.44	16.91	10	90
Ethnicity	American Indian/Alaska Native	4,217	36.33	14.35	10	90
	Asian	17,994	56.08	16.27	10	90
	Black/African American	38,832	38.93	16.01	10	90
	Hispanic/Latino	77,952	40.45	15.82	10	90
	Native Hawaiian or Pacific Islander	357	49.90	15.57	10	90
	Two or more races	8,340	48.03	17.14	10	90
	White	109,159	49.88	15.79	10	90
Economic Status*	Not Economically Disadvantaged	129,667	51.50	15.90	10	90
	Economically Disadvantaged	126,919	39.42	15.58	10	90
English Learner Status	Non-English Learner	218,282	47.44	16.61	10	90
	English Learner	38,373	34.62	13.89	10	90
Disabilities	Students without Disabilities	212,957	47.62	16.18	10	90
	Students with Disabilities	43,174	35.22	16.38	10	90
Writing Summative Score		**256,870**	**29.46**	**13.48**	**10**	**60**
Gender	Female	125,311	31.43	13.20	10	60
	Male	131,559	27.59	13.49	10	60
Ethnicity	American Indian/Alaska Native	4,217	23.77	12.65	10	59
	Asian	17,994	37.67	11.66	10	60
	Black/African American	38,832	25.52	13.56	10	60
	Hispanic/Latino	77,952	27.00	13.39	10	60
	Native Hawaiian or Pacific Islander	357	33.95	12.35	10	60
	Two or more races	8,340	29.67	13.77	10	60
	White	109,159	31.46	12.83	10	60
Economic Status*	Not Economically Disadvantaged	129,667	33.09	12.64	10	60
	Economically Disadvantaged	126,919	25.77	13.31	10	60
English Learner Status	Non-English Learner	218,282	30.47	13.34	10	60
	English Learner	38,373	23.78	12.88	10	60
Disabilities	Students without Disabilities	212,957	31.13	12.99	10	60
	Students with Disabilities	43,174	21.28	12.88	10	60

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

Grade 9 subgroup statistics for ELA/L, Reading, and Writing scale scores are presented in Table 12.6.[11] Mean scores were very similar to what was observed for grades 3 through 8. Mean scores were higher for female students than for male students. Mean scores were highest for Asian students and were lowest for American Indian/Alaska native students. Economically disadvantaged students performed less well than students who are not economically disadvantaged. English learners (EL) performed less well than non-EL students. Students with disabilities performed less well than students without disabilities. Similar patterns are observed in other high school assessments, with some small variations in the ordering of the ethnicity groups. Corresponding tables for grades 10 and 11 are presented in Appendix 12.5.

[11] Table A.12.54 in Appendix 12.5 is identical to Table 12.6.

Table 12.6 Subgroup Performance for ELA/L Scale Scores: Grade 9

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**121,619**	**748.83**	**40.72**	**650**	**850**
Gender	Female	59,248	755.81	39.13	650	850
	Male	62,371	742.20	41.09	650	850
Ethnicity	American Indian/Alaska Native	2,909	726.65	31.07	650	850
	Asian	10,492	783.32	36.13	650	850
	Black/African American	14,260	732.03	37.18	650	850
	Hispanic/Latino	42,054	733.28	37.91	650	850
	Native Hawaiian or Pacific Islander	239	759.91	41.41	650	850
	Two or more races	1,760	759.03	39.97	650	850
	White	49,897	760.38	36.58	650	850
Economic Status*	Not Economically Disadvantaged	73,488	760.42	38.92	650	850
	Economically Disadvantaged	48,075	731.16	36.85	650	850
English Learner Status	Non-English Learner	114,062	752.17	39.11	650	850
	English Learner	7,505	698.38	29.92	650	821
Disabilities	Students without Disabilities	99,057	755.24	38.70	650	850
	Students with Disabilities	22,510	720.74	37.30	650	850
Reading Summative Score		**121,619**	**49.70**	**16.45**	**10**	**90**
Gender	Female	59,248	51.46	16.01	10	90
	Male	62,371	48.03	16.69	10	90
Ethnicity	American Indian/Alaska Native	2,909	40.24	13.02	10	87
	Asian	10,492	62.27	14.70	10	90
	Black/African American	14,260	43.18	15.21	10	90
	Hispanic/Latino	42,054	43.65	15.39	10	90
	Native Hawaiian or Pacific Islander	239	52.93	16.43	10	90
	Two or more races	1,760	54.03	16.33	10	90
	White	49,897	54.40	14.95	10	90
Economic Status*	Not Economically Disadvantaged	73,488	54.24	15.76	10	90
	Economically Disadvantaged	48,075	42.77	15.01	10	90
English Learner Status	Non-English Learner	114,062	51.01	15.86	10	90
	English Learner	7,505	29.84	11.91	10	83
Disabilities	Students without Disabilities	99,057	52.07	15.75	10	90
	Students with Disabilities	22,510	39.28	15.39	10	90
Writing Summative Score		**121,619**	**32.95**	**12.21**	**10**	**60**
Gender	Female	59,248	35.66	11.05	10	60
	Male	62,371	30.37	12.69	10	60
Ethnicity	American Indian/Alaska Native	2,909	27.67	11.08	10	60
	Asian	10,492	42.24	9.58	10	60
	Black/African American	14,260	28.55	11.96	10	60
	Hispanic/Latino	42,054	28.84	12.12	10	60
	Native Hawaiian or Pacific Islander	239	36.95	11.34	10	60
	Two or more races	1,760	35.34	11.64	10	60
	White	49,897	35.92	10.86	10	60
Economic Status*	Not Economically Disadvantaged	73,488	36.03	11.28	10	60
	Economically Disadvantaged	48,075	28.24	12.06	10	60
English Learner Status	Non-English Learner	114,062	33.84	11.76	10	60
	English Learner	7,505	19.36	10.64	10	50
Disabilities	Students without Disabilities	99,057	34.90	11.27	10	60
	Students with Disabilities	22,510	24.37	12.47	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 .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. Figure 12.6 graphically represents the distributions of scale scores for Algebra I, Geometry, Algebra II, and Integrated Mathematics I, II, and III. Scale score distributions generally peaked between approximately 700 and the 750 Level 4 performance level cut score for Algebra I and Geometry. Algebra II distribution is flat from approximately 700 to 775. Integrated Mathematics I, II, and III distributions peaked slightly around or below 700.

12.4.5 Scale Score Cumulative Frequencies for Mathematics

The cumulative frequency distribution for the summative scale score is presented in Appendix 12.4 for mathematics assessments.

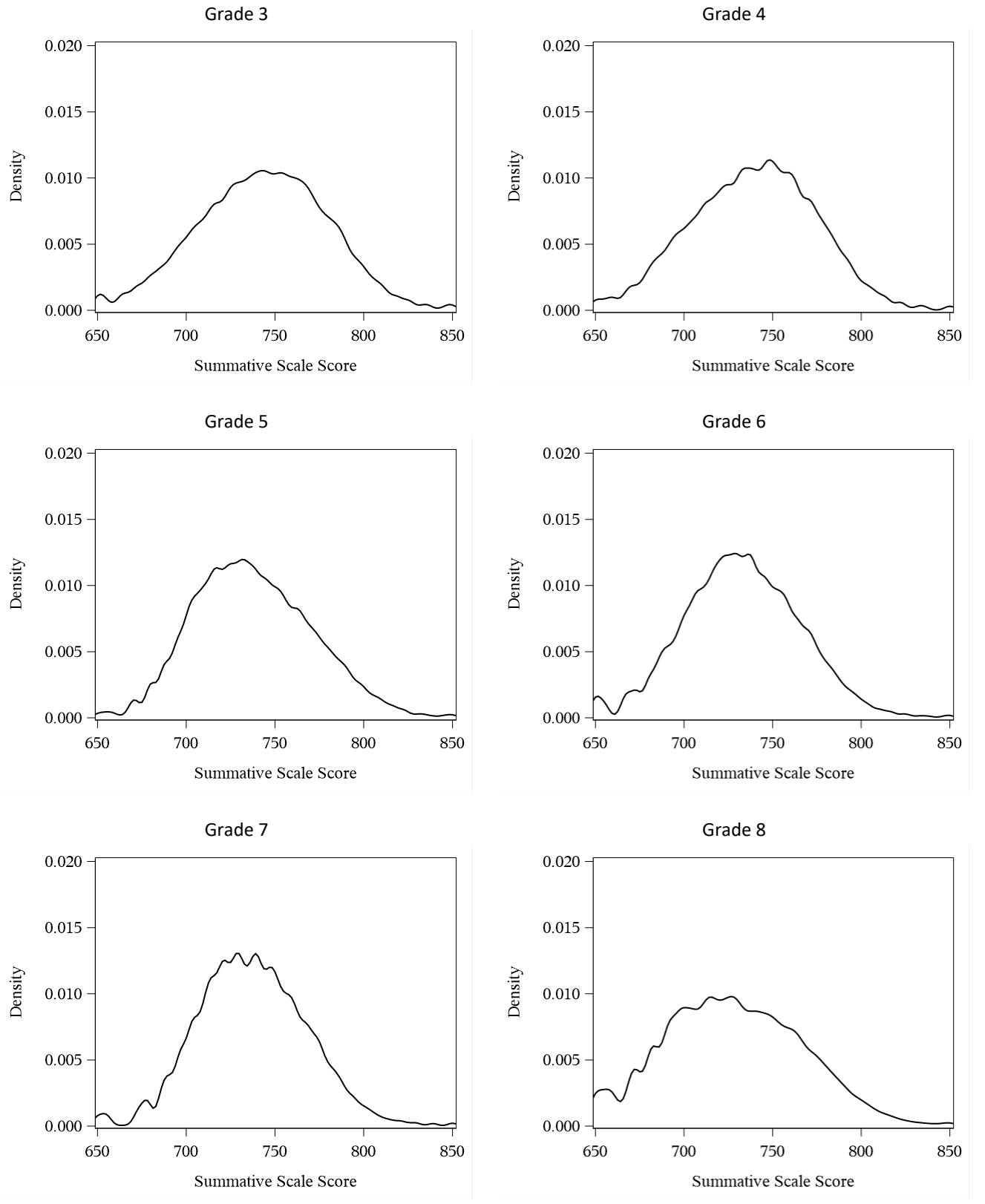

Figure 12.5 Distributions of Mathematics Scale Scores: Grades 3–8

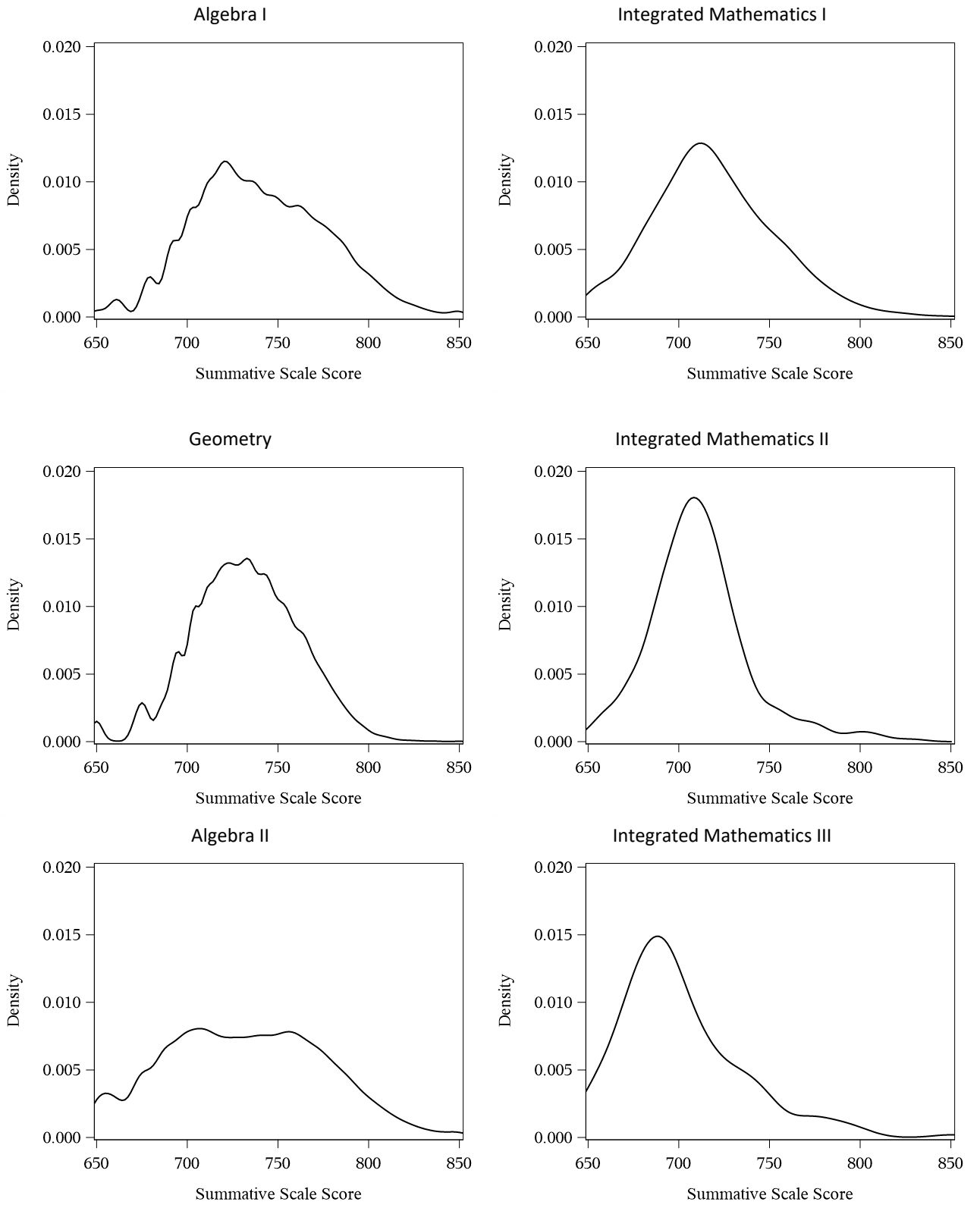

Figure 12.6 Distributions of Mathematics Scale Scores: High School

12.4.6 Summary Scale Score Statistics for Mathematics Groups

Subgroup statistics for mathematics scale scores are presented in Tables 12.7–12.9 [12] for grade 3, Algebra I, and Integrated Mathematics I, respectively. Grade 3 subgroup statistics are presented in Table 12.7. [13] Mean scores were similar for female and male students. Mean scores were highest for Asian students and were lowest for American Indian/Alaska native students. Economically disadvantaged students performed less well than students who are not economically disadvantaged. English learners (EL) performed less well than non-EL students. Students with disabilities performed less well than students without disabilities. Students using the Spanish Language form tended to have lower mean scores. Generally similar patterns were observed in grades 4 to 8, with some slight variations in the orderings of the ethnicity subgroups. Corresponding tables for all grades/courses are presented in Appendix 12.5.

Algebra I scale score statistics are presented in Table 12.8. [14] Mean scores were slightly higher for female students relative to male students. Mean scores were highest for Asian students and were lowest for American Indian/Alaska native students. Economically disadvantaged students performed less well than students who are not economically disadvantaged. English learners (EL) performed less well than non-EL students. Students with disabilities performed less well than students without disabilities. Students using the Spanish Language form tended to have lower mean scores. Similar patterns were observed in the other high school tests with some of the previously mentioned exceptions in the ordering of the ethnicities applying to these tests as well. In some instances, male students reported higher means than female students. Corresponding tables are presented in Appendix 12.5.

Integrated Mathematics I scale score statistics are presented in Table 12.9. [15] Mean scores were higher for female students relative to male students. Mean scores were highest for White students and were lowest for Hispanic/Latino students. Economically disadvantaged students performed less well than students who are not economically disadvantaged. English learners (EL) performed less well than non-EL students. Sample sizes for Integrated Mathematics I subgroups tended to be small, and some categories did not have sufficient sample sizes for reporting purposes in this table. Somewhat similar patterns were observed in Integrated Mathematics II and Integrated Mathematics III, but sample sizes for some subgroups are very small, and caution should be used in interpretations. Tables for these tests can be found in Appendix 12.5.

[12] Due to omitted demographic values, subgroup sample sizes in these tables may not sum to total sample size.
[13] Table A.12.57 in Appendix 12.5 is identical to Table 12.7.
[14] Table A.12.63 in Appendix 12.5 is identical to Table 12.8.
[15] Table A.12.66 in Appendix 12.5 is identical to Table 12.9.

Table 12.7 Subgroup Performance for Mathematics Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**258,807**	**743.16**	**36.43**	**650**	**850**
Gender	Female	126,222	742.86	35.48	650	850
	Male	132,585	743.45	37.31	650	850
Ethnicity	American Indian/Alaska Native	4,212	722.11	31.06	650	834
	Asian	18,134	773.87	33.90	650	850
	Black/African American	38,801	725.28	34.17	650	850
	Hispanic/Latino	79,715	733.21	33.33	650	850
	Native Hawaiian or Pacific Islander	357	753.43	35.90	650	850
	Two or more races	8,326	746.03	37.65	650	850
	White	109,242	752.24	33.69	650	850
Economic Status*	Not Economically Disadvantaged	130,196	756.72	34.10	650	850
	Economically Disadvantaged	128,322	729.47	33.44	650	850
English Learner Status	Non-English Learner	218,081	746.48	36.31	650	850
	English Learner	40,509	725.46	31.59	650	850
Disabilities	Students without Disabilities	214,859	747.35	34.86	650	850
	Students with Disabilities	43,202	722.52	37.02	650	850
Language Form	Spanish	4,812	714.60	31.61	650	825

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

Table 12.8 Subgroup Performance for Mathematics Scale Scores: Algebra I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**134,107**	**740.37**	**36.46**	**650**	**850**
Gender	Female	65,087	741.24	35.26	650	850
	Male	69,020	739.56	37.54	650	850
Ethnicity	American Indian/Alaska Native	3,093	716.67	26.37	650	850
	Asian	11,399	776.91	36.45	650	850
	Black/African American	16,464	724.66	30.36	650	850
	Hispanic/Latino	47,009	725.71	30.44	650	850
	Native Hawaiian or Pacific Islander	285	748.75	35.02	661	832
	Two or more races	2,117	750.40	38.71	650	850
	White	53,728	751.19	33.72	650	850
Economic Status*	Not Economically Disadvantaged	79,819	751.08	36.54	650	850
	Economically Disadvantaged	54,217	724.65	30.08	650	850
English Learner Status	Non-English Learner	124,436	742.77	36.05	650	850
	English Learner	9,604	709.61	26.33	650	850
Disabilities	Students without Disabilities	109,545	744.89	36.06	650	850
	Students with Disabilities	24,497	720.25	31.04	650	850
Language Form	Spanish	2,426	701.47	24.27	650	795

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

Table 12.9 Subgroup Performance for Mathematics Scale Scores: Integrated Mathematics I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**673**	**718.84**	**33.39**	**650**	**845**
Gender	Female	321	721.19	33.54	650	845
	Male	352	716.70	33.15	650	822
Ethnicity	American Indian/Alaska Native	28	718.50	31.86	659	799
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	27	714.56	24.14	659	750
	Hispanic/Latino	415	711.46	28.18	650	794
	Native Hawaiian or Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or More Races	n/r	n/r	n/r	n/r	n/r
	White	192	732.93	38.88	650	845
Economic Status*	Not Economically Disadvantaged	218	737.70	35.35	650	845
	Economically Disadvantaged	448	709.92	28.37	650	807
English Learner Status	Non-English Learner	554	721.88	34.31	650	845
	English Learner	112	704.83	24.44	650	778
Disabilities	Students without Disabilities	525	723.98	33.40	650	845
	Students with Disabilities	140	700.86	26.27	650	828
Language Form	Spanish	n/r	n/r	n/r	n/r	n/r

*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.

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

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

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

The formula for the standard error of measurement is:

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

Where σ_X is the standard deviation of the test raw score and $\rho_{xx'}$ is the reliability estimated by substitution of appropriate statistics for the parameters 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 .5 (Cortina, 1993; Schmitt, 1996). Estimates lower than .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,

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

involves the error variance, $\sigma^2(E)$ and the total score variance, $\sigma^2(X)$. Using Kolen et al.'s (1996) method, conditional raw score distributions are estimated using Lord and Wingersky's (1984) recursion formula. The conditional raw score distributions are transformed into conditional scale score distributions. Denote X as the raw sum score ranging from 0 to X, and S as a resulting scale score after transformation. The conditional distribution of scale scores is written as $P(X = x \mid \theta)$. 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

$$\sigma^2(Error_{scale}) = \int_{\theta} \sigma^2(s(X) \mid \theta) g(\theta) d\theta \tag{13-5}$$

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,

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

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

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 section includes only spring 2019 results. The fall 2018 results are located in the Addendum.[16] 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 11 ELA/L range from a low of .87 to a high of .91. The average reliability estimates are at least .90 except for grades 3, 4, 5, and 11, which are .88, .88, .88, and .87, respectively. The average raw score SEM is consistently between 6 percent and 7 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 N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
3	5	54	3.60	0.88	7,522	0.81	108,352	0.89
4	5	70	4.48	0.88	664	0.81	112,153	0.89
5	5	71	4.46	0.88	500	0.77	145,676	0.89
6	5	72	4.49	0.91	3,931	0.87	129,435	0.92
7	5	72	4.88	0.91	353	0.79	126,778	0.91
8	5	72	4.60	0.91	377	0.87	126,674	0.92
9	4	71	4.67	0.91	555	0.82	34,577	0.92
10	4	72	4.87	0.90	571	0.81	64,769	0.92
11	4	72	4.83	0.87	574	0.75	16,373	0.89

Mathematics

The average reliability estimates for mathematics assessments range from .90 to .93 except for Integrated Mathematics I, II, and III which range from .81 to .84. The raw score SEM consistently ranges from 4.5 percent to 6.5 percent of the maximum score.

[16] Addendum 13 provides a summary of reliability information for the fall 2018 administration.

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 N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
3	5	52	3.18	0.92	3,805	0.90	101,570	0.92
4	5	52	3.27	0.93	3,512	0.85	120,199	0.93
5	5	52	3.25	0.92	3,077	0.81	119,050	0.92
6	5	52	3.09	0.92	2,436	0.76	123,192	0.93
7	5	52	3.12	0.92	2,022	0.77	121,448	0.93
8	5	52	2.89	0.90	2,018	0.76	106,005	0.91
A1	5	55	2.89	0.91	2,213	0.54	59,757	0.92
GO	4	55	2.87	0.90	1,359	0.66	45,622	0.91
A2	4	55	2.94	0.90	544	0.58	26,115	0.91
M1	1	55	2.53	0.84	604	0.84	604	0.84
M2	1	55	2.54	0.82	522	0.82	522	0.82
M3	1	55	2.70	0.81	197	0.81	197	0.81

A1=Algebra I, GO=Geometry, A2=Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II, M3=Integrated Mathematics III.

13.3.2 Scale Score Reliability Results

Tables 13.3–13.5 summarize scale score reliability estimates for the total testing group for ELA/L and mathematics for spring 2019. The tables provide average reliabilities by grade/course, which are estimated by averaging the reliability estimates computed for all forms of the test within the grade/course 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. Since estimates of scale score reliability are sample independent, form-level results are included even for grades with low sample sizes; therefore, the number of forms listed in Tables 13.3 and 13.4 are larger than the number of forms listed in Tables 13.1 and 13.2.

English Language Arts/Literacy

Reliability estimates for ELA/L were calculated for both the post-equated and pre-equated scale scores, and are presented in Tables 13.3. and 13.4, respectively. The average post-equated scale score reliability estimates for grades 3 through11 ELA/L range from .86 to .89, while the average pre-equated scale score reliability estimates for grades 3 through 11 ELA/L range from .86 to .90. Pre- and post-equated scale score reliability estimates are at least .84 for all forms. The average SEM ranges from 10.75 to 15.04 for post-equated scale scores and between 10.74 and 14.93 for pre-equated scale scores.

Mathematics

The scale score reliability estimates for the mathematics assessments are presented in Table 13.5. Average scale score reliability estimates for the grades 3 through 8 mathematics assessments range from .88 to .91, with the exception of grade 8 at .85. For the high school assessments, these quantities range from .82 to .85. For grades 3–8, the average scale score SEM ranges from 9.37 to 10.12, with the exception of grade 8 at 13.97. For high school tests, the average scale score SEM ranges from 11.65 to 16.55.

Table 13.3 Summary of ELA/L Test Post-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	6	14.70	0.86	0.84	0.89
4	6	12.44	0.87	0.86	0.88
5	5	12.13	0.86	0.85	0.88
6	5	10.75	0.88	0.86	0.90
7	5	12.36	0.89	0.88	0.89
8	5	11.92	0.89	0.88	0.91
9	6	12.54	0.88	0.86	0.90
10	6	15.04	0.89	0.88	0.90
11	6	14.35	0.86	0.85	0.88

Table 13.4 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	6	14.88	0.86	0.84	0.88
	6	12.55	0.87	0.86	0.88
5	5	12.12	0.86	0.85	0.89
6	5	10.74	0.88	0.86	0.90
7	5	12.25	0.89	0.88	0.90
8	5	11.79	0.90	0.88	0.91
9	6	12.46	0.89	0.86	0.91
10	6	14.93	0.89	0.88	0.90
11	6	14.53	0.86	0.84	0.88

Table 13.5 Summary of Mathematics Test Scale Score Reliability Estimates for Total Group

Grade/Course Level	Number of Forms	Avg. Scale Score SEM	Avg. Scale Score Reliability	Min. Scale Score Reliability	Max. Scale Score Reliability
3	5	10.12	0.91	0.91	0.92
4	5	9.37	0.91	0.91	0.92
5	5	9.87	0.90	0.89	0.91
6	5	10.10	0.89	0.89	0.90
7	6	9.72	0.88	0.87	0.88
8	6	13.97	0.85	0.83	0.87
A1	6	14.09	0.84	0.83	0.85
GO	6	11.65	0.84	0.82	0.86
A2	6	15.63	0.85	0.83	0.86
M1	2	15.03	0.82	0.82	0.82
M2	2	13.91	0.82	0.82	0.82
M3	2	16.55	0.83	0.83	0.83

A1=Algebra I, GO=Geometry, A2=Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II, M3=Integrated Mathematics III.

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 13.6 and 13.7 summarize test reliability for groups of interest for ELA/L grade 3 and mathematics grade 3, respectively. Corresponding information is provided in Appendix 13.1 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 .01. The SEMs for females are slightly higher than for males for all ELA 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 .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 .01 to .03 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 .86 and .91. However, the average reliabilities for American Indian/Alaskan native students range from .81 to .86. 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 generally 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 .01 to .02 less than the average reliability for the total group of students. The majority of the average reliabilities range

from .85 to .90. The average reliabilities for English learner students are lower, ranging from .77 to .83. 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 .01 to .02 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. 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

Two of the four accommodation form types (closed caption and text-to-speech) had sufficient sample sizes to allow for estimation of reliability and SEM for grades 3 through 8. Grades 9 through 11 had only text-to-speech with a sufficient sample size. Within grades, the reliabilities and SEMs of the closed caption forms are similar to the average reliabilities for the total group. For the text-to-speech forms, the reliabilities and SEMs are somewhat lower than for the total group.

Mathematics

The text-to-speech forms had sufficient sample sizes for reliability and SEM estimation across grades/subjects, except for the Integrated Mathematics I, II, and III courses where the sample was not sufficient. 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

With the exception of Integrated Mathematics I, II, and III, there were sufficient numbers of students taking the Spanish-language form for reliability and SEM estimation. The average reliability ranged from .81 to .89 for grades 3 through 5, and .76 to .77 for grades 6 through 8. The average reliability ranged from .54 to .66 for the high school courses. The SEMs are generally lower for the students administered the Spanish-language forms.

Table 13.6 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 3

	Max. Raw Score	Avg. SEM	Average Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	54	3.60	0.88	7,522	0.81	108,352	0.89
Gender							
Male	54	3.52	0.88	4,828	0.82	54,968	0.89
Female	54	3.69	0.88	2,694	0.80	53,384	0.89
Ethnicity							
White	54	3.66	0.86	5,831	0.82	247	0.87
Black/African American	54	3.50	0.86	1,376	0.74	13,302	0.89
Asian/Pacific Islander	54	3.82	0.87	1,176	0.85	8,180	0.87
American Indian/Alaska Native	54	3.39	0.84	962	0.78	1,610	0.86
Hispanic/Latino	54	3.53	0.86	2,599	0.77	31,369	0.88
Multiple	54	3.59	0.88	211	0.78	3,822	0.89
Special Instruction Needs							
Economically Disadvantaged	54	3.49	0.86	4,294	0.76	49,138	0.88
Not Economically Disadvantaged	54	3.71	0.86	3,224	0.83	251	0.88
English Learner	54	3.39	0.83	916	0.72	14,687	0.85
Non-English Learner	54	3.64	0.87	6,606	0.81	93,607	0.89
Students with Disabilities	54	3.15	0.87	7,522	0.81	15,215	0.90
Students without Disabilities	54	3.69	0.87	27,177	0.83	92,866	0.88
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	54	3.41	0.89	119	0.89	119	0.89
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	54	2.81	0.80	7,403	0.80	7,403	0.80

n/r = not reported due to n<100.

Table 13.7 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 3

	Max. Raw Score	Avg. SEM	Average Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.18	0.92	3,805	0.90	101,570	0.92
Gender							
Male	52	3.17	0.92	1,959	0.90	52,216	0.92
Female	52	3.18	0.92	1,846	0.89	49,354	0.92
Ethnicity							
White	52	3.25	0.91	5,749	0.91	47,832	0.91
Black/African American	52	3.02	0.91	10,298	0.90	12,090	0.92
Asian/Pacific Islander	52	3.16	0.90	8,044	0.90	1,160	0.92
American Indian/Alaska Native	52	2.94	0.90	937	0.89	1,418	0.90
Hispanic/Latino	52	3.10	0.91	382	0.88	28,540	0.91
Multiple	52	3.19	0.92	505	0.92	3,542	0.93
Special Instruction Needs							
Economically Disadvantaged	52	3.06	0.91	810	0.89	44,854	0.91
Not Economically Disadvantaged	52	3.24	0.91	643	0.89	579	0.91
English Learner	52	3.02	0.89	336	0.88	13,114	0.90
Non-English Learner	52	3.20	0.92	1,053	0.91	88,413	0.92
Students with Disabilities	52	3.01	0.92	344	0.85	16,872	0.93
Students without Disabilities	52	3.20	0.91	3,459	0.90	84,465	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.05	0.92	1,313	0.90	19,441	0.93
Students Taking Translated Forms							
Spanish Language Form	52	2.82	0.89	738	0.88	3,805	0.90

n/r = not reported due to n<100.

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 13.8 for a summary of the ELA/L claims and subclaims.

Table 13.8 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 13.9 presents the average reliability estimates for all forms of the test at the specified grade and testing mode for the ELA/L tests. In order to assist in understanding the reliability estimates, range of maximum number of points for each major claim and subclaim is also provided.

The average reliabilities for the Reading claim for grades 3 through 11 range from .81 to .86 with a median of .85. They are based on maximum scores of 40–44 points per form, except for grade 3 (28–31 points). The Writing claim average reliabilities are based on a lower number of points than those for the Reading claim, and are slightly lower, ranging from .78 to .85 with a median of .82. The reliabilities for the Writing claim for grade 3 is based on a maximum raw score of 24 points, and the average reliabilities for grades 4 and 5 are based on between 27 and 30 points per form. The average reliabilities for the grades 5 through 11 Writing claims are based on a maximum score of 30 points.

The average reliabilities of the Reading Literature subclaim scores have a median of .68, and vary from .56 to .75. The maximum number of points per form ranges from 11 to 20. The average reliabilities of the Reading Information subclaim scores have a median of .67, and vary from .54 to .77, with 7–22 points per form. The average reliabilities of the Reading Vocabulary subclaim scores have a median of .55, and vary from .50 to .67. The maximum number of points per form for this subclaim ranges from 8 to 14.

The Writing Written Expression subclaim is based on 18 points for grade 3 and 21–24 points for grades 4 and 5. Grades 6 through 11 are based on 24 points for all forms. The median of the average reliabilities for the tests is .82

and the average reliabilities range from .71 to .86. The Writing Knowledge of Language and Conventions subclaims are all based on six points. The median average reliability is .85 and the reliabilities range from .80 to .87.

Table 13.9 Average ELA/L Reliability Estimates for Total Test and Subscores

	Reading: Total		Reading: Literature		Reading: Information		Reading: Vocabulary		Writing: Total		Writing Expression		Writing: Knowledge Language and Conventions	
Grade Level	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability
3	28-31	0.84	11-13	0.68	7-11	0.63	8-10	0.61	24-24	0.78	18-18	0.72	6-6	0.80
4	40-44	0.82	16-18	0.65	12-16	0.61	8-14	0.55	27-30	0.80	21-24	0.77	6-6	0.83
5	40-44	0.85	16-18	0.73	14-14	0.54	10-14	0.67	27-30	0.79	21-24	0.71	6-6	0.84
6	40-44	0.86	16-20	0.75	14-16	0.67	8-14	0.58	30-30	0.82	24-24	0.82	6-6	0.85
7	40-44	0.85	16-18	0.70	14-14	0.65	10-14	0.62	30-30	0.83	24-24	0.85	6-6	0.86
8	40-44	0.85	16-16	0.68	14-16	0.70	8-14	0.53	30-30	0.85	24-24	0.86	6-6	0.87
9	40-44	0.85	12-16	0.65	16-22	0.77	8-10	0.52	30-30	0.84	24-24	0.85	6-6	0.86
10	40-44	0.84	12-18	0.64	14-22	0.70	10-12	0.51	30-30	0.84	24-24	0.86	6-6	0.87
11	40-44	0.81	12-16	0.56	14-22	0.68	10-12	0.50	30-30	0.82	24-24	0.79	6-6	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/course level with connections to the Standards for Mathematical Practice.
- Subclaim B: Students solve problems involving the additional and supporting content for their grade/course level with connections to the Standards for Mathematical Practice.
- Subclaim C: Students express grade/course-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/course 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 13.10 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, Algebra I, Geometry, and Algebra II, the median of the average reliabilities for the Major Content subclaim is .82, with a range from .75 to .86. The maximum number of points per form range from 16 to 21.

The median of the average reliabilities for the Additional and Supporting Content subclaim for grades 3 through 8, Algebra I, Geometry, and Algebra II is .68, with a range from .58 to .71. The maximum number of points per form for this subclaim ranges from 9 to 12.

The average reliabilities for Mathematics Reasoning range from .51 to .75 for grades 3 through 8, Algebra I, Geometry, and Algebra II, with a median of .65. 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, Algebra I, Geometry, and Algebra II have a median of .69 and range from .62 to .75. The number of points is 12 for grades 3 through 8 and 15 for all high school courses.

The Integrated Mathematics assessments have low to moderate average reliabilities for Major Content (ranging from .58 to .66) and Modeling Practice (ranging from .60 to .61). In Table 13.10, four subclaim reliability estimates are less than .50, which prompts additional investigations. Scale scores tended to be low for the populations taking these assessments, with small standard deviations (see Appendix Tables A.12.45–A.12.47). Out of 850 possible scale score points, only one student on each of these tests received a score of 830 or above, and more than 98 percent of students received scores below 800 (see Appendix Tables A.12.66–A.12.68). The fact that the sample population for these tests did not include students at the full range of performance likely contributed to the lower reliability estimates.

Table 13.10 Average Mathematics Reliability Estimates for Total Test and Subscores

	Major Content		Additional & Supporting Content		Mathematics Reasoning		Modeling Practice	
Grade Level	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability
3	20-20	0.86	10-10	0.69	10-10	0.51	12-12	0.73
4	21-21	0.86	9-9	0.71	10-10	0.75	12-12	0.65
5	20-20	0.85	10-10	0.68	10-10	0.61	12-12	0.72
6	20-20	0.82	10-10	0.66	10-10	0.70	12-12	0.69
7	20-20	0.85	10-10	0.67	10-10	0.56	12-12	0.75
8	20-20	0.81	10-10	0.58	10-10	0.65	12-12	0.64
A1	17-17	0.75	9-9	0.68	10-10	0.74	15-15	0.73
GO	18-18	0.80	12-12	0.64	10-10	0.67	15-15	0.64
A2	16-18	0.77	12-12	0.68	10-10	0.60	15-15	0.62
M1	19-19	0.66	11-11	*	10-10	0.58	15-15	0.60
M2	18-18	0.59	12-12	*	10-10	0.56	15-15	0.61
M3	19-19	0.58	11-11	*	10-10	*	15-15	0.60

* Cronbach alpha below .50, further investigation summarized at the end of Section 13.6.

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

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 13.11 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 11 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 proportion accurately classified ranges from .66 to .74 with a median of .70; the proportion who would be consistently classified on two different test forms ranges from .56 to .64 with a median of .60. For classifying each student as being at Level 4 or higher vs. being at Level 3 or lower, the proportion accurately classified ranges from .89 to .91 with a median of .91; the proportion who would be consistently classified this way on two different test forms ranges from .85 to .87 with a median of .87.

Table 13.11 Reliability of Classification: Summary for ELA/L

Level	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
	Exact Level	Level 4 or higher vs. 3 or lower	Exact Level	Level 4 or higher vs. 3 or lower
3	0.69	0.90	0.60	0.86
4	0.68	0.89	0.57	0.85
5	0.70	0.89	0.60	0.85
6	0.74	0.91	0.64	0.87
7	0.70	0.91	0.60	0.87
8	0.71	0.91	0.61	0.87
9	0.72	0.91	0.62	0.87
10	0.69	0.91	0.59	0.87
11	0.66	0.89	0.56	0.85

Table 13.12 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 13.12. For "Level 4 and higher vs. 3 and lower" found in Table 13.11, the sum of the shaded values in Table 13.12 is approximately equal to the level of decision accuracy or consistency presented in Table 13.11. Note that the sums based on values in Table 13.12 may not match exactly to the values in Table 13.11 due to truncation and rounding.

Detailed information for all ELA/L spring results are provided in Appendix 13 Tables A.13.1 through A.13.9. Fall block results for ELA/L grades 9 through 11 are provided in the Addendum. The structure of these tables is the same as that of Table 13.12 and the values in the tables should be interpreted in the same manner. Table 13.12 includes the same information as Table A.13.1. The sum of the five bold values on the diagonal is approximately equal to the level of decision accuracy or consistency presented in Table 13.12. For "Level 4 and higher vs. 3 and lower" presented in Table 13.12, the sum of the shaded values in Table 13.12 is approximately equal to the level of decision accuracy or consistency presented in Table 13.12. Any differences between the sums based on values in Table 13.12 and the values in Table 13.12 are due to truncation and rounding.

Table 13.12 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	**0.15**	0.03	0.00	0.00	0.00	0.18
	700-724	0.04	**0.10**	0.05	0.00	0.00	0.19
	725-749	0.00	0.04	**0.12**	0.05	0.00	0.22
	750-809	0.00	0.00	0.05	**0.30**	0.03	0.38
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650-699	**0.15**	0.04	0.01	0.00	0.00	0.20
	700-724	0.04	**0.07**	0.05	0.01	0.00	0.18
	725-749	0.01	0.05	**0.09**	0.06	0.00	0.20
	750-809	0.00	0.01	0.07	**0.27**	0.03	0.37
	810-850	0.00	0.00	0.00	0.02	**0.02**	0.04

13.7.2 Mathematics

Table 13.13 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 .72 to .77 with a median of .75; the proportion who would be consistently classified on two different test forms ranges from .63 to .69 with a median of .65. For the six high school mathematics courses, the table shows that for classifying each student into one of the five performance levels, the proportion accurately classified ranges from .66 to .75 with a median of 73.; the proportion who would be consistently classified on two different test forms ranges from .55 to .67 with a median of .63.

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 is .91 for grades 3 and 8 and .92 for the grades in between; the proportion who would be consistently classified on two different test forms is .89 for grades 4 through 7 and .88 for grades 3 and 8. For the six high school mathematics courses, the proportion accurately classified as being at Level 4 or higher vs. being at Level 3 or lower ranges from .91 to .96 with a median of .92; the proportion who would be consistently classified on two different test forms ranges from .87 to .94 with a median of .88.

Appendix 13 Tables A.13.10 through A.13.21 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. Fall block results for Algebra I, Geometry, and Algebra II are provided in the Addendum.

Table 13.13 Reliability of Classification: Summary for Mathematics

Level	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
	Exact Level	Level 4 or higher vs. 3 or lower	Exact Level	Level 4 or higher vs. 3 or lower
3	0.74	0.91	0.65	0.88
4	0.77	0.92	0.69	0.89
5	0.75	0.92	0.65	0.89
6	0.75	0.92	0.65	0.89
7	0.76	0.92	0.66	0.89
8	0.72	0.91	0.63	0.88
A1	0.74	0.91	0.64	0.88
GO	0.73	0.91	0.63	0.88
A2	0.72	0.91	0.63	0.87
M1	0.68	0.92	0.58	0.89
M2	0.66	0.96	0.55	0.94
M3	0.75	0.96	0.67	0.94

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

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 13.14 displays both the expectations and the actual agreement percentages for perfect agreement and perfect plus adjacent agreement.

Table 13.14 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%	100%	100%
Mathematics	0–2	80%	97%	100%	100%
Mathematics	0–3	70%	95%	100%	99%
Mathematics	0–4	65%	94%	99%	99%
Mathematics	0–5	65%	93%	99%	98%
Mathematics	0–6	65%	95%	99%	98%
ELA/L	Multi-trait	65%	80%	100%	99%

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 13.14 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 handscoring.

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 2018–2019 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 the future development of additional items and tasks.

All item developers and item writers are provided 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, Reading claim score, and Writing claim score as well as ELA/L subclaim and 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 2019 administration.[17] Tables 14.1 through 14.9 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 subclaim reliabilities (from Section 13) are reported along the diagonal. 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 and reliability estimates for mathematics are provided in Tables 14.10 through 14.21. The shaded values along the diagonal are the reliabilities as reported in Section 13. 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 12.1 (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.

[17] Addendum 14 provides a summary of results for the fall 2018 administration.

Table 14.1 Average Intercorrelations and Reliability between Grade 3 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.84	256,761	256,761	256,761	256,761	256,761	256,761
RL	0.89	0.68	256,761	256,761	256,761	256,761	256,761
RI	0.85	0.64	0.63	256,761	256,761	256,761	256,761
RV	0.85	0.63	0.60	0.61	256,761	256,761	256,761
WR	0.70	0.62	0.68	0.51	0.78	256,761	256,761
WE	0.68	0.61	0.67	0.50	0.99	0.72	256,761
WKL	0.64	0.56	0.61	0.48	0.89	0.81	0.80

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	0.82	264,079	264,079	264,079	264,079	264,079	264,079
RL	0.89	0.65	264,079	264,079	264,079	264,079	264,079
RI	0.86	0.62	0.61	264,079	264,079	264,079	264,079
RV	0.81	0.59	0.56	0.55	264,079	264,079	264,079
WR	0.71	0.63	0.67	0.49	0.80	264,079	264,079
WE	0.70	0.62	0.67	0.49	0.99	0.77	264,079
WKL	0.66	0.58	0.62	0.46	0.92	0.86	0.83

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	0.85	271,471	271,471	271,471	271,471	271,471	271,471
RL	0.92	0.73	271,471	271,471	271,471	271,471	271,471
RI	0.80	0.61	0.54	271,471	271,471	271,471	271,471
RV	0.85	0.67	0.55	0.67	271,471	271,471	271,471
WR	0.71	0.67	0.62	0.54	0.79	271,471	271,471
WE	0.70	0.66	0.62	0.52	0.99	0.71	271,471
WKL	0.68	0.65	0.59	0.53	0.94	0.88	0.84

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	0.86	274,102	274,102	274,102	274,102	274,102	274,102
RL	0.92	0.75	274,102	274,102	274,102	274,102	274,102
RI	0.87	0.68	0.67	274,102	274,102	274,102	274,102
RV	0.81	0.66	0.60	0.58	274,102	274,102	274,102
WR	0.73	0.65	0.72	0.52	0.82	274,102	274,102
WE	0.72	0.64	0.71	0.51	1.00	0.82	274,102
WKL	0.72	0.64	0.70	0.52	0.96	0.94	0.85

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	0.85	267,676	267,676	267,676	267,676	267,676	267,676
RL	0.91	0.70	267,676	267,676	267,676	267,676	267,676
RI	0.88	0.68	0.65	267,676	267,676	267,676	267,676
RV	0.85	0.65	0.64	0.62	267,676	267,676	267,676
WR	0.73	0.65	0.74	0.52	0.83	267,676	267,676
WE	0.72	0.64	0.73	0.52	1.00	0.85	267,676
WKL	0.73	0.65	0.73	0.53	0.97	0.95	0.86

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	0.85	266,200	266,200	266,200	266,200	266,200	266,200
RL	0.90	0.68	266,200	266,200	266,200	266,200	266,200
RI	0.89	0.70	0.70	266,200	266,200	266,200	266,200
RV	0.81	0.62	0.60	0.53	266,200	266,200	266,200
WR	0.75	0.68	0.75	0.52	0.85	266,200	266,200
WE	0.74	0.67	0.74	0.52	1.00	0.86	266,200
WKL	0.74	0.67	0.73	0.52	0.98	0.96	0.87

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 9 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.85	120,414	120,414	120,414	120,414	120,414	120,414
RL	0.87	0.65	120,414	120,414	120,414	120,414	120,414
RI	0.91	0.68	0.77	120,414	120,414	120,414	120,414
RV	0.78	0.57	0.57	0.52	120,414	120,414	120,414
WR	0.78	0.67	0.78	0.50	0.84	120,414	120,414
WE	0.77	0.67	0.77	0.50	1.00	0.85	120,414
WKL	0.77	0.67	0.76	0.50	0.97	0.96	0.86

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

Table 14.8 Average Intercorrelations and Reliability between Grade 10 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.84	117,451	117,451	117,451	117,451	117,451	117,451
RL	0.87	0.64	117,451	117,451	117,451	117,451	117,451
RI	0.90	0.67	0.70	117,451	117,451	117,451	117,451
RV	0.79	0.58	0.58	0.51	117,451	117,451	117,451
WR	0.77	0.69	0.75	0.52	0.84	117,451	117,451
WE	0.76	0.69	0.75	0.52	1.00	0.86	117,451
WKL	0.76	0.68	0.74	0.52	0.97	0.96	0.87

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

Table 14.9 Average Intercorrelations and Reliability between Grade 11 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.81	34,482	34,482	34,482	34,482	34,482	34,482
RL	0.82	0.56	34,482	34,482	34,482	34,482	34,482
RI	0.88	0.58	0.68	34,482	34,482	34,482	34,482
RV	0.79	0.50	0.55	0.50	34,482	34,482	34,482
WR	0.70	0.60	0.69	0.44	0.82	34,482	34,482
WE	0.70	0.60	0.69	0.44	1.00	0.79	34,482
WKL	0.69	0.59	0.68	0.44	0.98	0.96	0.8

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

Table 14.10 Average Intercorrelations and Reliability between Grade 3 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.86	258,696	258,696	258,696
ASC	0.78	0.69	258,696	258,696
MR	0.65	0.59	0.51	258,696
MP	0.77	0.70	0.62	0.73

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 4 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.86	265,528	265,528	265,528
ASC	0.76	0.71	265,528	265,528
MR	0.77	0.69	0.75	265,528
MP	0.75	0.70	0.72	0.65

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 5 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.85	272,384	272,384	272,384
ASC	0.74	0.68	272,384	272,384
MR	0.68	0.62	0.61	272,384
MP	0.77	0.69	0.65	0.72

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

Table 14.13 Average Intercorrelations and Reliability between Grade 6 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.82	274,562	274,562	274,562
ASC	0.72	0.66	274,562	274,562
MR	0.77	0.66	0.70	274,562
MP	0.75	0.66	0.72	0.69

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

Table 14.14 Average Intercorrelations and Reliability between Grade 7 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.85	263,274	263,274	263,274
ASC	0.75	0.67	263,274	263,274
MR	0.71	0.64	0.56	263,274
MP	0.77	0.72	0.69	0.75

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

Table 14.15 Average Intercorrelations and Reliability between Grade 8 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.81	225,680	225,680	225,680
ASC	0.71	0.58	225,680	225,680
MR	0.70	0.62	0.65	225,680
MP	0.66	0.60	0.65	0.64

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

Table 14.16 Average Intercorrelations and Reliability between Algebra I Subclaims

	MC	ASC	MR	MP
MC	0.75	131,502	131,502	131,502
ASC	0.72	0.68	131,502	131,502
MR	0.73	0.72	0.74	131,502
MP	0.71	0.67	0.70	0.73

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

Table 14.17 Average Intercorrelations and Reliability between Geometry Subclaims

	MC	ASC	MR	MP
MC	0.80	103,760	103,760	103,760
ASC	0.71	0.64	103,760	103,760
MR	0.71	0.63	0.67	103,760
MP	0.73	0.66	0.70	0.64

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

Table 14.18 Average Intercorrelations and Reliability between Algebra II Subclaims

	MC	ASC	MR	MP
MC	0.77	66,387	66,387	66,387
ASC	0.73	0.68	66,387	66,387
MR	0.69	0.65	0.60	66,387
MP	0.64	0.61	0.60	0.62

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

Table 14.19 Average Intercorrelations and Reliability between Integrated Mathematics I Subclaims

	MC	ASC	MR	MP
MC	0.66	672	672	672
ASC	0.61	0.49	672	672
MR	0.62	0.56	0.58	672
MP	0.55	0.51	0.68	0.60

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

Table 14.20 Average Intercorrelations and Reliability between Integrated Mathematics II Subclaims

	MC	ASC	MR	MP
MC	0.59	531	531	531
ASC	0.52	0.41	531	531
MR	0.63	0.54	0.56	531
MP	0.59	0.50	0.69	0.61

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

Table 14.21 Average Intercorrelations and Reliability between Integrated Mathematics III Subclaims

	MC	ASC	MR	MP
MC	0.58	201	201	201
ASC	0.48	0.36	201	201
MR	0.60	0.43	0.45	201
MP	0.60	0.50	0.61	0.60

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 reliability estimates, computed using coefficient alpha (Cronbach, 1951), are presented in Tables 13.1 and 13.2 and are along the diagonals of Tables 14.1 through 14.18.[18] The average reliabilities for ELA/L and mathematics summative assessments range from .87 up to .93. Tables 13.5 through 13.14 summarize test reliability for groups of interest for ELA/L grades 3 through 11, and Tables 13.15 through 13.26 summarize test reliability for groups of interest for mathematics grades/courses. 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:

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

where l is the total number of items, $\sigma_{ii'}$ is the covariance of items i and i' $(i \neq i')$, and σ_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 σ_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.

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

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

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

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

where 0 is the observed score and $\hat{E}$ is the expected value of 0 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. 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, and
- 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 14.22.

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 .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 14.23 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 .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 14.23 centered around the expectation of -1/n-1.

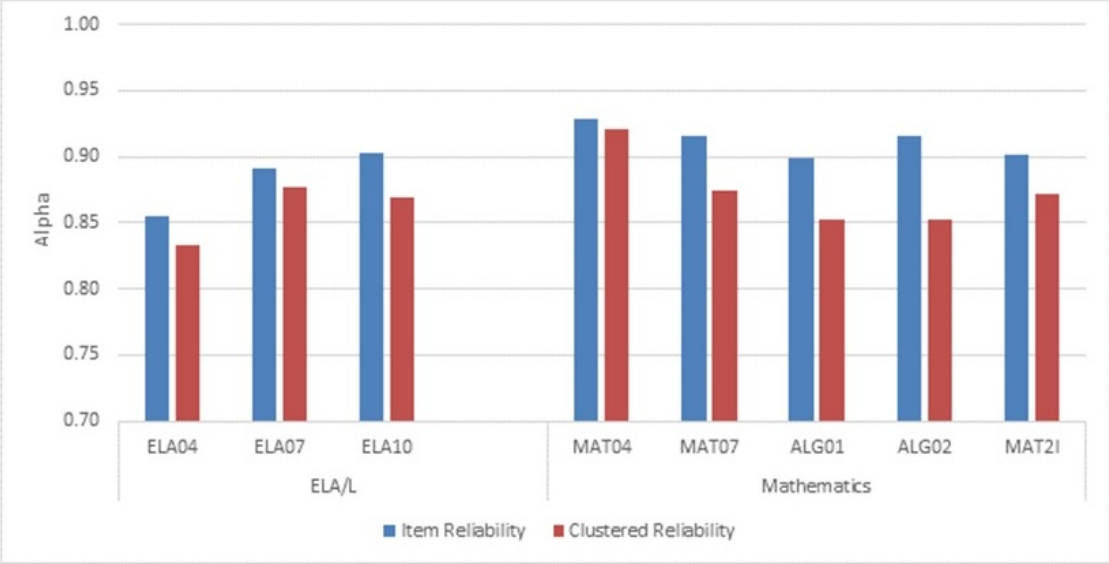

Figure 14.1 Comparison of Internal Consistency by Item and Cluster (Testlet)

Table 14.22 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
	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 14.23 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

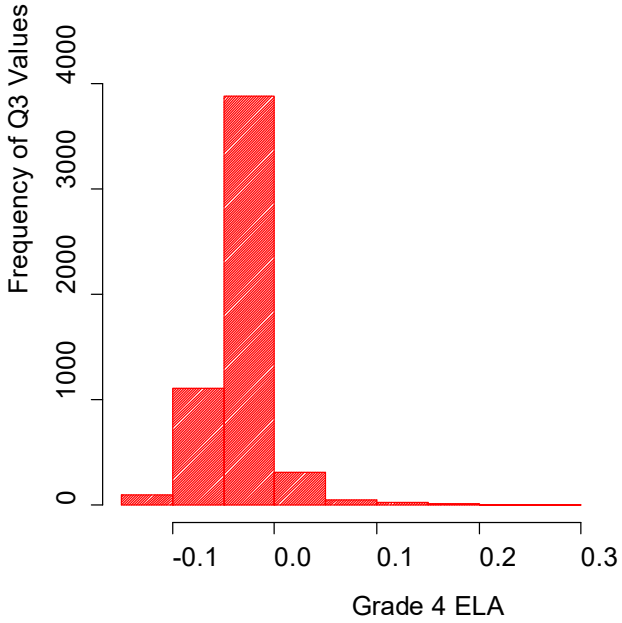

Figure 14.2 Distribution of Q3 Values for Grade 4 ELA/L (Spring 2015)

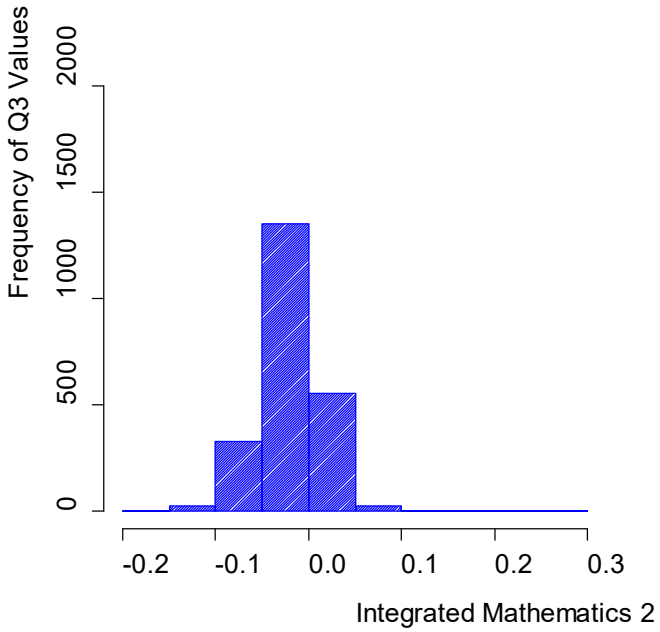

Figure 14.3 Distribution of Q3 Values for Integrated Mathematics II (Spring 2015)

14.4 Evidence Based on Relationships to Other Variables

Empirical results concerning the relationships between scores on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the AERA, APA, and NCME standards (2014), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, as well as demographic characteristics of students that are expected to be related and unrelated to test performance.

The relationship of the scores across the ELA/L and mathematics assessments was evaluated using correlational analyses. Tables 14.24 through 14.29 present the Pearson correlations observed between the ELA/L scale scores and the mathematics scale scores for each grade. For grades 3 through 8, students must have a valid test score for both ELA/L and mathematics at the same grade level to be included in the tables. These tables provide the correlation in the lower triangle and the sample size is provided in the upper triangle. In computing the correlations between a particular pair of ELA/L and mathematics tests, students must have taken both tests in spring 2019. ELA/L, Reading (RD), and Writing (WR) are moderately to highly correlated with mathematics; the correlations range from .60 up to .77 for grades 3 through 8. These correlations suggest that the ELA/L and mathematics tests are assessing different content. The higher intercorrelations between the ELA/L, Reading (RD), and Writing (WR) scores suggest stronger internal relationships when compared to the correlations with the mathematics content area.

The ELA/L and mathematics correlations for the high school tests are presented in Tables 14.30 through 14.32. Because students in high school can take the mathematics courses in different years (e.g., one student may take Algebra I in grade 9 while another student may take Algebra I in grade 10), the high school mathematics scores were correlated with several of the ELA/L grades (e.g., Algebra I correlated with both grades 9 and 10). Only correlations for pairings with total sample sizes of at least 100 are shown in the tables. Shaded cells indicate pairings with sample sizes less than 100. Across both modes of grades 8 through 11, ELA/L, Reading (RD), and Writing (WR) scores have correlations with high school mathematics tests that range from .29 to .77. Correlations between high school mathematics scores and corresponding ELA/L scores demonstrate low to moderate correlations.

Table 14.24 Correlations between ELA/L and Mathematics for Grade 3

	ELA/L	RD	WR	MA
ELA/L		256,111	256,111	256,111
RD	0.95		256,111	256,111
WR	0.86	0.70		256,111
MA	0.76	0.74	0.65	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.25 Correlations between ELA/L and Mathematics for Grade 4

	ELA/L	RD	WR	MA
ELA/L		263,299	263,299	263,299
RD	0.95		263,299	263,299
WR	0.88	0.70		263,299
MA	0.77	0.75	0.67	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.26 Correlations between ELA/L and Mathematics for Grade 5

	ELA/L	RD	WR	MA
ELA/L		270,656	270,656	270,656
RD	0.95		270,656	270,656
WR	0.85	0.70		270,656
MA	0.76	0.74	0.64	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.27 Correlations between ELA/L and Mathematics for Grade 6

	ELA/L	RD	WR	MA
ELA/L		272,726	272,726	272,726
RD	0.95		272,726	272,726
WR	0.86	0.71		272,726
MA	0.77	0.77	0.63	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.28 Correlations between ELA/L and Mathematics for Grade 7

	ELA/L	RD	WR	MA
ELA/L		261,360	261,360	261,360
RD	0.94		261,360	261,360
WR	0.90	0.72		261,360
MA	0.75	0.76	0.63	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.29 Correlations between ELA/L and Mathematics for Grade 8

	ELA/L	RD	WR	MA
ELA/L		223,718	223,718	223,718
RD	0.94		223,718	223,718
WR	0.88	0.69		223,718
MA	0.71	0.71	0.60	

Note: ELA/L = English language arts/literacy, RD = Reading, WR = Writing, MA = Mathematics.

Table 14.30 Correlations between ELA/L and Mathematics for High School

	Mathematics					
ELA/L	A1	GO	A2	M1	M2	M3
8	0.67	0.49	0.49			
	(36,122)	(4,012)	(429)			
9	0.63	0.67	0.63	0.73		
	(77,063)	(26,913)	(7,073)	(374)		
10	0.50	0.60	0.66	0.60	0.57	
	(7,478)	(64,261)	(32,831)	(262)	(307)	
11	0.31	0.50	0.53		0.70	0.77
	(813)	(3,772)	(20,828)		(173)	(178)

Note: ELA/L = English language arts/literacy, A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Table 14.31 Correlations between ELA/L Reading and Mathematics for High School

	Mathematics					
RD	A1	GO	A2	M1	M2	M3
8	0.67	0.49	0.52			
	(36,122)	(4,012)	(429)			
9	0.62	0.65	0.61	0.74		
	(77,063)	(26,913)	(7,073)	(374)		
10	0.49	0.59	0.65	0.60	0.52	
	(7,478)	(64,261)	(32,831)	(262)	(307)	
11	0.30	0.50	0.53		0.72	0.75
	(813)	(3,772)	(20,828)		(173)	(178)

Note: RD = Reading, A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Table 14.32 Correlations between ELA/L Writing and Mathematics for High School

	Mathematics					
WR	A1	GO	A2	M1	M2	M3
8	0.56	0.39	0.33			
	(36,122)	(4,012)	(429)			
9	0.53	0.58	0.53	0.58		
	(77,063)	(26,913)	(7,073)	(374)		
10	0.42	0.52	0.56	0.44	0.48	
	(7,478)	(64,261)	(32,831)	(262)	(307)	
11	0.29	0.40	0.42		0.59	0.68
	(813)	(3,772)	(20,828)		(173)	(178)

Note: WR = Writing, A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

14.5 Evidence from the 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:

- content alignment studies,
- a benchmarking study,
- a longitudinal study of external validity,
- a mode comparability study,
- a device comparability study, and
- Quality Testing Standards study.

The following paragraphs briefly describe each of these studies.

14.5.1 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, 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.

14.5.2 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.

14.5.3 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 .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 62.5 percent to 86.5 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 .509 to .886). However, students who met the benchmark on the external test had relatively low probabilities of meeting the mathematics benchmark (.097 to .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 67.3 percent to 79.7 percent. Students meeting the ELA/L benchmark had probabilities between .667 and .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 (.326 to .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 .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.

14.5.4 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.

14.5.5 Quality Testing Standards

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. Research conducted using 2017 (Boyd et al., 2018) and 2018 (Minchen et al., 2018) student data evaluated the effects of removing items from the original assessments to determine if scores arising from the two versions would be comparable. Research was conducted in several steps. First, subject matter experts identified item subsets from the original forms that maintained the integrity of the assessment and were approximately 65 to 80 percent of the original test length. Then, students were rescored on the item subsets, producing a set of hypothetical scores, as if the students had

only taken the subset of items. Finally, a series of analyses were conducted. While the research generally supported the comparability of the two versions, a limitation of the methodology was that the alternate blueprints were not actually administered as such. In this report, the shorter version of the blueprint is referred to as the current assessment and the original blueprint is referred to as the original assessment.

Through extensive research and guidance from the Technical Advisory Committee, the current blueprint was available in spring 2019 in addition to the original blueprint. In 2019, the option to administer either blueprint was made at the state or agency level. Since some states administered the current blueprint and some states administered the original blueprint, the following research evaluated the comparability between the two blueprints with respect to scale score comparability and performance level comparability.

The goal was to determine additional evidence to support scale score comparability and performance level comparability, according to the guidelines outlined in the Quality Testing Standards (QTS; The Center for Assessment, 2018). For the purpose of this work, scale score and performance level comparability have formal definitions. Scale score comparability is defined by The Center for Assessment (2018) as follows: If a student taking the current assessments with New Meridian content took the original assessment, would the student obtain a similar scale score? Performance level comparability is defined by The Center for Assessment (2018) as follows: If a student taking the current assessment with New Meridian content took the original assessment, would the student receive a similar designation in terms of college and career readiness or performance level 4 on the original blueprint?

For the spring 2019 assessments, the mathematics items on the current forms also appeared on the corresponding original forms; however, for ELA/L assessments, a small number of items were unique to the current forms. The scale scores were reported on the same scale regardless of the form and used the same performance level cut scores.

Three sets of analyses were conducted. Most of the analyses were conducted on a set of matched samples from the 2019 current and original forms, allowing for direct comparisons of assessment characteristics and outcomes to be made. Such samples were obtained through coarsened exact matching (CEM; Iacus et al., 2012), which used demographic information and prior achievement scores, where possible. Prior achievement scores were grouped into bands within each performance level, and students taking the current forms were matched with students who took the original forms who had identical information on all demographic and prior achievement variables. The prior assessments used in the matching process can be found in Tables 14.33 and 14.34. For grade 3 assessments, only demographic information is used in the matching process due to the lack of prior assessment data. Due to differences in high school assessment requirements across states and agencies, multiple prior assessments may have been used. For ELA/L grade 10, the prior assessment was ELA/L grade 8 for the matching process.

Table 14.33 Prior Grades Used in ELA/L Matching

Current Grade	Prior Grade	Prior Test Year
Grade 3	N/A	N/A
Grade 4	Grade 3	2018
Grade 5	Grade 4	2018
Grade 6	Grade 5	2018
Grade 7	Grade 6	2018
Grade 8	Grade 7	2018
Grade 10	Grade 8	2017

Table 14.34 Prior Grades/Courses Used in Mathematics Matching

Current Grade/ Course	Prior Grade /Course	Prior Test Year
Grade 3	N/A	N/A
Grade 4	Grade 3	2018
Grade 5	Grade 4	2018
Grade 6	Grade 5	2018
Grade 7	Grade 6	2018
Grade 8	Grade 7	2018
Algebra I	Grade 7 (44%), Grade 8 (56%)	2018
Geometry	Algebra I	2018
Algebra II	Algebra I (10%), Geometry (90%)	2018

Sample sizes before and after the matching process are listed in Table 14.35 for ELA/L and Table 14.36 for mathematics. ELA/L grade 9, Geometry, and Algebra II, matched samples were fairly small, ranging from 75 to 1,540. Due to the small sample for ELA/L grade 9, the comparability analyses were not conducted. Geometry and Algebra II were included in the comparability analyses; however, the results should be interpreted with caution given the small samples.

Table 14.35 ELA/L Matching Sample Size Results

ELA/L	Form	Unmatched		Matched	
		Current Forms N	Original Forms N	Current Forms N	Original Forms N
Grade 3	1	105,482	32,034	31,481	31,481
	2	105,309	31,861	31,272	31,272
Grade 4	1	105,826	28,153	27,695	27,695
	2	126,875	34,071	33,444	33,444
Grade 5	1	136,148	36,313	35,742	35,742
	2	101,869	27,272	26,721	26,721
Grade 6	1	119,838	31,031	30,667	30,667
	2	120,218	30,802	30,506	30,506
Grade 7	1	116,933	29,877	29,544	29,544
	2	117,757	29,835	29,593	29,593
Grade 8	1	118,198	29,638	29,312	29,312
	2	119,059	29,248	28,898	28,898
Grade 9	1	30,648	86	75	75
	2	71,029	116	102	102
Grade 10	1	55,046	27,951	22,970	22,970
	2	41,439	20,758	17,193	17,193

Table 14.36 Mathematics Matching Sample Size Results

		Unmatched		Matched	
	Form	Current Forms N	Original Forms N	Current Forms N	Original Forms N
Grade 3	1	88,858	26,531	25,970	25,970
	2	88,919	26,595	25,987	25,987
Grade 4	1	87,291	25,941	25,070	25,070
	2	87,488	26,192	25,207	25,207
Grade 5	1	91,136	27,333	26,377	26,377
	2	91,739	27,611	26,754	26,754
Grade 6	1	95,174	28,514	27,677	27,677
	2	94,800	28,342	27,665	27,665
Grade 7	1	93,777	24,547	23,855	23,855
	2	93,265	24,141	23,485	23,485
Grade 8	1	83,289	15,293	14,962	14,962
	2	76,135	13,973	13,695	13,695
Algebra I	1	43,232	21,530	16,926	16,926
	2	46,482	23,036	18,157	18,157
Geometry	1	40,673	3,252	1,540	1,540
	2	40,918	3,360	1,514	1,514
Algebra II	1	27,568	1,037	823	823
	2	27,527	1,066	753	753

Detailed matching results for select assessments can be found in the Appendix, Tables A.14.1 – A.14.3. ELA/L and mathematics for grade 6 and ELA/L grade 10 matching results are presented. Other grade levels had very similar results to grade 6, except for ELA/L grade 10.

The remaining analyses were conducted on assessment data from 2018 and 2019, rather than the matched samples. The second set of analyses was conducted at the grade level, using all available data from both 2018 and 2019, examining grade-level statistics over the course of two years, ensuring state participation was similar within each grade for both years. Finally, the last set of analyses used two-year student cohorts, examining students' scores over two years. Only students who completed assessments in both 2018 and 2019 were included; therefore, grade 3 student data from 2019 were not included.

Effect sizes were used throughout the research to determine the degree to which differences were practically significant. For differences between continuous distributions, such as scale score and claim score means, Cohen's (1988) D was used, and is calculated as

$$D = \frac{\overline{x_1} - \overline{x_2}}{S_p} \tag{14-4}$$

where $\overline{x_1}$ and $\overline{x_2}$ are the means of interest, and S_p is the pooled standard deviation of the scores in both distributions. For differences in proportions, Cohen's (1988) *h* was used, and is given by

$$h = 2\left(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_2}\right) \tag{14-5}$$

where p_1 and p_2 are the proportions of interest. And for differences in ordinal distributions, Cramer's (1946) V was used, which is given as

$$V = \sqrt{\frac{\chi^2}{n \times \min\left(r-1, c-1\right)}} \tag{14-6}$$

where χ^2 is the chi-squared value from the contingency table calculation, n is the total sample size, r is the number of rows in the contingency table, and c is the number of columns in the contingency table. Cohen (1988) defined effect sizes as .25, .5, and .8 as constituting small, medium, and large effects, respectively. A number of regression analyses are also performed, and the change in R^2 between the full and reduced models is examined; R^2 values of .01, .06, and .15 constitute the small, medium, and large effect sizes (Cohen, 1988).

Scale Score Comparability: Item-Level Analysis

Item-level evaluations (i.e., p-values, polyserial correlations, and DIF) were conducted separately for current and original forms on the matched sample for items that were common to both forms for each grade/course. First, p-values were compared. Scatterplots for the current form p-values and original form p-values for ELA/L grades 3 to 6 and mathematics grades 3 to 6 are presented in Figures 14.4 and 14.5, respectively.

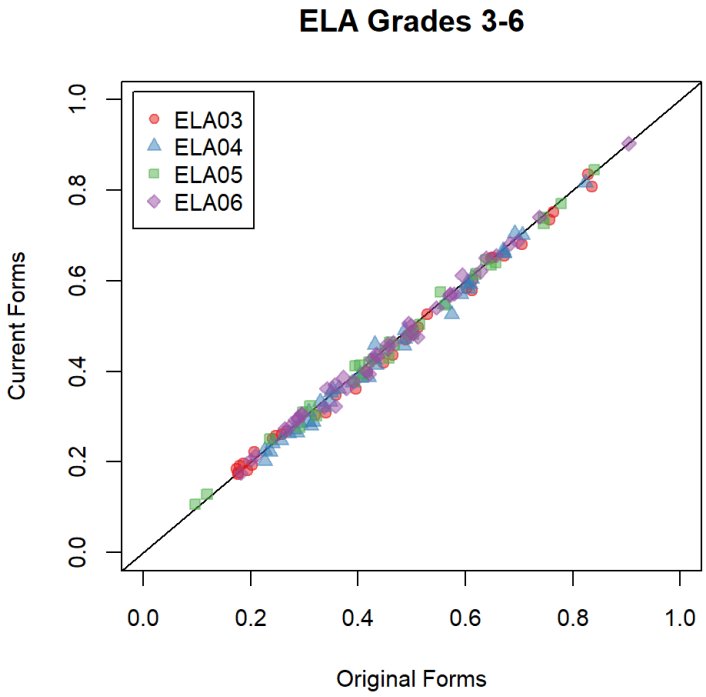

Figure 14.4 ELA/L Grades 3-6 P-Values

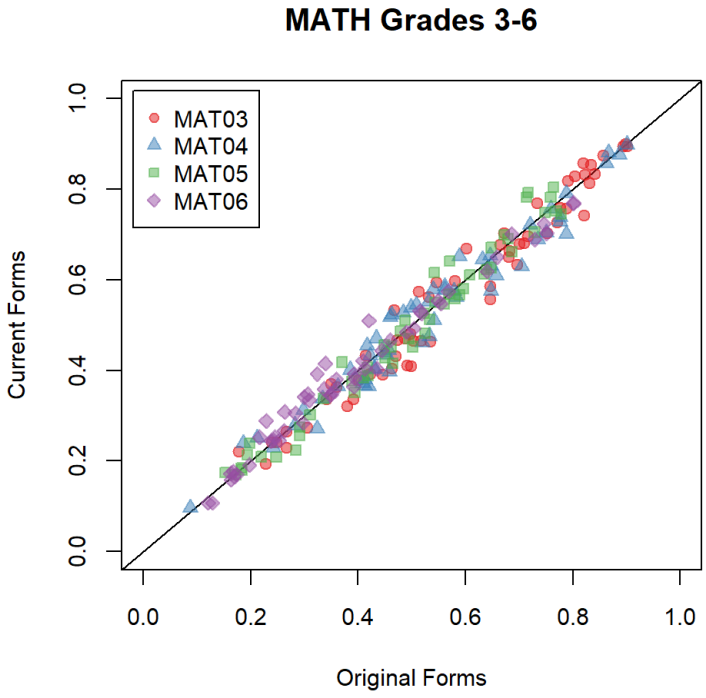

Figure 14.5 Mathematics Grades 3-6 P-Values

The scatterplots for all grades and courses are presented in Figures A.14.1 – A.14.6. Scatterplots show that most points cluster closely and evenly around the $y = x$ line, showing that items perform similarly on both forms with the matched samples, with the exception of ELA/L grade 10, Algebra II, and Geometry.

The distributions of p-value differences for all grades are presented in Tables A.14.4 and A.14.5. Differences tend to be small and center around zero, except for ELA/L grade 10, Algebra II, and Geometry. For ELA/L grades 3 through 8, differences in item difficulties range from -.049 to .070. For mathematics grades 3 through 8 and Algebra I, differences in item difficulties range from -.105 to .090. The high school assessments show larger differences. P-values for ELA/L grade 10 on the current forms were lower than on the original forms.

The polyserial correlations of common items on the current and original forms using the matched sample were also analyzed. Scatterplots, which are presented in Figures A.14.7 – A.14.12, show that most points cluster closely and evenly around the $y = x$ line, showing that items perform similarly on both forms with the matched sample, with the exception of Algebra I, Algebra II, and Geometry. The distributions of these differences, which are presented in Tables A.14.6 and A.14.7, tend to be small and center around 0, except for ELA/L grade 10, Algebra II, and Geometry. For ELA grades 3 through 8, differences in polyserial values range from -.058 to .043. For Mathematics grades 3 through 8, differences in polyserial values range from -.090 to 0.125. The high school assessments show larger differences.

Common items were checked for differential item functioning (DIF) on several categories separately for the current and original forms, using the matched samples. The resulting crosstabulation of DIF categories was examined. Percentages were computed for each possible combination of DIF categories and represented the total number of cross-tabulations divided by the total number of DIF calculations (items multiplied by categories for which the sample size was sufficient for DIF calculations) within a grade. For most tests, at least 90 percent of calculations displayed no DIF on both the current and original forms. DIF results summaries can be found in Tables A.14. 8 – A.14.10.

Scale Score Comparability: Test-Level Analysis

Test-level evaluations included analyzing reliability, scale score distributions, ELA/L claim score distributions, and subclaim distributions. Analyses showed that reliability, calculated as the stratified alpha, was slightly lower for current forms compared to their original form counterparts, as expected. For each assessment, the Spearman Brown (SB) Prophecy formula was used to predict the current form reliabilities based on the reduction in items. The current form reliability estimates tended to be generally similar to the Spearman-Brown prophecy values based on the corresponding reduction in points. This indicated that the loss of precision was approximately commensurate with the reduction in length. Similar results were found at the claim and subclaim levels.

Both raw score (RS) and scale score (SS) standard error of measurement (SEMs) are presented, as well as an adjusted raw score SEM that is simply the proportion of total points represented by the raw score SEM. The scale score and adjusted raw score SEMs were always slightly larger for the current forms, as expected. Reliability and SEM results at the summative level are available in Tables A.14.11 – A.14.16, while results for the claim and subclaim levels are available in and A.14.42 – A.14.52.

Scale score and subclaim distributions between the current and original forms tended to be similar, as evidenced by small effect sizes with respect to the difference in the means of the scale scores and distributions of the performance levels, except for ELA/L grade 10. The effect sizes, computed as Cohen's D, of the differences between the summative scale score current and original means were less than .20 in magnitude for all ELA/L and mathematics grades except ELA/L grade 10. Results are available in Tables A.14.17 and A.14.18. The effect sizes of

the differences between the current and original reading claim scale score means were also less than .20 in magnitude for all ELA/L grades except ELA/L grade 10. Results are presented in Table A.14.19. The effect sizes of the differences between the current and original writing claim scale score means were less than .20 in magnitude for all ELA/L grades except ELA/L grade 10. Results are available in Table A.14.20. Subclaim distributions for current and original forms using the matched sample were compared using Cramer's V effect size. All effect sizes were .20 or lower. Detailed results for ELA/L and mathematics grade 6 assessments are presented in Tables A14.21 and A.14.22, respectively, while results summaries for all grades and courses can be found in Tables A.14.23 and A.14.24.

Scale Score Comparability: Longitudinal Analysis

Longitudinal analyses generally revealed stability in scale score means when controlling for state participation. Effect sizes ranged in magnitude from 0 to .16, with all but two being smaller than .10. No clear directional pattern emerged. Detailed results can be found in Tables A.14.25 – A.14.28. Additionally, a regression analysis approach was used to examine the relationship between students' 2018 and 2019 scale scores. The full and reduced models are given below.

Full Model:

$$SS_{2019} = \beta_0 + \beta_1 \times SS_{2018} + \beta_2 \times C + \beta_3 \times SS_{2018} \times C \quad (14\text{-}7)$$

Reduced Model:

$$SS_{2019} = \beta_0 + \beta_1 \times SS_{2018} \quad (14\text{-}8)$$

where SS_{2019} is the scale score on the 2019 assessment, SS_{2018} is the scale score on the 2018 assessment, C is a categorical variable in which students taking the current assessment are indicated with a one and students taking the original assessment are indicated with a zero.

The changes in R^2 ranged from less than .0001 to .0260, demonstrating that the form choice for 2019 did not explain much additional variance in the 2019 scale scores. Regression results can be found in Tables A.14.29 and A.14.30.

As an additional component of the research, student growth percentiles (SGPs) were compared for students in the matched samples for grades 4 and higher who have prior achievement scores. Section 15 describes the SGP analyses conducted for spring 2019 administration. SGPs can be computed using either each individual state or the entire consortium as the peer group. For these analyses, SGPs are computed based on the consortium peer group.

The mean SGPs for students in the matched sample who were administered the current forms were compared with those in the sample who were administered the original forms. Means were computed across all students in the sample as well as for various subgroups. Similar means indicated that student growth can be measured similarly regardless of the type of form, providing additional evidence of comparability. SGP mean differences greater than 5 percentile points in magnitude, which corresponds to an effect size of approximately 0.18 (D. Betebenner, personal communication, September 10, 2019), may warrant further investigation.

For ELA/L and mathematics grades 4 – 8, differences between the mean SGPs were generally less than 5 percentile points in magnitude. At the overall level, mean differences (measured in percentile points and computed as the current form mean SGP minus original form mean SGP) ranged from -3.0 to 1.3 for ELA/L and from -2.7 to 3.5 for

mathematics. Subgroups evaluated were African American or Black, Asian, Hispanic, multiple races, Native American, white, economically disadvantaged, English learners, and students with disabilities. Except the Asian and Native American subgroups, the differences in the means were less than 5 in magnitude. For Asian students in mathematics grade 8, the difference in the means was 5.2. For Native American students, the differences for ELA/L grade 4, and mathematics grades 4, 6, and 8 were -5.3, -8.4, -9.1, and -6.5, respectively. Of note is that each of these exceptions occurs when the sample size is relatively small. For mathematics grade 8, there were only 730 Asian students administered each type of form; all Native American grades contained less than 200 students for each type of form. SGP mean differences for all students as well as for each of the subgroups for Algebra I tended to be slightly higher than 5 in absolute value, but always less than 10. Results for Geometry and Algebra 2 are not included due to small sample sizes.

These results provide additional evidence in support of comparability between the current and original scale scores at grades 4 – 8. For high school analyses, small samples, potential differences in course progressions, and possible differences in administration characteristics (e.g., graduation requirements) within each state complicate the interpretation of the results.

Performance Level Comparability: Test-Level Analyses

The performance level distributions for the current and original forms were compared using Cramer's V as the effect size measure. Summative performance level and college- and career-readiness (CCR), which is defined as students who attained performance levels 4 or 5, distributions tended to be similar across the current and original forms, with effect sizes of less than .10 in magnitude relative to the differences in their distributions, except for ELA/L grade 10. Detailed results for ELA/L and mathematics grade 3 can be found in Tables A.14.31 and A.14.32, respectively. A summary of the effect sizes for all assessments can be found in Table A.14.33. Additionally, the percentage of students attaining or exceeding the CCR indicator for Current and Original forms was calculated and compared using Cohen's *h* as the measure of effect size. All effect sizes were less than .10 in magnitude, except for ELA/L grade 10. These results can be found in Table A.14.34.

Performance Level Comparability: Classification Analyses

Classification accuracy and consistency were also computed using BB-Class (Brennan, 2004) in two ways: using all five performance levels and using only the CCR indicator. Both classification accuracy and consistency were always lower for current forms compared to the original forms, as expected, as there are differences in measurement precision discussed above. Effect sizes, as computed by Cohen's *h,* measuring the differences were small to moderate in magnitude, and ranged from -.04 to -.23 for performance level classification accuracy (Tables A.14.35 and A.14.37), from -.05 to -.25 for performance level classification consistency (Tables A.14.36 and A.14.38), from -.02 to -.10 for CCR classification accuracy (Tables A.14.35 and A.14.37), and from -.02 to -.12 for CCR classification consistency Tables (A.14.36 and A.14.38).

Performance Level Comparability: Longitudinal Analyses

Finally, a longitudinal evaluation of performance levels was conducted using all available data, rather than the matched samples. Performance level and CCR distributions were examined for each grade in 2018 and 2019, ensuring that data from both years represented the same states. Cramer's *V* and Cohen's *h* were used as the measures of effect size for the performance level and CCR comparisons, respectively. All effect sizes were .10 or less in magnitude. Detailed results for ELA/L and mathematics grade 6 can be found in Tables A.14.39 and A.14.40, while a summary of results across all assessments can be found in Table A.14.41.

Quality Testing Standards Summary

The purpose of the Quality Testing Standards study was to compare the results from the current and original assessments. Because states only administered one type, comparable samples were extracted from the data using coarsened exact matching. Using this data, a variety of analyses demonstrated that there appears to be broad comparability between the current and original scale scores and performance levels, that the current forms have less measurement precision than the original forms, and that the results from many of the high school tests were slightly less clear. Several factors limited the analysis of high school results. First, for ELA/L grade 10, the prior assessment used was ELA/L grade 8 from 2017. A test and results that are two years removed may be less than ideal. Second, high school tests tended to have smaller samples and were obtained from fewer states. Third, high school curriculum and course progressions may vary from state to state.

Additionally, several longitudinal analyses were conducted using assessment data from 2018 and 2019 rather than the matched sample. Although the analyses were limited in scope, the results support the findings from the matched analyses.

14.6 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 conducted 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.[19]

14.7 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.

14.8 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.

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

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 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.

14.9 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.

The technical report addendum provides the student characteristics and test reliability (total test score and for subclaims) for the 2018 fall block administration.

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. 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.

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

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 2018 *and* who had achieved a similar score as this student on the grade 6 ELA/L assessment in spring 2017 and the grade 5 ELA/L assessment in spring 2016.[20] 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 2018–2019 academic year is the fifth year of test administration. Students in states that participated in spring 2017 and spring 2018 generally received SGPs based on two prior scores. Students in states that participated in spring 2018 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 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 2017 and spring 2018 administrations. States administering traditional mathematics assessments in fall 2017 or fall 2018 may also have SGPs based on these prior scores.

[20] 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.

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.

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

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* ≥2), the τth quantile function for Y_t conditional on prior scores ($Y_{t-1},\ldots,Y_1$) is

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

where ϕ_{ju} (j =1,2,..., n students; u =1, ..., $t-1$ administrations) represent the B-spline basis functions. The SGP of each student i is the midpoint between the two consecutive τ whose quantile scores capture the student's

current score, multiplied by 100. For example, a student with a current score that lies between the fitted value for $\tau = .595$ and $\tau = .605$ would receive a SGP of 60.

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

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

		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
Male vs Female	52	1	2	.	.	50	96	1	2		
White vs Black	52	.	.	1	2	51	98	.	.		
White vs Hispanic	52	.	.	3	6	49	94	.	.		
White vs Asian	52	1	2	.	.	51	98	.	.		
White vs AmerIndian	52	.	.	.	.	52	100	.	.		
White vs Pacific Islander	52	.	.	1	2	50	96	1	2		
White vs Multiracial	52	.	.	.	.	52	100	.	.		
NoEcnDis vs EcnDis	52	.	.	.	.	52	100	.	.		
ELN vs ELY	52	.	.	2	4	50	96	.	.		
SWDN vs SWDY	52	.	.	.	.	52	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = 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
Male vs Female	68	.	.	5	7	62	91	1	1		
White vs Black	68	.	.	1	1	67	99	.	.		
White vs Hispanic	68	.	.	1	1	67	99	.	.		
White vs Asian	68	1	1	2	3	65	96	.	.		
White vs AmerIndian	68	.	.	2	3	66	97	.	.		
White vs Pacific Islander	68	.	.	1	1	67	99	.	.		
White vs Multiracial	68	1	1	.	.	67	99	.	.		
NoEcnDis vs EcnDis	68	.	.	.	.	68	100	.	.		
ELN vs ELY	68	.	.	1	1	67	99	.	.		
SWDN vs SWDY	68	.	.	3	4	65	96	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = 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
Male vs Female	61	2	3	2	3	56	92	1	2		
White vs Black	61	1	2	5	8	55	90	.	.		
White vs Hispanic	61	1	2	3	5	57	93	.	.		
White vs Asian	61	.	.	1	2	59	97	1	2		
White vs AmerIndian	61	.	.	.	.	61	100	.	.		
White vs Pacific Islander	61	1	2	1	2	58	95	1	2		
White vs Multiracial	61	.	.	.	.	61	100	.	.		
NoEcnDis vs EcnDis	61	.	.	4	7	57	93	.	.		
ELN vs ELY	61	5	8	3	5	53	87	.	.		
SWDN vs SWDY	61	1	2	3	5	57	93	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = 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
Male vs Female	71	1	1	8	11	61	86	1	1		
White vs Black	71	1	1	3	4	66	93	1	1		
White vs Hispanic	71	1	1	3	4	67	94	.	.		
White vs Asian	71	.	.	1	1	69	97	1	1		
White vs AmerIndian	71	.	.	6	8	64	90	1	1		
White vs Pacific Islander	71	.	.	1	1	70	99	.	.		
White vs Multiracial	71	.	.	.	.	71	100	.	.		
NoEcnDis vs EcnDis	71	.	.	.	.	71	100	.	.		
ELN vs ELY	71	1	1	2	3	68	96	.	.		
SWDN vs SWDY	71	.	.	.	.	71	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = 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
Male vs Female	57	1	2	5	9	51	89	.	.		
White vs Black	57	.	.	2	4	55	96	.	.		
White vs Hispanic	57	1	2	3	5	53	93	.	.		
White vs Asian	57	.	.	.	.	56	98	1	2		
White vs AmerIndian	57	3	5	.	.	53	93	1	2		
White vs Pacific Islander	57	.	.	1	2	55	96	1	2		
White vs Multiracial	57	.	.	.	.	57	100	.	.		
NoEcnDis vs EcnDis	57	.	.	1	2	56	98	.	.		
ELN vs ELY	57	2	4	6	11	48	84	1	2		
SWDN vs SWDY	57	.	.	.	.	57	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = 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
Male vs Female	67	.	.	3	4	62	93	2	3		
White vs Black	67	.	.	2	3	65	97	.	.		
White vs Hispanic	67	.	.	4	6	63	94	.	.		
White vs Asian	67	.	.	.	.	65	97	2	3		
White vs AmerIndian	67	.	.	5	7	62	93	.	.		
White vs Pacific Islander	67	.	.	.	.	67	100	.	.		
White vs Multiracial	67	.	.	.	.	67	100	.	.		
NoEcnDis vs EcnDis	67	.	.	.	.	67	100	.	.		
ELN vs ELY	67	2	3	6	9	59	88	.	.		
SWDN vs SWDY	67	.	.	.	.	67	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.7 Pre-Administration Differential Item Functioning for ELA/L Grade 9

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
Male vs Female	81	2	2	8	10	68	84	3	4		
White vs Black	81	.	.	4	5	77	95	.	.		
White vs Hispanic	81	.	.	3	4	78	96	.	.		
White vs Asian	81	1	1	.	.	79	98	1	1		
White vs AmerIndian	81	2	2	3	4	76	94	.	.		
White vs Pacific Islander	81	.	.	3	4	78	96	.	.		
White vs Multiracial	81	.	.	.	.	81	100	.	.		
NoEcnDis vs EcnDis	81	.	.	1	1	80	99	.	.		
ELN vs ELY	81	1	1	8	10	71	88	1	1		
SWDN vs SWDY	81	.	.	1	1	80	99	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.8 Pre-administration Differential Item Functioning for ELA/L Grade 10

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
Male vs Female	58	2	3	5	9	50	86	1	2		
White vs Black	58	.	.	.	.	57	98	1	2		
White vs Hispanic	58	.	.	1	2	57	98	.	.		
White vs Asian	58	1	2	.	.	57	98	.	.		
White vs AmerIndian	58	1	2	1	2	55	95	1	2		
White vs Pacific Islander	58	1	2	.	.	57	98	.	.		
White vs Multiracial	58	.	.	.	.	58	100	.	.		
NoEcnDis vs EcnDis	58	1	2	.	.	57	98	.	.		
ELN vs ELY	58	1	2	3	5	54	93	.	.		
SWDN vs SWDY	58	.	.	.	.	58	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.9 Pre-Administration Differential Item Functioning for ELA/L Grade 11

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
Male vs Female	55	1	2	4	7	49	89	1	2		
White vs Black	55	.	.	3	5	52	95	.	.		
White vs Hispanic	55	1	2	2	4	50	91	2	4		
White vs Asian	55	.	.	.	.	53	96	2	4		
White vs AmerIndian	55	3	5	1	2	50	91	1	2		
White vs Pacific Islander	55	.	.	.	.	55	100	.	.		
White vs Multiracial	55	.	.	.	.	55	100	.	.		
NoEcnDis vs EcnDis	55	1	2	.	.	54	98	.	.		
ELN vs ELY	55	3	5	4	7	48	87	.	.		
SWDN vs SWDY	55	.	.	.	.	55	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.10 Post-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
Male vs Female	52	1	2	.	.	51	98	.	.		
White vs Black	52	.	.	1	2	51	98	.	.		
White vs Hispanic	52	.	.	1	2	51	98	.	.		
White vs Asian	52	.	.	.	.	52	100	.	.		
White vs AmerIndian	52	.	.	1	2	51	98	.	.		
White vs Pacific Islander	52	1	2	3	6	47	90	1	2		
White vs Multiracial	52	.	.	.	.	52	100	.	.		
NoEcnDis vs EcnDis	52	.	.	.	.	52	100	.	.		
ELN vs ELY	52	.	.	3	6	49	94	.	.		
SWDN vs SWDY	52	.	.	1	2	51	98	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.11 Post-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
Male vs Female	68	.	.	3	4	62	91	3	4		
White vs Black	68	.	.	1	1	67	99	.	.		
White vs Hispanic	68	.	.	2	3	66	97	.	.		
White vs Asian	68	1	1	.	.	67	99	.	.		
White vs AmerIndian	68	.	.	1	1	67	99	.	.		
White vs Pacific Islander	68	.	.	2	3	65	96	1	1		
White vs Multiracial	68	.	.	.	.	68	100	.	.		
NoEcnDis vs EcnDis	68	.	.	.	.	68	100	.	.		
ELN vs ELY	68	.	.	.	.	68	100	.	.		
SWDN vs SWDY	68	.	.	.	.	68	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.12 Post-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
Male vs Female	61	.	.	4	7	53	87	4	7		
White vs Black	61	.	.	3	5	58	95	.	.		
White vs Hispanic	61	1	2	.	.	60	98	.	.		
White vs Asian	61	.	.	.	.	60	98	1	2		
White vs AmerIndian	61	1	2	3	5	57	93	.	.		
White vs Pacific Islander	61	.	.	1	2	60	98	.	.		
White vs Multiracial	61	.	.	.	.	61	100	.	.		
NoEcnDis vs EcnDis	61	.	.	.	.	61	100	.	.		
ELN vs ELY	61	2	3	5	8	54	89	.	.		
SWDN vs SWDY	61	1	2	.	.	60	98	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.13 Post-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
Male vs Female	71	1	1	7	10	63	89	.	.		
White vs Black	71	1	1	1	1	69	97	.	.		
White vs Hispanic	71	1	1	2	3	68	96	.	.		
White vs Asian	71	.	.	1	1	70	99	.	.		
White vs AmerIndian	71	2	3	7	10	61	86	1	1		
White vs Pacific Islander	71	1	1	2	3	68	96	.	.		
White vs Multiracial	71	.	.	.	.	71	100	.	.		
NoEcnDis vs EcnDis	71	.	.	.	.	71	100	.	.		
ELN vs ELY	71	1	1	10	14	60	85	.	.		
SWDN vs SWDY	71	.	.	1	1	70	99	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.14 Post-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
Male vs Female	57	.	.	4	7	53	93	.	.		
White vs Black	57	.	.	2	4	55	96	.	.		
White vs Hispanic	57	1	2	3	5	53	93	.	.		
White vs Asian	57	.	.	.	.	56	98	1	2		
White vs AmerIndian	57	2	4	4	7	51	89	.	.		
White vs Pacific Islander	57	.	.	.	.	57	100	.	.		
White vs Multiracial	57	.	.	.	.	57	100	.	.		
NoEcnDis vs EcnDis	57	.	.	1	2	56	98	.	.		
ELN vs ELY	57	4	7	7	12	46	81	.	.		
SWDN vs SWDY	57	.	.	.	.	57	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.15 Post-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
Male vs Female	67	.	.	3	4	63	94	1	1		
White vs Black	67	.	.	1	1	66	99	.	.		
White vs Hispanic	67	.	.	1	1	66	99	.	.		
White vs Asian	67	.	.	.	.	67	100	.	.		
White vs AmerIndian	67	1	1	4	6	62	93	.	.		
White vs Pacific Islander	67	.	.	.	.	66	99	1	1		
White vs Multiracial	67	.	.	.	.	67	100	.	.		
NoEcnDis vs EcnDis	67	.	.	.	.	67	100	.	.		
ELN vs ELY	67	2	3	7	10	58	87	.	.		
SWDN vs SWDY	67	.	.	.	.	67	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.16 Post-Administration Differential Item Functioning for ELA/L Grade 9

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
Male vs Female	81	1	1	8	10	68	84	4	5		
White vs Black	81	.	.	3	4	78	96	.	.		
White vs Hispanic	81	.	.	3	4	78	96	.	.		
White vs Asian	81	.	.	1	1	78	96	2	2		
White vs AmerIndian	81	4	5	2	2	75	93	.	.		
White vs Pacific Islander	81	.	.	4	5	77	95	.	.		
White vs Multiracial	81	.	.	.	.	81	100	.	.		
NoEcnDis vs EcnDis	81	.	.	1	1	80	99	.	.		
ELN vs ELY	81	3	4	10	12	68	84	.	.		
SWDN vs SWDY	81	.	.	.	.	81	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.17 Post-Administration Differential Item Functioning for ELA/L Grade 10

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
Male vs Female	58	1	2	6	10	51	88				
White vs Black	58	.	.	.	.	58	100				
White vs Hispanic	58	.	.	3	5	55	95				
White vs Asian	58	.	.	.	.	58	100				
White vs AmerIndian	58	3	5	4	7	51	88				
White vs Pacific Islander	58	.	.	5	9	53	91				
White vs Multiracial	58	.	.	.	.	58	100				
NoEcnDis vs EcnDis	58	.	.	1	2	57	98				
ELN vs ELY	58	5	9	7	12	46	79				
SWDN vs SWDY	58	.	.	.	.	58	100				

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.18 Post-Administration Differential Item Functioning for ELA/L Grade 11

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
Male vs Female	55	.	.	3	5	51	93	1	2		
White vs Black	55	1	2	3	5	51	93	.	.		
White vs Hispanic	55	.	.	2	4	53	96	.	.		
White vs Asian	55	.	.	2	4	52	95	1	2		
White vs AmerIndian	55	4	7	8	15	43	78	.	.		
White vs Pacific Islander	55	.	.	.	.	55	100	.	.		
White vs Multiracial	55	.	.	1	2	54	98	.	.		
NoEcnDis vs EcnDis	55	.	.	1	2	54	98	.	.		
ELN vs ELY	55	2	4	2	4	51	93	.	.		
SWDN vs SWDY	55	.	.	.	.	55	100	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.19 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
Male vs Female	77			1	1	75	97	1	1	.	.
White vs Black	77			6	8	68	88	3	4	.	.
White vs Hispanic	77			3	4	74	96	.	.	.	.
White vs Asian	77			.	.	67	87	9	12	1	1
White vs AmerIndian	77			2	3	75	97	.	.	.	.
White vs Pacific Islander	77			1	1	75	97	1	1	.	.
White vs Multiracial	77			1	1	75	97	1	1	.	.
NoEcnDis vs EcnDis	77			.	.	77	100	.	.	.	.
ELN vs ELY	77			1	1	76	99	.	.	.	.
SWDN vs SWDY	77			2	3	75	97	.	.	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.20 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
Male vs Female	72			1	1	71	99	.	.		
White vs Black	72			3	4	69	96	.	.		
White vs Hispanic	72			.	.	72	100	.	.		
White vs Asian	72			1	1	67	93	4	6		
White vs AmerIndian	72			4	6	68	94	.	.		
White vs Pacific Islander	72			1	1	70	97	1	1		
White vs Multiracial	72			.	.	72	100	.	.		
NoEcnDis vs EcnDis	72			.	.	72	100	.	.		
ELN vs ELY	72			4	6	67	93	1	1		
SWDN vs SWDY	72			2	3	70	97	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.21 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
Male vs Female	71			4	6	67	94	.	.	.	.
White vs Black	71			.	.	71	100	.	.	.	.
White vs Hispanic	71			.	.	70	99	1	1	.	.
White vs Asian	71			.	.	71	100	.	.	.	.
White vs AmerIndian	71			8	11	61	86	2	3	.	.
White vs Pacific Islander	71			.	.	71	100	.	.	.	.
White vs Multiracial	71			.	.	70	99	1	1	.	.
NoEcnDis vs EcnDis	71			.	.	71	100	.	.	.	.
ELN vs ELY	71			6	8	65	92	.	.	.	.
SWDN vs SWDY	71			1	1	69	97	.	.	1	1

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.22 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
Male vs Female	69	1	1	1	1	66	96	.	.	1	1
White vs Black	69	.	.	1	1	68	99	.	.	.	.
White vs Hispanic	69	.	.	.	.	69	100	.	.	.	.
White vs Asian	69	.	.	.	.	66	96	3	4	.	.
White vs AmerIndian	69	.	.	3	4	64	93	1	1	1	1
White vs Pacific Islander	69	.	.	.	.	69	100	.	.	.	.
White vs Multiracial	69	1	1	.	.	68	99	.	.	.	.
NoEcnDis vs EcnDis	69	.	.	.	.	69	100	.	.	.	.
ELN vs ELY	69	1	1	3	4	65	94	.	.	.	.
SWDN vs SWDY	69	.	.	3	4	63	91	3	4	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.23 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
Male vs Female	67	1	1	3	4	62	93	1	1	.	.
White vs Black	67	1	1	.	.	66	99	.	.	.	.
White vs Hispanic	67	.	.	1	1	66	99	.	.	.	.
White vs Asian	67	.	.	.	.	59	88	6	9	2	3
White vs AmerIndian	67	.	.	1	1	66	99	.	.	.	.
White vs Pacific Islander	67	.	.	1	1	66	99	.	.	.	.
White vs Multiracial	67	.	.	.	.	67	100	.	.	.	.
NoEcnDis vs EcnDis	67	.	.	.	.	67	100	.	.	.	.
ELN vs ELY	67	1	1	2	3	63	94	1	1	.	.
SWDN vs SWDY	67	.	.	.	.	67	100	.	.	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.24 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
Male vs Female	64			2	3	62	97	.	.		
White vs Black	64			2	3	61	95	1	2		
White vs Hispanic	64			.	.	64	100	.	.		
White vs Asian	64			.	.	62	97	2	3		
White vs AmerIndian	64			1	2	62	97	1	2		
White vs Pacific Islander	64			.	.	63	98	1	2		
White vs Multiracial	64			.	.	64	100	.	.		
NoEcnDis vs EcnDis	64			.	.	64	100	.	.		
ELN vs ELY	64			5	8	59	92	.	.		
SWDN vs SWDY	64			1	2	61	95	2	3		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.25 Differential Item Functioning for Algebra I

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
Male vs Female	111	.	.	2	2	108	97	1	1		
White vs Black	111	1	1	2	2	108	97	.	.		
White vs Hispanic	111	.	.	.	.	111	100	.	.		
White vs Asian	111	.	.	1	1	104	94	6	5		
White vs AmerIndian	111	.	.	3	3	107	96	1	1		
White vs Pacific Islander	111	.	.	.	.	111	100	.	.		
White vs Multiracial	111	.	.	1	1	109	98	1	1		
NoEcnDis vs EcnDis	111	.	.	.	.	111	100	.	.		
ELN vs ELY	111	1	1	4	4	102	92	4	4		
SWDN vs SWDY	111	1	1	.	.	110	99	.	.		

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.26 Differential Item Functioning for Geometry

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
Male vs Female	118	.	.	2	2	116	98	.	.	.	.
White vs Black	118	.	.	2	2	115	97	1	1	.	.
White vs Hispanic	118	.	.	2	2	116	98	.	.	.	.
White vs Asian	118	.	.	.	.	110	93	7	6	1	1
White vs AmerIndian	118	1	1	4	3	109	92	4	3	.	.
White vs Pacific Islander	118	.	.	.	.	118	100	.	.	.	.
White vs Multiracial	118	.	.	.	.	117	99	1	1	.	.
NoEcnDis vs EcnDis	118	.	.	1	1	117	99	.	.	.	.
NoEcnDis vs EcnDis	118	1	1	8	7	104	88	5	4	.	.
SWDN vs SWDY	118	1	1	2	2	115	97	.	.	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.27 Differential Item Functioning for Algebra II

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
Male vs Female	109	.	.	5	5	102	94	2	2	.	.
White vs Black	109	.	.	3	3	106	97	.	.	.	.
White vs Hispanic	109	.	.	1	1	108	99	.	.	.	.
White vs Asian	109	.	.	1	1	98	90	8	7	2	2
White vs AmerIndian	109	1	1	1	1	106	97	1	1	.	.
White vs Pacific Islander	109	.	.	.	.	108	99	1	1	.	.
White vs Multiracial	109	.	.	.	.	109	100	.	.	.	.
NoEcnDis vs EcnDis	109	.	.	.	.	109	100	.	.	.	.
ELN vs ELY	109	2	2	3	3	99	91	4	4	1	1
SWDN vs SWDY	109	.	.	4	4	105	96	.	.	.	.

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Table A.6.28 Differential Item Functioning for Integrated Mathematics I

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
Male vs Female	42			.	.	42	100				
White vs Black	42			.	.	42	100				
White vs Hispanic	42			.	.	42	100				
White vs Asian	42			.	.	42	100				
White vs AmerIndian	42			.	.	42	100				
White vs Pacific Islander	42			.	.	42	100				
White vs Multiracial	42			1	2	41	98				
NoEcnDis vs EcnDis	42			.	.	42	100				
NoEcnDis vs EcnDis	42			.	.	42	100				
SWDN vs SWDY	42			1	2	41	98				

Note: AmerIndian = American Indian/Alaska Native, Black = Black/African American, Hispanic = Hispanic/Latino, Pacific Islander = Native Hawaiian/Pacific Islander, Multiracial = Multiple Race Selected, NoEcnDis = not economically disadvantaged, EcnDis = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability.

Appendix 7.1: Post-Equated IRT Results for Spring 2019 English Language Arts/Literacy (ELA/L)

Table A.7.1 Post-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**
E03	All Items	128	58	0.32	0.91	-1.66	2.05	0.60	0.24	0.22	1.24
	Reading	92	46	0.00	0.74	-1.66	2.02	0.50	0.15	0.22	0.84
	Writing	36	12	1.53	0.25	1.26	2.05	0.96	0.12	0.72	1.24
E04	All Items	164	74	0.16	1.55	-9.56	2.35	0.45	0.23	0.12	0.99
	Reading	124	62	-0.03	1.61	-9.56	2.35	0.36	0.13	0.12	0.74
	Writing	40	12	1.18	0.37	0.81	1.83	0.89	0.06	0.81	0.99
E05	All Items	145	66	0.25	1.06	-5.38	2.63	0.49	0.21	0.10	0.96
	Reading	112	56	0.11	1.06	-5.38	2.06	0.42	0.15	0.10	0.75
	Writing	33	10	1.07	0.68	0.51	2.63	0.86	0.07	0.74	0.96
E06	All Items	172	77	0.27	0.87	-1.93	2.95	0.50	0.23	0.18	1.16
	Reading	130	65	0.08	0.81	-1.93	2.95	0.43	0.15	0.18	0.96
	Writing	42	12	1.26	0.42	0.68	1.88	0.91	0.15	0.63	1.16
E07	All Items	139	62	0.16	0.73	-1.34	2.37	0.48	0.25	0.17	1.13
	Reading	104	52	0.03	0.71	-1.34	2.37	0.39	0.12	0.17	0.71
	Writing	35	10	0.83	0.42	0.15	1.54	0.96	0.16	0.67	1.13
E08	All Items	159	72	0.13	0.82	-1.88	2.83	0.47	0.25	0.18	1.19
	Reading	124	62	0.03	0.82	-1.88	2.83	0.38	0.12	0.18	0.70
	Writing	35	10	0.78	0.36	0.37	1.23	1.02	0.16	0.73	1.19
E09	All Items	197	88	0.59	0.80	-1.36	2.95	0.51	0.29	0.14	1.23
	Reading	148	74	0.55	0.85	-1.36	2.95	0.40	0.15	0.14	0.76
	Writing	49	14	0.83	0.38	0.09	1.55	1.09	0.11	0.81	1.23
E10	All Items	141	63	0.59	0.79	-0.93	2.85	0.49	0.27	0.14	1.12
	Reading	106	53	0.56	0.85	-0.93	2.85	0.39	0.15	0.14	0.94
	Writing	35	10	0.73	0.32	0.30	1.24	1.02	0.07	0.93	1.12
E11	All Items	139	62	0.92	0.83	-0.67	4.55	0.46	0.24	0.08	1.10
	Reading	104	52	0.88	0.89	-0.67	4.55	0.38	0.15	0.08	0.84
	Writing	35	10	1.13	0.27	0.68	1.51	0.90	0.15	0.63	1.10

Table A.7.2 Post-Equated IRT Standard Errors of Item Parameter Estimates 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
E03	All Items	102	46	0.01	0.00	0.00	0.03	0.01	0.00	0.00	0.02
	Reading	72	36	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.01
	Writing	30	10	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02
E04	All Items	137	62	0.02	0.07	0.00	0.52	0.00	0.00	0.00	0.02
	Reading	104	52	0.02	0.07	0.00	0.52	0.00	0.00	0.00	0.01
	Writing	33	10	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.02
E05	All Items	141	64	0.01	0.01	0.00	0.06	0.00	0.00	0.00	0.01
	Reading	108	54	0.01	0.01	0.00	0.06	0.00	0.00	0.00	0.01
	Writing	33	10	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.01
E06	All Items	139	62	0.01	0.01	0.00	0.03	0.00	0.00	0.00	0.02
	Reading	104	52	0.01	0.01	0.00	0.03	0.00	0.00	0.00	0.01
	Writing	35	10	0.01	0.01	0.00	0.03	0.01	0.00	0.00	0.02
E07	All Items	135	60	0.01	0.00	0.00	0.02	0.01	0.00	0.00	0.02
	Reading	100	50	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.01
	Writing	35	10	0.01	0.00	0.00	0.02	0.01	0.01	0.00	0.02
E08	All Items	139	62	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.02
	Reading	104	52	0.01	0.01	0.00	0.03	0.00	0.00	0.00	0.01
	Writing	35	10	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.02
E09	All Items	77	34	0.01	0.01	0.00	0.04	0.01	0.00	0.00	0.02
	Reading	56	28	0.01	0.01	0.01	0.04	0.00	0.00	0.00	0.01
	Writing	21	6	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.02
E10	All Items	139	62	0.01	0.01	0.00	0.04	0.01	0.00	0.00	0.02
	Reading	104	52	0.01	0.01	0.01	0.04	0.00	0.00	0.00	0.01
	Writing	35	10	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.02
E11	All Items	100	44	0.03	0.06	0.01	0.38	0.01	0.01	0.01	0.03
	Reading	72	36	0.04	0.06	0.01	0.38	0.01	0.00	0.01	0.02
	Writing	28	8	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.03

Table A.7.3 Post-Equated IRT Item Model Fit for ELA/L by Grade

Grade	Item Grouping	No. of Score Points	No. of Items	G^2				Q_1			
				Mean	SD	Min	Max	Mean	SD	Min	Max
E03	All Items	102	46	2,732.6	2,016.8	385.7	10,703.0	2,574.6	2,037.9	360.1	11,583.4
	Reading	72	36	2,969.8	2,182.6	385.7	10,703.0	2,786.4	2,223.7	360.1	11,583.4
	Writing	30	10	1,878.6	880.8	416.8	2,880.3	1,812.3	842.8	385.8	2,758.7
E04	All Items	137	62	3,584.0	3,089.3	163.8	14,358.4	3,473.5	3,003.8	159.2	14,072.8
	Reading	104	52	3,697.0	3,305.7	163.8	14,358.4	3,591.1	3,211.4	159.2	14,072.8
	Writing	33	10	2,996.5	1,518.5	405.1	4,954.5	2,861.6	1,490.4	370.8	4,929.9
E05	All Items	141	64	2,920.3	3,540.3	151.3	18,025.6	2,806.2	3,505.6	142.6	17,306.2
	Reading	108	54	2,937.7	3,764.1	151.3	18,025.6	2,793.7	3,671.5	142.6	17,306.2
	Writing	33	10	2,826.2	2,070.3	406.3	5,952.4	2,873.7	2,576.5	365.8	8,622.0
E06	All Items	139	62	3,284.2	2,606.4	289.7	13,658.8	3,055.4	2,407.7	291.5	11,996.2
	Reading	104	52	3,319.6	2,785.0	289.7	13,658.8	3,110.2	2,574.1	291.5	11,996.2
	Writing	35	10	3,100.5	1,431.2	494.8	5,135.3	2,770.1	1,278.4	437.5	4,441.0
E07	All Items	135	60	3,436.0	4,207.6	148.1	24,499.4	3,263.3	4,170.4	140.0	26,003.2
	Reading	100	50	3,295.4	4,308.1	148.1	24,499.4	3,194.4	4,367.2	140.0	26,003.2
	Writing	35	10	4,139.1	3,788.2	474.8	10,342.1	3,607.9	3,165.3	418.5	8,867.1
E08	All Items	139	62	3,502.7	3,075.6	125.0	14,717.3	3,296.8	2,871.1	123.0	12,427.9
	Reading	104	52	3,262.2	3,178.3	125.0	14,717.3	3,140.1	3,016.9	123.0	12,427.9
	Writing	35	10	4,753.6	2,189.6	668.3	7,055.1	4,111.3	1,848.5	593.0	6,093.5
E09	All Items	77	34	2,394.4	2,548.5	252.1	13,398.9	2,225.1	2,452.2	226.2	12,715.7
	Reading	56	28	2,279.7	2,749.1	252.1	13,398.9	2,160.5	2,662.4	226.2	12,715.7
	Writing	21	6	2,929.2	1,279.9	1,419.3	4,383.7	2,526.5	1,130.7	1,203.2	3,824.6
E10	All Items	139	62	2,325.6	1,874.8	188.5	8,318.2	2,220.8	1,887.5	183.2	8,269.9
	Reading	104	52	2,307.3	2,024.2	188.5	8,318.2	2,247.8	2,041.9	183.2	8,269.9
	Writing	35	10	2,420.4	769.7	920.2	3,692.6	2,080.2	702.3	743.1	3,306.3
E11	All Items	100	44	565.9	320.9	105.9	1,718.9	514.5	294.2	104.4	1,666.5
	Reading	72	36	520.6	327.7	105.9	1,718.9	477.3	304.8	104.4	1,666.5
	Writing	28	8	770.1	193.0	428.4	1,063.0	682.3	166.6	369.9	902.5

Appendix 7.2: Pre-Equated IRT Results for Spring 2019 English Language Arts/Literacy (ELA/L)

Table A.7.4 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**
E03	All Items	128	58	0.37	0.97	-1.40	3.13	0.59	0.21	0.16	1.01
	Reading	92	46	0.05	0.81	-1.40	3.13	0.51	0.15	0.16	0.84
	Writing	36	12	1.59	0.39	1.16	2.32	0.90	0.10	0.72	1.01
E04	All Items	164	74	0.24	1.29	-6.48	2.29	0.45	0.22	0.17	1.02
	Reading	124	62	0.06	1.32	-6.48	2.29	0.37	0.12	0.17	0.75
	Writing	40	12	1.19	0.49	0.74	2.29	0.87	0.09	0.67	1.02
E05	All Items	145	66	0.28	1.15	-6.27	2.69	0.49	0.23	0.19	1.06
	Reading	112	56	0.14	1.16	-6.27	2.23	0.41	0.14	0.19	0.70
	Writing	33	10	1.06	0.71	0.47	2.69	0.91	0.09	0.71	1.06
E06	All Items	172	77	0.29	0.92	-1.97	4.45	0.51	0.23	0.20	1.13
	Reading	130	65	0.11	0.87	-1.97	4.45	0.45	0.17	0.20	1.10
	Writing	42	12	1.25	0.42	0.59	1.93	0.89	0.15	0.60	1.13
E07	All Items	139	62	0.22	0.70	-1.33	1.86	0.49	0.24	0.17	1.18
	Reading	104	52	0.10	0.68	-1.33	1.86	0.40	0.13	0.17	0.74
	Writing	35	10	0.84	0.44	0.30	1.70	0.95	0.15	0.66	1.18
E08	All Items	159	72	0.13	0.78	-2.03	2.68	0.47	0.23	0.19	1.12
	Reading	124	62	0.02	0.79	-2.03	2.68	0.39	0.12	0.19	0.69
	Writing	35	10	0.75	0.37	0.30	1.32	0.98	0.10	0.81	1.12
E09	All Items	197	88	0.63	0.79	-1.29	2.95	0.52	0.30	0.17	1.44
	Reading	148	74	0.59	0.84	-1.29	2.95	0.40	0.15	0.17	0.73
	Writing	49	14	0.85	0.38	0.12	1.55	1.12	0.16	0.86	1.44
E10	All Items	141	63	0.62	0.75	-0.54	2.81	0.50	0.28	0.13	1.24
	Reading	106	53	0.59	0.80	-0.54	2.81	0.40	0.16	0.13	0.93
	Writing	35	10	0.80	0.31	0.41	1.25	1.05	0.14	0.84	1.24
E11	All Items	139	62	0.88	0.68	-0.67	2.80	0.46	0.23	0.14	1.10
	Reading	104	52	0.82	0.71	-0.67	2.80	0.39	0.15	0.14	0.84
	Writing	35	10	1.20	0.33	0.61	1.74	0.85	0.17	0.56	1.10

Appendix 7.3: Pre-Equated IRT Results for Spring 2019 Mathematics

Table A.7.5 Pre-Equated IRT Summary Parameter Estimates for All Items for Mathematics by Grade/Subject

				b Estimates Summary				*a* Estimates Summary			
Grade	Item Grouping	No. of Score Points	No. of Items	Mean	SD	Min	Max	Mean	SD	Min	Max
M03	All Items	110	77	-0.28	0.98	-2.40	1.90	0.79	0.24	0.32	1.33
	SSMC	20	20	-0.77	0.88	-2.14	1.90	0.73	0.14	0.43	0.94
	CR	90	57	-0.11	0.96	-2.40	1.68	0.82	0.26	0.32	1.33
	Type I	74	67	-0.45	0.92	-2.40	1.90	0.83	0.22	0.43	1.33
	Type II	21	6	0.91	0.71	-0.07	1.68	0.43	0.06	0.32	0.51
	Type III	15	4	0.74	0.31	0.52	1.18	0.66	0.16	0.50	0.84
M04	All Items	112	72	-0.15	0.95	-2.61	2.54	0.74	0.20	0.38	1.32
	SSMC	18	18	-1.06	0.66	-2.01	0.37	0.70	0.22	0.40	1.24
	CR	94	54	0.15	0.83	-2.61	2.54	0.75	0.20	0.38	1.32
	Type I	74	61	-0.30	0.94	-2.61	2.54	0.76	0.20	0.40	1.32
	Type II	20	6	0.56	0.33	-0.11	0.80	0.59	0.17	0.38	0.81
	Type III	18	5	0.82	0.43	0.16	1.17	0.62	0.19	0.40	0.82
M05	All Items	116	71	0.02	0.91	-2.21	1.77	0.73	0.27	0.19	1.57
	SSMC	20	20	-0.58	0.77	-2.14	0.90	0.78	0.29	0.27	1.42
	CR	96	51	0.25	0.87	-2.21	1.77	0.70	0.26	0.19	1.57
	Type I	71	59	-0.16	0.87	-2.21	1.75	0.76	0.28	0.19	1.57
	Type II	24	7	0.91	0.62	0.05	1.77	0.53	0.18	0.27	0.73
	Type III	21	5	0.81	0.68	-0.17	1.69	0.64	0.15	0.45	0.80
M06	All Items	121	69	0.36	0.89	-3.02	1.98	0.72	0.24	0.20	1.30
	SSMC	15	15	-0.30	1.00	-3.02	0.74	0.64	0.25	0.20	1.19
	CR	106	54	0.54	0.77	-1.17	1.98	0.75	0.23	0.31	1.30
	Type I	75	57	0.23	0.89	-3.02	1.83	0.75	0.25	0.20	1.30
	Type II	25	7	0.90	0.52	-0.02	1.38	0.59	0.11	0.43	0.74
	Type III	21	5	1.09	0.70	0.13	1.98	0.59	0.11	0.45	0.70
M07	All Items	112	67	0.75	0.95	-1.03	3.36	0.69	0.29	0.19	1.38
	SSMC	20	20	0.41	1.17	-1.03	3.13	0.53	0.24	0.19	0.88
	CR	92	47	0.90	0.81	-0.67	3.36	0.76	0.28	0.25	1.38
	Type I	70	56	0.61	0.91	-1.03	3.13	0.73	0.30	0.19	1.38
	Type II	21	6	1.72	1.07	0.76	3.36	0.47	0.11	0.31	0.61
	Type III	21	5	1.19	0.44	0.60	1.76	0.58	0.09	0.50	0.74

Grade	Item Grouping	No. of Score Points	No. of Items	*b* Estimates Summary				*a* Estimates Summary			
				Mean	SD	Min	Max	Mean	SD	Min	Max
M08	All Items	115	64	0.91	0.98	-1.12	2.55	0.61	0.21	0.22	1.29
	SSMC	14	14	0.08	0.84	-1.12	1.74	0.47	0.15	0.22	0.78
	CR	101	50	1.15	0.88	-0.76	2.55	0.66	0.21	0.24	1.29
	Type I	73	52	0.70	0.94	-1.12	2.44	0.62	0.23	0.22	1.29
	Type II	24	7	1.65	0.45	1.08	2.45	0.65	0.13	0.55	0.91
	Type III	18	5	2.06	0.40	1.52	2.55	0.54	0.14	0.41	0.79
A1	All Items	209	111	1.27	1.03	-0.96	3.62	0.58	0.27	0.16	1.41
	SSMC	42	42	0.79	1.09	-0.96	3.62	0.45	0.18	0.16	0.85
	CR	167	69	1.56	0.88	-0.77	3.24	0.66	0.28	0.17	1.41
	Type I	131	91	1.11	1.07	-0.96	3.62	0.57	0.28	0.16	1.41
	Type II	39	11	1.89	0.29	1.55	2.51	0.68	0.17	0.38	0.91
	Type III	39	9	2.09	0.38	1.50	2.60	0.58	0.12	0.41	0.72
G1	All Items	223	118	1.16	0.94	-1.25	3.83	0.71	0.31	0.19	1.54
	SSMC	26	26	0.62	1.18	-1.25	3.83	0.47	0.19	0.19	0.77
	CR	197	92	1.31	0.80	-0.86	3.50	0.78	0.30	0.19	1.54
	Type I	130	95	0.99	0.95	-1.25	3.83	0.71	0.34	0.19	1.54
	Type II	42	12	1.94	0.53	1.17	2.79	0.75	0.08	0.63	0.89
	Type III	51	11	1.78	0.39	1.05	2.23	0.72	0.21	0.36	1.09
A2	All Items	218	109	1.41	0.92	-1.53	3.67	0.65	0.29	0.18	1.34
	SSMC	24	24	0.86	0.97	-1.53	2.48	0.49	0.20	0.18	0.89
	CR	194	85	1.57	0.85	-0.34	3.67	0.70	0.29	0.19	1.34
	Type I	133	88	1.24	0.88	-1.53	3.67	0.66	0.30	0.18	1.34
	Type II	34	10	1.83	0.85	0.48	3.29	0.63	0.20	0.40	0.96
	Type III	51	11	2.41	0.48	1.64	3.07	0.61	0.25	0.34	1.13
M1	All Items	81	42	1.02	0.88	-0.64	2.78	0.62	0.23	0.25	1.39
	SSMC	13	13	0.68	0.65	-0.64	1.86	0.49	0.17	0.25	0.84
	CR	68	29	1.16	0.93	-0.52	2.78	0.68	0.23	0.25	1.39
	Type I	49	34	0.75	0.73	-0.64	2.27	0.61	0.25	0.25	1.39
	Type II	14	4	2.02	0.69	1.12	2.78	0.72	0.10	0.57	0.78
	Type III	18	4	2.25	0.30	1.92	2.55	0.61	0.19	0.42	0.78

Grade	Item Grouping	No. of Score Points	No. of Items	*b* Estimates Summary				*a* Estimates Summary			
				Mean	SD	Min	Max	Mean	SD	Min	Max
M2	All Items	80	41	1.58	1.30	-0.67	4.68	0.67	0.31	0.17	1.30
	SSMC	11	11	-0.01	0.56	-0.67	1.44	0.67	0.22	0.31	1.00
	CR	69	30	2.17	0.96	0.41	4.68	0.68	0.35	0.17	1.30
	Type I	48	33	1.43	1.37	-0.67	4.68	0.68	0.34	0.17	1.30
	Type II	14	4	2.43	0.87	1.77	3.62	0.72	0.26	0.46	1.07
	Type III	18	4	1.97	0.58	1.32	2.54	0.62	0.11	0.46	0.71
M3	All Items	81	40	1.39	0.94	-0.35	3.32	0.57	0.27	0.17	1.27
	SSMC	7	7	1.02	0.69	-0.27	1.62	0.41	0.21	0.17	0.82
	CR	74	33	1.47	0.97	-0.35	3.32	0.60	0.27	0.24	1.27
	Type I	49	32	1.26	0.93	-0.35	3.32	0.59	0.29	0.17	1.27
	Type II	14	4	1.41	0.70	0.36	1.83	0.50	0.18	0.32	0.69
	Type III	18	4	2.40	0.72	1.57	3.06	0.47	0.09	0.37	0.57

Note: M03 through M08 = mathematics grades 3 through 8, A1 = Algebra I, GO = Geometry, A2 = Algebra II, M1 = Integrated Mathematics I, M2 = Integrated Mathematics II, M3 = Integrated Mathematics III.

Appendix 11: Students by Grade/Subject and Mode, for Each State

Table A.11.1 All ELA/L Students, by State, and Grade

		English Language Arts-Literacy									
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All States	N of Students	1,879,282	256,870	265,169	271,778	275,277	269,386	266,251	121,619	118,322	34,610
	N of CBT	1,825,655	224,957	259,642	267,807	271,346	265,686	263,370	121,061	117,751	34,035
	% of CBT	97.1	87.6	97.9	98.5	98.6	98.6	98.9	99.5	99.5	98.3
	N of PBT	53,627	31,913	5,527	3,971	3,931	3,700	2,881	558	571	575
	% of PBT	2.9	12.4	2.1	1.5	1.4	1.4	1.1	0.5	0.5	1.7
BIE	% of All Data	0.6	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0
	N of Students	8,485	1,322	1,336	1,391	1,298	1,108	1,065	190	193	582
	N of CBT	3,168	465	451	470	463	477	473	46	42	281
	% of CBT	37.3	35.2	33.8	33.8	35.7	43.1	44.4	24.2	21.8	48.3
	N of PBT	5,317	857	885	921	835	631	592	144	151	301
	% of PBT	62.7	64.8	66.2	66.2	64.3	56.9	55.6	75.8	78.2	51.7
IL	% of All Data	45.5	7.3	7.5	7.7	7.8	7.6	7.6	n/a	n/a	n/a
	N of Students	855,200	137,092	140,534	144,713	146,878	143,739	142,244	n/a	n/a	n/a
	N of CBT	810,389	106,583	136,238	141,983	144,199	141,046	140,340	n/a	n/a	n/a
	% of CBT	94.8	77.7	96.9	98.1	98.2	98.1	98.7	n/a	n/a	n/a
	N of PBT	44,811	30,509	4,296	2,730	2,679	2,693	1,904	n/a	n/a	n/a
	% of PBT	5.2	22.3	3.1	1.9	1.8	1.9	1.3	n/a	n/a	n/a
NJ	% of All Data	42.7	5.1	5.3	5.3	5.4	5.3	5.3	5.2	5.1	0.7
	N of Students	802,604	95,757	98,925	100,251	101,952	100,218	98,969	97,868	95,769	12,895
	N of CBT	801,390	95,668	98,816	100,147	101,825	100,106	98,862	97,654	95,517	12,795
	% of CBT	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.7	99.2
	N of PBT	1,214	89	109	104	127	112	107	214	252	100
	% of PBT	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.3	0.8

Table A.11.1 All ELA/L Students, by State, and Grade

		English Language Arts-Literacy									
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
NM	% of All Data	11.4	1.2	1.3	1.4	1.3	1.3	1.3	1.3	1.2	1.1
	N of Students	212,993	22,699	24,374	25,423	25,149	24,321	23,973	23,561	22,360	21,133
	N of CBT	210,708	22,241	24,137	25,207	24,859	24,057	23,695	23,361	22,192	20,959
	% of CBT	98.9	98.0	99.0	99.2	98.8	98.9	98.8	99.2	99.2	99.2
	N of PBT	2,285	458	237	216	290	264	278	200	168	174
	% of PBT	1.1	2.0	1.0	0.8	1.2	1.1	1.2	0.8	0.8	0.8

Note: BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; n/a=not applicable.

Table A.11.2 All Mathematics Students, by State, and Grade

		Mathematics												
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	A1	GO	A2	M1	M2	M3
All States	N of Students	1,871,889	258,807	266,629	272,714	275,732	264,960	225,726	134,107	105,010	66,789	673	541	201
	N of CBT	1,818,324	226,933	261,092	268,724	271,798	261,252	222,869	133,482	104,444	66,317	672	540	201
	% of CBT	97.1	88	98	99	99	99	99	100	100	99	100	100	100
	N of PBT	53,565	31,874	5,537	3,990	3,934	3,708	2,857	625	566	472	n/r	n/r	n/r
	% of PBT	2.9	12	2	2	1	1	1	1	1	1	n/r	n/r	n/r
BIE	% of All Data	0.6	0	0	0	0	0	0	0	0	0	0	0	n/r
	N of Students	8,446	1,321	1,341	1,390	1,280	1,099	1,048	230	232	493	8	4	n/r
	N of CBT	3,164	463	452	469	461	468	460	61	64	254	n/r	n/r	n/r
	% of CBT	37.5	35	34	34	36	43	44	27	28	52	n/r	n/r	n/r
	N of PBT	5,282	858	889	921	819	631	588	169	168	239	n/r	n/r	n/r
	% of PBT	62.5	65	66	66	64	57	56	74	72	49	n/r	n/r	n/r
IL	% of All Data	45.5	7	8	8	8	8	8	n/a	n/a	n/a	n/a	n/a	n/a
	N of Students	852,627	136,938	140,253	144,476	146,399	143,162	141,399	n/a	n/a	n/a	n/a	n/a	n/a
	N of CBT	807,882	106,468	135,959	141,751	143,719	140,475	139,510	n/a	n/a	n/a	n/a	n/a	n/a
	% of CBT	94.8	78	97	98	98	98	99	n/a	n/a	n/a	n/a	n/a	n/a
	N of PBT	44,745	30,470	4,294	2,725	2,680	2,687	1,889	n/a	n/a	n/a	n/a	n/a	n/a
	% of PBT	5.2	22	3	2	2	2	1	n/a	n/a	n/a	n/a	n/a	n/a
NJ	% of All Data	42.6	5	5	5	6	5	3	6	5	2	n/a	n/a	n/a
	N of Students	798,393	96,812	99,881	101,194	102,788	96,339	63,385	108,673	83,800	45,521	n/a	n/a	n/a
	N of CBT	797,135	96,726	99,764	101,066	102,641	96,212	63,268	108,437	83,587	45,434	n/a	n/a	n/a
	% of CBT	99.8	100	100	100	100	100	100	100	100	100	n/a	n/a	n/a
	N of PBT	1,258	86	117	128	147	127	117	236	213	87	n/a	n/a	n/a
	% of PBT	0.2	0	0	0	0	0	0	0	0	0	n/a	n/a	n/a
NM	% of All Data	11.2	1	1	1	1	1	1	1	1	1	0	0	0
	N of Students	212,423	23,736	25,154	25,654	25,265	24,360	19,894	25,204	20,978	20,775	665	537	201
	N of CBT	210,143	23,276	24,917	25,438	24,977	24,097	19,631	24,984	20,793	20,629	664	536	201
	% of CBT	98.9	98	99	99	99	99	99	99	99	99	100	100	100
	N of PBT	2,280	460	237	216	288	263	263	220	185	146	n/r	n/r	n/r

Table A.11.2 All Mathematics Students, by State, and Grade

		Mathematics												
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	A1	GO	A2	M1	M2	M3
	% of PBT	1.1	2	1	1	1	1	1	1	1	1	n/r	n/r	n/r

Note: BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico; A1=Algebra I, GO=Geometry, A2 = Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II, M3=Integrated Mathematics III. CBT=computer-based test; PBT=paper-based test; n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.3 All Spanish-Language Mathematics Students, by State, and Grade

		Mathematics												
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	A1	GO	A2	M1	M2	M3
All States	N of Students	23,534	4,812	3,691	3,270	2,642	2,255	2,118	2,426	1,728	558	n/r	n/r	n/r
	N of CBT	22,664	4,063	3,668	3,247	2,623	2,243	2,108	2,400	1,721	557	n/r	n/r	n/r
	% of CBT	96.3	84	99	99	99	100	100	99	100	100	n/r	n/r	n/r
	N of PBT	870	749	23	23	n/r	n/r	n/r	26	n/r	n/r	n/r	n/r	n/r
	% of PBT	3.7	16	1	1	n/r	n/r	n/r	1	n/r	n/r	n/r	n/r	n/r
BIE	% of All Data	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	N of Students	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	N of CBT	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	% of CBT	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	N of PBT	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	% of PBT	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
IL	% of All Data	30.9	10	6	7	4	2	2	n/a	n/a	n/a	n/a	n/a	n/a
	N of Students	7,275	2,244	1,477	1,726	942	448	438	n/a	n/a	n/a	n/a	n/a	n/a
	N of CBT	6,498	1,503	1,460	1,719	933	447	436	n/a	n/a	n/a	n/a	n/a	n/a
	% of CBT	89.3	67	99	100	99	100	100	n/a	n/a	n/a	n/a	n/a	n/a
	N of PBT	777	741	n/r	n/r	n/r	n/r	n/r	n/a	n/a	n/a	n/a	n/a	n/a
	% of PBT	10.7	33	n/r	n/r	n/r	n/r	n/r	n/a	n/a	n/a	n/a	n/a	n/a
NJ	% of All Data	53.6	7	6	5	6	7	6	9	7	2	n/a	n/a	n/a
	N of Students	12,600	1,525	1,374	1,267	1,407	1,549	1,425	2,141	1,534	378	n/a	n/a	n/a
	N of CBT	12,510	1,519	1,368	1,251	1,398	1,538	1,417	2,115	1,527	377	n/a	n/a	n/a
	% of CBT	99.3	100	100	99	99	99	99	99	100	100	n/a	n/a	n/a
	N of PBT	90	n/r	n/r	n/r	n/r	n/r	n/r	26	n/r	n/r	n/a	n/a	n/a
	% of PBT	0.7	n/r	n/r	n/r	n/r	n/r	n/r	1	n/r	n/r	n/a	n/a	n/a

Table A.11.3 All Spanish-Language Mathematics Students, by State, and Grade

		Mathematics												
State	Category	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	A1	GO	A2	M1	M2	M3
NM	% of All Data	15.6	4	4	1	1	1	1	1	1	1	n/r	n/r	n/r
	N of Students	3,659	1,043	840	277	293	258	255	285	194	180	n/r	n/r	n/r
	N of CBT	3,656	1,041	840	277	292	258	255	285	194	180	n/r	n/r	n/r
	% of CBT	99.9	100	100	100	100	100	100	100	100	100	n/r	n/r	n/r
	N of PBT	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r
	% of PBT	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r

Note: BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico; A1=Algebra I, GO=Geometry, A2=Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II, M3=Integrated Mathematics III. 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 All States Combined: ELA/L Students, by Grade, Mode, and Gender

			Female		Male	
Grade	Mode	Valid Cases	N	%	N	%
3	All	256,870	125,311	48.8	131,559	51.2
	CBT	224,957	109,598	48.7	115,359	51.3
	PBT	31,913	15,713	49.2	16,200	50.8
4	All	265,169	130,022	49.0	135,147	51.0
	CBT	259,642	127,379	49.1	132,263	50.9
	PBT	5,527	2,643	47.8	2,884	52.2
5	All	271,778	133,464	49.1	138,314	50.9
	CBT	267,807	131,529	49.1	136,278	50.9
	PBT	3,971	1,935	48.7	2,036	51.3
6	All	275,277	134,829	49.0	140,448	51.0
	CBT	271,346	132,988	49.0	138,358	51.0
	PBT	3,931	1,841	46.8	2,090	53.2
7	All	269,386	132,281	49.1	137,105	50.9
	CBT	265,686	130,478	49.1	135,208	50.9
	PBT	3,700	1,803	48.7	1,897	51.3
8	All	266,251	129,790	48.7	136,461	51.3
	CBT	263,370	128,462	48.8	134,908	51.2
	PBT	2,881	1,328	46.1	1,553	53.9
9	All	121,619	59,248	48.7	62,371	51.3
	CBT	121,061	58,987	48.7	62,074	51.3
	PBT	558	261	46.8	297	53.2
10	All	118,322	58,513	49.5	59,809	50.5
	CBT	117,751	58,262	49.5	59,489	50.5
	PBT	571	251	44.0	320	56.0
11	All	34,610	16,651	48.1	17,959	51.9
	CBT	34,035	16,401	48.2	17,634	51.8
	PBT	575	250	43.5	325	56.5

Note: BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test;

Table A.11.5 All States Combined: Mathematics Students, by Grade, Mode, and Gender

			Female		Male	
Grade	Mode	Valid Cases	N	%	N	%
3	All	258,807	126,222	48.8	132,585	51.2
	CBT	226,933	110,523	48.7	116,410	51.3
	PBT	31,874	15,699	49.3	16,175	50.7
4	All	266,629	130,725	49.0	135,904	51.0
	CBT	261,092	128,074	49.1	133,018	50.9
	PBT	5,537	2,651	47.9	2,886	52.1
5	All	272,714	133,881	49.1	138,833	50.9
	CBT	268,724	131,932	49.1	136,792	50.9
	PBT	3,990	1,949	48.8	2,041	51.2
6	All	275,732	135,084	49.0	140,648	51.0
	CBT	271,798	133,243	49.0	138,555	51.0
	PBT	3,934	1,841	46.8	2,093	53.2
7	All	264,960	130,334	49.2	134,626	50.8
	CBT	261,252	128,524	49.2	132,728	50.8
	PBT	3,708	1,810	48.8	1,898	51.2
8	All	225,726	109,096	48.3	116,630	51.7
	CBT	222,869	107,777	48.4	115,092	51.6
	PBT	2,857	1,319	46.2	1,538	53.8
A1	All	134,107	65,087	48.5	69,020	51.5
	CBT	133,482	64,799	48.5	68,683	51.5
	PBT	625	288	46.1	337	53.9
GO	All	105,010	52,124	49.6	52,886	50.4
	CBT	104,444	51,877	49.7	52,567	50.3
	PBT	566	247	43.6	319	56.4
A2	All	66,789	34,144	51.1	32,645	48.9
	CBT	66,317	33,923	51.2	32,394	48.8
	PBT	472	221	46.8	251	53.2
M1	All	673	321	47.7	352	52.3
	CBT	672	320	47.6	352	52.4
	PBT	n/r	n/r	n/r	n/r	n/r
M2	All	541	245	45.3	296	54.7
	CBT	540	244	45.2	296	54.8
	PBT	n/r	n/r	n/r	n/r	n/r
M3	All	201	108	53.7	93	46.3
	CBT	201	108	53.7	93	46.3
	PBT	n/r	n/r	n/r	n/r	n/r

Note: A1=Algebra I, GO=Geometry, A2=Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II, M3=Integrated Mathematics III. CBT=computer-based test; PBT=paper-based test; n/a=not applicable. and n/r=not reported due to n<20.

Table A.11.6 All States Combined: Spanish-Language Mathematics Students, by Grade, Mode, and Gender

			Female		Male	
Grade	Mode	Valid Cases	N	%	N	%
3	All	4,812	2,361	49.1	2,451	50.9
	CBT	4,063	1,981	48.8	2,082	51.2
	PBT	749	380	50.7	369	49.3
4	All	3,691	1,775	48.1	1,916	51.9
	CBT	3,668	1,766	48.1	1,902	51.9
	PBT	23	n/r	n/r	n/r	n/r
5	All	3,270	1,562	47.8	1,708	52.2
	CBT	3,247	1,552	47.8	1,695	52.2
	PBT	23	n/r	n/r	n/r	n/r
6	All	2,642	1,213	45.9	1,429	54.1
	CBT	2,623	1,203	45.9	1,420	54.1
	PBT	n/r	n/r	n/r	n/r	n/r
7	All	2,255	1,065	47.2	1,190	52.8
	CBT	2,243	1,062	47.3	1,181	52.7
	PBT	n/r	n/r	n/r	n/r	n/r
8	All	2,118	959	45.3	1,159	54.7
	CBT	2,108	953	45.2	1,155	54.8
	PBT	n/r	n/r	n/r	n/r	n/r
A1	All	2,426	1,079	44.5	1,347	55.5
	CBT	2,400	1,068	44.5	1,332	55.5
	PBT	26	n/r	n/r	n/r	n/r
GO	All	1,728	854	49.4	874	50.6
	CBT	1,721	851	49.4	870	50.6
	PBT	n/r	n/r	n/r	n/r	n/r
A2	All	558	279	50.0	279	50.0
	CBT	557	278	49.9	279	50.1
	PBT	n/r	n/r	n/r	n/r	n/r

Note: A1=Algebra I, GO=Geometry, A2=Algebra II. CBT=computer-based test; PBT=paper-based test; n/a=not applicable. and n/r=not reported due to n<20.

Integrated Mathematics I, Integrated Mathematics II, Integrated Mathematics III were not administered.

Table A.11.7 Demographic Information: Grade 3 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.4	43.6	52.1	39.0	77.4
Student with Disabilities	16.8	12.8	15.4	19.5	14.3
English learner	14.9	34.9	18.3	8.9	18.6
Male	51.2	52.1	51.3	51.0	51.2
Female	48.8	47.9	48.7	49.0	48.8
American Indian/Alaska Native	1.6	99.1	0.2	0.1	11.0
Asian	7.0	n/r	5.3	11.0	1.3
Black/African American	15.1	n/r	17.5	15.0	2.1
Hispanic/Latino	30.3	n/r	26.1	29.9	59.7
White/Caucasian	42.5	n/r	46.8	41.1	24.8
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.2	n/r	4.0	2.7	1.0
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.8 Demographic Information: Grade 4 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.5	43.9	52.2	38.9	77.5
Student with Disabilities	17.5	12.9	15.8	20.5	15.9
English learner	14.3	33.8	18.3	7.1	19.3
Male	51.0	51.0	50.9	51.1	50.5
Female	49.0	49.0	49.1	48.9	49.5
American Indian/Alaska Native	1.6	99.2	0.2	0.1	10.6
Asian	6.8	n/r	5.2	10.7	1.1
Black/African American	14.8	n/r	16.8	15.4	2.2
Hispanic/Latino	31.1	n/r	27.0	29.7	61.7
White/Caucasian	42.3	n/r	46.7	41.4	23.3
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.1	n/r	3.9	2.5	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.9 Demographic Information: Grade 5 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.4	43.9	52.3	38.2	77.9
Student with Disabilities	17.8	13.2	15.9	20.7	17.3
English learner	11.4	35.2	14.3	4.8	19.1
Male	50.9	47.9	50.8	51.0	51.2
Female	49.1	52.1	49.2	49.0	48.8
American Indian/Alaska Native	1.7	99.5	0.3	0.1	10.1
Asian	6.8	n/r	5.1	10.7	1.1
Black/African American	14.8	n/r	16.8	15.4	2.1
Hispanic/Latino	31.1	n/r	27.3	29.1	62.8
White/Caucasian	42.5	n/r	46.6	42.2	22.6
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.0	n/r	3.8	2.4	1.1
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.10 Demographic Information: Grade 6 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	48.9	43.6	52.1	37.8	75.3
Student with Disabilities	17.6	14.6	15.8	20.4	16.8
English learner	7.9	36.0	9.5	3.5	15.3
Male	51.0	49.8	50.9	51.1	51.4
Female	49.0	50.2	49.1	48.9	48.6
American Indian/Alaska Native	1.6	98.9	0.2	0.1	10.4
Asian	6.5	n/r	4.9	10.3	1.2
Black/African American	14.8	n/r	16.9	15.2	2.1
Hispanic/Latino	31.2	n/r	27.6	29.1	62.4
White/Caucasian	42.8	n/r	46.6	42.8	22.8
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	0.2	0.2
Two or More Races Reported	2.9	n/r	3.7	2.3	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.11 Demographic Information: Grade 7 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	47.3	51.1	50.4	36.3	73.8
Student with Disabilities	17.3	13.1	15.8	19.7	16.2
English learner	6.6	31.9	7.4	3.4	13.3
Male	50.9	47.0	50.9	51.0	50.6
Female	49.1	53.0	49.1	49.0	49.4
American Indian/Alaska Native	1.6	99.5	0.2	0.1	10.8
Asian	6.7	n/r	5.0	10.4	1.4
Black/African American	14.5	n/r	16.4	14.9	2.0
Hispanic/Latino	30.6	n/r	27.3	28.3	61.3
White/Caucasian	43.9	n/r	47.5	44.2	23.4
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	0.2	0.2
Two or More Races Reported	2.7	n/r	3.5	2.0	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.12 Demographic Information: Grade 8 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	46.0	50.0	49.2	34.7	73.1
Student with Disabilities	17.4	13.6	15.8	20.1	15.9
English learner	6.0	30.2	6.6	3.4	11.5
Male	51.3	48.0	51.4	51.2	50.6
Female	48.7	52.0	48.6	48.8	49.4
American Indian/Alaska Native	1.5	99.0	0.2	0.1	10.2
Asian	6.7	n/r	5.2	10.4	1.3
Black/African American	14.1	n/r	15.9	14.7	2.0
Hispanic/Latino	30.4	n/r	27.1	27.7	62.1
White/Caucasian	44.6	n/r	48.1	45.1	23.5
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.1
Two or More Races Reported	2.5	n/r	3.3	1.8	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.13 Demographic Information: Grade 9 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	39.5	62.1	n/a	32.8	67.4
Student with Disabilities	18.5	13.7	n/a	19.3	15.1
English learner	6.2	63.2	n/a	4.4	13.0
Male	51.3	44.2	n/a	51.3	51.1
Female	48.7	55.8	n/a	48.7	48.9
American Indian/Alaska Native	2.4	100.0	n/a	0.1	11.1
Asian	8.6	n/r	n/a	10.4	1.2
Black/African American	11.7	n/r	n/a	14.1	2.2
Hispanic/Latino	34.6	n/r	n/a	28.1	61.8
White/Caucasian	41.0	n/r	n/a	45.5	22.8
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.2	0.2
Two or More Races Reported	1.4	n/r	n/a	1.6	0.9
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.14 Demographic Information: Grade 10 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	37.1	66.3	n/a	30.7	63.9
Student with Disabilities	17.8	16.6	n/a	18.9	13.3
English learner	5.6	66.8	n/a	4.2	11.3
Male	50.5	48.7	n/a	50.7	49.9
Female	49.5	51.3	n/a	49.3	50.1
American Indian/Alaska Native	2.3	100.0	n/a	0.1	10.6
Asian	8.8	n/r	n/a	10.5	1.5
Black/African American	11.9	n/r	n/a	14.2	2.0
Hispanic/Latino	32.9	n/r	n/a	26.4	61.2
White/Caucasian	42.5	n/r	n/a	46.9	23.6
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.2	0.2
Two or More Races Reported	1.4	n/r	n/a	1.5	0.8
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.15 Demographic Information: Grade 11 ELA/L Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	59.2	66.0	n/a	52.3	63.3
Student with Disabilities	17.4	14.9	n/a	25.4	12.6
English learner	11.7	33.2	n/a	11.1	11.4
Male	51.9	48.5	n/a	55.6	49.7
Female	48.1	51.5	n/a	44.4	50.3
American Indian/Alaska Native	8.5	99.7	n/a	n/r	11.1
Asian	2.3	n/r	n/a	3.6	1.5
Black/African American	13.3	n/r	n/a	32.1	2.2
Hispanic/Latino	52.0	n/r	n/a	40.2	60.7
White/Caucasian	22.9	n/r	n/a	22.8	23.6
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.2	0.1
Two or More Races Reported	0.8	n/r	n/a	0.9	0.8
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.16 Demographic Information: Grade 3 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.6	43.5	52.1	39.0	78.3
Student with Disabilities	16.7	12.8	15.3	19.3	14.1
English learner	15.7	34.8	18.4	10.0	22.2
Male	51.2	52.3	51.3	51.1	51.3
Female	48.8	47.7	48.7	48.9	48.7
American Indian/Alaska Native	1.6	99.1	0.2	0.1	10.5
Asian	7.0	n/r	5.3	11.0	1.2
Black/African American	15.0	n/r	17.5	14.9	2.0
Hispanic/Latino	30.8	n/r	26.1	30.4	61.5
White/Caucasian	42.2	n/r	46.8	40.8	23.7
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.2	n/r	4.0	2.7	0.9
Unknown	0.0	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.17 Demographic Information: Grade 4 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.6	43.9	52.1	39.0	78.2
Student with Disabilities	17.4	12.8	15.7	20.2	15.6
English learner	14.9	33.6	18.4	8.0	21.9
Male	51.0	51.1	50.9	51.2	50.6
Female	49.0	48.9	49.1	48.8	49.4
American Indian/Alaska Native	1.6	99.2	0.2	0.1	10.2
Asian	6.8	n/r	5.2	10.7	1.1
Black/African American	14.7	n/r	16.8	15.2	2.1
Hispanic/Latino	31.4	n/r	27.0	30.1	62.8
White/Caucasian	42.1	n/r	46.7	41.1	22.6
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.1	n/r	3.9	2.5	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.18 Demographic Information: Grade 5 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.5	44.0	52.3	38.3	78.0
Student with Disabilities	17.7	13.2	15.9	20.4	17.2
English learner	11.8	35.2	14.3	5.8	20.0
Male	50.9	47.8	50.8	51.0	51.2
Female	49.1	52.2	49.2	49.0	48.8
American Indian/Alaska Native	1.7	99.5	0.3	0.1	10.0
Asian	6.8	n/r	5.1	10.7	1.1
Black/African American	14.7	n/r	16.7	15.2	2.1
Hispanic/Latino	31.4	n/r	27.3	29.5	63.1
White/Caucasian	42.3	n/r	46.6	41.9	22.4
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	3.0	n/r	3.8	2.3	1.1
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.19 Demographic Information: Grade 6 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	48.9	44.3	52.0	38.0	75.7
Student with Disabilities	17.5	14.8	15.8	20.2	16.8
English learner	8.4	36.5	9.5	4.5	16.3
Male	51.0	49.7	50.9	51.0	51.3
Female	49.0	50.3	49.1	49.0	48.7
American Indian/Alaska Native	1.6	98.9	0.2	0.1	10.4
Asian	6.6	n/r	4.9	10.3	1.2
Black/African American	14.7	n/r	16.8	15.1	2.1
Hispanic/Latino	31.4	n/r	27.6	29.6	62.8
White/Caucasian	42.7	n/r	46.7	42.4	22.4
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	0.2	0.2
Two or More Races Reported	2.9	n/r	3.7	2.2	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.20 Demographic Information: Grade 7 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	48.0	51.0	50.4	37.7	74.2
Student with Disabilities	17.4	13.1	15.7	20.3	16.2
English learner	7.1	31.8	7.4	4.6	14.3
Male	50.8	47.0	50.9	50.8	50.5
Female	49.2	53.0	49.1	49.2	49.5
American Indian/Alaska Native	1.6	99.5	0.2	0.1	10.7
Asian	6.0	n/r	5.0	8.8	1.3
Black/African American	14.5	n/r	16.3	15.3	2.0
Hispanic/Latino	31.2	n/r	27.3	29.6	61.8
White/Caucasian	43.9	n/r	47.6	44.1	23.0
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	0.2
Two or More Races Reported	2.7	n/r	3.5	1.9	0.9
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.21 Demographic Information: Grade 8 Mathematics Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	49.9	49.4	49.3	42.2	78.3
Student with Disabilities	19.1	13.7	15.8	26.8	18.6
English learner	7.3	30.4	6.7	6.0	14.5
Male	51.7	48.3	51.4	52.4	51.1
Female	48.3	51.7	48.6	47.6	48.9
American Indian/Alaska Native	1.6	99.0	0.2	0.1	11.1
Asian	4.7	n/r	5.2	4.9	0.8
Black/African American	15.1	n/r	15.8	17.9	2.0
Hispanic/Latino	32.0	n/r	27.2	33.1	65.0
White/Caucasian	43.8	n/r	48.2	42.2	20.3
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	0.2	n/r
Two or More Races Reported	2.6	n/r	3.3	1.6	0.8
Unknown	n/r	n/r	n/r	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/r=not reported due to n<20.

Table A.11.22 Demographic Information: Algebra I Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	40.4	62.6	n/a	33.8	68.6
Student with Disabilities	18.3	13.0	n/a	18.9	15.4
English learner	7.2	56.1	n/a	5.4	14.2
Male	51.5	43.5	n/a	51.5	51.4
Female	48.5	56.5	n/a	48.5	48.6
American Indian/Alaska Native	2.3	100.0	n/a	0.1	10.9
Asian	8.5	n/r	n/a	10.2	1.2
Black/African American	12.3	n/r	n/a	14.6	2.2
Hispanic/Latino	35.1	n/r	n/a	28.9	61.9
White/Caucasian	40.1	n/r	n/a	44.2	22.7
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.2	0.2
Two or More Races Reported	1.6	n/r	n/a	1.7	0.9
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.23 Demographic Information: Geometry Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	37.9	73.3	n/a	31.2	64.2
Student with Disabilities	16.6	22.8	n/a	17.4	13.4
English learner	6.2	64.7	n/a	4.7	11.5
Male	50.4	52.2	n/a	50.5	49.6
Female	49.6	47.8	n/a	49.5	50.4
American Indian/Alaska Native	2.6	100.0	n/a	0.1	11.5
Asian	8.3	n/r	n/a	10.0	1.6
Black/African American	11.7	n/r	n/a	14.1	1.9
Hispanic/Latino	34.0	n/r	n/a	27.5	60.6
White/Caucasian	41.8	n/r	n/a	46.5	23.5
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.2	0.2
Two or More Races Reported	1.4	n/r	n/a	1.5	0.8
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.24 Demographic Information: Algebra II Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	37.6	67.7	n/a	26.2	61.8
Student with Disabilities	10.8	12.2	n/a	11.1	10.1
English learner	5.8	33.1	n/a	3.1	11.1
Male	48.9	45.8	n/a	48.7	49.3
Female	51.1	54.2	n/a	51.3	50.7
American Indian/Alaska Native	4.0	99.6	n/a	0.1	10.3
Asian	12.1	n/r	n/a	17.0	1.7
Black/African American	10.3	n/r	n/a	14.1	2.2
Hispanic/Latino	34.0	n/r	n/a	21.9	61.3
White/Caucasian	38.0	n/r	n/a	45.0	23.6
Native Hawaiian/Pacific Islander	0.2	n/r	n/a	0.3	0.1
Two or More Races Reported	1.3	n/r	n/a	1.6	0.8
Unknown	n/r	n/r	n/a	n/r	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.25 Demographic Information: Integrated Mathematics I Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	66.6	n/r	n/a	n/a	67.2
Student with Disabilities	20.8	n/r	n/a	n/a	20.9
English learner	16.6	n/r	n/a	n/a	16.7
Male	52.3	n/r	n/a	n/a	52.3
Female	47.7	n/r	n/a	n/a	47.7
American Indian/Alaska Native	4.2	n/r	n/a	n/a	3.0
Asian	n/r	n/r	n/a	n/a	n/r
Black/African American	4.0	n/r	n/a	n/a	4.1
Hispanic/Latino	61.7	n/r	n/a	n/a	62.4
White/Caucasian	28.5	n/r	n/a	n/a	28.9
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a	n/r
Two or More Races Reported	n/r	n/r	n/a	n/a	n/r
Unknown	n/r	n/r	n/a	n/a	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.26 Demographic Information: Integrated Mathematics II Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	64.9	n/r	n/a	n/a	65.4
Student with Disabilities	17.7	n/r	n/a	n/a	17.9
English learner	22.2	n/r	n/a	n/a	22.3
Male	54.7	n/r	n/a	n/a	54.9
Female	45.3	n/r	n/a	n/a	45.1
American Indian/Alaska Native	n/r	n/r	n/a	n/a	n/r
Asian	n/r	n/r	n/a	n/a	n/r
Black/African American	3.7	n/r	n/a	n/a	3.7
Hispanic/Latino	68.9	n/r	n/a	n/a	69.5
White/Caucasian	22.7	n/r	n/a	n/a	22.9
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a	n/r
Two or More Races Reported	n/r	n/r	n/a	n/a	n/r
Unknown	n/r	n/r	n/a	n/a	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. n/a=not applicable; and n/r=not reported due to n<20.

Table A.11.27 Demographic Information: Integrated Mathematics III Students, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	NJ (%)	NM (%)
Economically Disadvantaged	83.1	n/r	n/a	n/a	83.1
Student with Disabilities	13.4	n/r	n/a	n/a	13.4
English learner	21.9	n/r	n/a	n/a	21.9
Male	46.3	n/r	n/a	n/a	46.3
Female	53.7	n/r	n/a	n/a	53.7
American Indian/Alaska Native	n/r	n/r	n/a	n/a	n/r
Asian	n/r	n/r	n/a	n/a	n/r
Black/African American	n/r	n/r	n/a	n/a	n/r
Hispanic/Latino	76.1	n/r	n/a	n/a	76.1
White/Caucasian	19.4	n/r	n/a	n/a	19.4
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a	n/r
Two or More Races Reported	n/r	n/r	n/a	n/a	n/r
Unknown	n/r	n/r	n/a	n/a	n/r

Note: All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, NJ=New Jersey, and NM=New Mexico. 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.12.7 Form Composition for ELA/L Grade 9

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.12.8 Form Composition for ELA/L Grade 10

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 ELA/L Grade 11

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.12.10 Form Composition for Mathematics 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 and Applications	3	12
TOTAL		33	52

Note: This table is identical to Table 12.3 in Section 12.

Table A.12.11 Form Composition for Mathematics Grade 4

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	17	21
	Additional & Supporting Content	8	9
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
TOTAL		31	52

Table A.12.12 Form Composition for Mathematics Grade 5

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	17	20
	Additional & Supporting Content	8	10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
TOTAL		31	52

Table A.12.13 Form Composition for Mathematics Grade 6

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	15	20
	Additional & Supporting Content	8	10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
TOTAL		29	52

Table A.12.14 Form Composition for Mathematics Grade 7

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	18	20
	Additional & Supporting Content	7	10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
TOTAL		31	52

Table A.12.15 Form Composition for Mathematics Grade 8

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	18	20
	Additional & Supporting Content	6	10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
TOTAL		30	52

Table A.12.16 Form Composition for Algebra I

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	12	17
	Additional & Supporting Content	8	10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		26	52

Table A.12.17 Form Composition for Geometry

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	15	18
	Additional & Supporting Content	9	12
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		30	55

Table A.12.18 Form Composition for Algebra II

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	14	17
	Additional & Supporting Content	9	12
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		29	54

Table A.12.19 Form Composition for Integrated Mathematics I

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	15	19
	Additional & Supporting Content	7	11
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		28	55

Table A.12.20 Form Composition for Integrated Mathematics II

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	13	17
	Additional & Supporting Content	10	13
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		29	55

Table A.12.21 Form Composition for Integrated Mathematics III

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	14	19
	Additional & Supporting Content	9	11
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
TOTAL		29	55

Appendix 12.2: Threshold Scores and Scaling Constants

Table A.12.22 Threshold Scores and Scaling Constants for ELA/L Grades 3 to 8

Assessment	Threshold Cut	Theta	Scale Score	A	B
Grade 3 ELA	Level 2 Cut	-0.9648	700	36.7227	735.4297
	Level 3 Cut	-0.2840	726		
	Level 4 Cut	0.3968	750		
	Level 5 Cut	2.0360	810		
Grade 4 ELA	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 ELA	Level 2 Cut	-1.3411	700	29.4580	739.5050
	Level 3 Cut	-0.4924	726		
	Level 4 Cut	0.3563	750		
	Level 5 Cut	2.0224	799		
Grade 6 ELA	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 ELA	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 ELA	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.12.23 Threshold Scores and Scaling Constants for Mathematics Grades 3 to 8

Assessment	Threshold Cut	Theta	Scale Score	A	B
Grade 3 Mathematics	Level 2 Cut	-1.4141	700	32.1135	745.4119
	Level 3 Cut	-0.6356	727		
	Level 4 Cut	0.1429	750		
	Level 5 Cut	1.3931	790		
Grade 4 Mathematics	Level 2 Cut	-1.3840	700	29.9167	741.4049
	Level 3 Cut	-0.5484	727		
	Level 4 Cut	0.2873	750		
	Level 5 Cut	1.8323	796		
Grade 5 Mathematics	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 Mathematics	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 Mathematics	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 Mathematics	Level 2 Cut	-0.8851	700	32.9505	729.1640
	Level 3 Cut	-0.1264	728		
	Level 4 Cut	0.6323	750		
	Level 5 Cut	2.1896	801		

Table A.12.24 Threshold Scores and Scaling Constants for High School ELA/L

Assessment	Threshold Cut	Theta	Scale Score	A	B
Grade 9 ELA/L	Level 2 Cut	-1.1635	700	34.2174	739.8124
	Level 3 Cut	-0.4329	726		
	Level 4 Cut	0.2977	750		
	Level 5 Cut	1.5065	791		
Grade 10 ELA/L	Level 2 Cut	-0.8909	700	43.1280	738.4223
	Level 3 Cut	-0.3112	725		
	Level 4 Cut	0.2684	750		
	Level 5 Cut	1.2858	794		
Grade 11 ELA/L	Level 2 Cut	-1.1017	700	34.9278	738.4801
	Level 3 Cut	-0.3859	726		
	Level 4 Cut	0.3298	750		
	Level 5 Cut	1.5206	792		

Table A.12.25 Threshold Scores and Scaling Constants for High School Mathematics

Assessment	Threshold Cut	Theta	Scale Score	A	B
Algebra I	Level 2 Cut	-1.1781	700	31.5325	737.1490
	Level 3 Cut	-0.3853	728		
	Level 4 Cut	0.4075	750		
	Level 5 Cut	2.1651	805		
Algebra II	Level 2 Cut	-0.5759	700	37.7676	721.7509
	Level 3 Cut	0.0860	726		
	Level 4 Cut	0.7480	750		
	Level 5 Cut	2.2728	808		
Geometry	Level 2 Cut	-1.3013	700	25.9775	733.8039
	Level 3 Cut	-0.3389	726		
	Level 4 Cut	0.6235	750		
	Level 5 Cut	1.8940	783		
Integrated Mathematics I	Level 2 Cut	-1.0919	700	32.0043	734.9446
	Level 3 Cut	-0.3107	726		
	Level 4 Cut	0.4704	750		
	Level 5 Cut	1.9934	799		
Integrated Mathematics II	Level 2 Cut	-0.9175	700	29.2865	726.8695
	Level 3 Cut	-0.0638	725		
	Level 4 Cut	0.7898	750		
	Level 5 Cut	1.9817	785		
Integrated Mathematics III	Level 2 Cut	-0.7076	700	37.3549	726.4336
	Level 3 Cut	-0.0384	726		
	Level 4 Cut	0.6309	750		
	Level 5 Cut	2.0689	804		

Table A.12.26 Scaling Constants for Reading and Writing Grades 3 to 11

	Reading		Writing	
	AR	BR	AW	BW
Grade 3 ELA/L	14.6891	44.1719	7.3445	32.0859
Grade 4 ELA/L	12.6184	46.4086	6.3093	33.2043
Grade 5 ELA/L	11.7832	45.8019	5.8916	32.9010
Grade 6 ELA/L	11.3264	45.4669	5.6632	32.7335
Grade 7 ELA/L	13.5664	46.9416	6.7832	33.4708
Grade 8 ELA/L	13.6472	47.3732	6.8237	33.6866
Grade 9 ELA/L	13.6870	45.9250	6.8435	32.9625
Grade 10 ELA/L	17.2512	45.3690	8.6256	32.6845
Grade 11 ELA/L	13.9712	45.3920	6.9856	32.6961

Appendix 12.3: IRT Test Characteristic Curves, Information Curves, and CSEM Curves

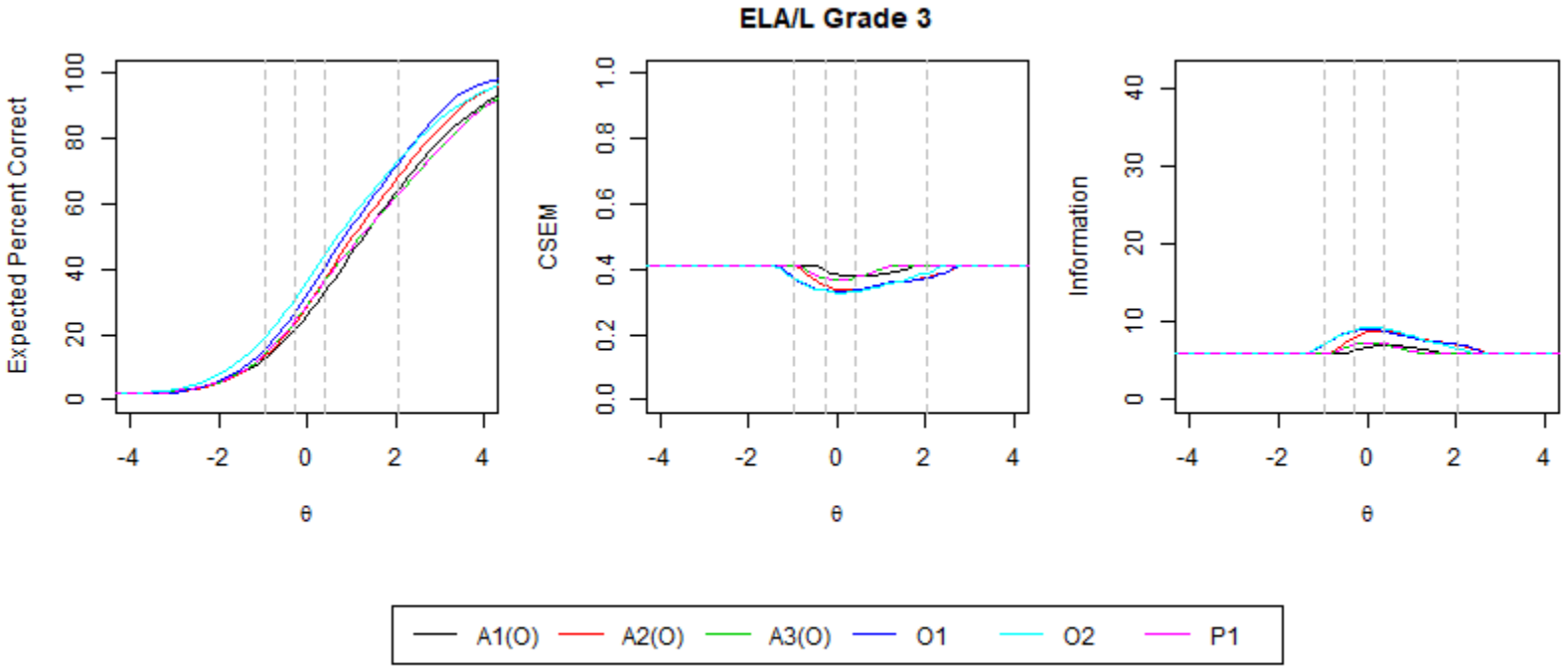

Figure A.12.1 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 3

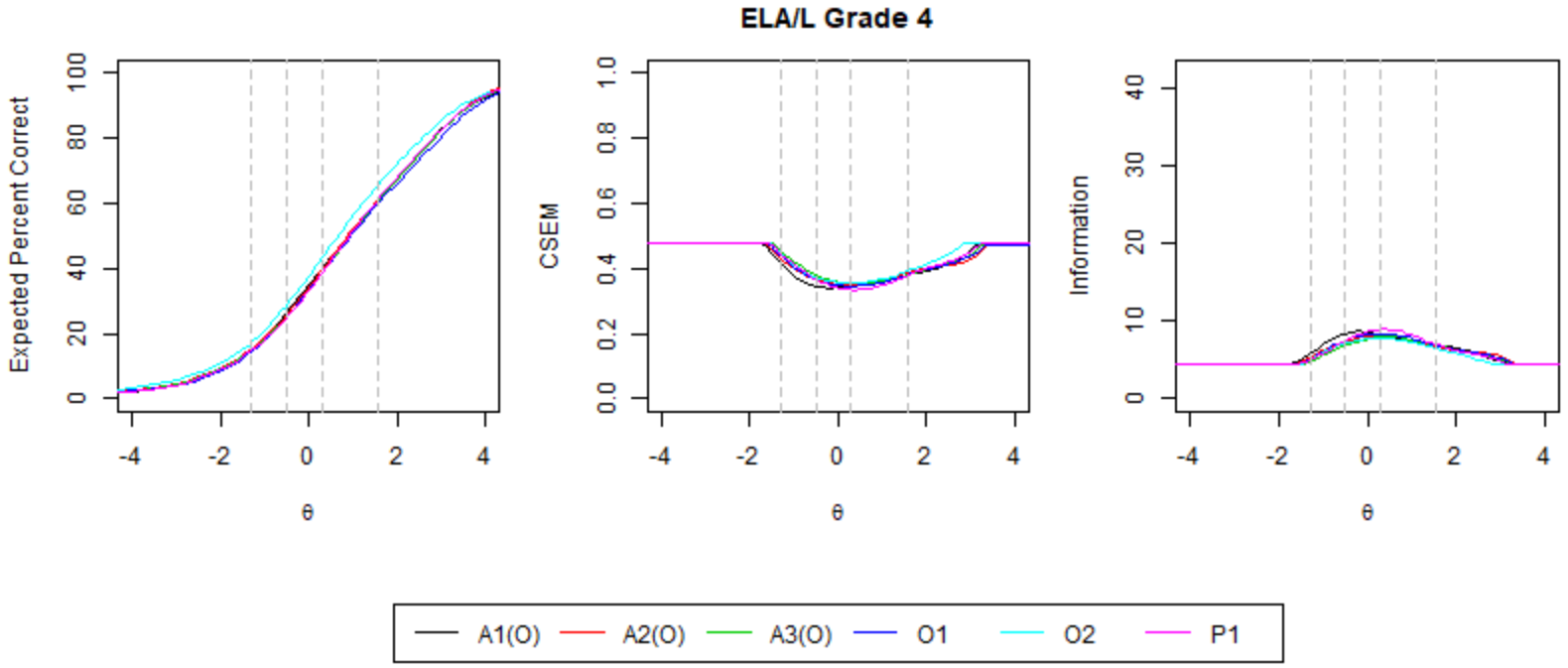

Figure A.12.2 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 4

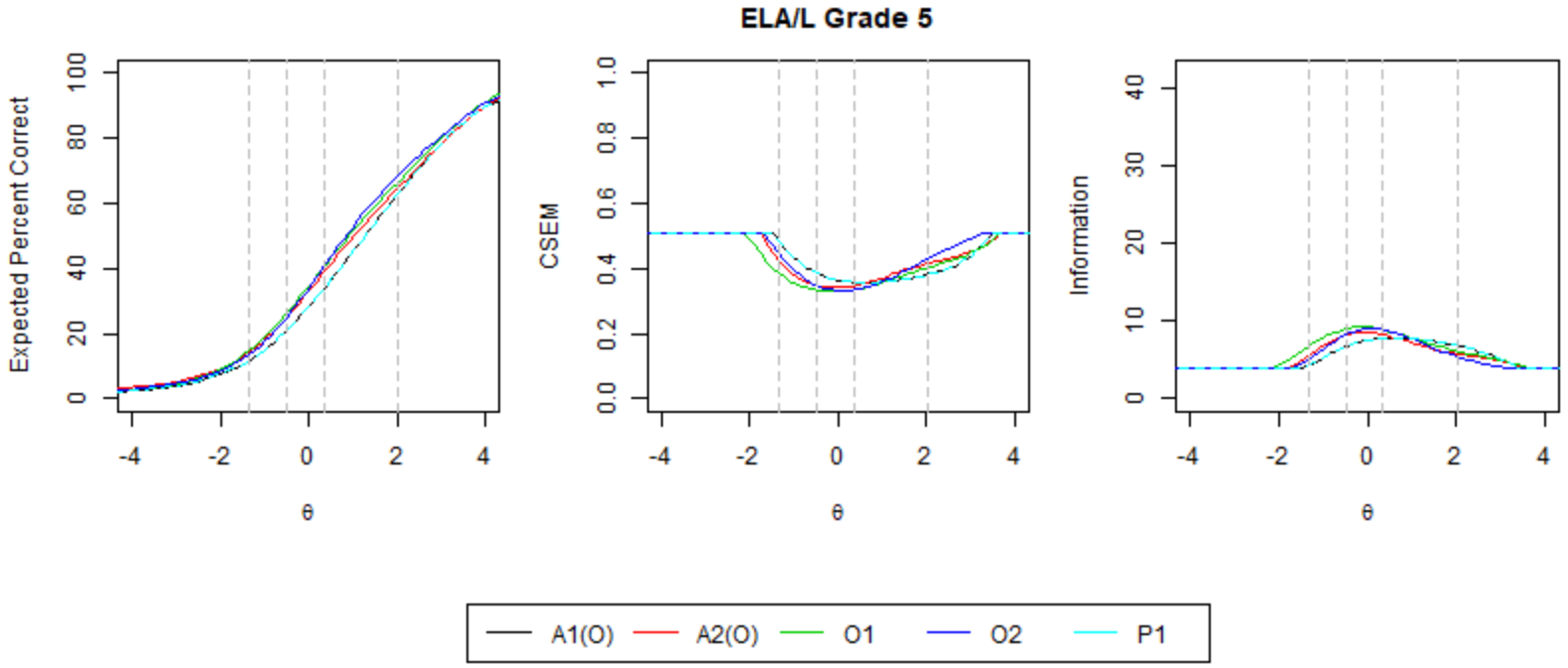

Figure A.12.3 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 5

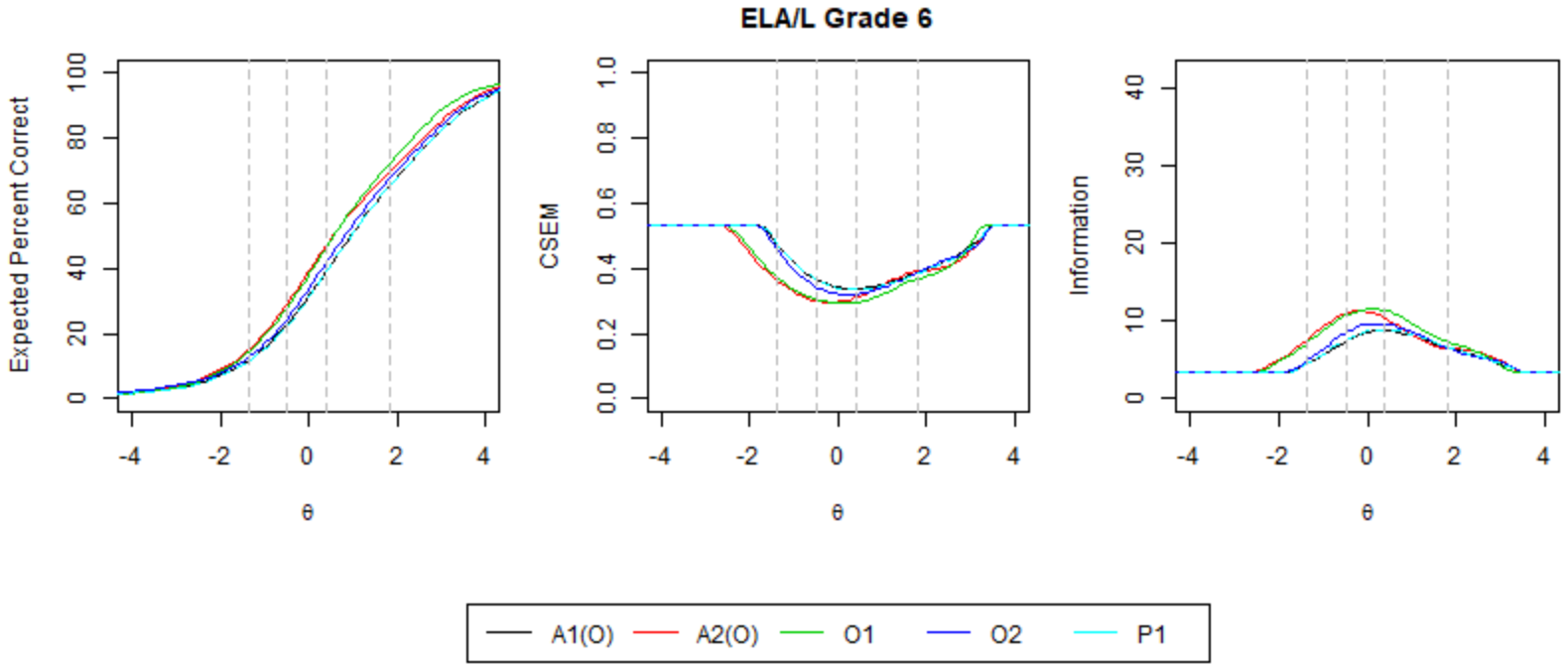

Figure A.12.4 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 6

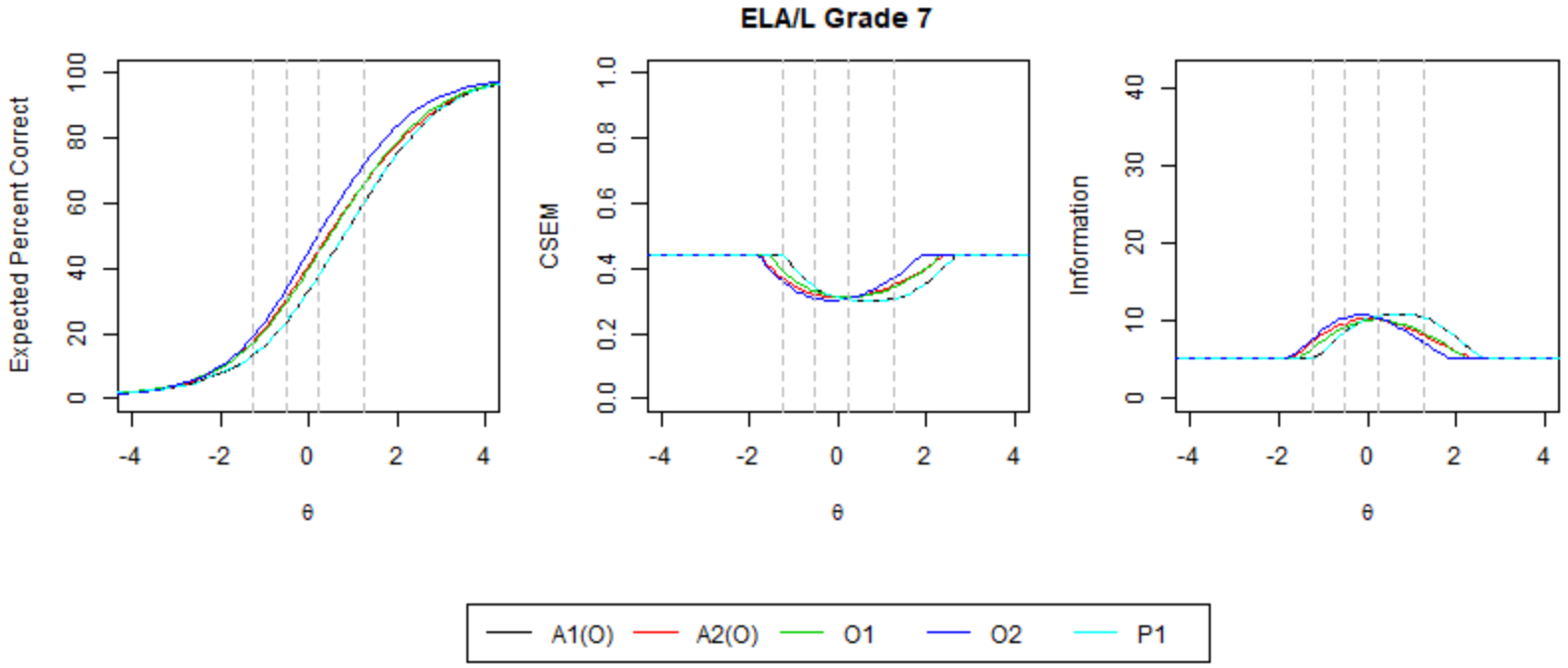

Figure A.12.5 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 7

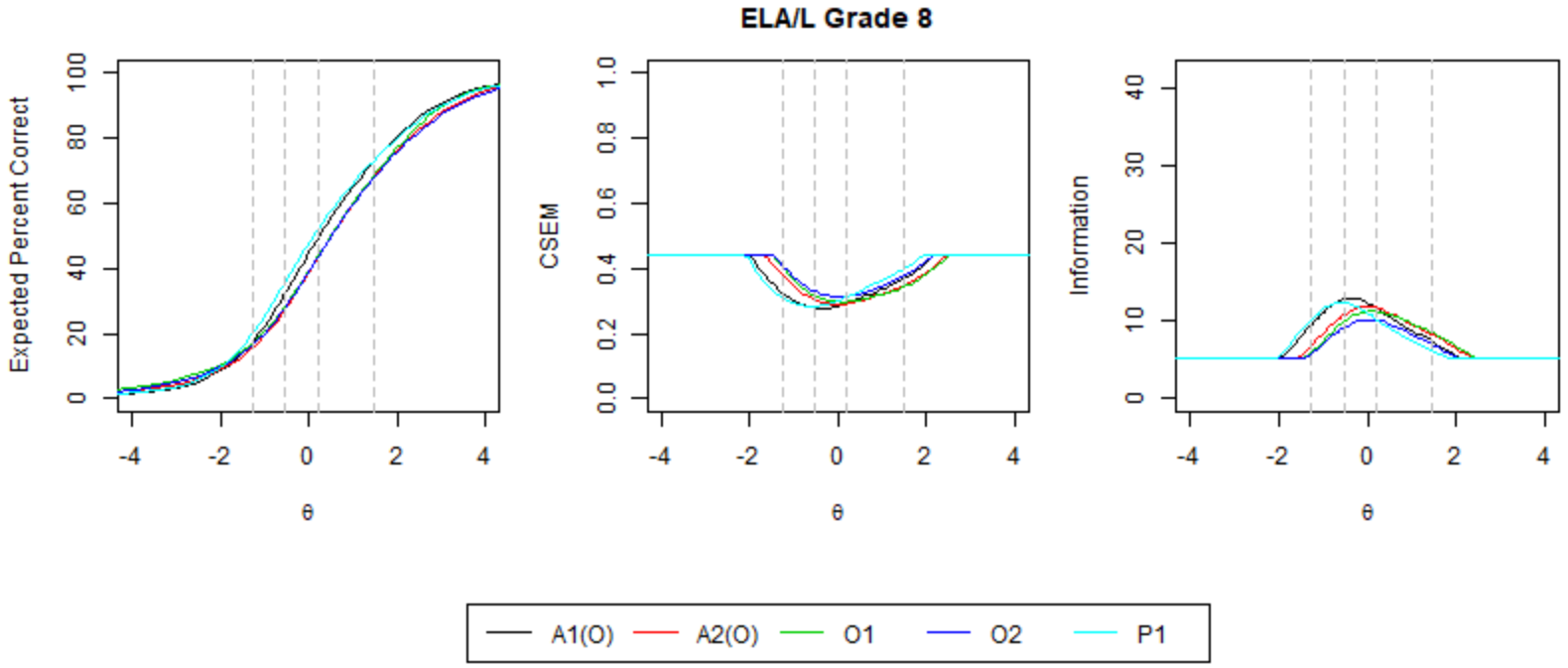

Figure A.12.6 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 8

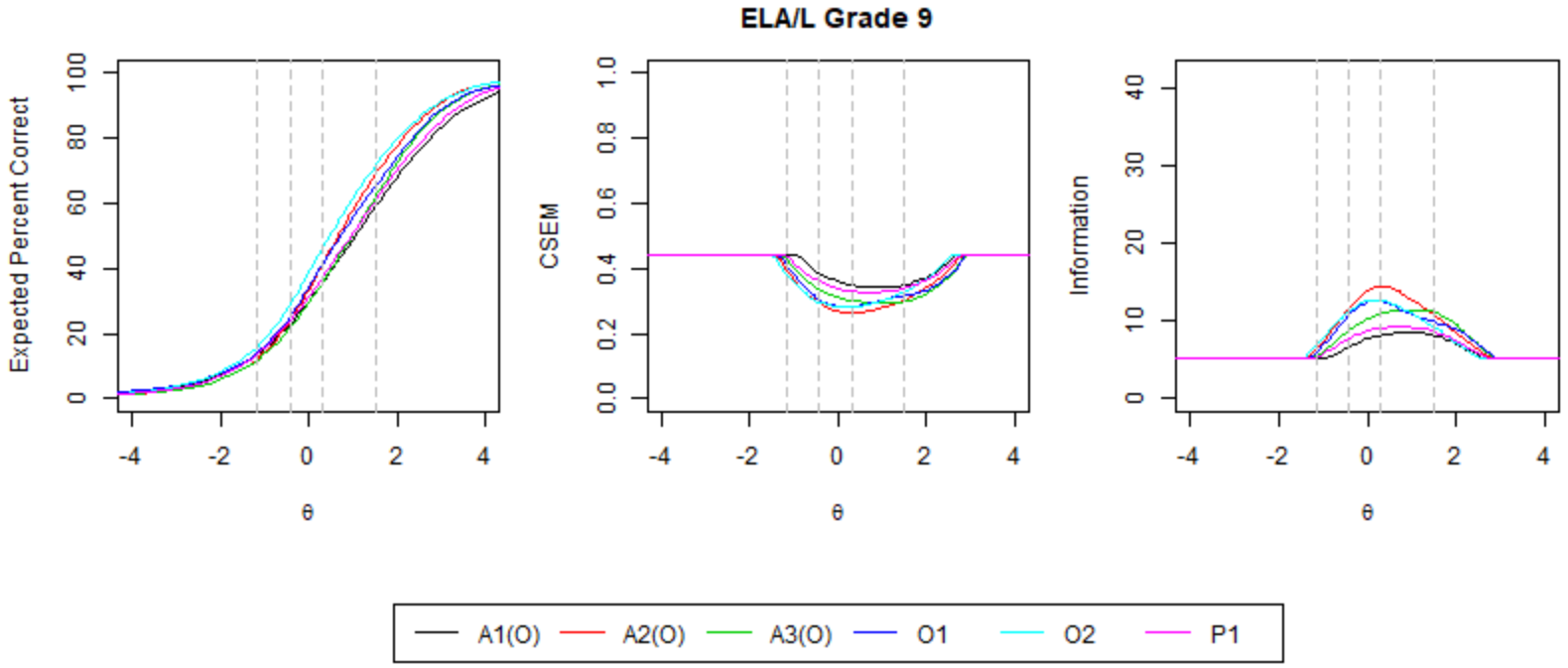

Figure A.12.7 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 9

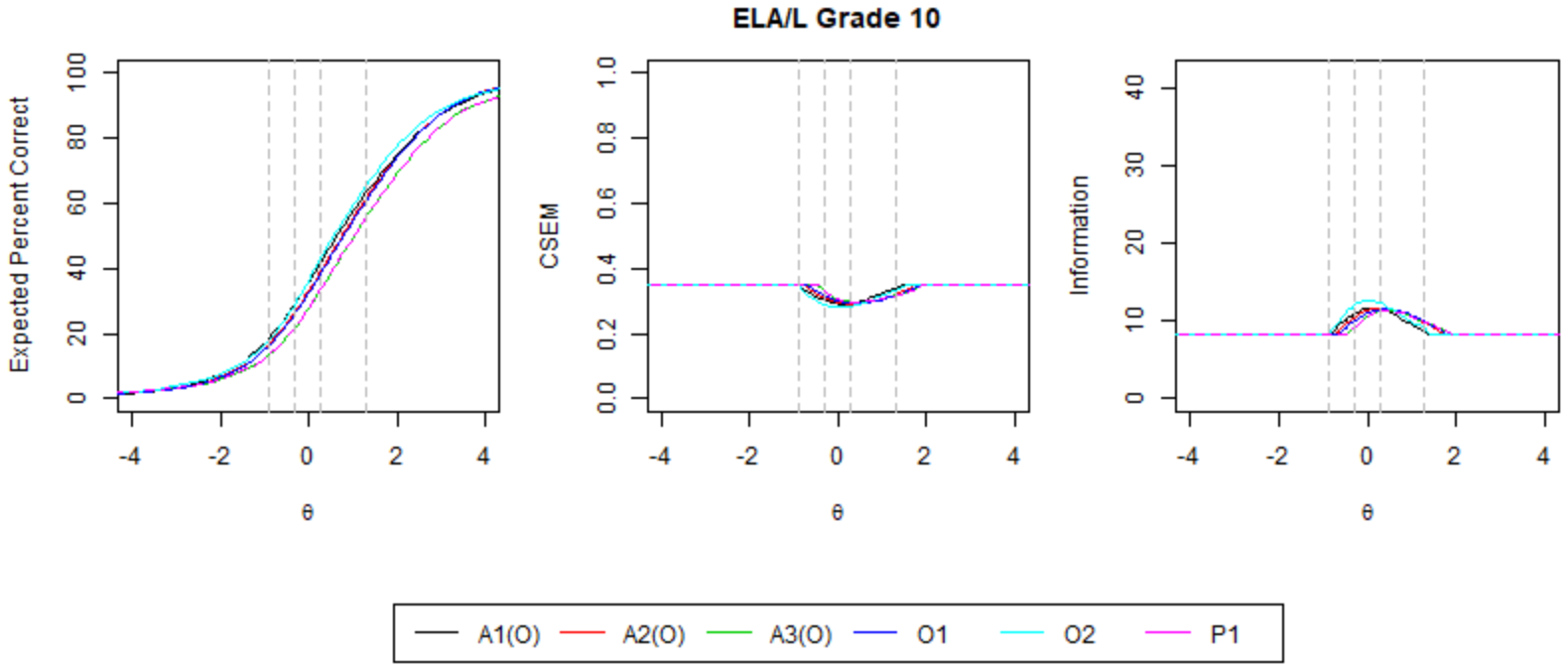

Figure A.12.8 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 10

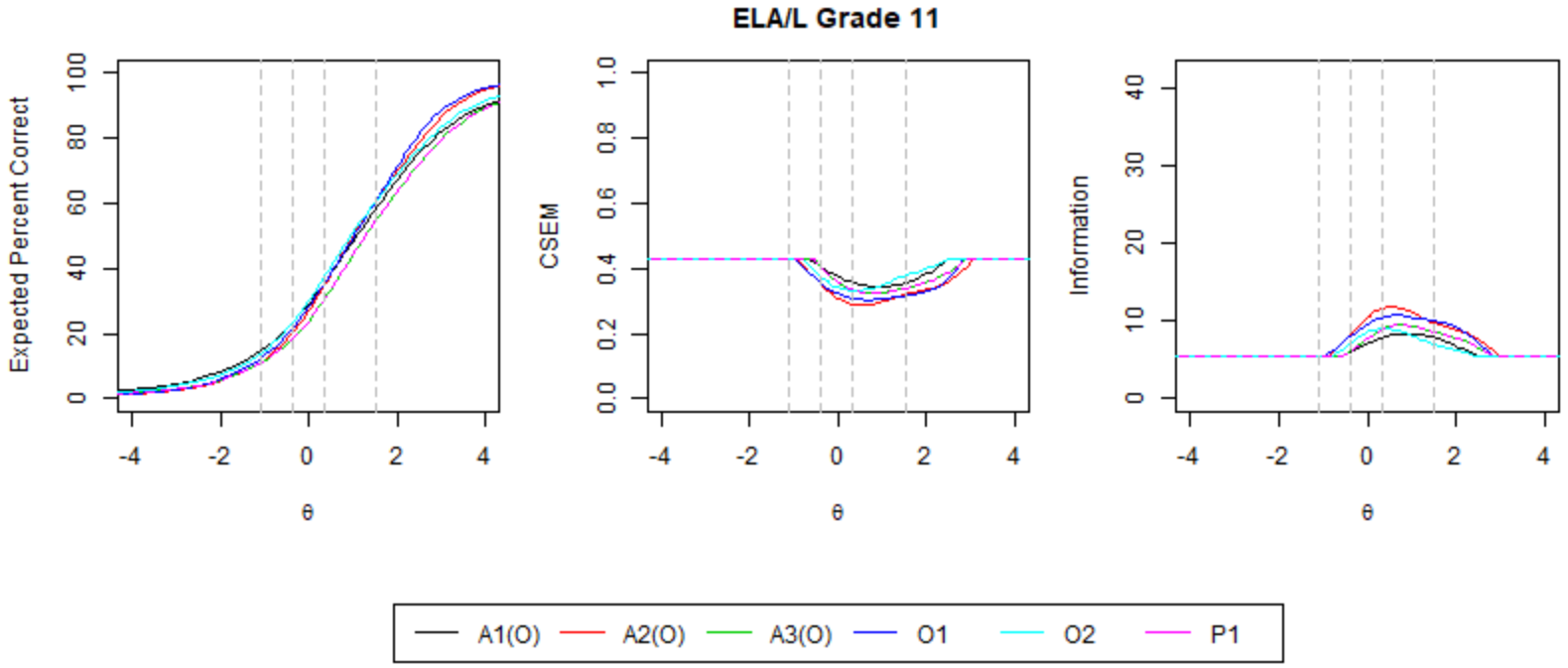

Figure A.12.9 Post-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 11

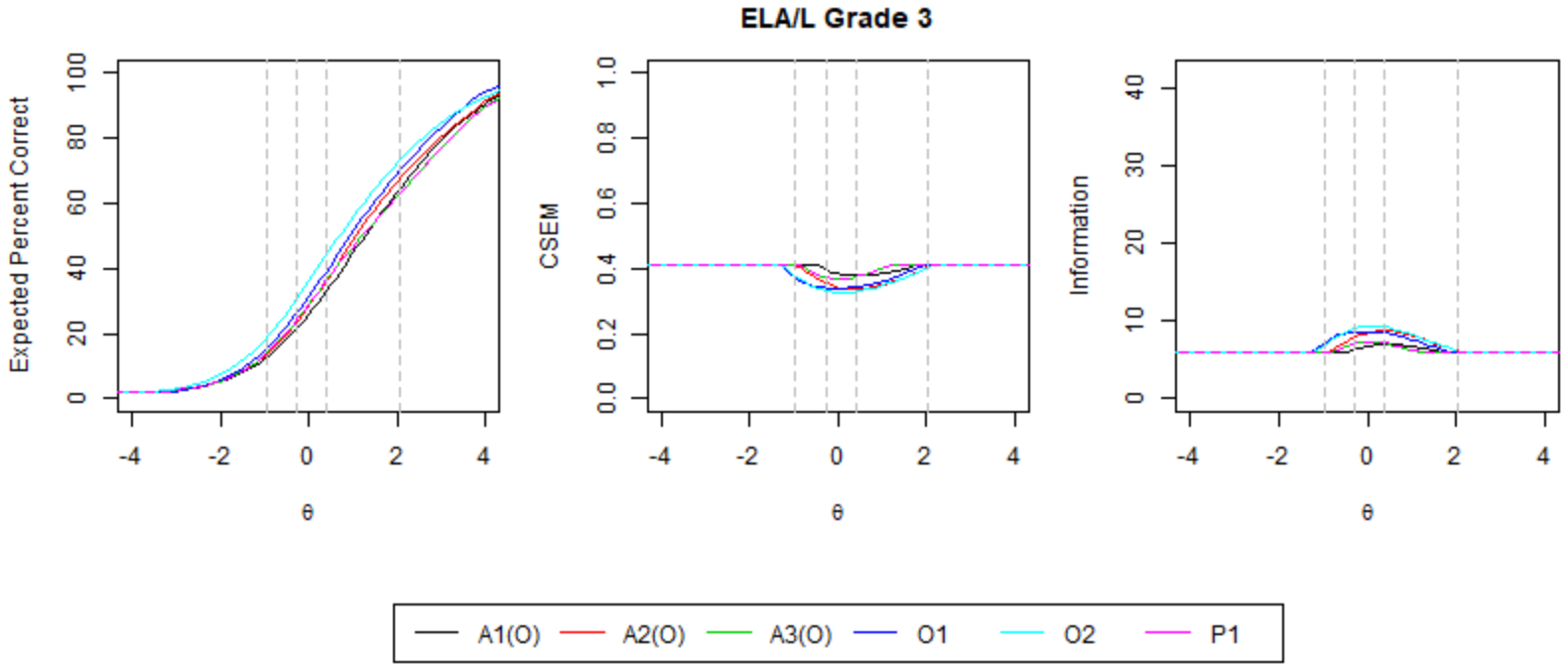

Figure A.12.10 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 3

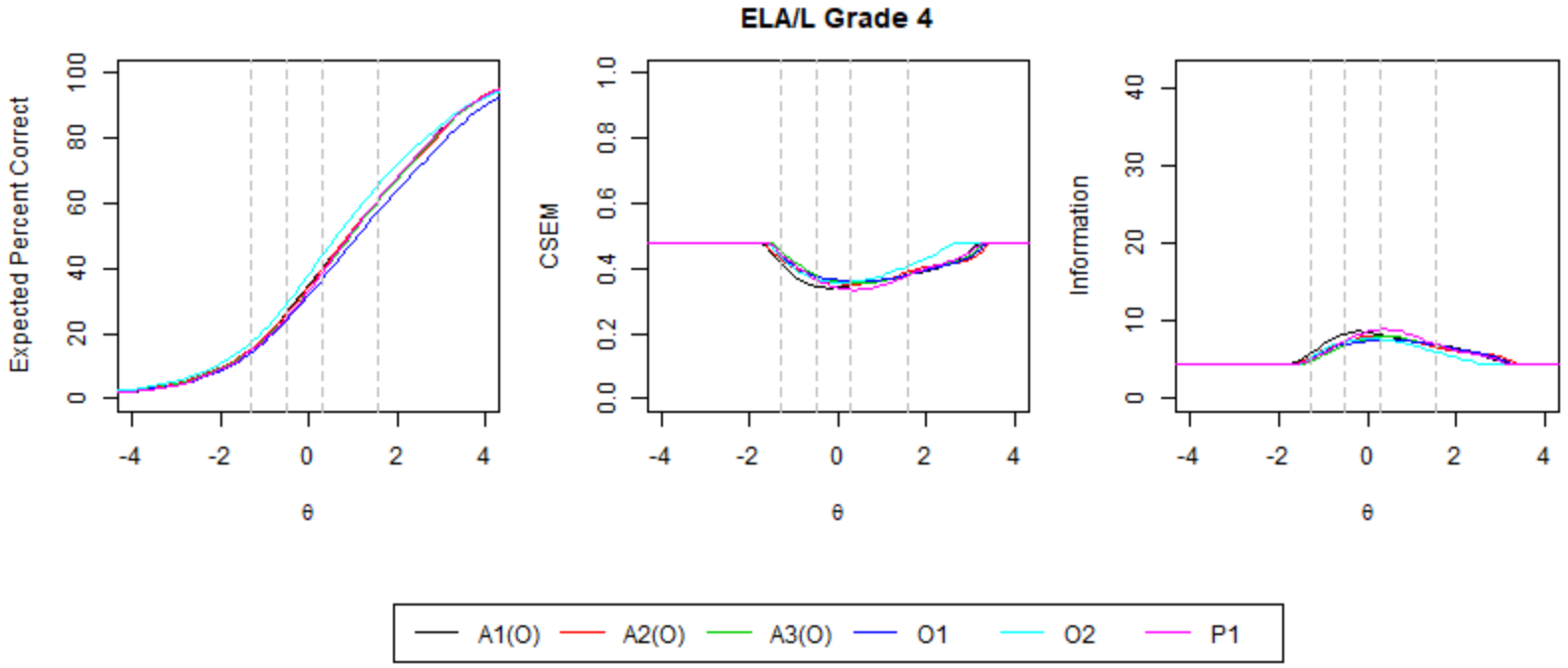

Figure A.12.11 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 4

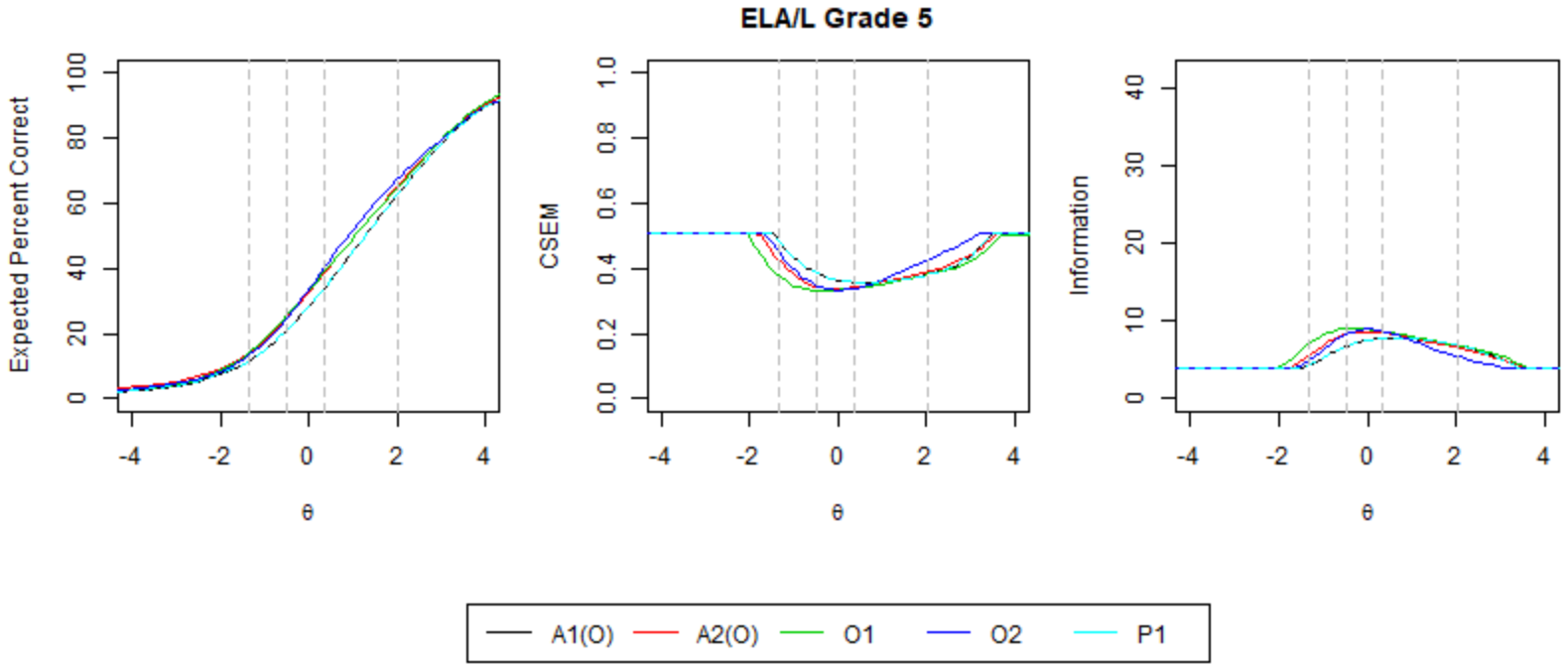

Figure A.12.12 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 5

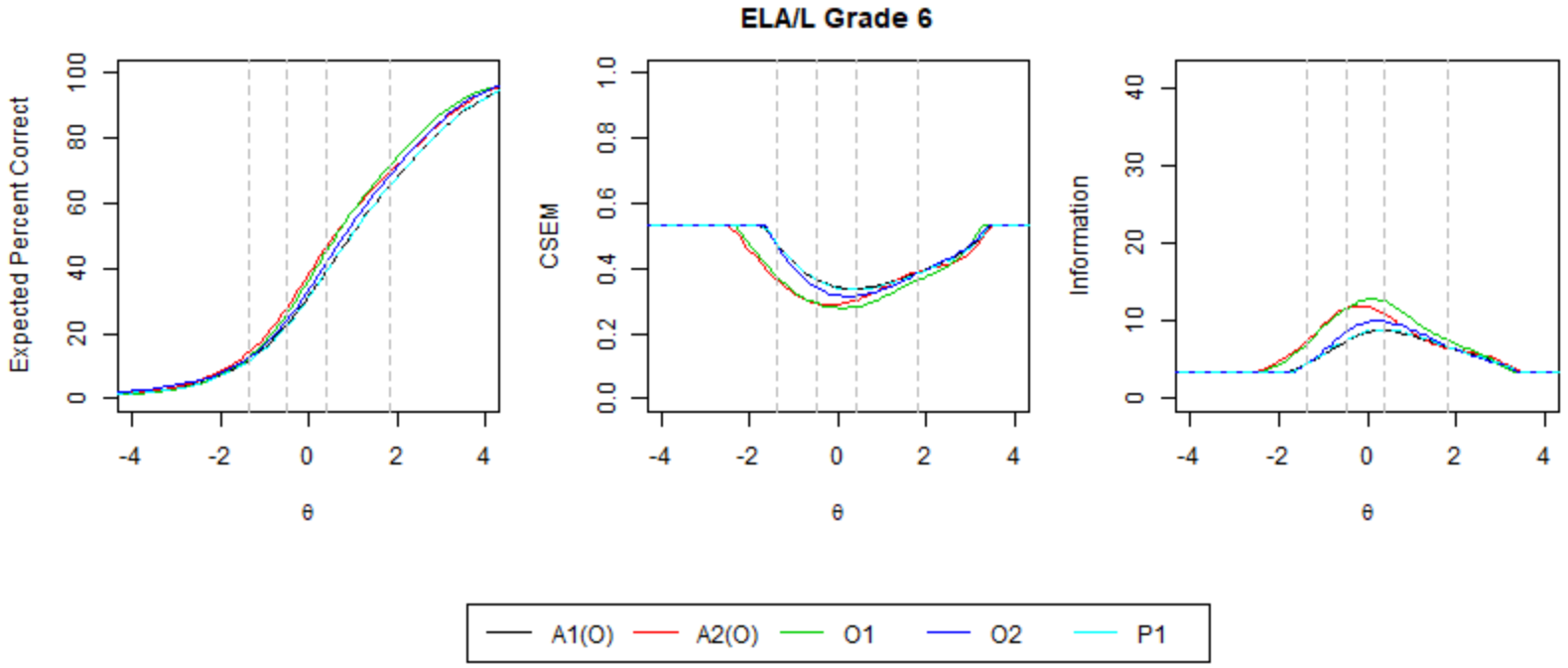

Figure A.12.13 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 6

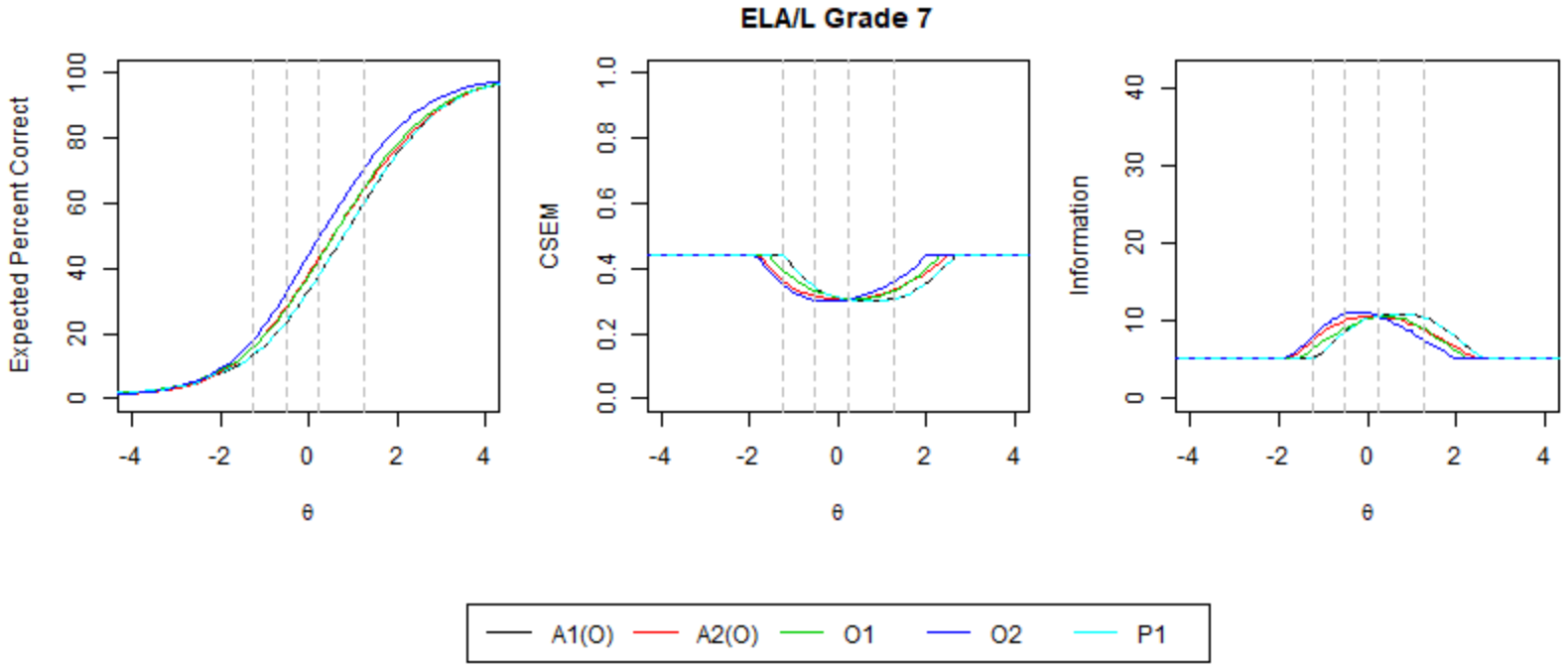

Figure A.12.14 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 7

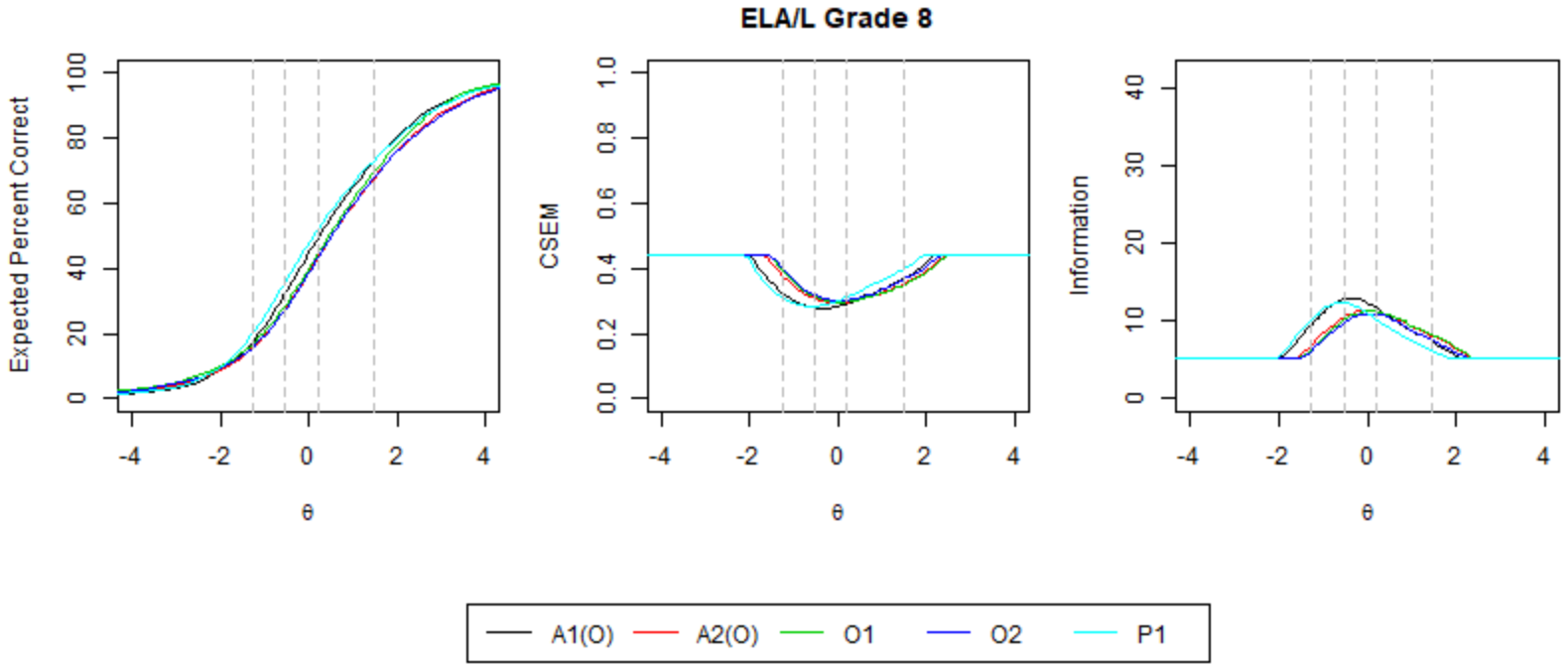

Figure A.12.15 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 8

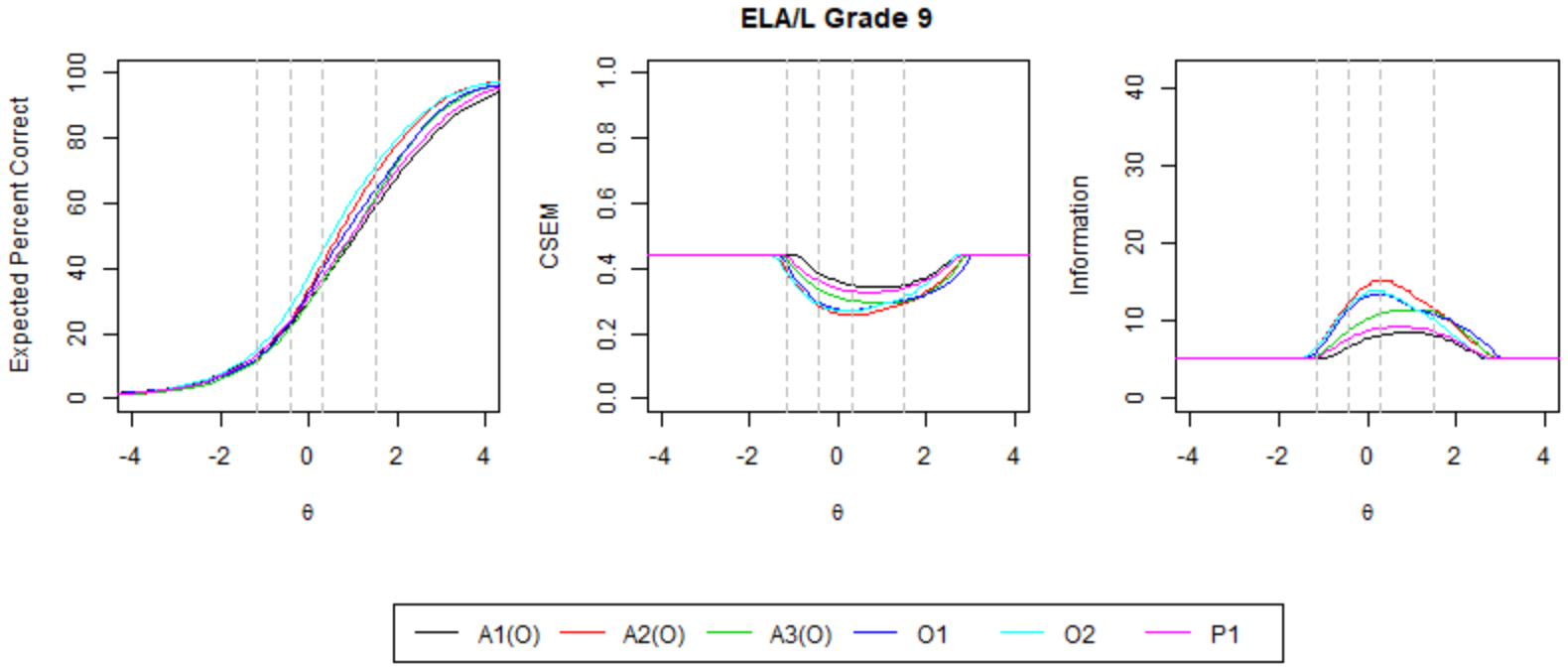

Figure A.12.16 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 9

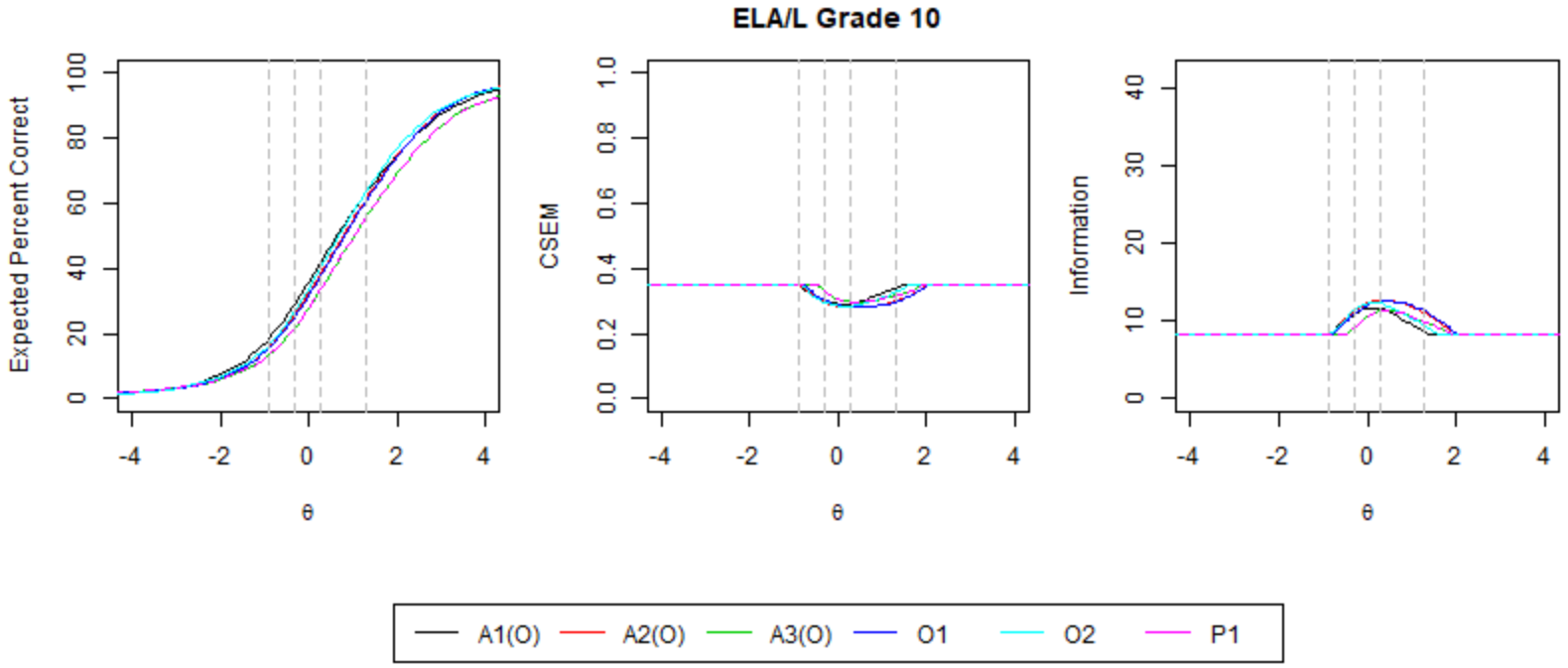

Figure A.12.17 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 10

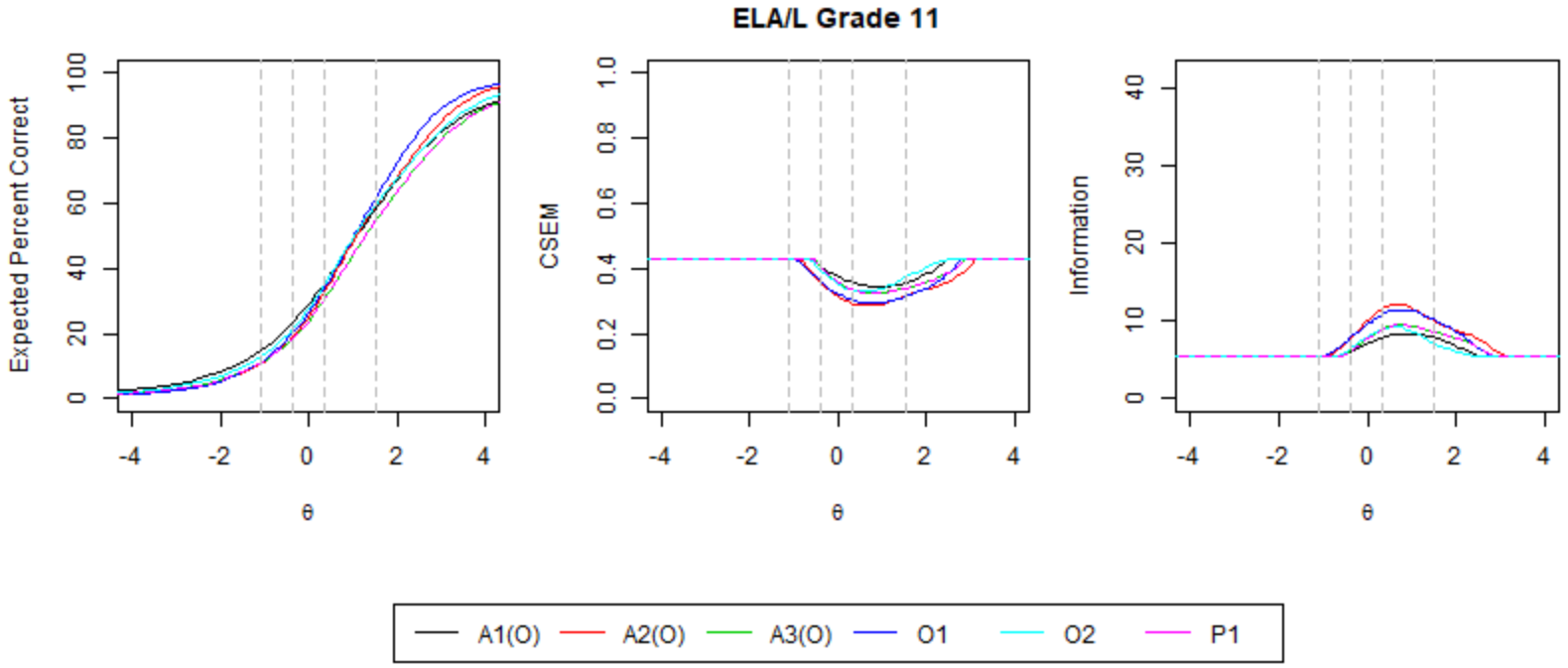

Figure A.12.18 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 11

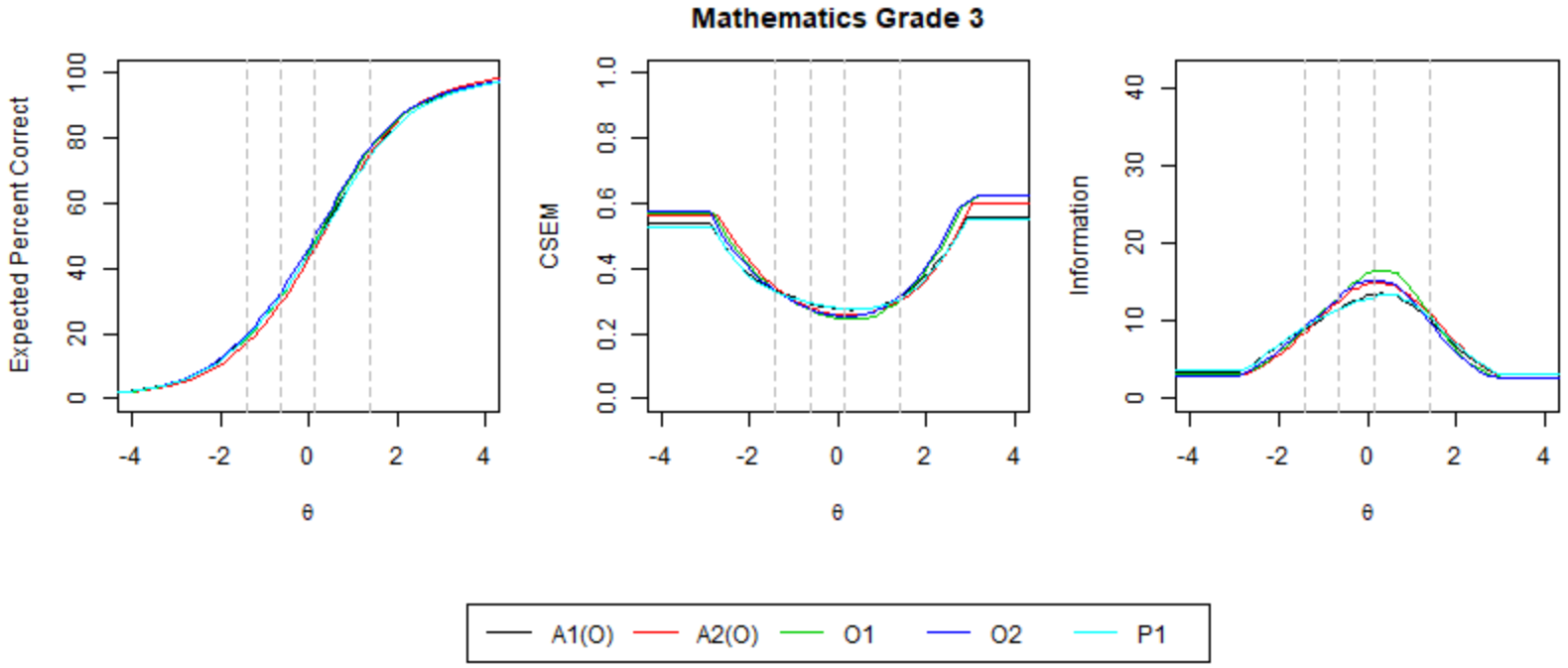

Figure A.12.19 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 3

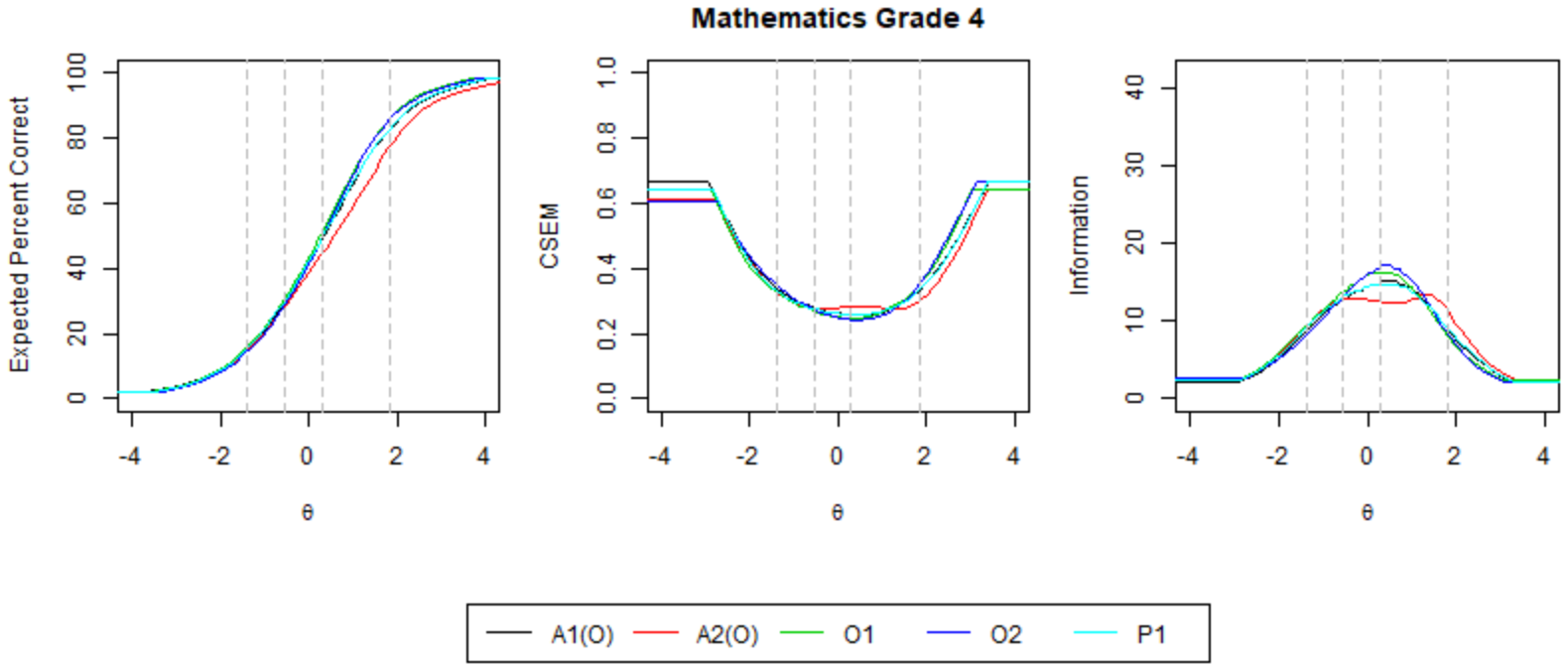

Figure A.12.20 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 4

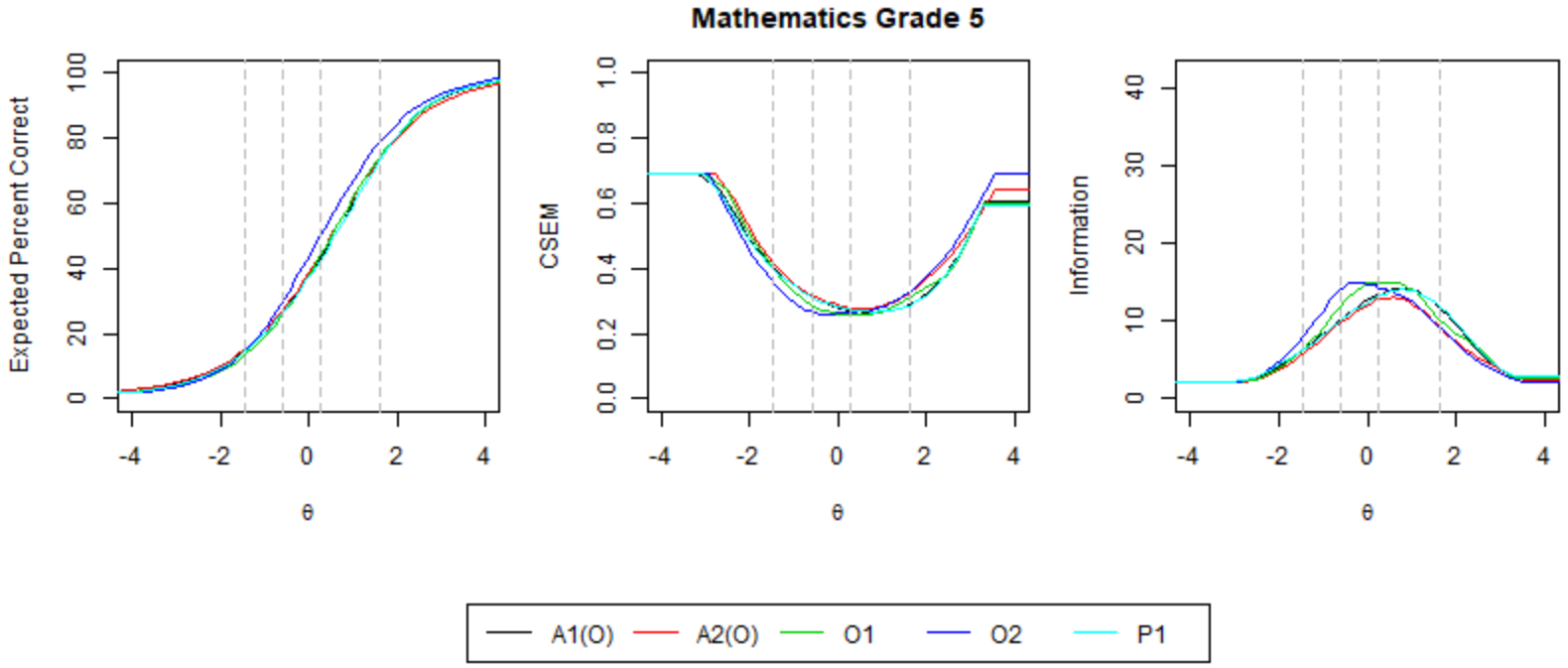

Figure A.12.21 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 5

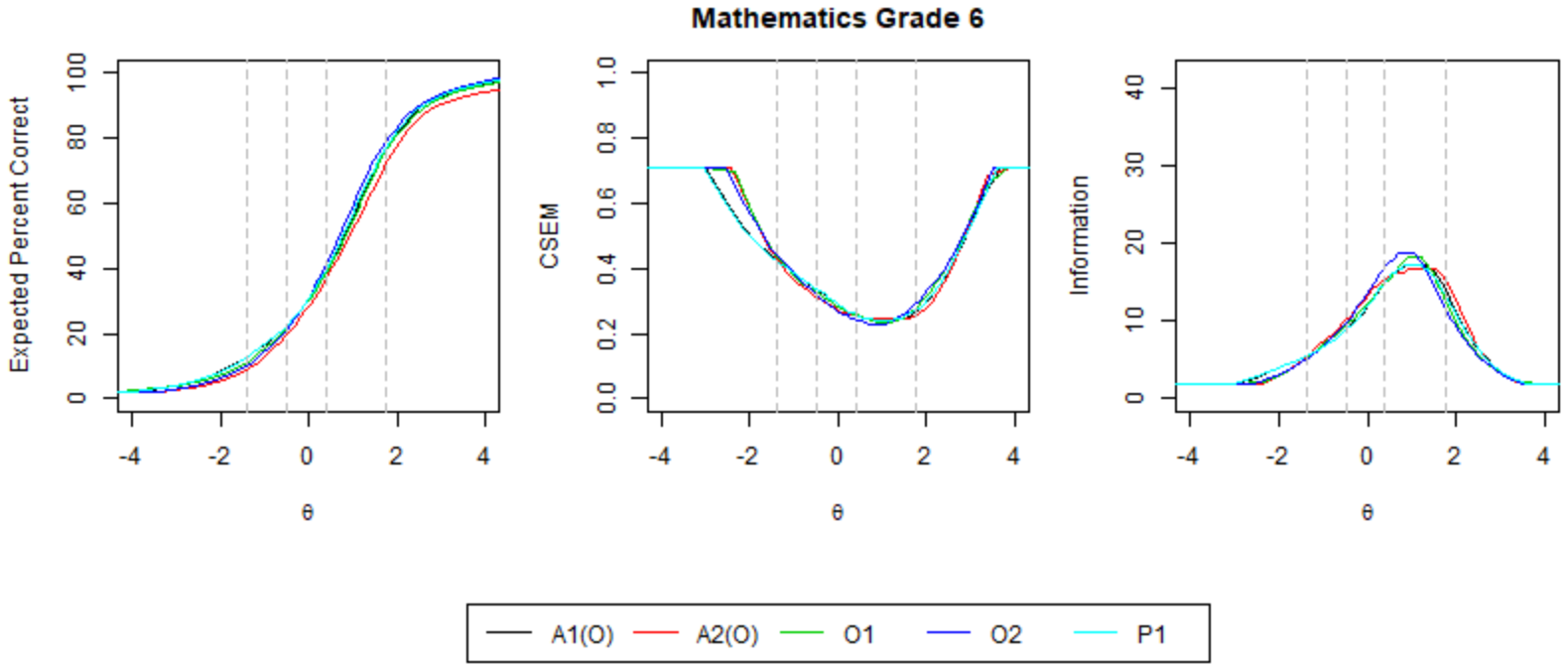

Figure A.12.22 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 6

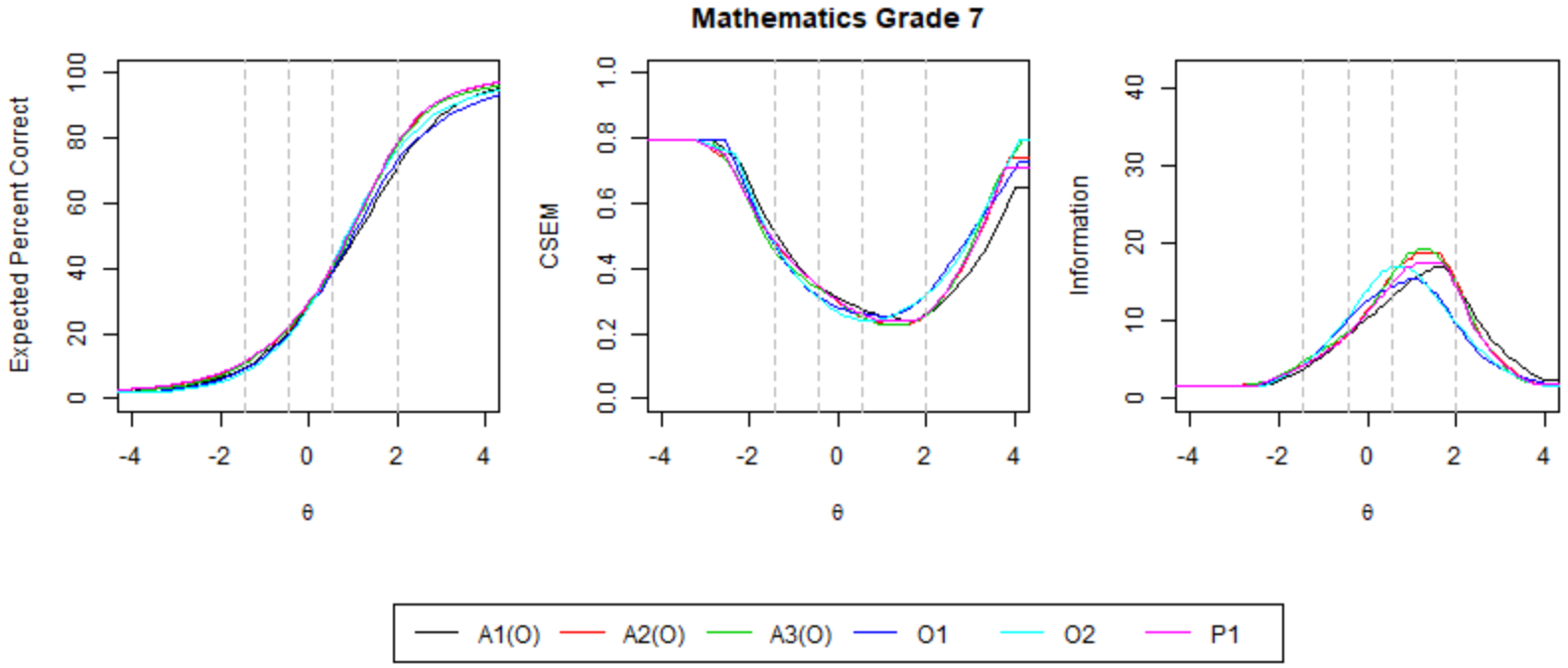

Figure A.12.23 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 7

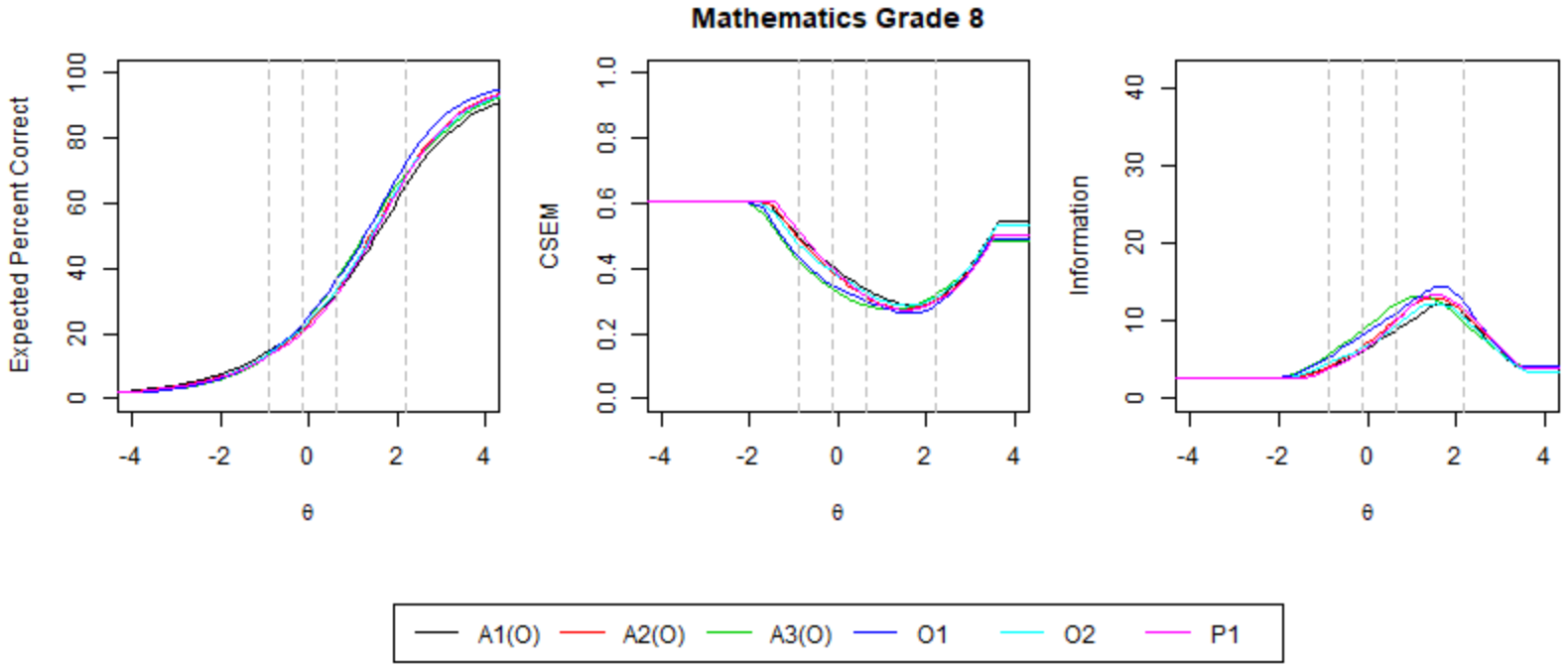

Figure A.12.24 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 8

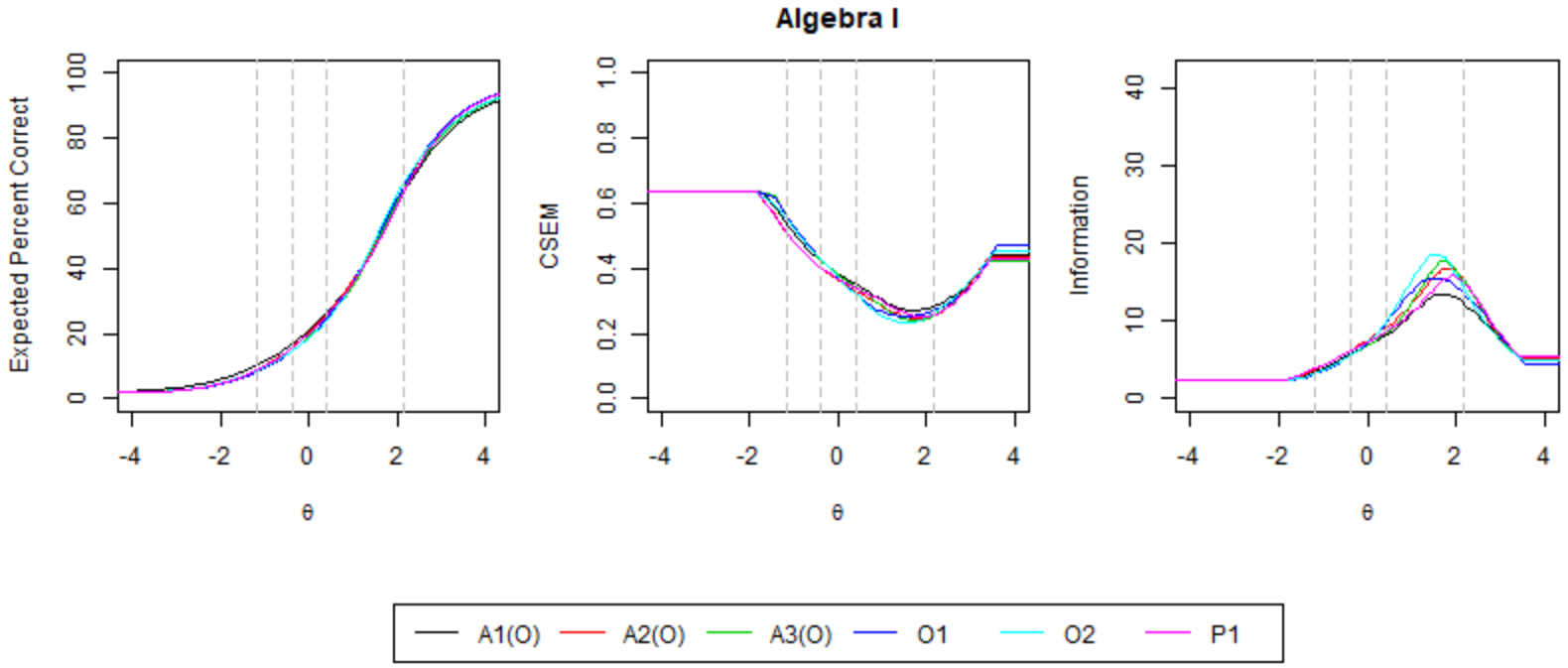

Figure A.12.25 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Algebra I

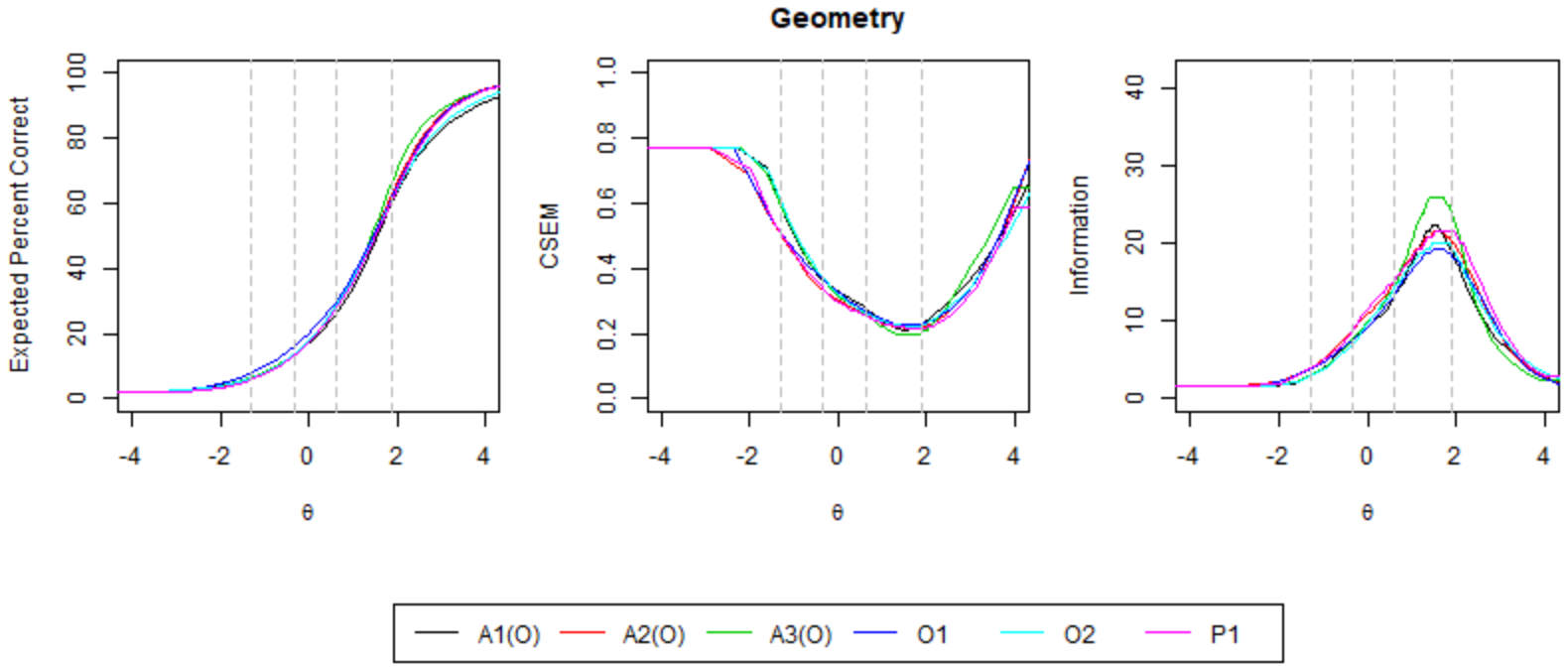

Figure A.12.26 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Geometry

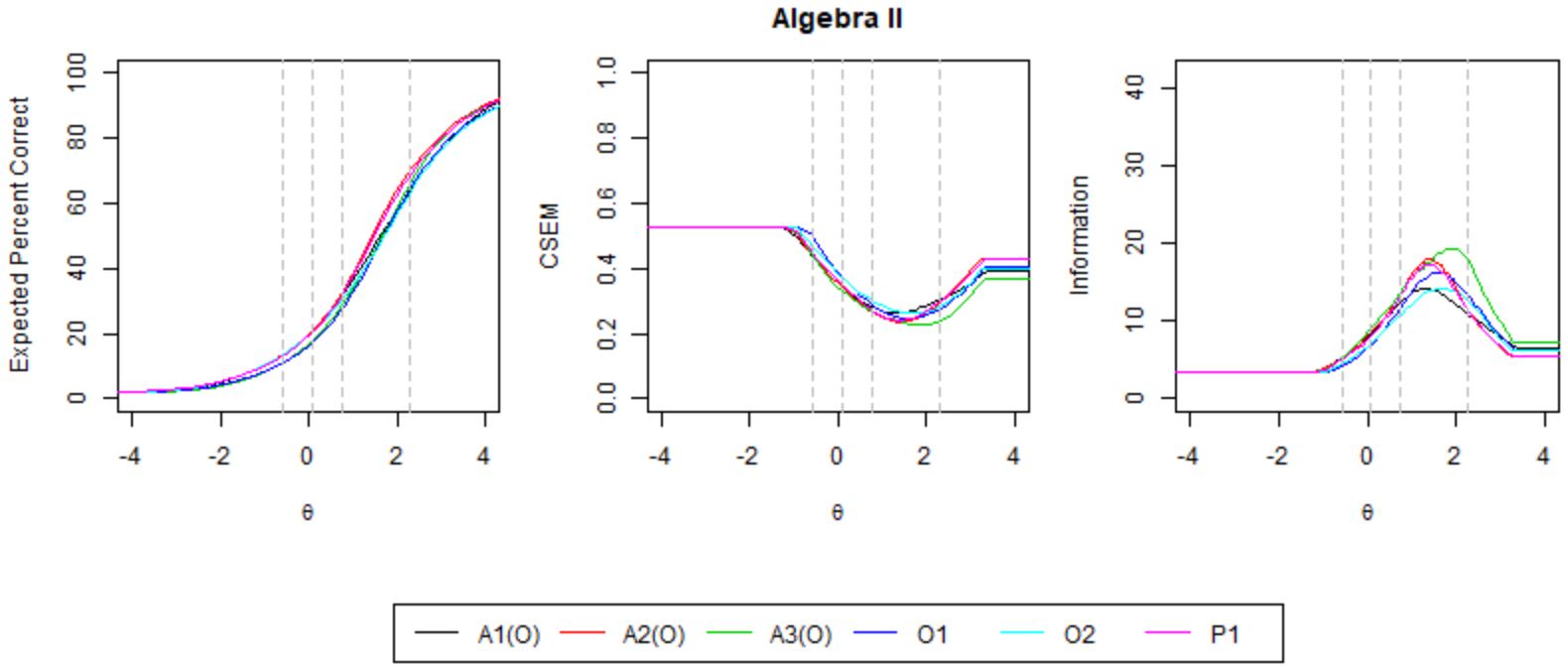

Figure A.12.27 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Algebra II

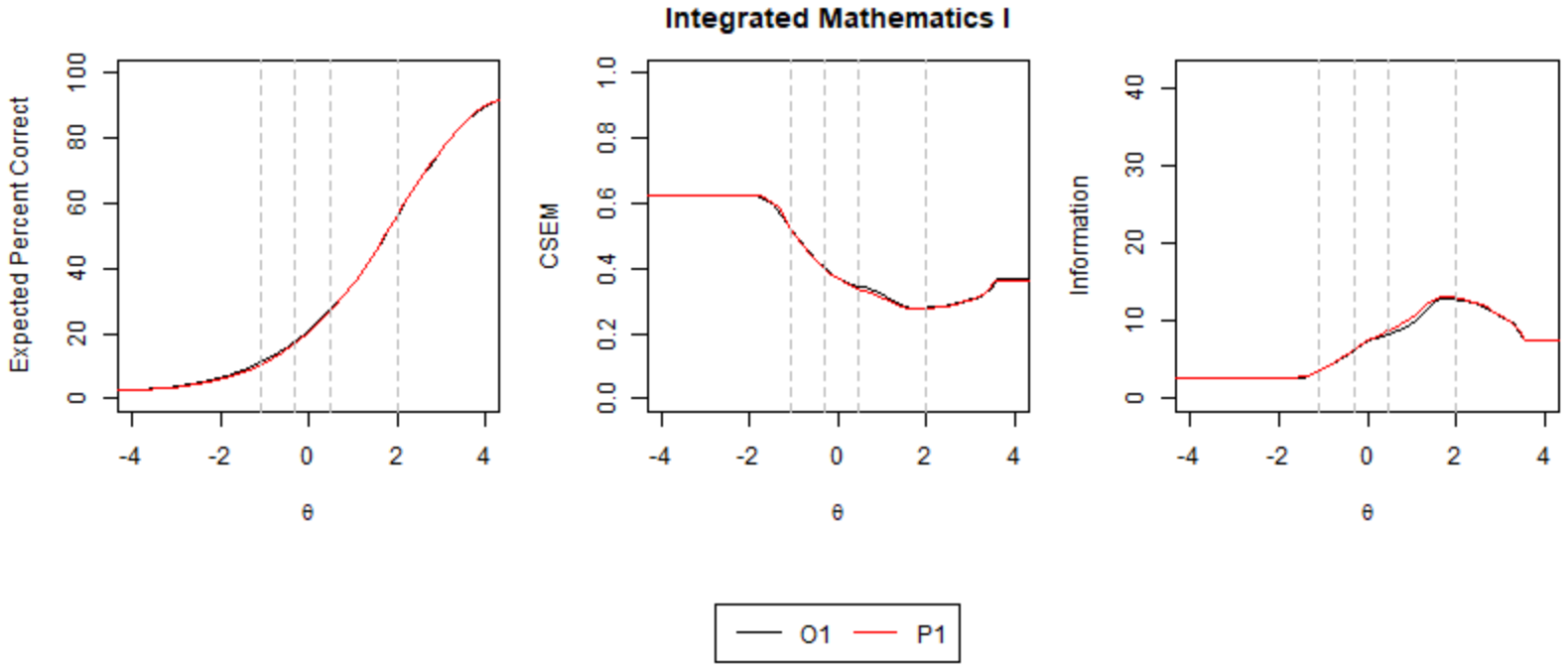

Figure A.12.28 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Integrated Mathematics I

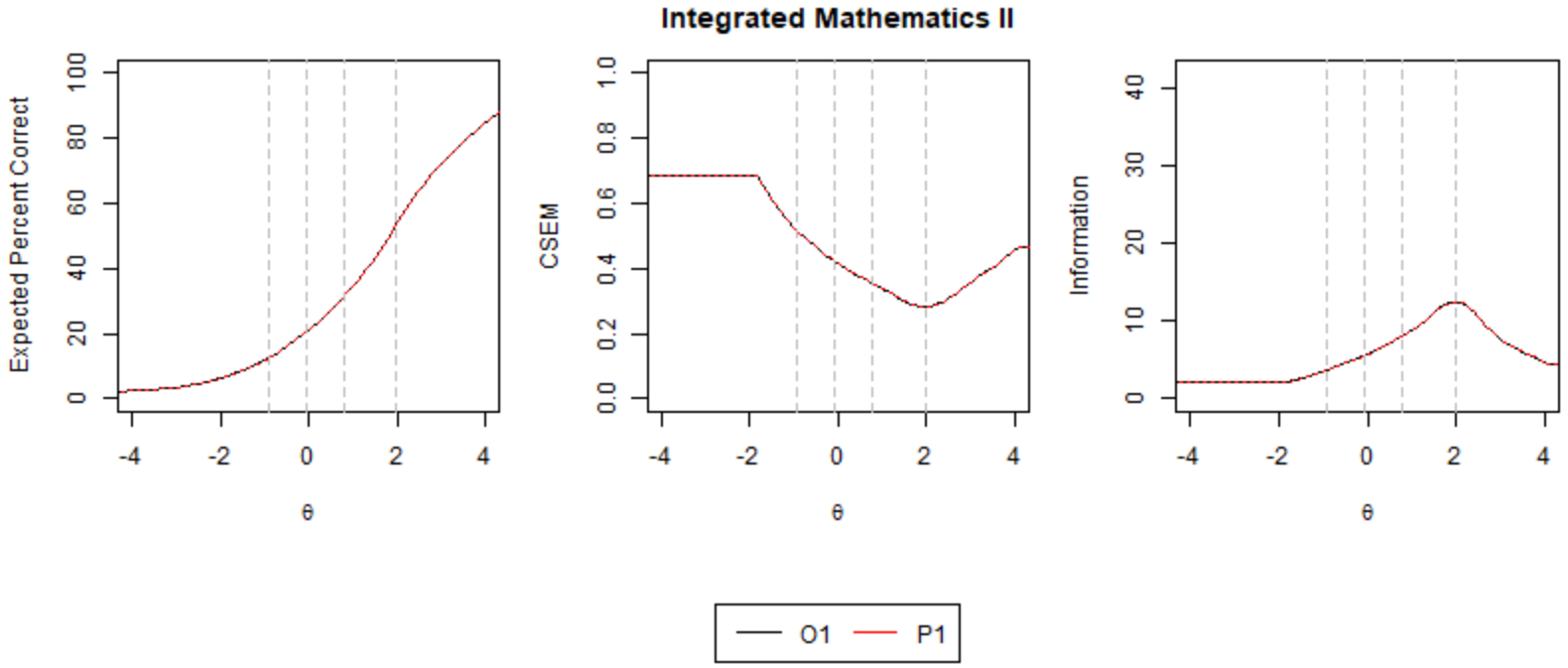

Figure A.12.29 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Integrated Mathematics II

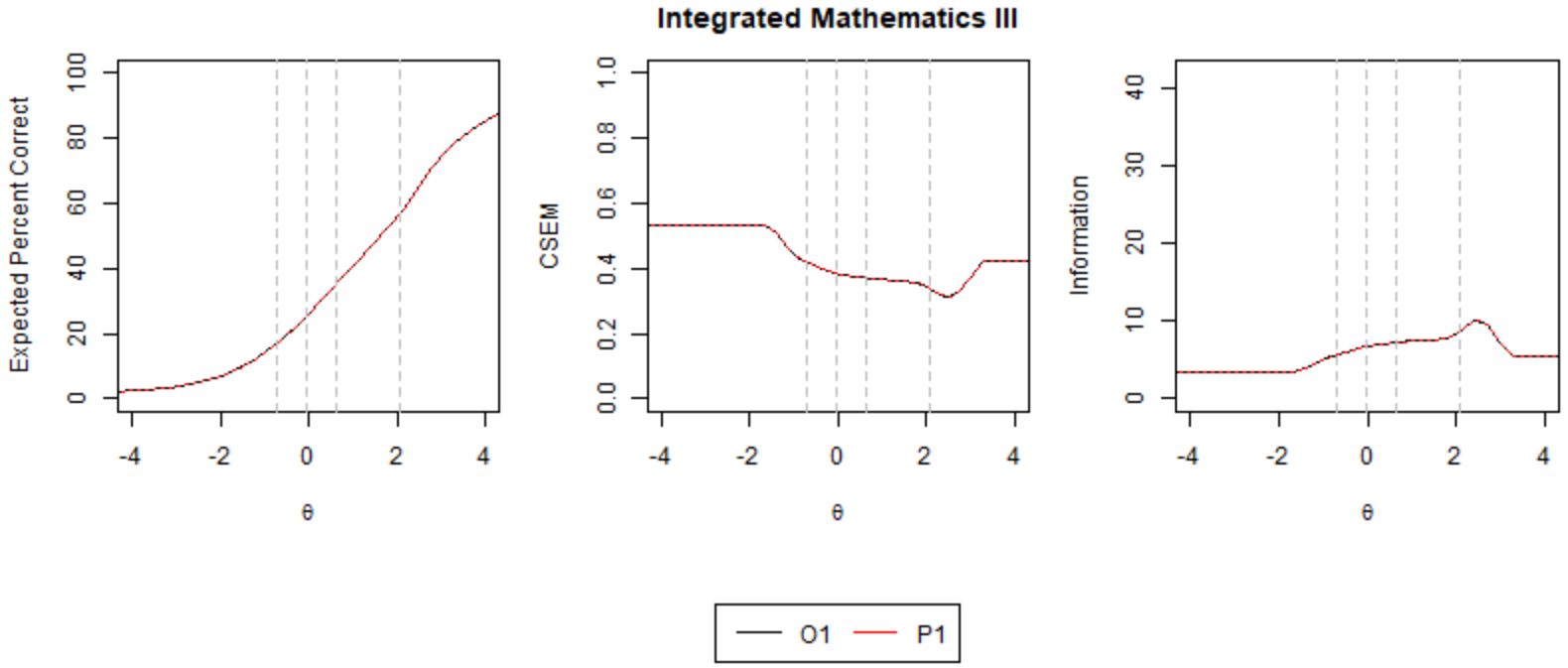

Figure A.12.30 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Integrated Mathematics III

Appendix 12.4: Scale Score Cumulative Frequencies

Table A.12.27 Scale Score Cumulative Frequencies: ELA/L Grade 3

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	7,238	2.82	7,238	2.82
655-659	358	0.14	7,596	2.96
660-664	3,502	1.36	11,098	4.32
665-669	3,705	1.44	14,803	5.76
670-674	2,959	1.15	17,762	6.91
675-679	4,166	1.62	21,928	8.54
680-684	5,657	2.20	27,585	10.74
685-689	7,831	3.05	35,416	13.79
690-694	7,749	3.02	43,165	16.80
695-699	7,544	2.94	50,709	19.74
700-704	7,277	2.83	57,986	22.57
705-709	7,652	2.98	65,638	25.55
710-714	8,978	3.50	74,616	29.05
715-719	9,731	3.79	84,347	32.84
720-724	10,809	4.21	95,156	37.04
725-729	11,196	4.36	106,352	41.40
730-734	9,041	3.52	115,393	44.92
735-739	14,392	5.60	129,785	50.53
740-744	9,345	3.64	139,130	54.16
745-749	13,269	5.17	152,399	59.33
750-754	10,472	4.08	162,871	63.41
755-759	13,037	5.08	175,908	68.48
760-764	9,145	3.56	185,053	72.04
765-769	10,666	4.15	195,719	76.19
770-774	10,512	4.09	206,231	80.29
775-779	7,470	2.91	213,701	83.19
780-784	9,695	3.77	223,396	86.97
785-789	6,409	2.50	229,805	89.46
790-794	4,312	1.68	234,117	91.14
795-799	4,787	1.86	238,904	93.01
800-804	3,030	1.18	241,934	94.19
805-809	2,621	1.02	244,555	95.21
810-814	3,380	1.32	247,935	96.52
815-819	2,012	0.78	249,947	97.30
820-824	1,325	0.52	251,272	97.82
825-829	1,506	0.59	252,778	98.41
830-834	967	0.38	253,745	98.78
835-839	755	0.29	254,500	99.08
840-844	635	0.25	255,135	99.32
845-850	1,735	0.68	256,870	100

Table A.12.28 Scale Score Cumulative Frequencies: ELA/L Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,421	0.91	2,421	0.91
655-659	1,318	0.50	3,739	1.41
660-664	1,263	0.48	5,002	1.89
665-669	1,903	0.72	6,905	2.60
670-674	2,166	0.82	9,071	3.42
675-679	4,836	1.82	13,907	5.24
680-684	5,678	2.14	19,585	7.39
685-689	6,167	2.33	25,752	9.71
690-694	6,116	2.31	31,868	12.02
695-699	6,131	2.31	37,999	14.33
700-704	5,946	2.24	43,945	16.57
705-709	9,532	3.59	53,477	20.17
710-714	7,838	2.96	61,315	23.12
715-719	12,407	4.68	73,722	27.8
720-724	10,647	4.02	84,369	31.82
725-729	12,158	4.59	96,527	36.40
730-734	13,265	5	109,792	41.40
735-739	13,607	5.13	123,399	46.54
740-744	11,974	4.52	135,373	51.05
745-749	13,641	5.14	149,014	56.20
750-754	11,502	4.34	160,516	60.53
755-759	13,437	5.07	173,953	65.60
760-764	12,866	4.85	186,819	70.45
765-769	11,987	4.52	198,806	74.97
770-774	11,216	4.23	210,022	79.20
775-779	10,339	3.90	220,361	83.10
780-784	9,063	3.42	229,424	86.52
785-789	6,040	2.28	235,464	88.80
790-794	5,702	2.15	241,166	90.95
795-799	4,874	1.84	246,040	92.79
800-804	5,411	2.04	251,451	94.83
805-809	3,290	1.24	254,741	96.07
810-814	2,599	0.98	257,340	97.05
815-819	1,929	0.73	259,269	97.78
820-824	1,671	0.63	260,940	98.41
825-829	1,250	0.47	262,190	98.88
830-834	764	0.29	262,954	99.16
835-839	727	0.27	263,681	99.44
840-844	453	0.17	264,134	99.61
845-850	1,035	0.39	265,169	100

Table A.12.29 Scale Score Cumulative Frequencies: ELA/L Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,962	0.72	1,962	0.72
655-659	121	0.04	2,083	0.77
660-664	1,806	0.66	3,889	1.43
665-669	721	0.27	4,610	1.70
670-674	1,151	0.42	5,761	2.12
675-679	4,219	1.55	9,980	3.67
680-684	5,357	1.97	15,337	5.64
685-689	3,773	1.39	19,110	7.03
690-694	5,487	2.02	24,597	9.05
695-699	6,534	2.40	31,131	11.45
700-704	8,563	3.15	39,694	14.61
705-709	8,822	3.25	48,516	17.85
710-714	10,012	3.68	58,528	21.54
715-719	9,504	3.50	68,032	25.03
720-724	12,860	4.73	80,892	29.76
725-729	13,329	4.90	94,221	34.67
730-734	13,659	5.03	107,880	39.69
735-739	13,828	5.09	121,708	44.78
740-744	13,825	5.09	135,533	49.87
745-749	14,777	5.44	150,310	55.31
750-754	15,008	5.52	165,318	60.83
755-759	14,900	5.48	180,218	66.31
760-764	12,528	4.61	192,746	70.92
765-769	12,028	4.43	204,774	75.35
770-774	11,399	4.19	216,173	79.54
775-779	9,577	3.52	225,750	83.06
780-784	9,627	3.54	235,377	86.61
785-789	6,282	2.31	241,659	88.92
790-794	6,608	2.43	248,267	91.35
795-799	6,273	2.31	254,540	93.66
800-804	4,440	1.63	258,980	95.29
805-809	3,695	1.36	262,675	96.65
810-814	1,813	0.67	264,488	97.32
815-819	2,030	0.75	266,518	98.06
820-824	1,829	0.67	268,347	98.74
825-829	791	0.29	269,138	99.03
830-834	928	0.34	270,066	99.37
835-839	554	0.20	270,620	99.57
840-844	428	0.16	271,048	99.73
845-850	730	0.27	271,778	100

Table A.12.30 Scale Score Cumulative Frequencies: ELA/L Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,055	0.38	1,055	0.38
655-659	1,087	0.39	2,142	0.78
660-664	638	0.23	2,780	1.01
665-669	1,881	0.68	4,661	1.69
670-674	1,289	0.47	5,950	2.16
675-679	4,235	1.54	10,185	3.70
680-684	989	0.36	11,174	4.06
685-689	5,293	1.92	16,467	5.98
690-694	6,254	2.27	22,721	8.25
695-699	8,169	2.97	30,890	11.22
700-704	6,648	2.42	37,538	13.64
705-709	9,927	3.61	47,465	17.24
710-714	8,290	3.01	55,755	20.25
715-719	11,481	4.17	67,236	24.42
720-724	12,982	4.72	80,218	29.14
725-729	14,412	5.24	94,630	34.38
730-734	13,795	5.01	108,425	39.39
735-739	15,592	5.66	124,017	45.05
740-744	17,000	6.18	141,017	51.23
745-749	17,863	6.49	158,880	57.72
750-754	13,760	5	172,640	62.72
755-759	15,715	5.71	188,355	68.42
760-764	14,295	5.19	202,650	73.62
765-769	12,238	4.45	214,888	78.06
770-774	11,435	4.15	226,323	82.22
775-779	10,127	3.68	236,450	85.90
780-784	8,937	3.25	245,387	89.14
785-789	6,779	2.46	252,166	91.60
790-794	4,653	1.69	256,819	93.29
795-799	4,719	1.71	261,538	95.01
800-804	3,909	1.42	265,447	96.43
805-809	2,471	0.90	267,918	97.33
810-814	1,606	0.58	269,524	97.91
815-819	1,832	0.67	271,356	98.58
820-824	1,091	0.40	272,447	98.97
825-829	914	0.33	273,361	99.30
830-834	463	0.17	273,824	99.47
835-839	483	0.18	274,307	99.65
840-844	153	0.06	274,460	99.70
845-850	817	0.30	275,277	100

Table A.12.31 Scale Score Cumulative Frequencies: ELA/L Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	3,877	1.44	3,877	1.44
655-659	2,018	0.75	5,895	2.19
660-664	908	0.34	6,803	2.53
665-669	3,501	1.30	10,304	3.82
670-674	3,006	1.12	13,310	4.94
675-679	3,643	1.35	16,953	6.29
680-684	4,465	1.66	21,418	7.95
685-689	4,468	1.66	25,886	9.61
690-694	5,446	2.02	31,332	11.63
695-699	4,493	1.67	35,825	13.30
700-704	7,981	2.96	43,806	16.26
705-709	5,931	2.20	49,737	18.46
710-714	8,572	3.18	58,309	21.65
715-719	9,871	3.66	68,180	25.31
720-724	9,561	3.55	77,741	28.86
725-729	9,937	3.69	87,678	32.55
730-734	12,671	4.70	100,349	37.25
735-739	11,525	4.28	111,874	41.53
740-744	13,043	4.84	124,917	46.37
745-749	13,946	5.18	138,863	51.55
750-754	12,540	4.66	151,403	56.20
755-759	13,253	4.92	164,656	61.12
760-764	14,178	5.26	178,834	66.39
765-769	12,002	4.46	190,836	70.84
770-774	11,619	4.31	202,455	75.15
775-779	9,667	3.59	212,122	78.74
780-784	9,068	3.37	221,190	82.11
785-789	9,062	3.36	230,252	85.47
790-794	6,352	2.36	236,604	87.83
795-799	5,804	2.15	242,408	89.99
800-804	6,030	2.24	248,438	92.22
805-809	3,505	1.30	251,943	93.52
810-814	3,314	1.23	255,257	94.76
815-819	2,809	1.04	258,066	95.80
820-824	2,238	0.83	260,304	96.63
825-829	1,775	0.66	262,079	97.29
830-834	2,098	0.78	264,177	98.07
835-839	1,371	0.51	265,548	98.58
840-844	944	0.35	266,492	98.93
845-850	2,894	1.07	269,386	100

Table A.12.32 Scale Score Cumulative Frequencies: ELA/L Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	3,877	1.44	3,877	1.44
655-659	2,018	0.75	5,895	2.19
660-664	908	0.34	6,803	2.53
665-669	3,501	1.30	10,304	3.82
670-674	3,006	1.12	13,310	4.94
675-679	3,643	1.35	16,953	6.29
680-684	4,465	1.66	21,418	7.95
685-689	4,468	1.66	25,886	9.61
690-694	5,446	2.02	31,332	11.63
695-699	4,493	1.67	35,825	13.30
700-704	7,981	2.96	43,806	16.26
705-709	5,931	2.20	49,737	18.46
710-714	8,572	3.18	58,309	21.65
715-719	9,871	3.66	68,180	25.31
720-724	9,561	3.55	77,741	28.86
725-729	9,937	3.69	87,678	32.55
730-734	12,671	4.70	100,349	37.25
735-739	11,525	4.28	111,874	41.53
740-744	13,043	4.84	124,917	46.37
745-749	13,946	5.18	138,863	51.55
750-754	12,540	4.66	151,403	56.20
755-759	13,253	4.92	164,656	61.12
760-764	14,178	5.26	178,834	66.39
765-769	12,002	4.46	190,836	70.84
770-774	11,619	4.31	202,455	75.15
775-779	9,667	3.59	212,122	78.74
780-784	9,068	3.37	221,190	82.11
785-789	9,062	3.36	230,252	85.47
790-794	6,352	2.36	236,604	87.83
795-799	5,804	2.15	242,408	89.99
800-804	6,030	2.24	248,438	92.22
805-809	3,505	1.30	251,943	93.52
810-814	3,314	1.23	255,257	94.76
815-819	2,809	1.04	258,066	95.80
820-824	2,238	0.83	260,304	96.63
825-829	1,775	0.66	262,079	97.29
830-834	2,098	0.78	264,177	98.07
835-839	1,371	0.51	265,548	98.58
840-844	944	0.35	266,492	98.93
845-850	2,894	1.07	269,386	100

Table A.12.33 Scale Score Cumulative Frequencies: ELA/L Grade 9

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,528	1.26	1,528	1.26
655-659	953	0.78	2,481	2.04
660-664	171	0.14	2,652	2.18
665-669	1,468	1.21	4,120	3.39
670-674	1,349	1.11	5,469	4.50
675-679	565	0.46	6,034	4.96
680-684	2,137	1.76	8,171	6.72
685-689	1,920	1.58	10,091	8.30
690-694	2,112	1.74	12,203	10.03
695-699	3,240	2.66	15,443	12.70
700-704	2,043	1.68	17,486	14.38
705-709	3,701	3.04	21,187	17.42
710-714	2,944	2.42	24,131	19.84
715-719	4,236	3.48	28,367	23.32
720-724	4,359	3.58	32,726	26.91
725-729	5,533	4.55	38,259	31.46
730-734	4,710	3.87	42,969	35.33
735-739	5,425	4.46	48,394	39.79
740-744	6,160	5.06	54,554	44.86
745-749	5,624	4.62	60,178	49.48
750-754	6,380	5.25	66,558	54.73
755-759	5,000	4.11	71,558	58.84
760-764	5,561	4.57	77,119	63.41
765-769	6,500	5.34	83,619	68.75
770-774	4,843	3.98	88,462	72.74
775-779	4,806	3.95	93,268	76.69
780-784	4,516	3.71	97,784	80.40
785-789	4,211	3.46	101,995	83.86
790-794	3,802	3.13	105,797	86.99
795-799	3,436	2.83	109,233	89.82
800-804	1,922	1.58	111,155	91.40
805-809	2,756	2.27	113,911	93.66
810-814	1,561	1.28	115,472	94.95
815-819	2,116	1.74	117,588	96.69
820-824	884	0.73	118,472	97.41
825-829	908	0.75	119,380	98.16
830-834	592	0.49	119,972	98.65
835-839	553	0.45	120,525	99.10
840-844	376	0.31	120,901	99.41
845-850	718	0.59	121,619	100

Table A.12.34 Scale Score Cumulative Frequencies: ELA/L Grade 10

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	5,095	4.31	5,095	4.31
655-659	1,052	0.89	6,147	5.20
660-664	938	0.79	7,085	5.99
665-669	1,106	0.93	8,191	6.92
670-674	1,492	1.26	9,683	8.18
675-679	1,458	1.23	11,141	9.42
680-684	2,010	1.70	13,151	11.11
685-689	1,160	0.98	14,311	12.09
690-694	2,640	2.23	16,951	14.33
695-699	1,951	1.65	18,902	15.98
700-704	1,981	1.67	20,883	17.65
705-709	3,166	2.68	24,049	20.33
710-714	2,809	2.37	26,858	22.70
715-719	3,293	2.78	30,151	25.48
720-724	3,573	3.02	33,724	28.50
725-729	3,483	2.94	37,207	31.45
730-734	4,377	3.70	41,584	35.14
735-739	3,389	2.86	44,973	38.01
740-744	4,614	3.90	49,587	41.91
745-749	4,623	3.91	54,210	45.82
750-754	4,842	4.09	59,052	49.91
755-759	3,729	3.15	62,781	53.06
760-764	4,978	4.21	67,759	57.27
765-769	4,905	4.15	72,664	61.41
770-774	4,838	4.09	77,502	65.50
775-779	3,775	3.19	81,277	68.69
780-784	4,573	3.86	85,850	72.56
785-789	4,316	3.65	90,166	76.20
790-794	3,188	2.69	93,354	78.90
795-799	3,085	2.61	96,439	81.51
800-804	2,943	2.49	99,382	83.99
805-809	3,499	2.96	102,881	86.95
810-814	1,596	1.35	104,477	88.30
815-819	2,423	2.05	106,900	90.35
820-824	2,014	1.70	108,914	92.05
825-829	2,007	1.70	110,921	93.75
830-834	1,073	0.91	111,994	94.65
835-839	1,088	0.92	113,082	95.57
840-844	929	0.79	114,011	96.36
845-850	4,311	3.64	118,322	100

Table A.12.35 Scale Score Cumulative Frequencies: ELA/L Grade 11

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	624	1.80	624	1.80
655-659	434	1.25	1,058	3.06
660-664	169	0.49	1,227	3.55
665-669	434	1.25	1,661	4.80
670-674	400	1.16	2,061	5.95
675-679	582	1.68	2,643	7.64
680-684	693	2.00	3,336	9.64
685-689	860	2.48	4,196	12.12
690-694	914	2.64	5,110	14.76
695-699	832	2.40	5,942	17.17
700-704	1,259	3.64	7,201	20.81
705-709	1,034	2.99	8,235	23.79
710-714	1,287	3.72	9,522	27.51
715-719	1,652	4.77	11,174	32.29
720-724	877	2.53	12,051	34.82
725-729	1,773	5.12	13,824	39.94
730-734	1,754	5.07	15,578	45.01
735-739	1,399	4.04	16,977	49.05
740-744	1,707	4.93	18,684	53.98
745-749	1,708	4.93	20,392	58.92
750-754	1,790	5.17	22,182	64.09
755-759	1,905	5.50	24,087	69.60
760-764	1,369	3.96	25,456	73.55
765-769	1,485	4.29	26,941	77.84
770-774	1,468	4.24	28,409	82.08
775-779	1,279	3.70	29,688	85.78
780-784	1,049	3.03	30,737	88.81
785-789	575	1.66	31,312	90.47
790-794	860	2.48	32,172	92.96
795-799	531	1.53	32,703	94.49
800-804	417	1.20	33,120	95.69
805-809	363	1.05	33,483	96.74
810-814	281	0.81	33,764	97.56
815-819	242	0.70	34,006	98.25
820-824	142	0.41	34,148	98.67
825-829	139	0.40	34,287	99.07
830-834	105	0.30	34,392	99.37
835-839	52	0.15	34,444	99.52
840-844	61	0.18	34,505	99.70
845-850	105	0.30	34,610	100

Table A.12.36 Scale Score Cumulative Frequencies: Mathematics Grade 3

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,735	1.06	2,735	1.06
655-659	360	0.14	3,095	1.20
660-664	993	0.38	4,088	1.58
665-669	1,666	0.64	5,754	2.22
670-674	3,276	1.27	9,030	3.49
675-679	1,781	0.69	10,811	4.18
680-684	4,273	1.65	15,084	5.83
685-689	3,122	1.21	18,206	7.03
690-694	5,494	2.12	23,700	9.16
695-699	8,487	3.28	32,187	12.44
700-704	6,438	2.49	38,625	14.92
705-709	8,093	3.13	46,718	18.05
710-714	10,276	3.97	56,994	22.02
715-719	10,421	4.03	67,415	26.05
720-724	11,805	4.56	79,220	30.61
725-729	10,705	4.14	89,925	34.75
730-734	15,060	5.82	104,985	40.56
735-739	14,834	5.73	119,819	46.30
740-744	10,862	4.20	130,681	50.49
745-749	11,335	4.38	142,016	54.87
750-754	14,158	5.47	156,174	60.34
755-759	13,850	5.35	170,024	65.70
760-764	12,512	4.83	182,536	70.53
765-769	12,702	4.91	195,238	75.44
770-774	12,113	4.68	207,351	80.12
775-779	8,851	3.42	216,202	83.54
780-784	10,168	3.93	226,370	87.47
785-789	7,544	2.91	233,914	90.38
790-794	6,515	2.52	240,429	92.90
795-799	4,171	1.61	244,600	94.51
800-804	3,399	1.31	247,999	95.82
805-809	2,919	1.13	250,918	96.95
810-814	2,498	0.97	253,416	97.92
815-819	1,922	0.74	255,338	98.66
820-824	78	0.03	255,416	98.69
825-829	1,335	0.52	256,751	99.21
830-834	512	0.20	257,263	99.40
835-839	468	0.18	257,731	99.58
840-844	30	0.01	257,761	99.6
845-850	1,046	0.40	258,807	100

Table A.12.37 Scale Score Cumulative Frequencies: Mathematics Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,352	0.51	1,352	0.51
655-659	1,033	0.39	2,385	0.89
660-664	1,558	0.58	3,943	1.48
665-669	1,677	0.63	5,620	2.11
670-674	2,491	0.93	8,111	3.04
675-679	2,549	0.96	10,660	4
680-684	5,835	2.19	16,495	6.19
685-689	6,483	2.43	22,978	8.62
690-694	7,055	2.65	30,033	11.26
695-699	7,260	2.72	37,293	13.99
700-704	7,286	2.73	44,579	16.72
705-709	7,608	2.85	52,187	19.57
710-714	14,243	5.34	66,430	24.91
715-719	10,867	4.08	77,297	28.99
720-724	10,790	4.05	88,087	33.04
725-729	14,023	5.26	102,110	38.30
730-734	13,779	5.17	115,889	43.46
735-739	13,550	5.08	129,439	48.55
740-744	13,273	4.98	142,712	53.52
745-749	18,790	7.05	161,502	60.57
750-754	12,502	4.69	174,004	65.26
755-759	14,383	5.39	188,387	70.66
760-764	13,955	5.23	202,342	75.89
765-769	10,651	3.99	212,993	79.88
770-774	12,358	4.63	225,351	84.52
775-779	9,341	3.50	234,692	88.02
780-784	8,542	3.20	243,234	91.23
785-789	3,977	1.49	247,211	92.72
790-794	7,027	2.64	254,238	95.35
795-799	2,967	1.11	257,205	96.47
800-804	2,692	1.01	259,897	97.48
805-809	2,148	0.81	262,045	98.28
810-814	1,802	0.68	263,847	98.96
815-819	1	0	263,848	98.96
820-824	1,297	0.49	265,145	99.44
825-829	0	0	265,145	99.44
830-834	371	0.14	265,516	99.58
835-839	477	0.18	265,993	99.76
840-844	2	0	265,995	99.76
845-850	634	0.24	266,629	100

Table A.12.38 Scale Score Cumulative Frequencies: Mathematics Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	948	0.35	948	0.35
655-659	568	0.21	1,516	0.56
660-664	134	0.05	1,650	0.61
665-669	107	0.04	1,757	0.64
670-674	2,794	1.02	4,551	1.67
675-679	208	0.08	4,759	1.75
680-684	4,904	1.80	9,663	3.54
685-689	7,066	2.59	16,729	6.13
690-694	4,296	1.58	21,025	7.71
695-699	9,336	3.42	30,361	11.13
700-704	9,995	3.67	40,356	14.8
705-709	14,525	5.33	54,881	20.12
710-714	16,018	5.87	70,899	26
715-719	14,197	5.21	85,096	31.2
720-724	18,320	6.72	103,416	37.92
725-729	12,818	4.70	116,234	42.62
730-734	16,357	6	132,591	48.62
735-739	15,181	5.57	147,772	54.19
740-744	14,558	5.34	162,330	59.52
745-749	13,461	4.94	175,791	64.46
750-754	15,628	5.73	191,419	70.19
755-759	11,450	4.20	202,869	74.39
760-764	13,107	4.81	215,976	79.20
765-769	9,789	3.59	225,765	82.78
770-774	8,995	3.30	234,760	86.08
775-779	6,202	2.27	240,962	88.36
780-784	7,658	2.81	248,620	91.17
785-789	5,235	1.92	253,855	93.08
790-794	4,751	1.74	258,606	94.83
795-799	2,800	1.03	261,406	95.85
800-804	3,501	1.28	264,907	97.14
805-809	2,046	0.75	266,953	97.89
810-814	1,674	0.61	268,627	98.50
815-819	1,393	0.51	270,020	99.01
820-824	1,071	0.39	271,091	99.40
825-829	305	0.11	271,396	99.52
830-834	408	0.15	271,804	99.67
835-839	230	0.08	272,034	99.75
840-844	0	0	272,034	99.75
845-850	680	0.25	272,714	100

Table A.12.39 Scale Score Cumulative Frequencies: Mathematics Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,716	0.99	2,716	0.99
655-659	1,776	0.64	4,492	1.63
660-664	0	0	4,492	1.63
665-669	3,024	1.10	7,516	2.73
670-674	3,639	1.32	11,155	4.05
675-679	264	0.10	11,419	4.14
680-684	4,378	1.59	15,797	5.73
685-689	10,976	3.98	26,773	9.71
690-694	6,332	2.30	33,105	12.01
695-699	6,150	2.23	39,255	14.24
700-704	13,342	4.84	52,597	19.08
705-709	13,455	4.88	66,052	23.96
710-714	13,301	4.82	79,353	28.78
715-719	12,805	4.64	92,158	33.42
720-724	24,251	8.80	116,409	42.22
725-729	11,083	4.02	127,492	46.24
730-734	20,224	7.33	147,716	53.57
735-739	17,701	6.42	165,417	59.99
740-744	11,042	4	176,459	64
745-749	15,867	5.75	192,326	69.75
750-754	14,084	5.11	206,410	74.86
755-759	11,750	4.26	218,160	79.12
760-764	12,212	4.43	230,372	83.55
765-769	8,562	3.11	238,934	86.65
770-774	9,212	3.34	248,146	90
775-779	6,690	2.43	254,836	92.42
780-784	4,931	1.79	259,767	94.21
785-789	5,045	1.83	264,812	96.04
790-794	2,747	1	267,559	97.04
795-799	2,304	0.84	269,863	97.87
800-804	1,326	0.48	271,189	98.35
805-809	1,171	0.42	272,360	98.78
810-814	1,026	0.37	273,386	99.15
815-819	816	0.30	274,202	99.45
820-824	316	0.11	274,518	99.56
825-829	339	0.12	274,857	99.68
830-834	238	0.09	275,095	99.77
835-839	215	0.08	275,310	99.85
840-844	0	0	275,310	99.85
845-850	422	0.15	275,732	100

Table A.12.40 Scale Score Cumulative Frequencies: Mathematics Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,136	0.43	1,136	0.43
655-659	1,438	0.54	2,574	0.97
660-664	79	0.03	2,653	1
665-669	4	0	2,657	1
670-674	1,904	0.72	4,561	1.72
675-679	3,129	1.18	7,690	2.90
680-684	166	0.06	7,856	2.96
685-689	3,476	1.31	11,332	4.28
690-694	4,941	1.86	16,273	6.14
695-699	10,381	3.92	26,654	10.06
700-704	11,558	4.36	38,212	14.42
705-709	6,457	2.44	44,669	16.86
710-714	17,877	6.75	62,546	23.61
715-719	11,646	4.40	74,192	28
720-724	21,330	8.05	95,522	36.05
725-729	15,201	5.74	110,723	41.79
730-734	12,955	4.89	123,678	46.68
735-739	19,824	7.48	143,502	54.16
740-744	14,343	5.41	157,845	59.57
745-749	18,859	7.12	176,704	66.69
750-754	13,630	5.14	190,334	71.83
755-759	14,851	5.60	205,185	77.44
760-764	9,107	3.44	214,292	80.88
765-769	12,411	4.68	226,703	85.56
770-774	7,527	2.84	234,230	88.40
775-779	8,337	3.15	242,567	91.55
780-784	5,906	2.23	248,473	93.78
785-789	5,078	1.92	253,551	95.69
790-794	3,280	1.24	256,831	96.93
795-799	1,783	0.67	258,614	97.60
800-804	2,238	0.84	260,852	98.45
805-809	1,200	0.45	262,052	98.90
810-814	404	0.15	262,456	99.05
815-819	824	0.31	263,280	99.37
820-824	411	0.16	263,691	99.52
825-829	230	0.09	263,921	99.61
830-834	301	0.11	264,222	99.72
835-839	349	0.13	264,571	99.85
840-844	0	0	264,571	99.85
845-850	389	0.15	264,960	100

Table A.12.41 Scale Score Cumulative Frequencies: Mathematics Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,136	0.43	1,136	0.43
655-659	1,438	0.54	2,574	0.97
660-664	79	0.03	2,653	1
665-669	4	0	2,657	1
670-674	1,904	0.72	4,561	1.72
675-679	3,129	1.18	7,690	2.90
680-684	166	0.06	7,856	2.96
685-689	3,476	1.31	11,332	4.28
690-694	4,941	1.86	16,273	6.14
695-699	10,381	3.92	26,654	10.06
700-704	11,558	4.36	38,212	14.42
705-709	6,457	2.44	44,669	16.86
710-714	17,877	6.75	62,546	23.61
715-719	11,646	4.4	74,192	28
720-724	21,330	8.05	95,522	36.05
725-729	15,201	5.74	110,723	41.79
730-734	12,955	4.89	123,678	46.68
735-739	19,824	7.48	143,502	54.16
740-744	14,343	5.41	157,845	59.57
745-749	18,859	7.12	176,704	66.69
750-754	13,630	5.14	190,334	71.83
755-759	14,851	5.60	205,185	77.44
760-764	9,107	3.44	214,292	80.88
765-769	12,411	4.68	226,703	85.56
770-774	7,527	2.84	234,230	88.40
775-779	8,337	3.15	242,567	91.55
780-784	5,906	2.23	248,473	93.78
785-789	5,078	1.92	253,551	95.69
790-794	3,280	1.24	256,831	96.93
795-799	1,783	0.67	258,614	97.60
800-804	2,238	0.84	260,852	98.45
805-809	1,200	0.45	262,052	98.90
810-814	404	0.15	262,456	99.05
815-819	824	0.31	263,280	99.37
820-824	411	0.16	263,691	99.52
825-829	230	0.09	263,921	99.61
830-834	301	0.11	264,222	99.72
835-839	349	0.13	264,571	99.85
840-844	0	0	264,571	99.85
845-850	389	0.15	264,960	100

Table A.12.42 Scale Score Cumulative Frequencies: Algebra I

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	604	0.45	604	0.45
655-659	0	0	604	0.45
660-664	1,833	1.37	2,437	1.82
665-669	2	0	2,439	1.82
670-674	0	0	2,439	1.82
675-679	2,224	1.66	4,663	3.48
680-684	1,918	1.43	6,581	4.91
685-689	0	0	6,581	4.91
690-694	7,119	5.31	13,700	10.22
695-699	0	0	13,700	10.22
700-704	9,681	7.22	23,381	17.43
705-709	87	0.06	23,468	17.50
710-714	10,563	7.88	34,031	25.38
715-719	10,249	7.64	44,280	33.02
720-724	9,102	6.79	53,382	39.81
725-729	144	0.11	53,526	39.91
730-734	7,623	5.68	61,149	45.60
735-739	9,239	6.89	70,388	52.49
740-744	5,368	4	75,756	56.49
745-749	6,938	5.17	82,694	61.66
750-754	5,900	4.40	88,594	66.06
755-759	5,232	3.90	93,826	69.96
760-764	5,900	4.40	99,726	74.36
765-769	5,139	3.83	104,865	78.20
770-774	4,413	3.29	109,278	81.49
775-779	3,829	2.86	113,107	84.34
780-784	4,184	3.12	117,291	87.46
785-789	3,770	2.81	121,061	90.27
790-794	2,939	2.19	124,000	92.46
795-799	2,086	1.56	126,086	94.02
800-804	1,838	1.37	127,924	95.39
805-809	1,526	1.14	129,450	96.53
810-814	1,289	0.96	130,739	97.49
815-819	788	0.59	131,527	98.08
820-824	880	0.66	132,407	98.73
825-829	369	0.28	132,776	99.01
830-834	393	0.29	133,169	99.30
835-839	236	0.18	133,405	99.48
840-844	183	0.14	133,588	99.61
845-850	519	0.39	134,107	100

Table A.12.43 Scale Score Cumulative Frequencies: Geometry

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,291	1.23	1,291	1.23
655-659	35	0.03	1,326	1.26
660-664	0	0	1,326	1.26
665-669	0	0	1,326	1.26
670-674	1,010	0.96	2,336	2.22
675-679	1,955	1.86	4,291	4.09
680-684	67	0.06	4,358	4.15
685-689	1,960	1.87	6,318	6.02
690-694	3,338	3.18	9,656	9.20
695-699	2,893	2.75	12,549	11.95
700-704	7,543	7.18	20,092	19.13
705-709	74	0.07	20,166	19.20
710-714	7,681	7.31	27,847	26.52
715-719	7,135	6.79	34,982	33.31
720-724	6,494	6.18	41,476	39.50
725-729	8,505	8.10	49,981	47.6
730-734	7,323	6.97	57,304	54.57
735-739	4,382	4.17	61,686	58.74
740-744	7,575	7.21	69,261	65.96
745-749	6,229	5.93	75,490	71.89
750-754	5,021	4.78	80,511	76.67
755-759	5,326	5.07	85,837	81.74
760-764	4,164	3.97	90,001	85.71
765-769	4,058	3.86	94,059	89.57
770-774	3,094	2.95	97,153	92.52
775-779	2,390	2.28	99,543	94.79
780-784	1,816	1.73	101,359	96.52
785-789	1,376	1.31	102,735	97.83
790-794	647	0.62	103,382	98.45
795-799	766	0.73	104,148	99.18
800-804	262	0.25	104,410	99.43
805-809	201	0.19	104,611	99.62
810-814	161	0.15	104,772	99.77
815-819	59	0.06	104,831	99.83
820-824	59	0.06	104,890	99.89
825-829	45	0.04	104,935	99.93
830-834	18	0.02	104,953	99.95
835-839	32	0.03	104,985	99.98
840-844	0	0	104,985	99.98
845-850	25	0.02	105,010	100

Table A.12.44 Scale Score Cumulative Frequencies: Algebra II

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,915	2.87	1,915	2.87
655-659	1,280	1.92	3,195	4.78
660-664	1,284	1.92	4,479	6.71
665-669	0	0	4,479	6.71
670-674	1,992	2.98	6,471	9.69
675-679	1,901	2.85	8,372	12.53
680-684	0	0	8,372	12.53
685-689	4,632	6.94	13,004	19.47
690-694	0	0	13,004	19.47
695-699	4,721	7.07	17,725	26.54
700-704	1,956	2.93	19,681	29.47
705-709	2,284	3.42	21,965	32.89
710-714	3,670	5.49	25,635	38.38
715-719	1,582	2.37	27,217	40.75
720-724	2,944	4.41	30,161	45.16
725-729	2,571	3.85	32,732	49.01
730-734	2,384	3.57	35,116	52.58
735-739	2,338	3.50	37,454	56.08
740-744	2,168	3.25	39,622	59.32
745-749	3,042	4.55	42,664	63.88
750-754	1,935	2.90	44,599	66.78
755-759	3,549	5.31	48,148	72.09
760-764	1,682	2.52	49,830	74.61
765-769	3,006	4.50	52,836	79.11
770-774	2,006	3	54,842	82.11
775-779	2,338	3.50	57,180	85.61
780-784	1,549	2.32	58,729	87.93
785-789	1,802	2.70	60,531	90.63
790-794	1,151	1.72	61,682	92.35
795-799	938	1.40	62,620	93.76
800-804	1,110	1.66	63,730	95.42
805-809	684	1.02	64,414	96.44
810-814	536	0.80	64,950	97.25
815-819	490	0.73	65,440	97.98
820-824	350	0.52	65,790	98.50
825-829	210	0.31	66,000	98.82
830-834	182	0.27	66,182	99.09
835-839	146	0.22	66,328	99.31
840-844	118	0.18	66,446	99.49
845-850	343	0.51	66,789	100

Table A.12.45 Scale Score Cumulative Frequencies: Integrated Mathematics I

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	11	1.63	11	1.63
655-659	25	3.71	36	5.35
660-664	0	0	36	5.35
665-669	0	0	36	5.35
670-674	0	0	36	5.35
675-679	50	7.43	86	12.78
680-684	0	0	86	12.78
685-689	62	9.21	148	21.99
690-694	0	0	148	21.99
695-699	59	8.77	207	30.76
700-704	0	0	207	30.76
705-709	74	11	281	41.75
710-714	71	10.55	352	52.30
715-719	0	0	352	52.30
720-724	49	7.28	401	59.58
725-729	33	4.90	434	64.49
730-734	38	5.65	472	70.13
735-739	28	4.16	500	74.29
740-744	23	3.42	523	77.71
745-749	22	3.27	545	80.98
750-754	32	4.75	577	85.74
755-759	17	2.53	594	88.26
760-764	12	1.78	606	90.04
765-769	17	2.53	623	92.57
770-774	8	1.19	631	93.76
775-779	11	1.63	642	95.39
780-784	8	1.19	650	96.58
785-789	6	0.89	656	97.47
790-794	5	0.74	661	98.22
795-799	3	0.45	664	98.66
800-804	1	0.15	665	98.81
805-809	2	0.30	667	99.11
810-814	1	0.15	668	99.26
815-819	1	0.15	669	99.41
820-824	2	0.30	671	99.70
825-829	1	0.15	672	99.85
830-834	0	0	672	99.85
835-839	0	0	672	99.85
840-844	0	0	672	99.85
845-850	1	0.15	673	100

Table A.12.46 Scale Score Cumulative Frequencies: Integrated Mathematics II

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	4	0.74	4	0.74
655-659	0	0	4	0.74
660-664	16	2.96	20	3.70
665-669	0	0	20	3.70
670-674	30	5.55	50	9.24
675-679	0	0	50	9.24
680-684	39	7.21	89	16.45
685-689	0	0	89	16.45
690-694	59	10.91	148	27.36
695-699	0	0	148	27.36
700-704	60	11.09	208	38.45
705-709	65	12.01	273	50.46
710-714	59	10.91	332	61.37
715-719	52	9.61	384	70.98
720-724	40	7.39	424	78.37
725-729	21	3.88	445	82.26
730-734	23	4.25	468	86.51
735-739	14	2.59	482	89.09
740-744	5	0.92	487	90.02
745-749	9	1.66	496	91.68
750-754	6	1.11	502	92.79
755-759	9	1.66	511	94.45
760-764	1	0.18	512	94.64
765-769	4	0.74	516	95.38
770-774	6	1.11	522	96.49
775-779	5	0.92	527	97.41
780-784	2	0.37	529	97.78
785-789	1	0.18	530	97.97
790-794	0	0	530	97.97
795-799	4	0.74	534	98.71
800-804	0	0	534	98.71
805-809	3	0.55	537	99.26
810-814	2	0.37	539	99.63
815-819	0	0	539	99.63
820-824	1	0.18	540	99.82
825-829	0	0	540	99.82
830-834	1	0.18	541	100
835-839	0	0	541	100
840-844	0	0	541	100
845-850	0	0	541	100

Table A.12.47 Scale Score Cumulative Frequencies: Integrated Mathematics III

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	12	5.97	12	5.97
655-659	0	0	12	5.97
660-664	14	6.97	26	12.94
665-669	0	0	26	12.94
670-674	24	11.94	50	24.88
675-679	0	0	50	24.88
680-684	23	11.44	73	36.32
685-689	0	0	73	36.32
690-694	26	12.94	99	49.25
695-699	21	10.45	120	59.70
700-704	12	5.97	132	65.67
705-709	9	4.48	141	70.15
710-714	9	4.48	150	74.63
715-719	8	3.98	158	78.61
720-724	2	1	160	79.60
725-729	5	2.49	165	82.09
730-734	7	3.48	172	85.57
735-739	4	1.99	176	87.56
740-744	3	1.49	179	89.05
745-749	5	2.49	184	91.54
750-754	4	1.99	188	93.53
755-759	0	0	188	93.53
760-764	0	0	188	93.53
765-769	1	0.50	189	94.03
770-774	3	1.49	192	95.52
775-779	2	1	194	96.52
780-784	1	0.50	195	97.01
785-789	3	1.49	198	98.51
790-794	0	0	198	98.51
795-799	1	0.50	199	99
800-804	1	0.50	200	99.50
805-809	0	0	200	99.50
810-814	0	0	200	99.50
815-819	0	0	200	99.50
820-824	0	0	200	99.50
825-829	0	0	200	99.50
830-834	0	0	200	99.50
835-839	0	0	200	99.50
840-844	0	0	200	99.50
845-850	1	0.50	201	100

Appendix 12.5: Subgroup Scale Score Performance

Table A.12.48 Subgroup Performance for ELA/L Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		256,870	738.54	42.05	650	850
Gender	Female	125,311	743.09	42.23	650	850
	Male	131,559	734.21	41.43	650	850
Ethnicity	American Indian/Alaska Native	4,217	715.63	36.47	650	850
	Asian	17,994	767.23	40.36	650	850
	Black/African American	38,832	722.71	41.08	650	850
	Hispanic/Latino	77,952	727.08	40.23	650	850
	Native Hawaiian/Pacific Islander	357	751.86	39.65	650	850
	Two or more races	8,340	743.03	42.71	650	850
	White	109,159	748.14	38.83	650	850
Economic Status*	Not Economically Disadvantaged	129,667	753.06	39.30	650	850
	Economically Disadvantaged	126,919	723.77	39.53	650	850
English Learner Status	Non English Learner	218,282	743.05	41.39	650	850
	English Learner	38,373	713.08	36.30	650	850
Disabilities	Students without Disabilities	212,957	744.14	40.21	650	850
	Students with Disabilities	43,174	711.14	40.16	650	850
Reading Summative Score		256,870	45.51	16.86	10	90
Gender	Female	125,311	46.64	16.74	10	90
	Male	131,559	44.44	16.91	10	90
Ethnicity	American Indian/Alaska Native	4,217	36.33	14.35	10	90
	Asian	17,994	56.08	16.27	10	90
	Black/African American	38,832	38.93	16.01	10	90
	Hispanic/Latino	77,952	40.45	15.82	10	90
	Native Hawaiian/Pacific Islander	357	49.90	15.57	10	90
	Two or more races	8,340	48.03	17.14	10	90
	White	109,159	49.88	15.79	10	90
Economic Status*	Not Economically Disadvantaged	129,667	51.50	15.90	10	90
	Economically Disadvantaged	126,919	39.42	15.58	10	90
English Learner Status	Non English Learner	218,282	47.44	16.61	10	90
	English Learner	38,373	34.62	13.89	10	90
Disabilities	Students without Disabilities	212,957	47.62	16.18	10	90
	Students with Disabilities	43,174	35.22	16.38	10	90

Group Type	Group	N	Mean	SD	Min	Max
Writing Summative Score		256,870	29.46	13.48	10	60
Gender	Female	125,311	31.43	13.20	10	60
	Male	131,559	27.59	13.49	10	60
Ethnicity	American Indian/Alaska Native	4,217	23.77	12.65	10	59
	Asian	17,994	37.67	11.66	10	60
	Black/African American	38,832	25.52	13.56	10	60
	Hispanic/Latino	77,952	27.00	13.39	10	60
	Native Hawaiian/Pacific Islander	357	33.95	12.35	10	60
	Two or more races	8,340	29.67	13.77	10	60
	White	109,159	31.46	12.83	10	60
Economic Status*	Not Economically Disadvantaged	129,667	33.09	12.64	10	60
	Economically Disadvantaged	126,919	25.77	13.31	10	60
English Learner Status	Non English Learner	218,282	30.47	13.34	10	60
	English Learner	38,373	23.78	12.88	10	60
Disabilities	Students without Disabilities	212,957	31.13	12.99	10	60
	Students with Disabilities	43,174	21.28	12.88	10	60

Note: This table is identical to Table 12.5 in Section 12. *Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table A.12.49 Subgroup Performance for ELA/L Scale Scores: Grade 4

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		265,169	742.91	38.37	650	850
Gender	Female	130,022	747.52	38.23	650	850
	Male	135,147	738.48	37.98	650	850
Ethnicity	American Indian/Alaska Native	4,359	721.74	32.81	650	850
	Asian	18,129	771.70	36.89	650	850
	Black/African American	39,358	726.36	35.67	650	850
	Hispanic/Latino	82,384	732.16	35.78	650	850
	Native Hawaiian/Pacific Islander	389	756.65	38.75	650	850
	Two or more races	8,286	746.68	39.14	650	850
	White	112,249	752.45	35.88	650	850
Economic Status*	Not Economically Disadvantaged	133,575	757.26	36.05	650	850
	Economically Disadvantaged	131,320	728.38	35.03	650	850
English Learner Status	Non English Learner	227,020	747.25	37.81	650	850
	English Learner	37,962	717.11	30.74	650	850
Disabilities	Students without Disabilities	218,058	748.60	36.19	650	850
	Students with Disabilities	46,416	716.36	37.16	650	850
Reading Summative Score		265,169	47.40	15.46	10	90
Gender	Female	130,022	48.43	15.28	10	90
	Male	135,147	46.41	15.57	10	90
Ethnicity	American Indian/Alaska Native	4,359	38.24	12.70	10	90
	Asian	18,129	58.28	15.17	10	90
	Black/African American	39,358	41.14	14.10	10	90
	Hispanic/Latino	82,384	42.91	14.18	10	90
	Native Hawaiian/Pacific Islander	389	52.21	15.32	10	90
	Two or more races	8,286	49.46	15.88	10	90
	White	112,249	51.33	14.76	10	90
Economic Status*	Not Economically Disadvantaged	133,575	53.05	14.85	10	90
	Economically Disadvantaged	131,320	41.69	13.88	10	90
English Learner Status	Non English Learner	227,020	49.16	15.30	10	90
	English Learner	37,962	36.99	11.92	10	90
Disabilities	Students without Disabilities	218,058	49.50	14.75	10	90
	Students with Disabilities	46,416	37.59	14.94	10	90
Writing Summative Score		265,169	31.79	11.69	10	60
Gender	Female	130,022	33.76	11.18	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	135,147	29.90	11.85	10	60
Ethnicity	American Indian/Alaska Native	4,359	27.42	11.22	10	60
	Asian	18,129	39.30	9.89	10	60
	Black/African American	39,358	27.12	11.92	10	60
	Hispanic/Latino	82,384	29.34	11.61	10	60
	Native Hawaiian/Pacific Islander	389	35.60	11.54	10	60
	Two or more races	8,286	32.15	11.83	10	60
	White	112,249	34.15	10.72	10	60
Economic Status*	Not Economically Disadvantaged	133,575	35.50	10.48	10	60
	Economically Disadvantaged	131,320	28.03	11.65	10	60
English Learner Status	Non English Learner	227,020	32.85	11.44	10	60
	English Learner	37,962	25.50	11.14	10	60
Disabilities	Students without Disabilities	218,058	33.50	10.84	10	60
	Students with Disabilities	46,416	23.82	12.23	10	60

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

Table A.12.50 Subgroup Performance for ELA/L Scale Scores: Grade 5

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		271,778	744.04	36.42	650	850
Gender	Female	133,464	749.34	36.41	650	850
	Male	138,314	738.93	35.68	650	850
Ethnicity	American Indian/Alaska Native	4,531	724.39	31.49	650	850
	Asian	18,361	773.18	35.27	650	850
	Black/African American	40,237	727.49	33.60	650	850
	Hispanic/Latino	84,649	733.81	33.93	650	850
	Native Hawaiian/Pacific Islander	397	755.47	38.30	650	850
	Two or more races	8,124	747.75	36.13	650	850
	White	115,461	753.15	33.81	650	850
Economic Status*	Not Economically Disadvantaged	137,097	757.73	34.31	650	850
	Economically Disadvantaged	134,366	730.15	33.06	650	850
English Learner Status	Non English Learner	240,709	748.08	35.52	650	850
	English Learner	30,849	712.74	26.87	650	847
Disabilities	Students without Disabilities	222,630	749.83	34.17	650	850
	Students with Disabilities	48,371	717.58	34.69	650	850
Reading Summative Score		271,778	47.69	14.67	10	90
Gender	Female	133,464	48.94	14.63	10	90
	Male	138,314	46.48	14.61	10	90
Ethnicity	American Indian/Alaska Native	4,531	39.29	12.59	10	90
	Asian	18,361	58.71	14.50	10	90
	Black/African American	40,237	41.12	13.21	10	90
	Hispanic/Latino	84,649	43.43	13.50	10	90
	Native Hawaiian/Pacific Islander	397	51.75	14.98	11	90
	Two or more races	8,124	49.60	14.69	10	90
	White	115,461	51.54	13.83	10	90
Economic Status*	Not Economically Disadvantaged	137,097	53.13	14.03	10	90
	Economically Disadvantaged	134,366	42.17	13.17	10	90
English Learner Status	Non English Learner	240,709	49.30	14.36	10	90
	English Learner	30,849	35.19	10.52	10	90
Disabilities	Students without Disabilities	222,630	49.84	13.90	10	90
	Students with Disabilities	48,371	37.87	14.13	10	90
Writing Summative Score		271,778	30.92	12.55	10	60
Gender	Female	133,464	33.50	11.74	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	138,314	28.42	12.81	10	60
Ethnicity	American Indian/Alaska Native	4,531	26.67	11.86	10	57
	Asian	18,361	39.26	10.04	10	60
	Black/African American	40,237	25.99	12.79	10	60
	Hispanic/Latino	84,649	28.35	12.51	10	60
	Native Hawaiian/Pacific Islander	397	33.97	12.97	10	57
	Two or more races	8,124	31.31	12.61	10	60
	White	115,461	33.32	11.62	10	60
Economic Status*	Not Economically Disadvantaged	137,097	34.81	11.29	10	60
	Economically Disadvantaged	134,366	26.96	12.54	10	60
English Learner Status	Non English Learner	240,709	32.06	12.21	10	60
	English Learner	30,849	22.07	11.66	10	57
Disabilities	Students without Disabilities	222,630	32.86	11.67	10	60
	Students with Disabilities	48,371	22.03	12.65	10	60

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

Table A.12.51 Subgroup Performance for ELA/L Scale Scores: Grade 6

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		275,277	743.02	34.47	650	850
Gender	Female	134,829	749.50	33.87	650	850
	Male	140,448	736.80	33.90	650	850
Ethnicity	American Indian/Alaska Native	4,387	725.43	29.34	650	850
	Asian	18,028	771.46	33.91	650	850
	Black/African American	40,795	727.02	32.13	650	850
	Hispanic/Latino	85,881	733.85	32.05	650	850
	Native Hawaiian/Pacific Islander	425	759.11	32.29	650	843
	Two or more races	7,914	744.82	34.45	650	850
	White	117,838	751.38	31.98	650	850
Economic Status*	Not Economically Disadvantaged	140,393	755.45	32.56	650	850
	Economically Disadvantaged	134,549	730.10	31.53	650	850
English Learner Status	Non English Learner	253,116	746.04	33.53	650	850
	English Learner	21,837	708.24	24.84	650	845
Disabilities	Students without Disabilities	225,755	748.66	32.33	650	850
	Students with Disabilities	48,406	716.80	32.06	650	850
Reading Summative Score		275,277	47.43	13.64	10	90
Gender	Female	134,829	49.20	13.41	10	90
	Male	140,448	45.73	13.65	10	90
Ethnicity	American Indian/Alaska Native	4,387	39.47	11.29	10	90
	Asian	18,028	58.15	13.69	10	90
	Black/African American	40,795	41.42	12.54	10	90
	Hispanic/Latino	85,881	43.63	12.56	10	90
	Native Hawaiian/Pacific Islander	425	52.53	12.49	10	86
	Two or more races	7,914	48.68	13.75	10	90
	White	117,838	50.84	12.81	10	90
Economic Status*	Not Economically Disadvantaged	140,393	52.32	13.04	10	90
	Economically Disadvantaged	134,549	42.35	12.33	10	90
English Learner Status	Non English Learner	253,116	48.63	13.30	10	90
	English Learner	21,837	33.61	9.35	10	90
Disabilities	Students without Disabilities	225,755	49.55	12.89	10	90
	Students with Disabilities	48,406	37.53	12.70	10	90
Writing Summative Score		275,277	30.53	12.42	10	60
Gender	Female	134,829	33.40	11.44	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	140,448	27.77	12.70	10	60
Ethnicity	American Indian/Alaska Native	4,387	27.00	11.92	10	60
	Asian	18,028	38.92	10.04	10	60
	Black/African American	40,795	25.13	12.72	10	60
	Hispanic/Latino	85,881	28.13	12.35	10	60
	Native Hawaiian/Pacific Islander	425	35.82	11.24	10	60
	Two or more races	7,914	30.35	12.58	10	60
	White	117,838	32.99	11.40	10	60
Economic Status*	Not Economically Disadvantaged	140,393	34.26	11.15	10	60
	Economically Disadvantaged	134,549	26.65	12.49	10	60
English Learner Status	Non English Learner	253,116	31.42	12.11	10	60
	English Learner	21,837	20.22	11.36	10	60
Disabilities	Students without Disabilities	225,755	32.38	11.61	10	60
	Students with Disabilities	48,406	21.90	12.47	10	60

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

Table A.12.52 Subgroup Performance for ELA/L Scale Scores: Grade 7

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		269,386	746.87	41.47	650	850
Gender	Female	132,281	754.63	40.62	650	850
	Male	137,105	739.39	40.91	650	850
Ethnicity	American Indian/Alaska Native	4,210	725.29	35.29	650	850
	Asian	17,922	779.94	39.06	650	850
	Black/African American	38,949	727.84	38.94	650	850
	Hispanic/Latino	82,439	734.88	38.90	650	850
	Native Hawaiian/Pacific Islander	406	759.62	43.19	650	850
	Two or more races	7,192	749.20	41.48	650	850
	White	118,262	757.07	38.34	650	850
Economic Status*	Not Economically Disadvantaged	141,686	761.52	38.77	650	850
	Economically Disadvantaged	127,411	730.64	38.20	650	850
English Learner Status	Non English Learner	251,349	750.05	40.27	650	850
	English Learner	17,734	702.12	31.01	650	841
Disabilities	Students without Disabilities	221,724	753.69	38.68	650	850
	Students with Disabilities	46,561	714.39	38.98	650	850
Reading Summative Score		269,386	48.95	16.57	10	90
Gender	Female	132,281	50.95	16.28	10	90
	Male	137,105	47.02	16.62	10	90
Ethnicity	American Indian/Alaska Native	4,210	39.27	14.11	10	90
	Asian	17,922	60.71	15.73	10	90
	Black/African American	38,949	41.61	15.37	10	90
	Hispanic/Latino	82,439	43.87	15.43	10	90
	Native Hawaiian/Pacific Islander	406	53.33	17.27	10	90
	Two or more races	7,192	50.74	16.62	10	90
	White	118,262	53.35	15.48	10	90
Economic Status*	Not Economically Disadvantaged	141,686	54.81	15.57	10	90
	Economically Disadvantaged	127,411	42.46	15.16	10	90
English Learner Status	Non English Learner	251,349	50.26	16.08	10	90
	English Learner	17,734	30.55	11.72	10	90
Disabilities	Students without Disabilities	221,724	51.55	15.56	10	90
	Students with Disabilities	46,561	36.53	15.59	10	90
Writing Summative Score		269,386	32.66	12.55	10	60
Gender	Female	132,281	35.60	11.69	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	137,105	29.81	12.69	10	60
Ethnicity	American Indian/Alaska Native	4,210	28.33	11.60	10	60
	Asian	17,922	41.80	10.55	10	60
	Black/African American	38,949	27.43	12.63	10	60
	Hispanic/Latino	82,439	29.98	12.25	10	60
	Native Hawaiian/Pacific Islander	406	36.18	12.75	10	60
	Two or more races	7,192	32.44	12.65	10	60
	White	118,262	35.01	11.69	10	60
Economic Status*	Not Economically Disadvantaged	141,686	36.39	11.50	10	60
	Economically Disadvantaged	127,411	28.52	12.37	10	60
English Learner Status	Non English Learner	251,349	33.42	12.29	10	60
	English Learner	17,734	21.94	11.26	10	60
Disabilities	Students without Disabilities	221,724	34.57	11.67	10	60
	Students with Disabilities	46,561	23.55	12.62	10	60

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

Table A.12.53 Subgroup Performance for ELA/L Scale Scores: Grade 8

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		266,251	746.55	42.10	650	850
Gender	Female	129,790	755.34	40.92	650	850
	Male	136,461	738.18	41.49	650	850
Ethnicity	American Indian/Alaska Native	3,944	724.49	34.52	650	850
	Asian	17,899	782.67	40.27	650	850
	Black/African American	37,562	727.27	38.67	650	850
	Hispanic/Latino	80,900	734.43	38.66	650	850
	Native Hawaiian/Pacific Islander	396	767.30	40.52	656	850
	Two or more races	6,751	747.51	42.17	650	850
	White	118,789	756.06	39.60	650	850
Economic Status*	Not Economically Disadvantaged	143,534	760.65	40.44	650	850
	Economically Disadvantaged	122,419	730.06	37.77	650	850
English Learner Status	Non English Learner	250,095	749.36	41.22	650	850
	English Learner	15,855	702.50	29.37	650	850
Disabilities	Students without Disabilities	218,947	753.02	39.98	650	850
	Students with Disabilities	46,260	715.98	38.33	650	850
Reading Summative Score		266,251	48.88	16.94	10	90
Gender	Female	129,790	51.44	16.63	10	90
	Male	136,461	46.44	16.87	10	90
Ethnicity	American Indian/Alaska Native	3,944	39.02	13.75	10	90
	Asian	17,899	62.38	16.49	10	90
	Black/African American	37,562	41.60	15.52	10	90
	Hispanic/Latino	80,900	43.89	15.45	10	90
	Native Hawaiian/Pacific Islander	396	56.17	16.49	10	90
	Two or more races	6,751	50.03	17.10	10	90
	White	118,789	52.77	16.12	10	90
Economic Status*	Not Economically Disadvantaged	143,534	54.40	16.45	10	90
	Economically Disadvantaged	122,419	42.42	15.12	10	90
English Learner Status	Non English Learner	250,095	50.01	16.62	10	90
	English Learner	15,855	31.17	11.16	10	87
Disabilities	Students without Disabilities	218,947	51.38	16.20	10	90
	Students with Disabilities	46,260	37.02	15.30	10	90
Writing Summative Score		266,251	31.99	13.04	10	60
Gender	Female	129,790	35.24	11.92	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	136,461	28.90	13.30	10	60
Ethnicity	American Indian/Alaska Native	3,944	27.76	11.87	10	60
	Asian	17,899	41.68	10.67	10	60
	Black/African American	37,562	26.34	13.02	10	60
	Hispanic/Latino	80,900	29.00	12.77	10	60
	Native Hawaiian/Pacific Islander	396	37.97	11.47	10	60
	Two or more races	6,751	31.53	13.22	10	60
	White	118,789	34.50	12.11	10	60
Economic Status*	Not Economically Disadvantaged	143,534	35.80	12.00	10	60
	Economically Disadvantaged	122,419	27.54	12.79	10	60
English Learner Status	Non English Learner	250,095	32.74	12.78	10	60
	English Learner	15,855	20.30	11.36	10	60
Disabilities	Students without Disabilities	218,947	33.85	12.29	10	60
	Students with Disabilities	46,260	23.25	12.96	10	60

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

Table A.12.54 Subgroup Performance for ELA/L Scale Scores: Grade 9

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		121,619	748.83	40.72	650	850
Gender	Female	59,248	755.81	39.13	650	850
	Male	62,371	742.20	41.09	650	850
Ethnicity	American Indian/Alaska Native	2,909	726.65	31.07	650	850
	Asian	10,492	783.32	36.13	650	850
	Black/African American	14,260	732.03	37.18	650	850
	Hispanic/Latino	42,054	733.28	37.91	650	850
	Native Hawaiian/Pacific Islander	239	759.91	41.41	650	850
	Two or more races	1,760	759.03	39.97	650	850
	White	49,897	760.38	36.58	650	850
Economic Status*	Not Economically Disadvantaged	73,488	760.42	38.92	650	850
	Economically Disadvantaged	48,075	731.16	36.85	650	850
English Learner Status	Non English Learner	114,062	752.17	39.11	650	850
	English Learner	7,505	698.38	29.92	650	821
Disabilities	Students without Disabilities	99,057	755.24	38.70	650	850
	Students with Disabilities	22,510	720.74	37.30	650	850
Reading Summative Score		121,619	49.70	16.45	10	90
Gender	Female	59,248	51.46	16.01	10	90
	Male	62,371	48.03	16.69	10	90
Ethnicity	American Indian/Alaska Native	2,909	40.24	13.02	10	87
	Asian	10,492	62.27	14.70	10	90
	Black/African American	14,260	43.18	15.21	10	90
	Hispanic/Latino	42,054	43.65	15.39	10	90
	Native Hawaiian/Pacific Islander	239	52.93	16.43	10	90
	Two or more races	1,760	54.03	16.33	10	90
	White	49,897	54.40	14.95	10	90
Economic Status*	Not Economically Disadvantaged	73,488	54.24	15.76	10	90
	Economically Disadvantaged	48,075	42.77	15.01	10	90
English Learner Status	Non English Learner	114,062	51.01	15.86	10	90
	English Learner	7,505	29.84	11.91	10	83
Disabilities	Students without Disabilities	99,057	52.07	15.75	10	90
	Students with Disabilities	22,510	39.28	15.39	10	90
Writing Summative Score		121,619	32.95	12.21	10	60
Gender	Female	59,248	35.66	11.05	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	62,371	30.37	12.69	10	60
Ethnicity	American Indian/Alaska Native	2,909	27.67	11.08	10	60
	Asian	10,492	42.24	9.58	10	60
	Black/African American	14,260	28.55	11.96	10	60
	Hispanic/Latino	42,054	28.84	12.12	10	60
	Native Hawaiian/Pacific Islander	239	36.95	11.34	10	60
	Two or more races	1,760	35.34	11.64	10	60
	White	49,897	35.92	10.86	10	60
Economic Status*	Not Economically Disadvantaged	73,488	36.03	11.28	10	60
	Economically Disadvantaged	48,075	28.24	12.06	10	60
English Learner Status	Non English Learner	114,062	33.84	11.76	10	60
	English Learner	7,505	19.36	10.64	10	50
Disabilities	Students without Disabilities	99,057	34.90	11.27	10	60
	Students with Disabilities	22,510	24.37	12.47	10	60

Note: This table is identical to Table 12.6 in Section 12. *Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table A.12.55 Subgroup Performance for ELA/L Scale Scores: Grade 10

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		118,322	752.65	50.15	650	850
Gender	Female	58,513	761.55	48.29	650	850
	Male	59,809	743.93	50.40	650	850
Ethnicity	American Indian/Alaska Native	2,673	727.25	40.09	650	850
	Asian	10,443	791.96	44.26	650	850
	Black/African American	14,094	732.54	47.11	650	850
	Hispanic/Latino	38,972	734.45	46.87	650	850
	Native Hawaiian/Pacific Islander	270	761.29	48.95	650	850
	Two or more races	1,632	763.59	49.21	650	850
	White	50,232	765.18	46.10	650	850
Economic Status*	Not Economically Disadvantaged	74,425	765.20	48.09	650	850
	Economically Disadvantaged	43,853	731.37	46.25	650	850
English Learner Status	Non English Learner	111,641	756.26	48.56	650	850
	English Learner	6,641	692.04	35.47	650	839
Disabilities	Students without Disabilities	97,165	760.07	47.71	650	850
	Students with Disabilities	21,115	718.54	46.84	650	850
Reading Summative Score		118,322	50.86	19.87	10	90
Gender	Female	58,513	53.07	19.28	10	90
	Male	59,809	48.70	20.21	10	90
Ethnicity	American Indian/Alaska Native	2,673	39.24	16.50	10	90
	Asian	10,443	64.93	18.04	10	90
	Black/African American	14,094	43.56	18.78	10	90
	Hispanic/Latino	38,972	43.78	18.67	10	90
	Native Hawaiian/Pacific Islander	270	53.86	19.86	10	90
	Two or more races	1,632	55.70	19.36	10	90
	White	50,232	55.92	18.33	10	90
Economic Status*	Not Economically Disadvantaged	74,425	55.81	19.05	10	90
	Economically Disadvantaged	43,853	42.46	18.36	10	90
English Learner Status	Non English Learner	111,641	52.30	19.24	10	90
	English Learner	6,641	26.68	13.86	10	90
Disabilities	Students without Disabilities	97,165	53.60	19.01	10	90
	Students with Disabilities	21,115	38.27	18.90	10	90
Writing Summative Score		118,322	34.37	13.67	10	60
Gender	Female	58,513	37.44	12.76	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	59,809	31.36	13.87	10	60
Ethnicity	American Indian/Alaska Native	2,673	29.72	11.56	10	60
	Asian	10,443	44.44	11.32	10	60
	Black/African American	14,094	29.12	13.42	10	60
	Hispanic/Latino	38,972	30.04	13.23	10	60
	Native Hawaiian/Pacific Islander	270	36.90	12.69	10	60
	Two or more races	1,632	36.45	13.65	10	60
	White	50,232	37.27	12.66	10	60
Economic Status*	Not Economically Disadvantaged	74,425	37.35	13.03	10	60
	Economically Disadvantaged	43,853	29.32	13.24	10	60
English Learner Status	Non English Learner	111,641	35.21	13.34	10	60
	English Learner	6,641	20.17	11.11	10	58
Disabilities	Students without Disabilities	97,165	36.31	12.95	10	60
	Students with Disabilities	21,115	25.46	13.35	10	60

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

Table A.12.56 Subgroup Performance for ELA/L Scale Scores: Grade 11

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		34,610	738.46	39.39	650	850
Gender	Female	16,651	746.21	38.23	650	850
	Male	17,959	731.27	39.09	650	850
Ethnicity	American Indian/Alaska Native	2,935	735.21	34.83	650	837
	Asian	784	755.77	44.40	650	850
	Black/African American	4,606	726.16	37.07	650	850
	Hispanic/Latino	18,013	736.82	38.18	650	850
	Native Hawaiian/Pacific Islander	58	752.72	37.69	664	844
	Two or more races	287	743.33	42.14	650	850
	White	7,921	748.56	41.38	650	850
Economic Status*	Not Economically Disadvantaged	14,074	745.44	40.53	650	850
	Economically Disadvantaged	20,497	733.72	37.85	650	850
English Learner Status	Non English Learner	30,538	742.43	38.52	650	850
	English Learner	4,038	708.61	32.43	650	827
Disabilities	Students without Disabilities	28,539	744.71	37.55	650	850
	Students with Disabilities	6,032	709.04	34.21	650	850
Reading Summative Score		34,610	45.49	15.62	10	90
Gender	Female	16,651	47.49	15.25	10	90
	Male	17,959	43.63	15.72	10	90
Ethnicity	American Indian/Alaska Native	2,935	42.55	13.61	10	89
	Asian	784	51.70	17.92	10	90
	Black/African American	4,606	40.70	14.30	10	90
	Hispanic/Latino	18,013	44.81	14.95	10	90
	Native Hawaiian/Pacific Islander	58	51.60	15.93	10	83
	Two or more races	287	47.65	17.09	10	86
	White	7,921	50.18	16.78	10	90
Economic Status*	Not Economically Disadvantaged	14,074	48.65	16.25	10	90
	Economically Disadvantaged	20,497	43.34	14.78	10	90
English Learner Status	Non English Learner	30,538	47.11	15.33	10	90
	English Learner	4,038	33.31	11.97	10	80
Disabilities	Students without Disabilities	28,539	47.79	15.05	10	90
	Students with Disabilities	6,032	34.67	13.56	10	90
Writing Summative Score		34,610	28.92	13.67	10	60
Gender	Female	16,651	32.23	12.83	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	17,959	25.84	13.70	10	60
Ethnicity	American Indian/Alaska Native	2,935	29.58	12.79	10	60
	Asian	784	34.16	13.77	10	60
	Black/African American	4,606	25.52	13.43	10	60
	Hispanic/Latino	18,013	28.47	13.55	10	60
	Native Hawaiian/Pacific Islander	58	32.84	12.85	10	58
	Two or more races	287	30.28	13.78	10	60
	White	7,921	31.06	13.84	10	60
Economic Status*	Not Economically Disadvantaged	14,074	30.55	13.67	10	60
	Economically Disadvantaged	20,497	27.80	13.56	10	60
English Learner Status	Non English Learner	30,538	29.95	13.50	10	60
	English Learner	4,038	21.15	12.36	10	53
Disabilities	Students without Disabilities	28,539	30.90	13.11	10	60
	Students with Disabilities	6,032	19.56	12.30	10	60

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

Table A.12.57 Subgroup Performance for Mathematics Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		258,807	743.16	36.43	650	850
Gender	Female	126,222	742.86	35.48	650	850
	Male	132,585	743.45	37.31	650	850
Ethnicity	American Indian/Alaska Native	4,212	722.11	31.06	650	834
	Asian	18,134	773.87	33.90	650	850
	Black/African American	38,801	725.28	34.17	650	850
	Hispanic/Latino	79,715	733.21	33.33	650	850
	Native Hawaiian/Pacific Islander	357	753.43	35.90	650	850
	Two or more races	8,326	746.03	37.65	650	850
	White	109,242	752.24	33.69	650	850
Economic Status*	Not Economically Disadvantaged	130,196	756.72	34.10	650	850
	Economically Disadvantaged	128,322	729.47	33.44	650	850
English Learner Status	Non English Learner	218,081	746.48	36.31	650	850
	English Learner	40,509	725.46	31.59	650	850
Disabilities	Students without Disabilities	214,859	747.35	34.86	650	850
	Students with Disabilities	43,202	722.52	37.02	650	850
Language Form	Spanish	4,812	714.60	31.61	650	825

Note: This table is identical to Table 12.7 in Section 12. *Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table A.12.58 Subgroup Performance for Mathematics Scale Scores: Grade 4

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		266,629	739.36	34.87	650	850
Gender	Female	130,725	739.06	34.00	650	850
	Male	135,904	739.65	35.68	650	850
Ethnicity	American Indian/Alaska Native	4,359	719.29	30.54	650	850
	Asian	18,252	770.85	32.46	650	850
	Black/African American	39,261	721.10	32.35	650	850
	Hispanic/Latino	83,788	729.47	31.58	650	850
	Native Hawaiian/Pacific Islander	389	751.87	31.89	650	850
	Two or more races	8,276	741.35	35.72	650	850
	White	112,286	748.60	31.92	650	850
Economic Status*	Not Economically Disadvantaged	133,984	752.80	32.52	650	850
	Economically Disadvantaged	132,367	725.82	31.76	650	850
English Learner Status	Non English Learner	226,704	742.88	34.63	650	850
	English Learner	39,735	719.41	28.96	650	850
Disabilities	Students without Disabilities	219,546	743.89	33.17	650	850
	Students with Disabilities	46,386	718.03	34.81	650	850
Language Form	Spanish	3,691	707.84	24.13	650	818

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

Table A.12.59 Subgroup Performance for Mathematics Scale Scores: Grade 5

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		272,714	737.90	33.04	650	850
Gender	Female	133,881	738.11	31.70	650	850
	Male	138,833	737.69	34.29	650	850
Ethnicity	American Indian/Alaska Native	4,529	719.67	27.11	650	850
	Asian	18,489	770.39	33.28	650	850
	Black/African American	40,145	720.70	27.70	650	850
	Hispanic/Latino	85,519	728.32	28.55	650	850
	Native Hawaiian/Pacific Islander	396	747.06	34.61	658	850
	Two or more races	8,123	740.37	34.13	650	850
	White	115,494	746.28	31.50	650	850
Economic Status*	Not Economically Disadvantaged	137,489	750.77	32.45	650	850
	Economically Disadvantaged	134,911	724.84	28.13	650	850
English Learner Status	Non English Learner	240,393	741.03	32.88	650	850
	English Learner	32,102	714.65	23.64	650	850
Disabilities	Students without Disabilities	223,663	742.03	32.28	650	850
	Students with Disabilities	48,276	718.87	29.69	650	850
Language Form	Spanish	3,270	705.77	25.74	650	798

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

Table A.12.60 Subgroup Performance for Mathematics Scale Scores: Grade 6

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		275,732	732.71	32.62	650	850
Gender	Female	135,084	733.68	31.70	650	850
	Male	140,648	731.77	33.45	650	850
Ethnicity	American Indian/Alaska Native	4,365	715.05	27.20	650	847
	Asian	18,072	765.13	32.78	650	850
	Black/African American	40,584	714.90	28.72	650	850
	Hispanic/Latino	86,697	723.15	28.62	650	850
	Native Hawaiian/Pacific Islander	421	745.71	32.57	650	850
	Two or more races	7,901	734.33	33.26	650	850
	White	117,682	741.42	30.13	650	850
Economic Status*	Not Economically Disadvantaged	140,461	745.18	31.32	650	850
	Economically Disadvantaged	134,955	719.76	28.64	650	850
English Learner Status	Non English Learner	252,346	735.28	32.05	650	850
	English Learner	23,079	704.72	24.67	650	833
Disabilities	Students without Disabilities	226,403	737.25	31.21	650	850
	Students with Disabilities	48,234	711.36	30.60	650	850
Language Form	Spanish	2,642	704.24	23.78	650	787

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

Table A.12.61 Subgroup Performance for Mathematics Scale Scores: Grade 7

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		264,960	737.26	30.54	650	850
Gender	Female	130,334	737.79	29.65	650	850
	Male	134,626	736.76	31.37	650	850
Ethnicity	American Indian/Alaska Native	4,176	719.79	24.88	650	830
	Asian	15,958	765.59	31.36	650	850
	Black/African American	38,511	721.29	26.47	650	850
	Hispanic/Latino	82,618	728.38	27.02	650	850
	Native Hawaiian/Pacific Islander	391	745.16	32.35	650	839
	Two or more races	7,039	739.44	32.33	650	850
	White	116,261	745.45	28.67	650	850
Economic Status*	Not Economically Disadvantaged	137,571	748.20	29.67	650	850
	Economically Disadvantaged	127,101	725.46	26.84	650	850
English Learner Status	Non English Learner	245,777	739.30	30.09	650	850
	English Learner	18,879	710.96	23.32	650	850
Disabilities	Students without Disabilities	217,645	741.88	28.88	650	850
	Students with Disabilities	46,217	715.48	28.77	650	850
Language Form	Spanish	2,255	707.48	20.80	650	781

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

Table A.12.62 Subgroup Performance for Mathematics Scale Scores: Grade 8

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		225,726	728.24	38.46	650	850
Gender	Female	109,096	730.49	37.24	650	850
	Male	116,630	726.14	39.46	650	850
Ethnicity	American Indian/Alaska Native	3,663	708.55	31.76	650	848
	Asian	10,576	761.90	40.41	650	850
	Black/African American	34,129	709.76	33.27	650	850
	Hispanic/Latino	72,286	718.85	34.72	650	850
	Native Hawaiian/Pacific Islander	277	743.73	39.08	650	850
	Two or more races	5,896	730.02	40.95	650	850
	White	98,890	738.47	37.06	650	850
Economic Status*	Not Economically Disadvantaged	112,889	741.01	37.87	650	850
	Economically Disadvantaged	112,539	715.46	34.61	650	850
English Learner Status	Non English Learner	208,985	730.55	38.18	650	850
	English Learner	16,439	699.06	28.90	650	829
Disabilities	Students without Disabilities	181,576	733.94	37.15	650	850
	Students with Disabilities	43,104	704.09	34.30	650	850
Language Form	Spanish	2,118	695.86	25.51	650	827

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

Table A.12.63 Subgroup Performance for Mathematics Scale Scores: Algebra I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		134,107	740.37	36.46	650	850
Gender	Female	65,087	741.24	35.26	650	850
	Male	69,020	739.56	37.54	650	850
Ethnicity	American Indian/Alaska Native	3,093	716.67	26.37	650	850
	Asian	11,399	776.91	36.45	650	850
	Black/African American	16,464	724.66	30.36	650	850
	Hispanic/Latino	47,009	725.71	30.44	650	850
	Native Hawaiian/Pacific Islander	285	748.75	35.02	661	832
	Two or more races	2,117	750.40	38.71	650	850
	White	53,728	751.19	33.72	650	850
Economic Status*	Not Economically Disadvantaged	79,819	751.08	36.54	650	850
	Economically Disadvantaged	54,217	724.65	30.08	650	850
English Learner Status	Non English Learner	124,436	742.77	36.05	650	850
	English Learner	9,604	709.61	26.33	650	850
Disabilities	Students without Disabilities	109,545	744.89	36.06	650	850
	Students with Disabilities	24,497	720.25	31.04	650	850
Language Form	Spanish	2,426	701.47	24.27	650	795

Note: This table is identical to Table 12.8 in Section 12. *Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table A.12.64 Subgroup Performance for Mathematics Scale Scores: Geometry

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		105,010	732.24	29.24	650	850
Gender	Female	52,124	733.09	28.02	650	850
	Male	52,886	731.40	30.36	650	850
Ethnicity	American Indian/Alaska Native	2,743	717.38	24.46	650	795
	Asian	8,727	759.47	28.81	650	850
	Black/African American	12,246	718.41	25.90	650	831
	Hispanic/Latino	35,728	721.24	25.49	650	828
	Native Hawaiian/Pacific Islander	235	738.56	29.78	650	850
	Two or more races	1,437	740.74	29.68	650	828
	White	43,888	740.25	26.70	650	850
Economic Status*	Not Economically Disadvantaged	65,206	739.82	28.82	650	850
	Economically Disadvantaged	39,764	719.84	25.42	650	850
English Learner Status	Non English Learner	98,467	733.89	28.88	650	850
	English Learner	6,508	707.33	22.52	650	825
Disabilities	Students without Disabilities	87,509	735.71	28.44	650	850
	Students with Disabilities	17,463	714.89	26.92	650	850
Language Form	Spanish	1,728	704.01	21.00	650	778

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

Table A.12.65 Subgroup Performance for Mathematics Scale Scores: Algebra II

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		66,789	732.16	43.25	650	850
Gender	Female	34,144	731.04	40.67	650	850
	Male	32,645	733.33	45.77	650	850
Ethnicity	American Indian/Alaska Native	2,691	703.59	29.45	650	837
	Asian	8,084	772.25	36.91	650	850
	Black/African American	6,882	706.91	35.67	650	848
	Hispanic/Latino	22,694	712.22	35.61	650	850
	Native Hawaiian/Pacific Islander	140	744.49	40.28	650	826
	Two or more races	892	747.78	44.01	650	850
	White	25,401	746.48	38.71	650	850
Economic Status*	Not Economically Disadvantaged	41,685	745.65	42.15	650	850
	Economically Disadvantaged	25,087	709.78	34.98	650	850
English Learner Status	Non English Learner	62,905	734.80	42.70	650	850
	English Learner	3,872	689.52	26.62	650	850
Disabilities	Students without Disabilities	59,543	735.44	42.39	650	850
	Students with Disabilities	7,228	705.28	40.80	650	850
Language Form	Spanish	558	680.16	24.11	650	763

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

Table A.12.66 Subgroup Performance for Mathematics Scale Scores: Integrated Mathematics I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		673	718.84	33.39	650	845
Gender	Female	321	721.19	33.54	650	845
	Male	352	716.70	33.15	650	822
Ethnicity	American Indian/Alaska Native	28	718.50	31.86	659	799
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	27	714.56	24.14	659	750
	Hispanic/Latino	415	711.46	28.18	650	794
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or More Races	n/r	n/r	n/r	n/r	n/r
	White	192	732.93	38.88	650	845
Economic Status*	Not Economically Disadvantaged	218	737.70	35.35	650	845
	Economically Disadvantaged	448	709.92	28.37	650	807
English Learner Status	Non English Learner	554	721.88	34.31	650	845
	English Learner	112	704.83	24.44	650	778
Disabilities	Students without Disabilities	525	723.98	33.40	650	845
	Students with Disabilities	140	700.86	26.27	650	828
Language Form	Spanish	n/r	n/r	n/r	n/r	n/r

Note: This table is identical to Table 12.9 in Section 12. *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.

Table A.12.67 Subgroup Performance for Mathematics Scale Scores: Integrated Mathematics II

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		541	711.28	27.80	650	832
Gender	Female	245	712.42	28.44	650	812
	Male	296	710.33	27.27	650	832
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	n/r	n/r	n/r	n/r	n/r
	Hispanic/Latino	373	706.85	21.43	650	805
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	n/r	n/r	n/r	n/r	n/r
	White	123	722.27	36.70	650	832
Economic Status*	Not Economically Disadvantaged	186	720.77	35.19	650	832
	Economically Disadvantaged	351	706.37	21.52	650	799
English Learner Status	Non English Learner	417	714.09	29.41	650	832
	English Learner	120	701.85	18.95	650	741
Disabilities	Students without Disabilities	441	713.17	28.00	650	824
	Students with Disabilities	96	703.05	25.83	660	832
Language Form	Spanish	n/r	n/r	n/r	n/r	n/r

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.

Table A.12.68 Subgroup Performance for Mathematics Scale Scores: Integrated Mathematics III

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		201	700.94	33.77	650	850
Gender	Female	108	701.78	31.77	650	802
	Male	93	699.96	36.11	650	850
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	n/r	n/r	n/r	n/r	n/r
	Hispanic/Latino	153	693.18	26.24	650	789
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or More Races	n/r	n/r	n/r	n/r	n/r
	White	39	731.97	39.91	652	850
Economic Status*	Not Economically Disadvantaged	34	744.53	36.25	674	850
	Economically Disadvantaged	167	692.06	25.39	650	789
English Learner Status	Non English Learner	157	703.78	36.59	650	850
	English Learner	44	690.80	17.66	664	733
Disabilities	Students without Disabilities	174	702.61	32.04	650	802
	Students with Disabilities	27	690.15	42.46	650	850
Language Form	Spanish	n/r	n/r	n/r	n/r	n/r

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.

Appendix 13.1: Reliability of Classification by Content and Grade/Subject

Table A.13.1 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 3

	Max. Raw Score	Avg. SEM	Avg. Reliability	Min. Sample Size	Min. Reliability	Max. Sample Size	Max. Reliability
Total Group	54	3.60	0.88	7,522	0.81	108,352	0.89
Gender							
Male	54	3.52	0.88	4,828	0.82	54,968	0.89
Female	54	3.69	0.88	2,694	0.80	53,384	0.89
Ethnicity							
White	54	3.66	0.86	5,831	0.82	247	0.87
Black/African American	54	3.50	0.86	1,376	0.74	13,302	0.89
Asian/Pacific Islander	54	3.82	0.87	1,176	0.85	8,180	0.87
American Indian/Alaska Native	54	3.39	0.84	962	0.78	1,610	0.86
Hispanic/Latino	54	3.53	0.86	2,599	0.77	31,369	0.88
Multiple	54	3.59	0.88	211	0.78	3,822	0.89
Special Instruction Needs							
Economically Disadvantaged	54	3.49	0.86	4,294	0.76	49,138	0.88
Not Economically Disadvantaged	54	3.71	0.86	3,224	0.83	251	0.88
English Learner	54	3.39	0.83	916	0.72	14,687	0.85
Non-English Learner	54	3.64	0.87	6,606	0.81	93,607	0.89
Students with Disabilities	54	3.15	0.87	7,522	0.81	15,215	0.90
Students without Disabilities	54	3.69	0.87	27,177	0.83	92,866	0.88
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	54	3.41	0.89	119	0.89	119	0.89
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	54	2.81	0.80	7,403	0.80	7,403	0.80

n/r = not reported due to n<100.

Table A.13.2 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 4

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	70	4.48	0.88	664	0.81	112,153	0.89
Gender							
Male	70	4.39	0.88	415	0.82	56,746	0.89
Female	70	4.56	0.88	249	0.80	55,407	0.89
Ethnicity							
White	70	4.57	0.86	208	0.80	47,855	0.87
Black/African American	70	4.31	0.86	102	0.72	16,192	0.87
Asian/Pacific Islander	70	4.63	0.87	9,705	0.86	249	0.89
American Indian/Alaska Native	71	4.39	0.82	922	0.73	1,505	0.86
Hispanic/Latino	70	4.37	0.86	2,883	0.80	34,951	0.87
Multiple	70	4.51	0.88	223	0.81	3,494	0.89
Special Instruction Needs							
Economically Disadvantaged	70	4.34	0.85	4,883	0.77	55,033	0.86
Not Economically Disadvantaged	70	4.60	0.86	223	0.81	57,063	0.87
English Learner	70	4.19	0.80	981	0.66	16,253	0.81
Non-English Learner	70	4.52	0.87	543	0.81	95,862	0.88
Students with Disabilities	71	4.01	0.87	530	0.80	16,048	0.90
Students without Disabilities	70	4.57	0.87	134	0.80	95,863	0.88
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	74	4.47	0.89	153	0.89	153	0.89
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.71	0.81	8,293	0.81	8,293	0.81

n/r = not reported due to n<100.

Table A.13.3 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 5

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	71	4.46	0.88	500	0.77	145,676	0.89
Gender							
Male	71	4.35	0.88	325	0.78	73,486	0.89
Female	71	4.53	0.88	175	0.75	72,190	0.89
Ethnicity							
White	71	4.48	0.87	175	0.80	62,261	0.88
Black/African American	71	4.34	0.86	1,878	0.77	20,969	0.87
Asian/Pacific Islander	71	4.48	0.88	7,733	0.86	10,280	0.89
American Indian/Alaska Native	72	4.36	0.82	953	0.72	1,968	0.85
Hispanic/Latino	71	4.40	0.86	201	0.74	45,543	0.87
Multiple	71	4.48	0.88	240	0.81	4,425	0.89
Special Instruction Needs							
Economically Disadvantaged	71	4.36	0.85	306	0.72	71,184	0.86
Not Economically Disadvantaged	71	4.49	0.87	194	0.80	74,421	0.88
English Learner	71	4.00	0.77	118	0.56	16,660	0.78
Non-English Learner	71	4.48	0.88	382	0.79	128,964	0.89
Students with Disabilities	72	4.10	0.87	482	0.76	21,058	0.89
Students without Disabilities	71	4.50	0.88	2,364	0.79	124,291	0.88
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	74	4.57	0.91	216	0.91	216	0.91
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.95	0.82	8,980	0.82	8,980	0.82

n/r = not reported due to n<100.

Table A.13.4 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 6

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	72	4.49	0.91	3,931	0.87	129,435	0.92
Gender							
Male	72	4.34	0.91	2,090	0.86	65,215	0.92
Female	72	4.63	0.90	1,841	0.86	64,220	0.91
Ethnicity							
White	72	4.58	0.89	1,705	0.87	162	0.91
Black/African American	72	4.26	0.89	531	0.82	18,682	0.91
Asian/Pacific Islander	72	4.60	0.90	8,909	0.88	8,765	0.91
American Indian/Alaska Native	72	4.50	0.86	866	0.79	1,781	0.88
Hispanic/Latino	72	4.42	0.89	660	0.83	40,522	0.91
Multiple	72	4.47	0.91	114	0.86	3,736	0.92
Special Instruction Needs							
Economically Disadvantaged	72	4.35	0.89	5,181	0.84	62,734	0.90
Not Economically Disadvantaged	72	4.60	0.89	1,669	0.86	160	0.92
English Learner	72	3.85	0.80	619	0.75	10,010	0.83
Non-English Learner	72	4.53	0.90	8,388	0.87	119,316	0.91
Students with Disabilities	72	3.97	0.90	1,219	0.82	18,561	0.92
Students without Disabilities	72	4.59	0.90	2,585	0.84	110,370	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	74	4.28	0.91	193	0.91	193	0.91
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.87	0.87	8,928	0.87	8,928	0.87

n/r = not reported due to n<100.

Table A.13.5 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 7

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	72	4.88	0.91	353	0.79	126,778	0.91
Gender							
Male	72	4.77	0.91	225	0.80	64,020	0.91
Female	72	4.98	0.90	128	0.77	62,758	0.91
Ethnicity							
White	72	4.92	0.89	131	0.76	55,706	0.90
Black/African American	72	4.77	0.89	1,736	0.83	18,032	0.90
Asian/Pacific Islander	72	4.80	0.89	8,847	0.88	186	0.92
American Indian/Alaska Native	72	4.98	0.85	666	0.78	1,740	0.87
Hispanic/Latino	72	4.86	0.89	132	0.78	38,892	0.90
Multiple	72	4.85	0.91	149	0.85	3,474	0.92
Special Instruction Needs							
Economically Disadvantaged	72	4.83	0.89	246	0.79	59,400	0.89
Not Economically Disadvantaged	72	4.92	0.89	107	0.80	67,256	0.90
English Learner	72	4.37	0.81	648	0.72	8,199	0.83
Non-English Learner	72	4.91	0.90	289	0.81	118,422	0.91
Students with Disabilities	72	4.50	0.90	346	0.80	18,095	0.91
Students without Disabilities	72	4.95	0.90	2,550	0.84	108,176	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	74	4.43	0.92	227	0.92	227	0.92
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	4.29	0.87	8,124	0.87	8,124	0.87

n/r = not reported due to n<100.

Table A.13.6 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 8

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	72	4.60	0.91	377	0.87	126,674	0.92
Gender							
Male	72	4.49	0.91	5,147	0.87	64,297	0.92
Female	72	4.69	0.90	118	0.86	62,377	0.91
Ethnicity							
White	72	4.67	0.90	57,492	0.89	56,778	0.91
Black/African American	72	4.44	0.89	1,763	0.82	516	0.91
Asian/Pacific Islander	72	4.52	0.90	8,856	0.89	8,816	0.92
American Indian/Alaska Native	72	4.64	0.86	612	0.85	1,668	0.87
Hispanic/Latino	72	4.54	0.89	134	0.82	38,702	0.90
Multiple	72	4.61	0.91	159	0.88	3,185	0.92
Special Instruction Needs							
Economically Disadvantaged	72	4.50	0.89	238	0.84	57,541	0.90
Not Economically Disadvantaged	72	4.66	0.90	69,445	0.89	69,019	0.91
English Learner	72	3.96	0.80	560	0.70	439	0.86
Non-English Learner	72	4.63	0.91	322	0.88	119,136	0.92
Students with Disabilities	72	4.17	0.90	363	0.87	18,329	0.91
Students without Disabilities	72	4.67	0.90	1,789	0.86	107,865	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	74	4.34	0.94	255	0.94	255	0.94
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.81	0.87	7,642	0.87	7,642	0.87

n/r = not reported due to n<100.

Table A.13.7 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 9

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	71	4.67	0.91	555	0.82	34,577	0.92
Gender							
Male	71	4.55	0.91	295	0.83	17,515	0.92
Female	71	4.79	0.90	260	0.79	17,062	0.91
Ethnicity							
White	71	4.77	0.89	131	0.82	13,746	0.90
Black/African American	71	4.63	0.89	802	0.84	3,445	0.89
Asian/Pacific Islander	71	4.56	0.89	7,043	0.89	112	0.92
American Indian/Alaska Native	72	4.45	0.85	151	0.77	1,582	0.86
Hispanic/Latino	71	4.56	0.90	198	0.81	12,441	0.90
Multiple	71	4.82	0.91	1,186	0.90	501	0.92
Special Instruction Needs							
Economically Disadvantaged	71	4.55	0.89	347	0.81	13,927	0.90
Not Economically Disadvantaged	71	4.74	0.90	176	0.84	20,646	0.91
English Learner	71	3.79	0.81	111	0.75	2,172	0.82
Non-English Learner	71	4.72	0.90	356	0.83	32,401	0.91
Students with Disabilities	72	4.31	0.90	226	0.87	5,120	0.91
Students without Disabilities	71	4.75	0.91	297	0.79	29,454	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.85	0.87	3,915	0.87	3,915	0.87

n/r = not reported due to n<100.

Table A.13.8 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 10

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	72	4.87	0.90	571	0.81	64,769	0.92
Gender							
Male	72	4.77	0.91	320	0.81	32,425	0.92
Female	72	4.95	0.90	251	0.78	32,344	0.91
Ethnicity							
White	72	4.95	0.89	166	0.86	27,692	0.90
Black/African American	72	4.82	0.88	630	0.77	7,519	0.89
Asian/Pacific Islander	72	4.79	0.89	4,484	0.86	5,811	0.91
American Indian/Alaska Native	72	4.65	0.85	159	0.73	1,406	0.87
Hispanic/Latino	72	4.75	0.89	179	0.80	21,342	0.90
Multiple	72	5.03	0.90	713	0.87	849	0.91
Special Instruction Needs							
Economically Disadvantaged	72	4.74	0.89	331	0.79	23,781	0.89
Not Economically Disadvantaged	72	4.93	0.90	210	0.86	40,982	0.91
English Learner	72	3.93	0.81	159	0.74	3,598	0.81
Non-English Learner	72	4.91	0.90	382	0.84	61,166	0.91
Students with Disabilities	73	4.50	0.89	275	0.86	10,119	0.91
Students without Disabilities	72	4.93	0.90	266	0.75	54,645	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	4.08	0.85	2,948	0.85	2,948	0.85

n/r = not reported due to n<100.

Table A.13.9 Summary of Test Reliability Estimates for Subgroups: ELA/L Grade 11

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	72	4.83	0.87	574	0.75	16,373	0.89
Gender							
Male	72	4.61	0.87	324	0.72	8,473	0.89
Female	72	5.06	0.86	250	0.76	7,900	0.88
Ethnicity							
White	72	4.91	0.88	238	0.81	3,661	0.91
Black/African American	72	4.75	0.83	174	0.78	2,124	0.85
Asian/Pacific Islander	72	4.98	0.90	393	0.90	374	0.91
American Indian/Alaska Native	72	4.81	0.81	306	0.67	1,250	0.86
Hispanic/Latino	72	4.81	0.86	134	0.80	8,801	0.88
Multiple	72	4.87	0.88	143	0.85	131	0.92
Special Instruction Needs							
Economically Disadvantaged	72	4.76	0.85	332	0.77	9,689	0.88
Not Economically Disadvantaged	72	4.92	0.87	218	0.70	6,676	0.90
English Learner	72	4.07	0.78	197	0.62	1,881	0.81
Non-English Learner	72	4.92	0.86	353	0.80	14,486	0.89
Students with Disabilities	72	3.95	0.83	178	0.58	2,468	0.86
Students without Disabilities	72	5.00	0.86	372	0.76	13,897	0.88
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.60	0.77	695	0.77	695	0.77

n/r = not reported due to n<100.

Table A.13.10 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 3

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.18	0.92	3,805	0.90	101,570	0.92
Gender							
Male	52	3.17	0.92	1,959	0.90	52,216	0.92
Female	52	3.18	0.92	1,846	0.89	49,354	0.92
Ethnicity							
White	52	3.25	0.91	5,749	0.91	47,832	0.91
Black/African American	52	3.02	0.91	10,298	0.90	12,090	0.92
Asian/Pacific Islander	52	3.16	0.90	8,044	0.90	1,160	0.92
American Indian/Alaska Native	52	2.94	0.90	937	0.89	1,418	0.90
Hispanic/Latino	52	3.10	0.91	382	0.88	28,540	0.91
Multiple	52	3.19	0.92	505	0.92	3,542	0.93
Special Instruction Needs							
Economically Disadvantaged	52	3.06	0.91	810	0.89	44,854	0.91
Not Economically Disadvantaged	52	3.24	0.91	643	0.89	579	0.91
English Learner	52	3.02	0.89	336	0.88	13,114	0.90
Non-English Learner	52	3.20	0.92	1,053	0.91	88,413	0.92
Students with Disabilities	52	3.01	0.92	344	0.85	16,872	0.93
Students without Disabilities	52	3.20	0.91	3,459	0.90	84,465	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.05	0.92	1,313	0.90	19,441	0.93
Students Taking Translated Forms							
Spanish Language Form	52	2.82	0.89	738	0.88	3,805	0.90

n/r = not reported due to n<100.

Table A.13.11 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 4

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.27	0.93	3,512	0.85	120,199	0.93
Gender							
Male	52	3.24	0.93	1,818	0.85	61,771	0.94
Female	52	3.30	0.93	1,694	0.86	58,428	0.93
Ethnicity							
White	52	3.39	0.92	542	0.88	51,986	0.92
Black/African American	52	3.05	0.92	233	0.88	17,196	0.92
Asian/Pacific Islander	52	3.30	0.92	8,886	0.92	8,833	0.92
American Indian/Alaska Native	52	2.98	0.91	899	0.90	1,493	0.92
Hispanic/Latino	52	3.18	0.91	3,465	0.85	36,624	0.92
Multiple	52	3.28	0.94	141	0.92	3,861	0.94
Special Instruction Needs							
Economically Disadvantaged	52	3.13	0.92	2,974	0.85	57,881	0.92
Not Economically Disadvantaged	52	3.39	0.92	538	0.85	62,266	0.92
English Learner	52	3.01	0.89	3,512	0.85	16,523	0.90
Non-English Learner	52	3.32	0.93	1,057	0.89	103,638	0.93
Students with Disabilities	52	2.96	0.93	319	0.81	20,475	0.94
Students without Disabilities	52	3.33	0.92	3,193	0.86	99,467	0.93
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.09	0.93	1,265	0.85	33,300	0.93
Students Taking Translated Forms							
Spanish Language Form	52	2.45	0.85	3,512	0.85	3,512	0.85

n/r = not reported due to n<100.

Table A.13.12 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 5

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.25	0.92	3,077	0.81	119,050	0.92
Gender							
Male	52	3.21	0.93	1,597	0.83	61,449	0.93
Female	52	3.29	0.91	1,480	0.78	57,601	0.92
Ethnicity							
White	52	3.32	0.91	465	0.86	52,203	0.92
Black/African American	52	3.08	0.89	175	0.81	16,897	0.89
Asian/Pacific Islander	52	3.22	0.92	8,708	0.92	8,603	0.92
American Indian/Alaska Native	52	3.04	0.87	890	0.84	1,587	0.88
Hispanic/Latino	52	3.20	0.89	2,944	0.81	35,906	0.90
Multiple	52	3.26	0.93	3,567	0.93	3,749	0.93
Special Instruction Needs							
Economically Disadvantaged	52	3.16	0.89	2,561	0.81	57,275	0.89
Not Economically Disadvantaged	52	3.31	0.92	516	0.82	61,714	0.92
English Learner	52	2.99	0.82	697	0.80	12,907	0.84
Non-English Learner	52	3.28	0.92	781	0.86	106,098	0.92
Students with Disabilities	52	3.02	0.90	293	0.68	21,576	0.91
Students without Disabilities	52	3.29	0.92	2,784	0.81	97,196	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.13	0.91	1,055	0.82	32,567	0.91
Students Taking Translated Forms							
Spanish Language Form	52	2.69	0.81	3,077	0.81	3,077	0.81

n/r = not reported due to n<100.

Table A.13.13 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 6

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.09	0.92	2,436	0.76	123,192	0.93
Gender							
Male	52	3.06	0.93	1,313	0.77	63,093	0.93
Female	52	3.11	0.92	1,123	0.76	60,099	0.93
Ethnicity							
White	52	3.27	0.91	330	0.85	54,767	0.92
Black/African American	52	2.67	0.89	497	0.82	17,429	0.90
Asian/Pacific Islander	52	3.48	0.93	8,821	0.92	8,493	0.93
American Indian/Alaska Native	52	2.67	0.87	794	0.83	1,586	0.89
Hispanic/Latino	52	2.88	0.89	2,395	0.76	37,137	0.90
Multiple	52	3.13	0.93	109	0.89	3,602	0.93
Special Instruction Needs							
Economically Disadvantaged	52	2.80	0.89	2,012	0.76	58,180	0.90
Not Economically Disadvantaged	52	3.32	0.92	423	0.76	64,912	0.93
English Learner	52	2.37	0.82	163	0.76	9,149	0.84
Non-English Learner	52	3.14	0.92	638	0.84	113,925	0.93
Students with Disabilities	52	2.59	0.90	222	0.75	21,989	0.92
Students without Disabilities	52	3.18	0.92	2,213	0.76	100,740	0.93
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	2.78	0.92	811	0.77	30,158	0.93
Students Taking Translated Forms							
Spanish Language Form	52	2.34	0.76	2,436	0.76	2,436	0.76

n/r = not reported due to n<100.

Table A.13.14 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 7

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	3.12	0.92	2,022	0.77	121,448	0.93
Gender							
Male	52	3.09	0.93	1,060	0.78	62,060	0.93
Female	52	3.14	0.92	962	0.76	59,388	0.92
Ethnicity							
White	52	3.27	0.92	213	0.83	55,056	0.92
Black/African American	52	2.76	0.90	108	0.80	16,631	0.91
Asian/Pacific Islander	52	3.41	0.92	7,667	0.92	7,697	0.92
American Indian/Alaska Native	52	2.70	0.89	611	0.85	1,608	0.89
Hispanic/Latino	52	2.94	0.90	1,999	0.77	36,799	0.91
Multiple	52	3.15	0.93	3,260	0.93	3,330	0.93
Special Instruction Needs							
Economically Disadvantaged	52	2.87	0.90	1,583	0.77	56,542	0.91
Not Economically Disadvantaged	52	3.30	0.92	439	0.75	64,793	0.92
English Learner	52	2.46	0.84	100	0.64	7,658	0.86
Non-English Learner	52	3.16	0.92	421	0.82	113,669	0.93
Students with Disabilities	52	2.61	0.91	183	0.65	21,651	0.92
Students without Disabilities	52	3.21	0.92	1,838	0.77	99,310	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	2.83	0.93	660	0.79	27,303	0.93
Students Taking Translated Forms							
Spanish Language Form	52	2.36	0.77	2,022	0.77	2,022	0.77

n/r = not reported due to n<100.

Table A.13.15 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 8

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	52	2.89	0.90	2,018	0.76	106,005	0.91
Gender							
Male	52	2.82	0.91	1,105	0.75	54,945	0.92
Female	52	2.95	0.90	913	0.76	51,060	0.91
Ethnicity							
White	52	3.02	0.90	241	0.84	48,867	0.91
Black/African American	52	2.58	0.86	100	0.80	14,889	0.87
Asian/Pacific Islander	52	3.25	0.92	4,651	0.92	5,176	0.92
American Indian/Alaska Native	52	2.51	0.84	553	0.82	1,425	0.86
Hispanic/Latino	52	2.74	0.87	232	0.69	32,576	0.89
Multiple	52	2.88	0.92	2,511	0.91	2,930	0.93
Special Instruction Needs							
Economically Disadvantaged	52	2.68	0.87	488	0.75	50,963	0.88
Not Economically Disadvantaged	52	3.05	0.91	487	0.71	54,923	0.91
English Learner	52	2.36	0.80	2,018	0.76	6,616	0.82
Non-English Learner	52	2.92	0.91	572	0.81	99,253	0.91
Students with Disabilities	52	2.46	0.87	119	0.62	20,430	0.88
Students without Disabilities	52	2.97	0.90	1,896	0.76	85,074	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	2.65	0.90	507	0.76	22,793	0.91
Students Taking Translated Forms							
Spanish Language Form	52	2.29	0.76	2,018	0.76	2,018	0.76

n/r = not reported due to n<100.

Table A.13.16 Summary of Test Reliability Estimates for Subgroups: Algebra I

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.89	0.91	2,213	0.54	59,757	0.92
Gender							
Male	55	2.86	0.92	1,214	0.56	30,649	0.93
Female	55	2.91	0.90	999	0.52	29,108	0.92
Ethnicity							
White	55	3.08	0.91	132	0.88	25,561	0.91
Black/African American	55	2.57	0.87	7,199	0.86	6,773	0.87
Asian/Pacific Islander	55	3.36	0.93	5,348	0.92	5,463	0.93
American Indian/Alaska Native	55	2.37	0.79	162	0.71	1,246	0.82
Hispanic/Latino	55	2.59	0.85	2,200	0.53	19,633	0.87
Multiple	55	3.06	0.93	1,016	0.93	952	0.93
Special Instruction Needs							
Economically Disadvantaged	55	2.57	0.85	1,512	0.56	22,880	0.86
Not Economically Disadvantaged	55	3.07	0.92	700	0.48	36,872	0.92
English Learner	55	2.22	0.75	2,213	0.54	3,113	0.82
Non-English Learner	55	2.93	0.92	117	0.79	56,639	0.92
Students with Disabilities	55	2.47	0.88	212	0.86	11,292	0.88
Students without Disabilities	55	2.97	0.91	2,160	0.54	48,462	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	2.67	0.90	549	0.50	7,062	0.92
Students Taking Translated Forms							
Spanish Language Form	55	2.09	0.54	2,213	0.54	2,213	0.54

n/r = not reported due to n<100.

Table A.13.17 Summary of Test Reliability Estimates for Subgroups: Geometry

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.87	0.90	1,359	0.66	45,622	0.91
Gender							
Male	55	2.86	0.91	688	0.59	22,758	0.92
Female	55	2.88	0.90	671	0.70	22,864	0.91
Ethnicity							
White	55	3.02	0.89	21,105	0.89	135	0.90
Black/African American	55	2.50	0.85	5,375	0.84	4,766	0.87
Asian/Pacific Islander	55	3.38	0.92	4,201	0.92	4,077	0.92
American Indian/Alaska Native	55	2.42	0.81	156	0.61	1,246	0.83
Hispanic/Latino	55	2.55	0.84	1,352	0.66	14,843	0.86
Multiple	55	3.07	0.92	658	0.91	647	0.92
Special Instruction Needs							
Economically Disadvantaged	55	2.51	0.84	920	0.65	16,538	0.86
Not Economically Disadvantaged	55	3.03	0.91	439	0.66	29,078	0.91
English Learner	55	2.13	0.76	1,359	0.66	1,920	0.82
Non-English Learner	55	2.91	0.91	327	0.87	43,698	0.91
Students with Disabilities	55	2.43	0.87	8,333	0.86	229	0.89
Students without Disabilities	55	2.94	0.90	1,327	0.66	37,968	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	2.61	0.88	235	0.34	4,207	0.90
Students Taking Translated Forms							
Spanish Language Form	55	1.95	0.66	1,359	0.66	1,359	0.66

n/r = not reported due to n<100.

Table A.13.18 Summary of Test Reliability Estimates for Subgroups: Algebra II

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.94	0.90	544	0.58	26,115	0.91
Gender							
Male	55	2.96	0.91	273	0.65	12,935	0.92
Female	55	2.92	0.88	271	0.42	13,180	0.90
Ethnicity							
White	55	3.14	0.88	12,023	0.88	10,595	0.89
Black/African American	55	2.54	0.84	2,784	0.84	2,282	0.85
Asian/Pacific Islander	55	3.40	0.89	3,269	0.89	3,867	0.89
American Indian/Alaska Native	55	2.42	0.73	220	0.64	998	0.75
Hispanic/Latino	55	2.58	0.83	540	0.57	8,563	0.84
Multiple	55	3.17	0.92	386	0.91	358	0.92
Special Instruction Needs							
Economically Disadvantaged	55	2.56	0.83	357	0.60	9,373	0.83
Not Economically Disadvantaged	55	3.12	0.90	187	0.54	16,735	0.90
English Learner	55	2.12	0.65	159	0.53	1,328	0.67
Non-English Learner	55	2.98	0.90	274	0.77	24,808	0.91
Students with Disabilities	55	2.51	0.89	139	0.73	3,205	0.90
Students without Disabilities	55	2.99	0.90	536	0.58	23,068	0.91
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	2.68	0.87	154	0.34	1,781	0.90
Students Taking Translated Forms							
Spanish Language Form	55	1.88	0.58	544	0.58	544	0.58

n/r = not reported due to n<100.

Table A.13.19 Summary of Test Reliability Estimates for Subgroups: Integrated Mathematics I

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.53	0.84	604	0.84	604	0.84
Gender							
Male	55	2.51	0.83	318	0.83	318	0.83
Female	55	2.56	0.86	286	0.86	286	0.86
Ethnicity							
White	55	2.79	0.89	165	0.89	165	0.89
Black/African American	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Asian/Pacific Islander	n/r	n/r	n/r	n/r	n/r	n/r	n/r
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	55	2.40	0.74	381	0.74	381	0.74
Multiple	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Special Instruction Needs							
Economically Disadvantaged	55	2.39	0.75	405	0.75	405	0.75
Not Economically Disadvantaged	55	2.79	0.87	192	0.87	192	0.87
English Learner	55	n/r	n/r	100	n/r	100	n/r
Non-English Learner	55	2.58	0.85	497	0.85	497	0.85
Students with Disabilities	55	2.17	0.58	128	0.58	128	0.58
Students without Disabilities	55	2.61	0.85	468	0.85	468	0.85
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students Taking Translated Forms							
Spanish Language Form	n/r	n/r	n/r	n/r	n/r	n/r	n/r

n/r = not reported due to n<100.

Table A.13.20 Summary of Test Reliability Estimates for Subgroups: Integrated Mathematics II

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.54	0.82	522	0.82	522	0.82
Gender							
Male	55	2.52	0.83	283	0.83	283	0.83
Female	55	2.57	0.82	239	0.82	239	0.82
Ethnicity							
White	55	2.78	0.90	117	0.90	117	0.90
Black/African American	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Asian/Pacific Islander	n/r	n/r	n/r	n/r	n/r	n/r	n/r
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	55	2.45	0.61	363	0.61	363	0.61
Multiple	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Special Instruction Needs							
Economically Disadvantaged	55	2.44	0.59	340	0.59	340	0.59
Not Economically Disadvantaged	55	2.72	0.90	178	0.90	178	0.90
English Learner	55	2.40	0.22	117	0.22	117	0.22
Non-English Learner	55	2.58	0.85	401	0.85	401	0.85
Students with Disabilities	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students without Disabilities	55	2.58	0.83	429	0.83	429	0.83
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students Taking Translated Forms							
Spanish Language Form	n/r	n/r	n/r	n/r	n/r	n/r	n/r

n/r = not reported due to n<100.

Table A.13.21 Summary of Test Reliability Estimates for Subgroups: Integrated Mathematics III

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	55	2.70	0.81	197	0.81	197	0.81
Gender							
Male	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Female	55	2.69	0.78	105	0.78	105	0.78
Ethnicity							
White	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Black/African American	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Asian/Pacific Islander	n/r	n/r	n/r	n/r	n/r	n/r	n/r
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	55	2.53	0.67	152	0.67	152	0.67
Multiple	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Special Instruction Needs							
Economically Disadvantaged	55	2.52	0.59	164	0.59	164	0.59
Not Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
English Learner	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Non-English Learner	55	2.75	0.84	153	0.84	153	0.84
Students with Disabilities	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students without Disabilities	55	2.73	0.76	170	0.76	170	0.76
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students Taking Translated Forms							
Spanish Language Form	n/r	n/r	n/r	n/r	n/r	n/r	n/r

n/r = not reported due to n<100.

Appendix 13.2: Reliability of Classification by Content and Grade/Subject

Table A.13.22 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	**0.15**	0.03	0.00	0.00	0.00	0.18
	700-724	0.04	**0.10**	0.05	0.00	0.00	0.19
	725-749	0.00	0.04	**0.12**	0.05	0.00	0.22
	750-809	0.00	0.00	0.05	**0.30**	0.03	0.38
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650-699	**0.15**	0.04	0.01	0.00	0.00	0.20
	700-724	0.04	**0.07**	0.05	0.01	0.00	0.18
	725-749	0.01	0.05	**0.09**	0.06	0.00	0.20
	750-809	0.00	0.01	0.07	**0.27**	0.03	0.37
	810-850	0.00	0.00	0.00	0.02	**0.02**	0.04

Table A.13.23 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	**0.11**	0.02	0.00	0.00	0.00	0.13
	700-724	0.04	**0.11**	0.04	0.00	0.00	0.19
	725-749	0.00	0.05	**0.15**	0.05	0.00	0.25
	750-809	0.00	0.00	0.05	**0.25**	0.04	0.34
	810-850	0.00	0.00	0.00	0.02	**0.08**	0.10
Decision Consistency	650-699	**0.10**	0.04	0.01	0.00	0.00	0.14
	700-724	0.04	**0.08**	0.06	0.01	0.00	0.19
	725-749	0.01	0.05	**0.11**	0.07	0.00	0.23
	750-809	0.00	0.01	0.07	**0.21**	0.04	0.32
	810-850	0.00	0.00	0.00	0.04	**0.07**	0.12

Table A.13.24 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	**0.08**	0.02	0.00	0.00	0.00	0.10
	700-724	0.03	**0.11**	0.04	0.00	0.00	0.19
	725-749	0.00	0.05	**0.16**	0.06	0.00	0.26
	750-809	0.00	0.00	0.05	**0.31**	0.03	0.39
	810-850	0.00	0.00	0.00	0.01	**0.04**	0.05
Decision Consistency	650-699	**0.08**	0.03	0.01	0.00	0.00	0.12
	700-724	0.03	**0.09**	0.06	0.01	0.00	0.19
	725-749	0.00	0.05	**0.12**	0.07	0.00	0.25
	750-809	0.00	0.01	0.07	**0.27**	0.03	0.38
	810-850	0.00	0.00	0.00	0.03	**0.04**	0.07

Table A.13.25 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	**0.08**	0.02	0.00	0.00	0.00	0.10
	700-724	0.03	**0.12**	0.04	0.00	0.00	0.19
	725-749	0.00	0.04	**0.20**	0.05	0.00	0.28
	750-809	0.00	0.00	0.05	**0.27**	0.02	0.35
	810-850	0.00	0.00	0.00	0.02	**0.06**	0.08
Decision Consistency	650-699	**0.08**	0.03	0.00	0.00	0.00	0.11
	700-724	0.03	**0.10**	0.06	0.00	0.00	0.19
	725-749	0.00	0.05	**0.16**	0.06	0.00	0.27
	750-809	0.00	0.00	0.07	**0.24**	0.03	0.34
	810-850	0.00	0.00	0.00	0.03	**0.06**	0.09

Table A.13.26 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	**0.10**	0.02	0.00	0.00	0.00	0.12
	700-724	0.03	**0.10**	0.04	0.00	0.00	0.17
	725-749	0.00	0.04	**0.14**	0.05	0.00	0.23
	750-809	0.00	0.00	0.05	**0.22**	0.04	0.31
	810-850	0.00	0.00	0.00	0.03	**0.14**	0.18
Decision Consistency	650-699	**0.10**	0.03	0.00	0.00	0.00	0.14
	700-724	0.03	**0.08**	0.05	0.01	0.00	0.17
	725-749	0.00	0.04	**0.11**	0.06	0.00	0.22
	750-809	0.00	0.01	0.06	**0.18**	0.04	0.29
	810-850	0.00	0.00	0.00	0.05	**0.14**	0.19

Table A.13.27 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	**0.11**	0.02	0.00	0.00	0.00	0.13
	700-724	0.03	**0.10**	0.04	0.00	0.00	0.17
	725-749	0.00	0.04	**0.14**	0.05	0.00	0.23
	750-809	0.00	0.00	0.05	**0.27**	0.03	0.35
	810-850	0.00	0.00	0.00	0.02	**0.10**	0.13
Decision Consistency	650-699	**0.10**	0.03	0.01	0.00	0.00	0.14
	700-724	0.03	**0.08**	0.05	0.01	0.00	0.17
	725-749	0.00	0.04	**0.11**	0.06	0.00	0.21
	750-809	0.00	0.01	0.06	**0.23**	0.04	0.33
	810-850	0.00	0.00	0.00	0.04	**0.10**	0.14

Table A.13.28 Reliability of Classification: Grade 9 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.10**	0.02	0.00	0.00	0.00	0.12
	700-724	0.03	**0.09**	0.04	0.00	0.00	0.15
	725-749	0.00	0.04	**0.14**	0.05	0.00	0.22
	750-809	0.00	0.00	0.05	**0.27**	0.04	0.36
	810-850	0.00	0.00	0.00	0.03	**0.12**	0.15
Decision Consistency	650-699	**0.09**	0.03	0.00	0.00	0.00	0.13
	700-724	0.03	**0.07**	0.05	0.01	0.00	0.16
	725-749	0.00	0.04	**0.11**	0.06	0.00	0.21
	750-809	0.00	0.01	0.06	**0.23**	0.04	0.34
	810-850	0.00	0.00	0.00	0.05	**0.12**	0.17

Table A.13.29 Reliability of Classification: Grade 10 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.13**	0.02	0.00	0.00	0.00	0.15
	700-724	0.03	**0.06**	0.04	0.00	0.00	0.13
	725-749	0.00	0.03	**0.09**	0.05	0.00	0.17
	750-809	0.00	0.00	0.05	**0.23**	0.04	0.32
	810-850	0.00	0.00	0.00	0.04	**0.18**	0.22
Decision Consistency	650-699	**0.12**	0.03	0.01	0.00	0.00	0.16
	700-724	0.03	**0.05**	0.04	0.01	0.00	0.13
	725-749	0.01	0.03	**0.06**	0.06	0.00	0.16
	750-809	0.00	0.01	0.06	**0.19**	0.05	0.30
	810-850	0.00	0.00	0.00	0.07	**0.17**	0.24

Table A.13.30 Reliability of Classification: Grade 11 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.13**	0.02	0.00	0.00	0.00	0.16
	700-724	0.04	**0.10**	0.05	0.00	0.00	0.19
	725-749	0.00	0.05	**0.14**	0.06	0.00	0.24
	750-809	0.00	0.00	0.06	**0.25**	0.04	0.34
	810-850	0.00	0.00	0.00	0.02	**0.05**	0.07
Decision Consistency	650-699	**0.13**	0.04	0.01	0.00	0.00	0.18
	700-724	0.04	**0.08**	0.06	0.01	0.00	0.19
	725-749	0.01	0.05	**0.10**	0.07	0.00	0.22
	750-809	0.00	0.01	0.07	**0.20**	0.04	0.32
	810-850	0.00	0.00	0.00	0.04	**0.05**	0.09

Table A.13.31 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	**0.10**	0.02	0.00	0.00	0.00	0.12
	700-724	0.03	**0.12**	0.04	0.00	0.00	0.19
	725-749	0.00	0.04	**0.16**	0.05	0.00	0.25
	750-809	0.00	0.00	0.04	**0.29**	0.03	0.36
	810-850	0.00	0.00	0.00	0.02	**0.07**	0.09
Decision Consistency	650-699	**0.09**	0.03	0.00	0.00	0.00	0.13
	700-724	0.03	**0.10**	0.05	0.00	0.00	0.18
	725-749	0.00	0.05	**0.13**	0.06	0.00	0.24
	750-809	0.00	0.00	0.06	**0.25**	0.03	0.35
	810-850	0.00	0.00	0.00	0.04	**0.07**	0.10

Table A.13.32 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	**0.11**	0.02	0.00	0.00	0.00	0.13
	700-724	0.03	**0.14**	0.04	0.00	0.00	0.20
	725-749	0.00	0.04	**0.20**	0.04	0.00	0.27
	750-809	0.00	0.00	0.04	**0.30**	0.02	0.36
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650-699	**0.11**	0.03	0.00	0.00	0.00	0.14
	700-724	0.03	**0.12**	0.05	0.00	0.00	0.20
	725-749	0.00	0.04	**0.16**	0.05	0.00	0.26
	750-809	0.00	0.00	0.06	**0.27**	0.02	0.35
	810-850	0.00	0.00	0.00	0.02	**0.03**	0.05

Table A.13.33 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	**0.08**	0.02	0.00	0.00	0.00	0.11
	700-724	0.03	**0.20**	0.04	0.00	0.00	0.27
	725-749	0.00	0.05	**0.19**	0.04	0.00	0.28
	750-809	0.00	0.00	0.04	**0.23**	0.02	0.29
	810-850	0.00	0.00	0.00	0.01	**0.05**	0.06
Decision Consistency	650-699	**0.08**	0.04	0.00	0.00	0.00	0.12
	700-724	0.03	**0.16**	0.06	0.00	0.00	0.25
	725-749	0.00	0.06	**0.15**	0.05	0.00	0.27
	750-809	0.00	0.00	0.05	**0.21**	0.02	0.28
	810-850	0.00	0.00	0.00	0.02	**0.05**	0.07

Table A.13.34 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	**0.11**	0.03	0.00	0.00	0.00	0.14
	700-724	0.03	**0.20**	0.04	0.00	0.00	0.27
	725-749	0.00	0.05	**0.20**	0.04	0.00	0.29
	750-809	0.00	0.00	0.04	**0.21**	0.02	0.26
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650-699	**0.11**	0.04	0.00	0.00	0.00	0.15
	700-724	0.03	**0.17**	0.06	0.00	0.00	0.26
	725-749	0.00	0.06	**0.16**	0.05	0.00	0.28
	750-809	0.00	0.00	0.05	**0.18**	0.02	0.26
	810-850	0.00	0.00	0.00	0.02	**0.03**	0.05

Table A.13.35 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	**0.08**	0.02	0.00	0.00	0.00	0.09
	700-724	0.02	**0.19**	0.04	0.00	0.00	0.26
	725-749	0.00	0.05	**0.23**	0.04	0.00	0.32
	750-809	0.00	0.00	0.04	**0.22**	0.02	0.28
	810-850	0.00	0.00	0.00	0.01	**0.04**	0.05
Decision Consistency	650-699	**0.07**	0.03	0.00	0.00	0.00	0.11
	700-724	0.03	**0.16**	0.06	0.00	0.00	0.25
	725-749	0.00	0.06	**0.19**	0.06	0.00	0.31
	750-809	0.00	0.00	0.06	**0.20**	0.02	0.28
	810-850	0.00	0.00	0.00	0.02	**0.04**	0.06

Table A.13.36 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	**0.22**	0.03	0.00	0.00	0.00	0.25
	700-724	0.05	**0.15**	0.04	0.00	0.00	0.24
	725-749	0.00	0.05	**0.13**	0.05	0.00	0.22
	750-809	0.00	0.00	0.04	**0.22**	0.02	0.27
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.02
Decision Consistency	650-699	**0.21**	0.05	0.01	0.00	0.00	0.26
	700-724	0.05	**0.11**	0.05	0.01	0.00	0.22
	725-749	0.01	0.06	**0.10**	0.05	0.00	0.21
	750-809	0.00	0.01	0.05	**0.19**	0.02	0.27
	810-850	0.00	0.00	0.00	0.02	**0.02**	0.03

Table A.13.37 Reliability of Classification: Algebra I

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.08**	0.03	0.00	0.00	0.00	0.11
	700-724	0.03	**0.20**	0.04	0.00	0.00	0.27
	725-749	0.00	0.06	**0.14**	0.04	0.00	0.25
	750-809	0.00	0.00	0.04	**0.28**	0.01	0.34
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650-699	**0.07**	0.05	0.00	0.00	0.00	0.13
	700-724	0.03	**0.16**	0.05	0.00	0.00	0.25
	725-749	0.00	0.07	**0.11**	0.06	0.00	0.25
	750-809	0.00	0.01	0.05	**0.26**	0.02	0.33
	810-850	0.00	0.00	0.00	0.02	**0.03**	0.05

Table A.13.38 Reliability of Classification: Geometry

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.09**	0.02	0.00	0.00	0.00	0.12
	700-724	0.03	**0.20**	0.05	0.00	0.00	0.27
	725-749	0.00	0.05	**0.23**	0.05	0.00	0.33
	750-809	0.00	0.00	0.04	**0.19**	0.02	0.25
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650-699	**0.09**	0.04	0.00	0.00	0.00	0.13
	700-724	0.03	**0.17**	0.07	0.00	0.00	0.27
	725-749	0.00	0.07	**0.19**	0.06	0.00	0.32
	750-809	0.00	0.00	0.06	**0.16**	0.02	0.25
	810-850	0.00	0.00	0.00	0.02	**0.02**	0.04

Table A.13.39 Reliability of Classification: Algebra II

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.22**	0.03	0.00	0.00	0.00	0.25
	700-724	0.04	**0.11**	0.04	0.00	0.00	0.19
	725-749	0.00	0.04	**0.11**	0.05	0.00	0.20
	750-809	0.00	0.00	0.04	**0.27**	0.02	0.33
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision Consistency	650-699	**0.21**	0.05	0.01	0.00	0.00	0.26
	700-724	0.05	**0.08**	0.05	0.01	0.00	0.19
	725-749	0.01	0.05	**0.08**	0.05	0.00	0.19
	750-809	0.00	0.01	0.05	**0.24**	0.02	0.32
	810-850	0.00	0.00	0.00	0.02	**0.02**	0.04

Table A.13.40 Reliability of Classification: Integrated Mathematics I

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.24**	0.05	0.00	0.00	0.00	0.30
	700-724	0.06	**0.18**	0.06	0.00	0.00	0.30
	725-749	0.00	0.06	**0.12**	0.05	0.00	0.23
	750-809	0.00	0.00	0.03	**0.12**	0.01	0.16
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01
Decision Consistency	650-699	**0.23**	0.08	0.01	0.00	0.00	0.31
	700-724	0.07	**0.14**	0.06	0.01	0.00	0.28
	725-749	0.01	0.07	**0.09**	0.05	0.00	0.22
	750-809	0.00	0.01	0.05	**0.11**	0.01	0.18
	810-850	0.00	0.00	0.00	0.01	**0.01**	0.01

Table A.13.41 Reliability of Classification: Integrated Mathematics II

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.23**	0.15	0.00	0.00	0.00	0.38
	700-724	0.05	**0.31**	0.04	0.00	0.00	0.40
	725-749	0.00	0.06	**0.08**	0.02	0.00	0.15
	750-809	0.00	0.00	0.01	**0.04**	0.01	0.06
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.02
Decision Consistency	650-699	**0.20**	0.17	0.01	0.00	0.00	0.38
	700-724	0.07	**0.25**	0.04	0.00	0.00	0.36
	725-749	0.00	0.08	**0.06**	0.02	0.00	0.17
	750-809	0.00	0.01	0.02	**0.03**	0.01	0.07
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.02

Table A.13.42 Reliability of Classification: Integrated Mathematics III

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.54**	0.07	0.00	0.00	0.00	0.62
	700-724	0.05	**0.10**	0.04	0.00	0.00	0.20
	725-749	0.00	0.03	**0.06**	0.02	0.00	0.11
	750-809	0.00	0.00	0.02	**0.05**	0.00	0.07
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision Consistency	650-699	**0.50**	0.08	0.01	0.00	0.00	0.60
	700-724	0.08	**0.07**	0.04	0.01	0.00	0.20
	725-749	0.01	0.03	**0.04**	0.02	0.00	0.11
	750-809	0.00	0.01	0.02	**0.05**	0.00	0.08
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.01

Appendix 14: Quality Testing Standards

Table A.14.1 ELA/L Grade 6 Form 1 Matching Results

ELA/L Grade 6 Form 1	Unmatched		DIFF*	Matched		DIFF*
	Current Form 1	Original Form 1		Current Form 1	Original Form 1	
Sample Size	119,838	31,031		30,667	30,667	
American Indian/Alaska Native	1.3	0.3	1	0.3	0.3	0
Asian	6.8	6.7	0.1	6.7	6.7	0
Black/African American	14.1	32.8	-18.6	32.2	32.2	0
Hispanic/Latino Ethnicity	31.4	18.9	12.5	19.1	19.1	0
Hawaiian/Pacific Islander	0.2	0.2	0	0.1	0.1	0
White	43.4	36.5	6.9	37	37	0
Two or More Races	2.9	4.7	-1.8	4.7	4.7	0
Female	49.7	49.4	0.3	49.4	49.4	0
Economic Disadvantage	48.3	44.1	4.2	44.5	44.5	0
English Learner	7.2	5.7	1.4	5.6	5.6	0
Students with Disabilities	14.4	13.9	0.5	13.7	13.7	0
Grade 6	100	100	0	100	100	0
Prior Year Scale Score	745	742.3	2.7	742.7	742.7	0
Prior Performance Level 1	10.2	11.7	-1.5	11.4	11.4	0
Prior Performance Level 2	18	19	-1	18.8	18.8	0
Prior Performance Level 3	26.4	26.3	0.1	26.4	26.4	0
Prior Performance Level 4	39.3	38.5	0.9	38.8	38.8	0
Prior Performance Level 5	6.1	4.6	1.5	4.6	4.6	0

*DIFF = Current Percent – Original Percent

Table A.14.2 Mathematics Grade 6 Form 1 Matching Results

Mathematics Grade 6 Form 1	Unmatched			Matched		
	Current Form 1	Original Form 1	DIFF*	Current Form 1	Original Form 1	DIFF*
Sample Size	95,174	28,514		27,677	27,677	
American Indian/Alaska Native	1.1	0.2	0.9	0.2	0.2	0
Asian	7.6	7	0.6	7.1	7.1	0
Black/African American	11.5	33.4	-21.9	31.6	31.6	0
Hispanic/Latino Ethnicity	28	17.9	10.1	18.5	18.5	0
Hawaiian/Pacific Islander	0.1	0.2	0	0.1	0.1	0
White	48.4	36.5	11.9	37.6	37.6	0
Two or More Races	3.2	4.8	-1.6	4.9	4.9	0
Female	50.2	50	0.2	50.1	50.1	0
Economic Disadvantage	42.6	42.4	0.3	43.2	43.2	0
English Learner	4.6	3.7	0.9	3.5	3.5	0
Students with Disabilities	9.8	11	-1.2	10.6	10.6	0
Grade 6	100	100	0	100	100	0
Prior Year Scale Score	743.9	741.1	2.8	741.7	741.7	0
Prior Performance Level 1	9	12.6	-3.6	12	12	0
Prior Performance Level 2	18.9	20.3	-1.4	20	20	0
Prior Performance Level 3	28.6	25.6	3	25.8	25.8	0
Prior Performance Level 4	35.7	33.8	1.9	34.3	34.3	0
Prior Performance Level 5	7.8	7.8	0	7.8	7.8	0

*DIFF = Current Percent – Original Percent

Table A.14.3 ELA/L Grade 10 Form 1 Matching Results

ELA/L Grade 10 Form 1	Unmatched			Matched		
	Current Form 1	Original Form 1	DIFF*	Current Form 1	Original Form 1	DIFF*
Sample Size	55,046	27,951		22,970	22,970	
American Indian/Alaska Native	2	0.3	1.7	0.3	0.3	0
Asian	9.3	7.5	1.8	8.6	8.6	0
Black/African American	11.1	33.2	-22	24.1	24.1	0
Hispanic/Latino Ethnicity	32.1	14.9	17.2	17.5	17.5	0
Hawaiian/Pacific Islander	0.2	0.1	0.1	0.1	0.1	0
White	44	39.5	4.5	46.9	46.9	0
Two or More Races	1.3	4.6	-3.3	2.6	2.6	0
Female	50.2	50.5	-0.2	50.5	50.5	0
Economic Disadvantage	35.8	35	0.9	32.6	32.6	0
English Learner	3.2	3.2	0	2.9	2.9	0
Students with Disabilities	15.6	14.7	0.9	14.4	14.4	0
Grade 9	1.3	3.5	-2.2	1.8	1.8	0
Grade 10	98.6	96.5	2.2	98.2	98.2	0
2017 Scale Score	755.5	740	15.5	746.3	746.2	0.1
2017 Performance Level 1	8.8	15.8	-7	11.3	11.3	0
2017 Performance Level 2	13	18.8	-5.8	15.9	15.9	0
2017 Performance Level 3	21.4	23.7	-2.2	24.5	24.5	0
2017 Performance Level 4	39.6	34	5.6	39.1	39.1	0
2017 Performance Level 5	17.3	7.7	9.5	9.3	9.3	0

*DIFF = Current Percent – Original Percent

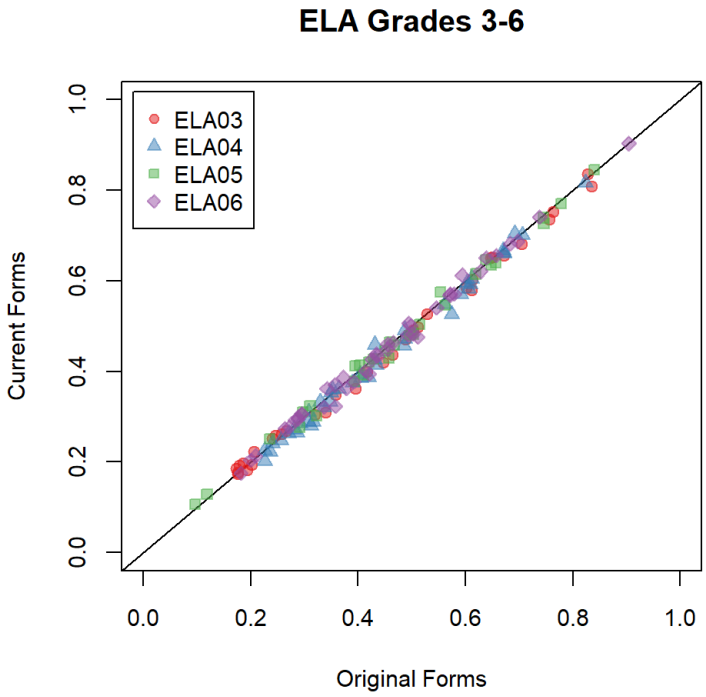

Figure A.14.1 ELA/L Grades 3-6 P-Values

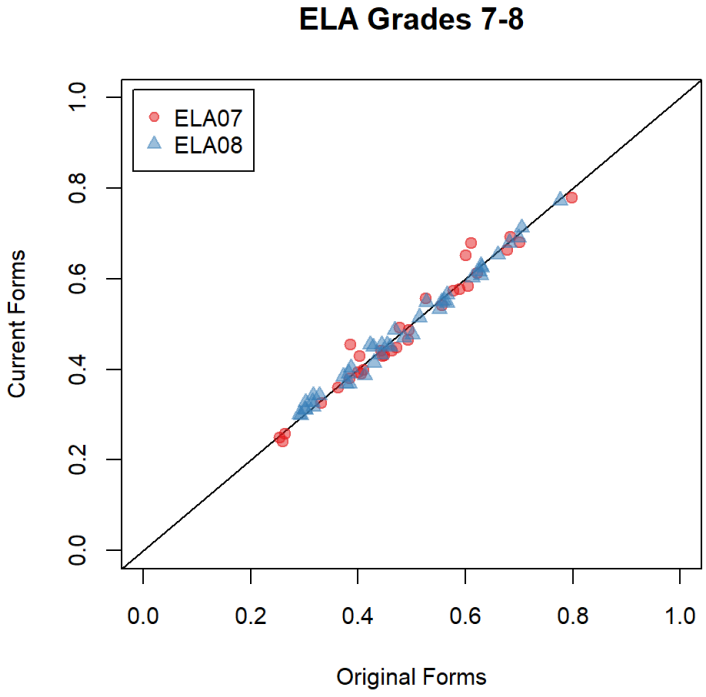

Figure A.14.2 ELA/L Grades 7-8 P-Values

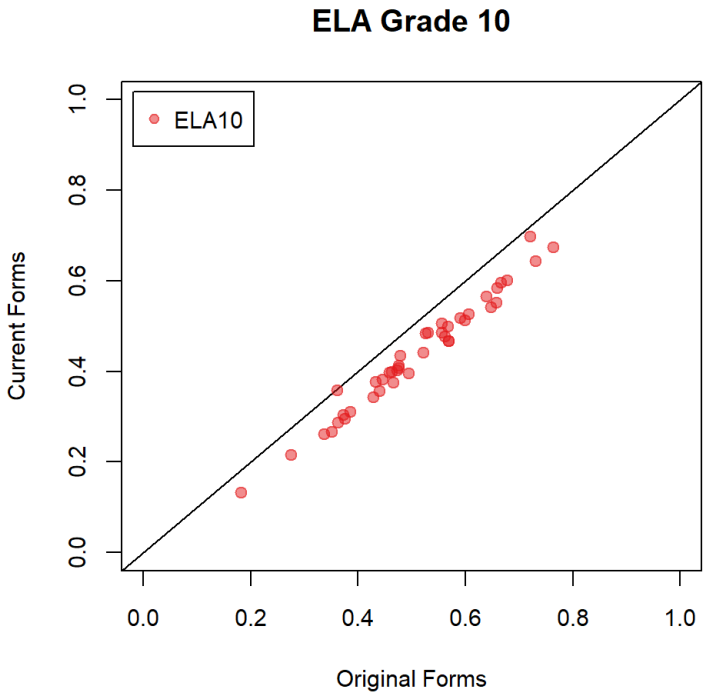

Figure A.14.3 ELA/L Grade 10 P-Values

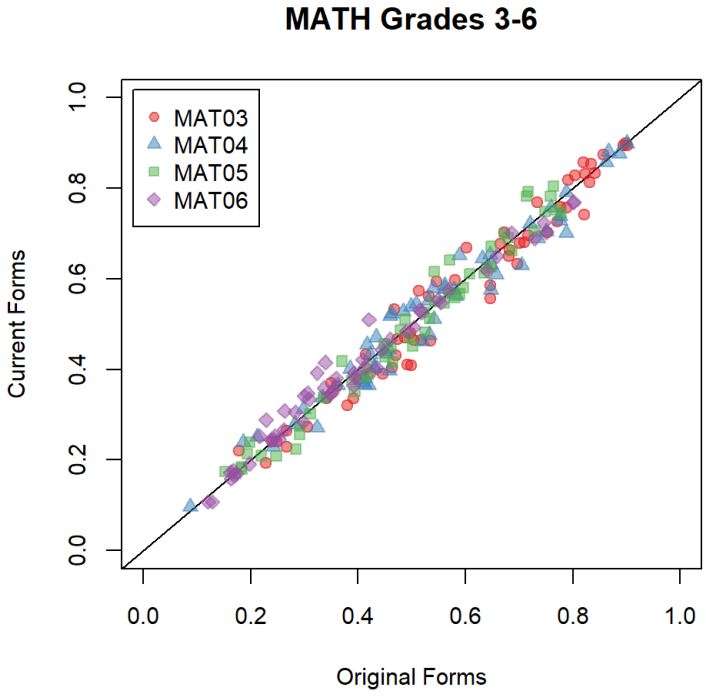

Figure A.14.4 Mathematics Grades 3-6 P-Values

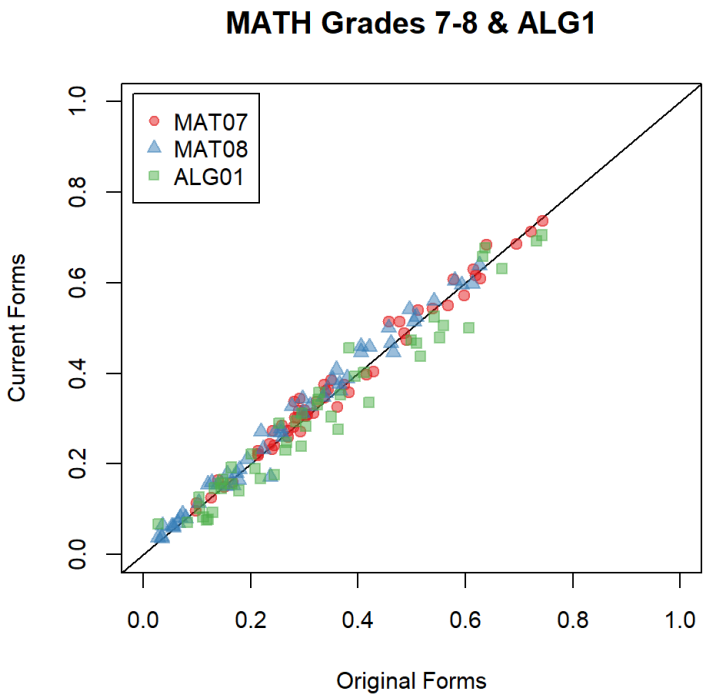

Figure A.14.5 Mathematics Grade 7-8 and Algebra I P-Values

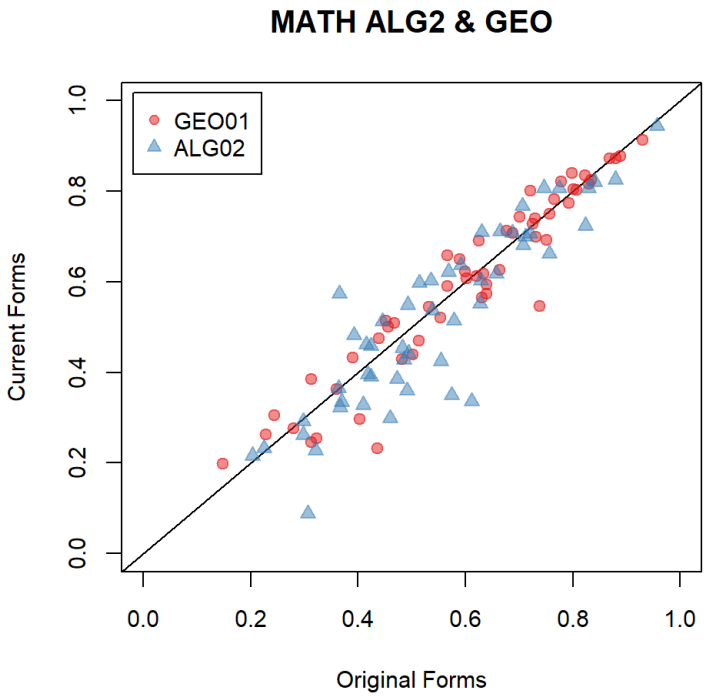

Figure A.14.6 Algebra II and Geometry P-Values

Table A.14.4 Distributions of P-Value Differences* for ELA/L

Grade	N	Min	25%	Median	75%	Max
3	34	-0.034	-0.017	-0.01	0.004	0.016
4	42	-0.049	-0.019	-0.01	-0.004	0.028
5	31	-0.029	-0.016	-0.006	0.009	0.021
6	42	-0.035	-0.008	-0.001	0.008	0.02
7	31	-0.026	-0.016	-0.006	0	0.07
8	42	-0.025	-0.01	0	0.011	0.032
10	42	-0.106	-0.085	-0.073	-0.062	-0.003

*Difference = Current P-value – Original P-value

Table A.14.5 Distributions of P-Value Differences* for Mathematics

Grade/ Course	N	Min	25%	Median	75%	Max
3	59	-0.088	-0.038	-0.017	0.018	0.068
4	56	-0.086	-0.036	-0.003	0.016	0.064
5	54	-0.06	-0.023	-0.01	0.011	0.075
6	52	-0.048	-0.009	0	0.015	0.09
7	55	-0.034	-0.006	0.006	0.022	0.057
8	54	-0.065	0.005	0.013	0.025	0.054
Algebra I	48	-0.105	-0.042	-0.019	0.014	0.073
Geometry	55	-0.204	-0.031	0.004	0.04	0.094
Algebra II	51	-0.275	-0.062	-0.022	0.04	0.209

*Difference = Current P-value – Original P-value

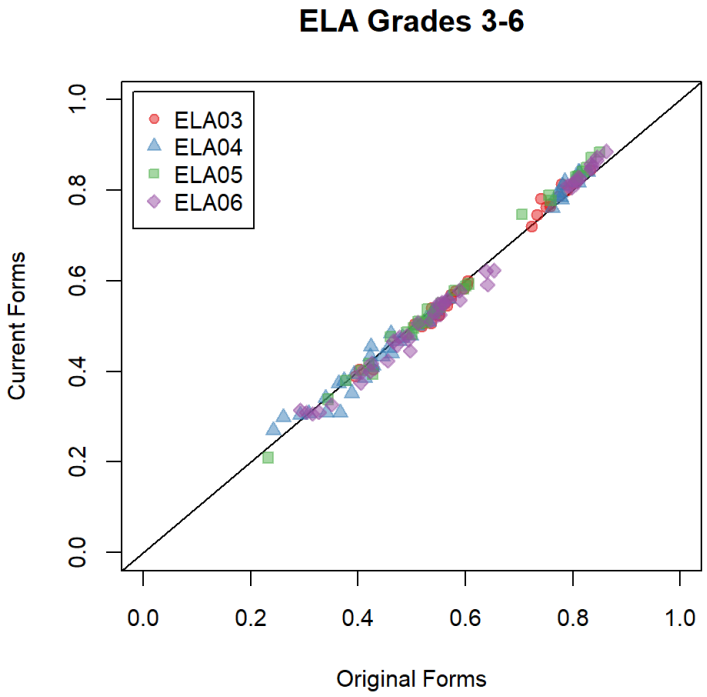

Figure A.14.7 Polyserial Correlations ELA/L Grades 3-6

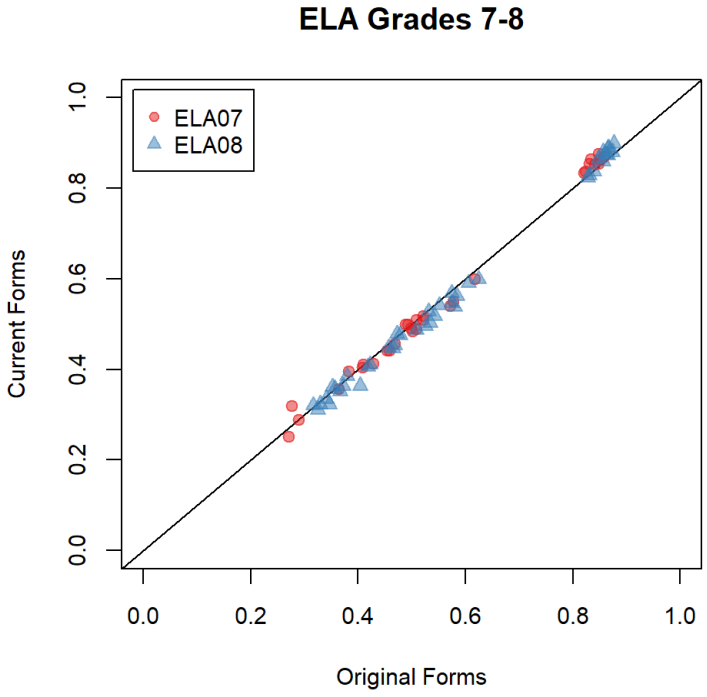

Figure A.14.8 Polyserial Correlations ELA/L Grades 7-8

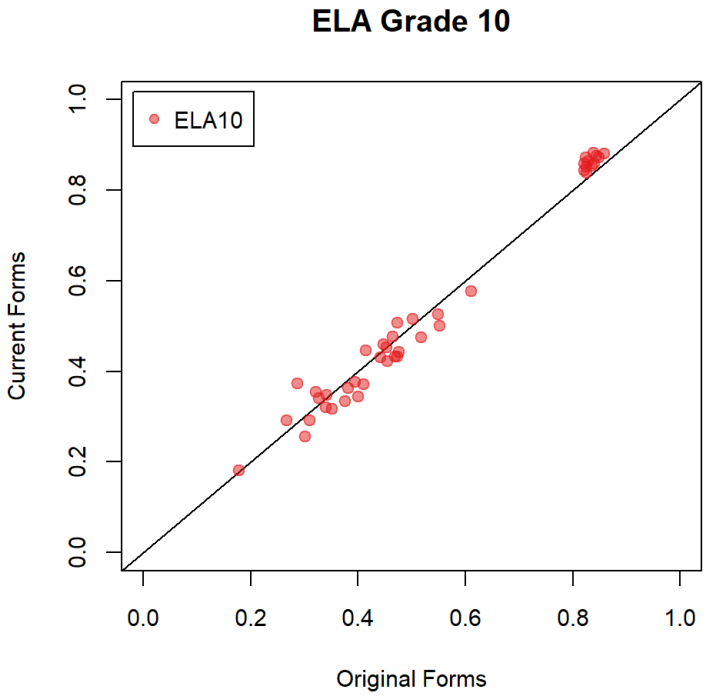

Figure A.14.9 Polyserial Correlations ELA/L Grade 10

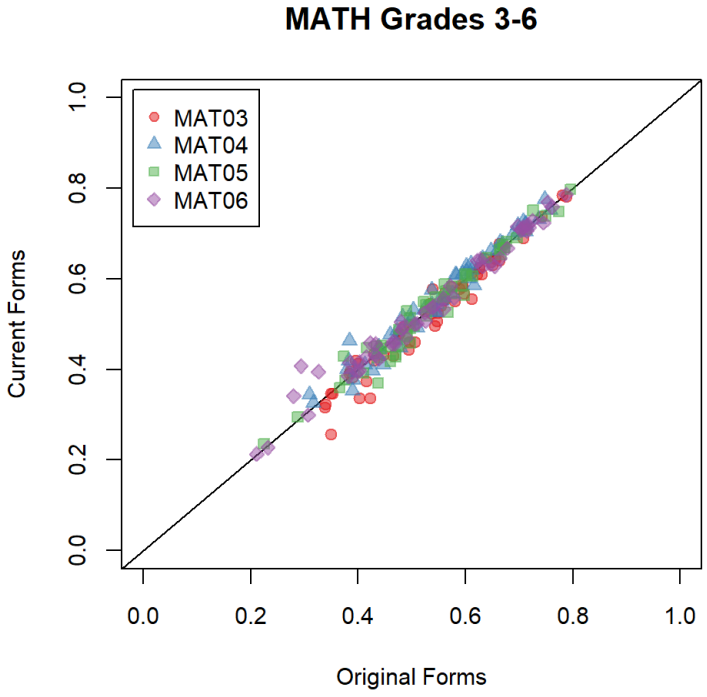

Figure A.14.10 Polyserial Correlations Mathematics Grades 3-6

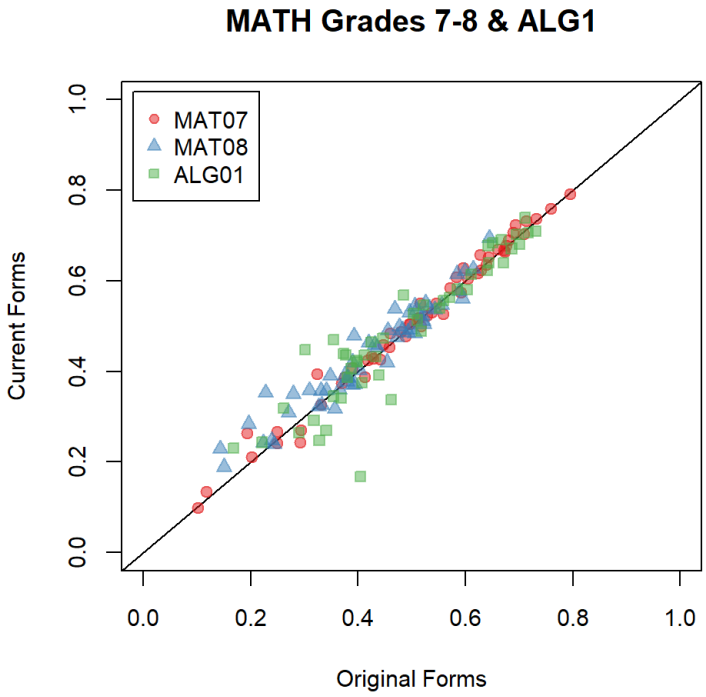

Figure A.14.11 Polyserial Correlations Mathematics Grades 7-8 and Algebra I

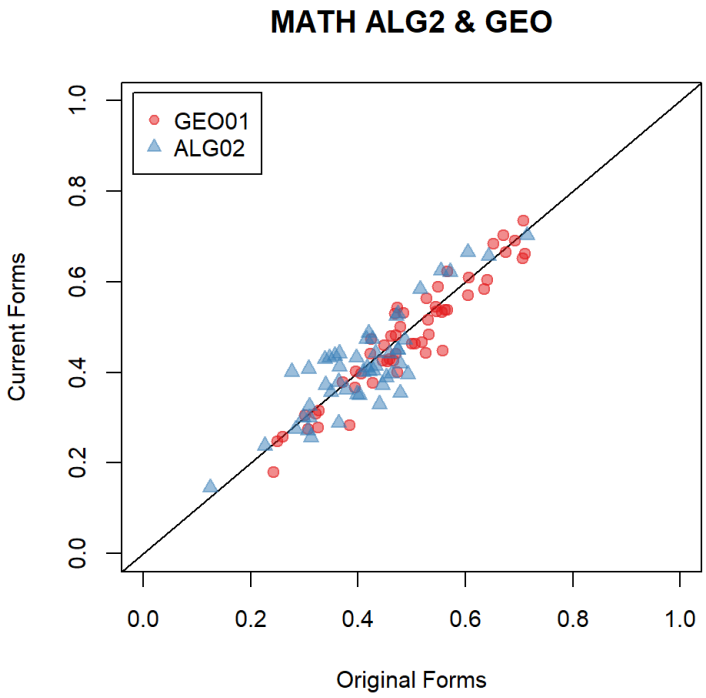

Figure A.14.12 Polyserial Correlations Algebra II and Geometry

Table A.14.6 Distributions of Polyserial Differences* for ELA/L

Grade	N	Min	25%	Median	75%	Max
3	34	-0.029	-0.015	-0.004	0.012	0.041
4	42	-0.058	-0.011	0	0.017	0.037
5	31	-0.034	-0.013	-0.003	0.020	0.042
6	42	-0.052	-0.022	-0.008	0.013	0.028
7	31	-0.031	-0.015	0	0.012	0.043
8	42	-0.042	-0.017	-0.007	0.005	0.023
10	42	-0.055	-0.032	0.010	0.026	0.088

*Difference = Current Polyserial – Original Polyserial

Table A.14.7 Distributions of Polyserial Differences* for Mathematics

Grade/ Course	N	Min	25%	Median	75%	Max
3	59	-0.092	-0.022	-0.01	0.004	0.040
4	56	-0.036	-0.004	0.008	0.018	0.079
5	54	-0.067	-0.011	-0.002	0.010	0.056
6	52	-0.026	-0.008	-0.001	0.012	0.113
7	55	-0.050	-0.005	0.005	0.012	0.070
8	54	-0.040	-0.006	0.014	0.034	0.125
Algebra I	48	-0.238	-0.022	0.001	0.025	0.145
Geometry	55	-0.108	-0.037	-0.011	0.012	0.072
Algebra II	51	-0.125	-0.025	0.002	0.052	0.125

*Difference = Current Polyserial – Original Polyserial

Table A.14.8 DIF Category Crosstabulations for ELA/L

	Percent of DIF Calculations		
ELA/L Grades 3-8 & 10	None	B DIF (Current)	C DIF (Current)
None	89.9% – 96.7%	0% – 2.7%	0% – 0.4%
B DIF (Original)	0.6% – 4.8%	1.2% – 2.4%	0%
C DIF (Original)	0% – 0.4%	0% – 1.8%	0% – 1.6%

Table A.14.9 DIF Category Crosstabulations for Mathematics Grades 3-8 and Algebra I

Mathematics Grades 3 – 8 & Algebra I	Percent of DIF Calculations		
	None	B DIF (Current)	C DIF (Current)
None	94.5% – 97.3%	0.2% – 2.1%	0% – 0.3%
B DIF (Original)	1.4% – 2.5%	0.2% – 2.2%	0% – 0.5%
C DIF (Original)	0% – 0.5%	0 %– 0.5%	0% – 0.2%

Table A.14.10 DIF Category Crosstabulations for Algebra II and Geometry

Geometry & Algebra II	Percent of DIF Calculations		
	None	B DIF (Current)	C DIF (Current)
None	73.2% – 77.5%	8.6% – 12.7%	0% – 1.4%
B DIF (Original)	5.9% – 7.3%	2% – 3.2%	0% – 0.5%
C DIF (Original)	1.8% – 2.0%	0% – 0.9%	0% – 3.2%

Table A.14.11 ELA/L Reliability

	Original		Current Form 1				Current Form 2			
Grade	Pts	Alpha**	Pts	Alpha	SB	Diff*	Pts	Alpha	SB	Diff*
3	82	0.92	54	0.90	0.89	0.01	55	0.89	0.89	0
4	106	0.92	74	0.89	0.89	0	67	0.88	0.88	0
5	106	0.93	74	0.89	0.89	0	67	0.88	0.89	-0.01
6	109	0.94	74	0.92	0.92	0	70	0.90	0.90	0
7	109	0.94	74	0.91	0.91	0	70	0.90	0.91	-0.01
8	109	0.94	74	0.92	0.92	0	70	0.90	0.91	-0.01
10	109	0.93	74	0.90	0.89	0.01	70	0.88	0.89	-0.01

*DIFF = Current Alpha – Spearman Brown (SB) Prophecy

**Alpha = Weighted average of the stratified alphas from Original form 1 and Original form 2

Table A.14.12 ELA/L Raw Score Standard Error of Measurement

	Original			Current Form 1			Current Form 2		
Grade	RS Points	RS SEM	SEM/ Points	RS Points	RS SEM	SEM/ Points	RS Points	RS SEM	SEM/ Points
3	82	4.42	0.054	54	3.54	0.066	55	3.58	0.065
4	106	5.41	0.051	74	4.46	0.06	67	4.51	0.067
5	106	5.46	0.052	74	4.48	0.061	67	4.48	0.067
6	109	5.53	0.051	74	4.50	0.061	70	4.49	0.064
7	109	5.93	0.054	74	4.71	0.064	70	5.06	0.072
8	109	5.63	0.052	74	4.52	0.061	70	4.69	0.067
10	109	5.95	0.044	74	4.71	0.05	70	5.20	0.06

Table A.14.13 ELA/L Scale Score Standard Error of Measurement

	Original Form 1		Original Form 2		Current Form 1		Current Form 2	
Grade	SS Points	SS SEM	SS Points	SS SEM	SS Points	SS SEM	SS Points	SS SEM
3	82	11.6	82	11.8	54	13.8	55	13.9
4	106	10.6	106	10.6	74	12.9	67	13.3
5	106	9.7	106	9.5	74	11.9	67	12.6
6	109	8	109	8.4	74	9.7	70	10.9
7	109	9.7	109	9.7	74	11.9	70	12.9
8	109	9.8	109	9.7	74	11.8	70	12.9
10	109	11.4	109	11.6	74	14.6	70	16.3

Table A.14.14 Mathematics Reliability

Grade/ Course	Original		Current Form 1 and Form 2			
	Points	Alpha**	Points	Alpha**	SB	Diff*
3	66	0.94	52	0.92	0.93	-0.01
4	66	0.94	52	0.93	0.93	0
5	66	0.94	52	0.93	0.93	0
6	66	0.95	52	0.93	0.94	-0.01
7	66	0.93	52	0.92	0.91	0.01
8	66	0.87	52	0.86	0.84	0.02
Algebra I	81	0.93	55	0.90	0.90	0
Geometry	81	0.93	55	0.89	0.90	-0.01
Algebra II	81	0.89	55	0.84	0.85	-0.01

**Alpha = Weighted average of the stratified alphas from form 1 and form 2

Table A.14.15 Mathematics Raw Score Standard Error of Measurement

	Original			Current		
Grade/Course	RS Points	RS SEM	SEM/ Points	RS Points	RS SEM	SEM/ Points
3	66	3.58	0.054	52	3.20	0.062
4	66	3.74	0.057	52	3.32	0.064
5	66	3.69	0.056	52	3.29	0.063
6	66	3.49	0.053	52	3.14	0.060
7	66	3.50	0.053	52	3.10	0.060
8	66	2.96	0.045	52	2.71	0.052
Algebra I	81	3.61	0.045	55	2.88	0.052
Geometry	81	4.21	0.052	55	3.51	0.064
Algebra II	81	4.25	0.052	55	3.50	0.064

Table A.14.16 Mathematics Scale Score Standard Error of Measurement

	Original				Current			
	Form 1		Form 2		Form 1		Form 2	
Grade/Course	SS Points	SS SEM	SS Points	SS SEM	SS Points	SS SEM	SS Points	SS SEM
3	66	8.8	66	8.8	52	9.9	52	10.3
4	66	7.9	66	8.4	52	8.9	52	9.2
5	66	8.2	66	7.9	52	9.3	52	9.3
6	66	7.6	66	7.3	52	9.1	52	8.6
7	66	7.5	66	7.3	52	8.3	52	8.1
8	66	11.0	66	11.5	52	12.0	52	13.0
Algebra I	80	8.9	81	8.7	55	10.8	55	10.4
Geometry	81	6.4	81	6.4	55	7.9	55	8.0
Algebra II	81	9.7	81	9.8	55	11.4	55	12.2

Table A.14.17 ELA/L Scale Score Descriptive Statistics

		Current			Original				
Grade	N	Mean	Median	SD	Mean	Median	SD	Diff*	D
3	62,753	737.6	739	41.9	739.2	740	42.3	-1.6	-0.04
4	61,139	742.3	742	38.5	744.7	746	37.3	-2.5	-0.06
5	62,463	744.3	743	36.2	744.6	745	35.0	-0.4	-0.01
6	61,173	743.2	744	33.9	742.6	744	32.7	0.6	0.02
7	59,137	746	747	40.8	747.4	749	39.2	-1.4	-0.04
8	58,210	746.6	748	41.5	745.1	746	40.5	1.5	0.04
10	40,163	749	752	46.9	767.1	770	42.7	-18.1	-0.40

*Diff = Current mean – Original mean

Table A.14.18 Mathematics Scale Score Descriptive Statistics

		Current			Original				
Grade	N	Mean	Median	SD	Mean	Median	SD	Diff*	D
3	51,957	746.6	747	35.5	748.4	749	36.8	-1.8	-0.05
4	50,277	745.1	747	34.8	746.7	748	34.0	-1.65	-0.05
5	53,131	743.6	743	33.6	744.9	744	33.8	-1.33	-0.04
6	55,342	735.8	736	32.7	736.1	735	32.2	-0.33	-0.01
7	47,340	735.3	735	28.4	735	734	27.7	0.35	0.01
8	28,657	717	715	33.1	713.7	713	31.8	3.27	0.10
Algebra I	35,083	739.7	739	33.4	743.5	742	32.9	-3.82	-0.12
Geometry	3,054	773.4	776.5	24.9	772.6	775	24.7	0.81	0.03
Algebra II	1,576	778.2	779	29.6	782.3	782	28.9	-4.09	-0.14

*Diff = Current mean – Original mean

Table A.14.19 ELA/L Writing Claim Score Descriptive Statistics

		Current			Original				
Grade	N	Mean	Median	SD	Mean	Median	SD	Diff*	D
3	62,753	45.3	45	16.8	46.7	47	17.3	-1.4	-0.08
4	61,139	47.2	47	15.5	48.2	48	15.1	-1	-0.07
5	62,463	47.7	47	14.6	48.3	49	14.3	-0.6	-0.04
6	61,173	47.5	47	13.4	47.5	47	13.3	0	0
7	59,137	48.6	49	16.3	49.3	50	16.0	-0.7	-0.04
8	58,210	48.9	48	16.8	48.8	49	16.4	0.1	0.01
10	40,163	49.3	49	18.6	57.2	57	17.8	-7.8	-0.43

*Diff = Current mean – Original mean

Table A.14.20 Reading Claim Score Descriptive Statistics

		Current			Original				
Grade	N	Mean	Median	SD	Mean	Median	SD	Diff*	D
3	62,753	29	33	13.5	29.8	32	12.7	-0.8	-0.06
4	61,139	31.6	34	11.7	32.5	34	10.6	-0.9	-0.08
5	62,463	31.0	33	12.6	31.8	33	10.9	-0.8	-0.07
6	61,173	30.5	34	12.4	30.8	33	11.2	-0.3	-0.02
7	59,137	32.4	34	12.4	32.8	35	11.5	-0.4	-0.03
8	58,210	32.0	33	12.9	31.6	34	12.2	0.3	0.03
10	40,163	33.6	35	13.0	37.7	39	11.0	-4.1	-0.34

*Diff = Current mean – Original mean

Table A.14.21 ELA/L Subclaim Distributions

		Percent of Students by Subclaim Performance Level				
Form	Level	RL	RI	RV	WE	WKL
Current	1	45	42.2	44.9	39.5	38.2
	2	26.3	24.7	23.7	27.3	28.3
	3	28.7	33.1	31.4	33.1	33.4
Original	1	44.5	45.6	44.1	41.9	40
	2	25.2	22.4	24.7	25.4	26.1
	3	30.3	32.1	31.2	32.7	33.9
ES	-	0.02	0.04	0.01	0.03	0.03

Table A.14.22 Mathematics Subclaim Distributions

		Percent of Students by Subclaim Performance Level			
Form	Level	A (MC)	C (MR)	D (MP)	B (ASC)
Current	1	33.5	36.7	31	33.5
	2	30.5	27.1	26.4	33.9
	3	36	36.1	42.5	32.6
Original	1	32.6	37.5	32.1	33
	2	29	24.4	25.6	28.3
	3	38.4	38.1	42.2	38.7
ES	-	0.03	0.03	0.01	0.07

Table A.14.23 ELA/L Subclaim Distribution Comparison: Effect Size

	Subclaim Distribution Effect Size				
Grade	RL	RI	RV	WE	WKL
3	0.01	0.03	0.1	0.14	0.1
4	0.03	0.03	0.08	0.11	0.04
5	0.03	0.03	0.03	0.11	0.08
6	0.02	0.04	0.01	0.03	0.03
7	0.04	0.06	0.05	0.1	0.08
8	0.02	0.05	0.07	0.03	0.04
10	0.19	0.2	0.15	0.15	0.14

Table A.14.24 Mathematics Subclaim Distribution Comparison: Effect Size

Grade/	Subclaim Distribution Effect Size			
Course	A (MC)	C (MR)	D (MP)	B (ASC)
3	0.03	0.01	0.06	0.09
4	0.03	0.02	0.03	0.02
5	0.04	0.11	0.03	0.01
6	0.03	0.03	0.01	0.07
7	0.03	0.19	0.01	0.05
8	0.04	0.13	0.03	0.06
Algebra I	0.05	0.11	0.11	0.06
Geometry	0.03	0.05	0.04	0.02
Algebra II	0.06	0.04	0.16	0.09

Table A.14.25 ELA/L Longitudinal Scale Score Comparison: Original to Current

	2018 Original SS			2019 Current SS			2019-2018		
Grade	N**	Mean	SD	N**	Mean	SD	DIFF*	SD	D
3	265,192	739.7	42.5	257,201	738.5	42.1	-1.2	42.3	-0.03
4	270,283	744.4	37.2	265,584	742.8	38.4	-1.6	37.8	-0.04
5	274,435	743.0	35.3	272,234	744.0	36.5	1.0	35.9	0.03
6	269,341	742.6	33.5	275,880	742.9	34.6	0.3	34.1	0.01
7	266,380	745.5	40.4	270,119	746.7	41.6	1.2	41.0	0.03
8	267,861	744.1	40.5	267,281	746.3	42.2	2.3	41.4	0.05
9	123,153	746.9	39.8	122,200	748.5	40.9	1.6	40.4	0.04
10	118,486	744.2	48.6	118,902	752.3	50.3	8.1	49.5	0.16

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.14.26 ELA/L Longitudinal Scale Score Comparison: Original to Original

	2018 Original			2019 Original			2019-2018		
Grade	N**	Mean	SD	N**	Mean	SD	DIFF*	SD	D
3	74,206	735.3	43.4	72,606	737.1	42.5	1.8	43	0.04
4	75,608	741.8	37.9	74,281	741.8	38.2	0	38.1	0
5	74,695	740.4	35.4	75,575	741.8	35.9	1.4	35.7	0.04
6	76,094	739.3	33	79,034	740.6	33.1	1.4	33.1	0.04
7	73,574	742.8	39.8	75,398	745.2	39.6	2.3	39.7	0.06
8	72,661	739.6	40.3	72,976	743	40.8	3.3	40.5	0.08
9	3,449	728.5	39.9	3,468	731.7	40.9	3.2	40.4	0.08
10	72,150	744.2	49.4	74,517	747.8	48.6	3.6	49	0.07

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.14.27 Mathematics Longitudinal Scale Score Comparison: Original to Current

	2018 Original			2019 Current			2019-2018		
Grade	N**	Mean	SD	N**	Mean	SD	DIFF*	SD	D
3	267,990	742.6	36.7	259,115	743.1	36.5	0.5	36.6	0.01
4	272,625	738.1	33.6	267,191	739.3	34.9	1.2	34.3	0.03
5	275,716	738.2	33.6	273,312	737.8	33.1	-0.4	33.4	-0.01
6	270,735	734.7	31.9	276,652	732.6	32.7	-2.1	32.3	-0.07
7	262,841	736.6	29.5	265,978	737.2	30.6	0.6	30.1	0.02
8	224,120	727.5	37.3	226,912	728.0	38.5	0.6	37.9	0.02
A1***	136,154	742.5	37.1	134,975	740.0	36.7	-2.6	36.9	-0.07
GE***	112,873	732.6	27.4	105,676	731.9	29.5	-0.7	28.4	-0.02
A2***	20,658	714.8	33.2	21,414	712.4	34.8	-2.4	34.0	-0.07

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

***A1: Algebra I, GE: Geometry, A2: Algebra II

Table A.14.28 Mathematics Longitudinal Scale Score Comparison: Original to Original

	2018 Original			2019 Original			2019-2018		
Grade	N**	Mean	SD	N**	Mean	SD	DIFF*	SD	D
3	80,700	741.9	39.1	79,361	741.7	38.2	-0.2	38.7	0
4	82,028	737.9	34.8	80,844	739.5	35.8	1.6	35.3	0.05
5	80,953	738	34.9	81,733	738.7	34.4	0.7	34.6	0.02
6	76,153	732.9	32.4	79,141	731.6	32.8	-1.4	32.7	-0.04
7	62,141	731.5	28.9	63,242	731.3	28.7	-0.1	28.8	0
8	41,129	714.6	34.4	40,263	710.2	32.8	-4.3	33.6	-0.13
A1***	82,923	736.5	36.3	86,205	734.3	35	-2.1	35.7	-0.06
GE***	7,110	726.1	24.6	6,967	727.5	27.2	1.5	25.9	0.06
A2***	2,841	727.6	33.6	2,943	725.5	34.1	-2.2	33.9	-0.06

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

***A1: Algebra I, GE: Geometry, A2: Algebra II

Table A.14.29 ELA/L Longitudinal Regression

	Sample Size		R2			
Grade (Prior Grade)	Original-Current	Original-Original	All	Full	Reduced	Change
4 (3)	251,957	70,459	322,416	0.6486	0.648	0.0007
5 (4)	258,568	71,980	330,548	0.6948	0.6948	0
6 (5)	261,213	69,545	330,758	0.6967	0.6966	0.0001
7 (6)	255,849	70,466	326,315	0.7093	0.709	0.0004
8 (7)	253,432	68,542	321,974	0.7263	0.7261	0.0002
9 (8)	109,156	3,015	112,171	0.7306	0.7306	0.0001
10 (8)	103,001	53,963	156,964	0.6598	0.6338	0.026

Table A.14.30 Mathematics Longitudinal Regression

	Sample Size		R2			
Grade (Prior Grade)	Original-Current	Original-Original	All	Full	Reduced	Change
4 (3)	254,114	75,024	329,138	0.7335	0.7332	0.0003
5 (4)	260,243	76,369	336,612	0.7286	0.7283	0.0003
6 (5)	261,817	73,544	335,361	0.7121	0.712	0.0001
7 (6)	251,850	59,342	311,192	0.7391	0.7388	0.0003
8 (7)	213,821	37,357	251,178	0.6821	0.6795	0.0026
A1 (7,8) ***	105,010	50,900	155,910	0.6443	0.642	0.0023
GE (A1) ***	92,531	11,117	103,648	0.6769	0.6707	0.0062
A2 (A1,GE) ***	60,547	4,136	64,683	0.6793	0.6766	0.0027

***A1: Algebra I, GE: Geometry, A2: Algebra II

Table A.14.31 ELA/L Grade 3 Performance Level Comparison

	N Count		Percent		
Level	Current	Original	Current	Original	DIFF
1	12,869	12,533	20.5	20	0.5
2	11,212	10,901	17.9	17.4	0.5
3	13,896	12,699	22.1	20.2	1.9
4	21,847	23,625	34.8	37.6	-2.8
5	2,929	2,995	4.7	4.8	-0.1
Cramer's V Effect Size = .03					

Table A.14.32 Mathematics Grade 3 Performance Level Comparison

	N Count		Percent		
Level	Current	Original	Current	Original	DIFF
1	5,315	5,430	10.2	10.5	-0.2
2	8,385	7,462	16.1	14.4	1.8
3	12,854	13,100	24.7	25.2	-0.5
4	19,894	19,503	38.3	37.5	0.8
5	5,509	6,462	10.6	12.4	-1.8
Cramer's V Effect Size = .04					

Table A.14.33 Performance Level Comparison Summary: Effect Sizes

ELA/L		Mathematics	
Grade	Cramer's V Effect Size	Grade/ Course	Cramer's V Effect Size
3	0.03	3	0.04
4	0.04	4	0.03
5	0.04	5	0.03
6	0.02	6	0.02
7	0.02	7	0.02
8	0.04	8	0.06
10	0.20	Algebra I	0.09
		Geometry	0.04
		Algebra II	0.07

Table A.14.34 College and Career Readiness Comparison Summary: Effect Sizes

Proportion of Students at or Above the CCR Cut							
ELA/L				Mathematics			
Grade	Current	Original	Cohen's *h***	Grade/Course	Current	Original	Cohen's *h* **
3	0.39	0.42	-0.06	3	0.49	0.50	-0.02
4	0.43	0.46	-0.05	4	0.46	0.48	-0.03
5	0.45	0.46	-0.03	5	0.43	0.44	-0.02
6	0.43	0.43	-0.01	6	0.34	0.34	0
7	0.48	0.50	-0.04	7	0.30	0.30	0
8	0.48	0.47	0.01	8	0.18	0.14	0.09
10	0.51	0.68	-0.35	Algebra I	0.38	0.42	-0.09
				Geometry	0.87	0.86	0.03
				Algebra II	0.86	0.89	-0.09

**Computed as Current proportion – Original proportion

Table A.14.35 ELA/L Classification Accuracy

	Performance Level Classification			College and Career Readiness* Classification		
Grade	Current	Original	Cohen's *h*	Current	Original	Cohen's *h*
3	0.71	0.75	-0.10	0.90	0.92	-0.05
4	0.68	0.74	-0.13	0.89	0.91	-0.06
5	0.72	0.78	-0.15	0.90	0.92	-0.08
6	0.74	0.79	-0.13	0.91	0.92	-0.06
7	0.71	0.77	-0.13	0.91	0.93	-0.06
8	0.71	0.77	-0.13	0.91	0.93	-0.07
10	0.67	0.77	-0.23	0.90	0.93	-0.10

Table A.14.36 ELA/L Classification Consistency

	Performance Level Classification			College and Career Readiness* Classification		
Grade	Current	Original	Cohen's *h*	Current	Original	Cohen's *h*
3	0.61	0.66	-0.10	0.86	0.88	-0.06
4	0.57	0.64	-0.15	0.85	0.88	-0.07
5	0.62	0.70	-0.17	0.86	0.89	-0.09
6	0.64	0.71	-0.15	0.87	0.89	-0.08
7	0.60	0.67	-0.15	0.87	0.90	-0.07
8	0.62	0.69	-0.15	0.87	0.90	-0.08
10	0.57	0.69	-0.25	0.86	0.90	-0.12

Table A.14.37 Mathematics Classification Accuracy

Grade/ Course	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	Cohen's *h*	Current	Original	Cohen's *h*
3	0.75	0.78	-0.06	0.91	0.93	-0.05
4	0.78	0.80	-0.05	0.92	0.92	-0.02
5	0.77	0.79	-0.04	0.92	0.93	-0.02
6	0.77	0.81	-0.10	0.92	0.94	-0.05
7	0.77	0.79	-0.04	0.92	0.93	-0.03
8	0.71	0.73	-0.04	0.92	0.93	-0.06
Algebra I	0.74	0.79	-0.11	0.91	0.92	-0.06
Geometry	0.81	0.85	-0.11	0.96	0.96	-0.03
Algebra II	0.82	0.86	-0.1	0.92	0.95	-0.10

Table A.14.38 Mathematics Classification Consistency

Grade/ Course	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	h	Current	Original	h
3	0.66	0.69	-0.07	0.88	0.90	-0.06
4	0.69	0.72	-0.06	0.89	0.89	-0.03
5	0.68	0.70	-0.05	0.89	0.90	-0.02
6	0.68	0.73	-0.12	0.89	0.91	-0.06
7	0.68	0.70	-0.05	0.89	0.90	-0.04
8	0.61	0.63	-0.05	0.88	0.90	-0.07
Algebra I	0.65	0.70	-0.13	0.87	0.89	-0.07
Geometry	0.73	0.78	-0.13	0.94	0.94	-0.04
Algebra II	0.74	0.79	-0.12	0.89	0.92	-0.12

Table A.14.39 ELA/L Grade 6 Performance Level Comparison

	Original to Current			Original to Original		
Level	Current States 2018	Current States 2019	DIFF	Original States 2018	Original States 2019	DIFF
1	10.2	11.3	1.1	12.4	12.6	0.2
2	20.1	17.9	-2.2	21.3	18.8	-2.5
3	28	28.5	0.5	27.7	27.5	-0.2
4	33.3	33.8	0.5	32.1	34.3	2.2
5	8.3	8.4	0.1	6.6	6.8	0.2
	Cramer's V Effect Size = .03			Cramer's V Effect Size = .03		

Table A.14.40 Mathematics Grade 6 Performance Level Comparison

	Original to Current			Original to Original		
Level	Current States 2018	Current States 2019	DIFF	Original States 2018	Original States 2019	DIFF
1	13.4	14.4	1	15.7	17.5	1.8
2	25.9	28.0	2.1	26.1	25.9	-0.2
3	28.4	27.4	-0.9	26.8	26.8	0
4	27.4	25.5	-1.9	26.9	25.4	-1.5
5	5	4.7	-0.3	4.5	4.3	-0.2
	Cramer's V Effect Size = .03			Cramer's V Effect Size = .03		

Table A.14.41 Performance Level Comparison Summary: Effect Sizes

ELA/L			Mathematics		
Grade	Original to Current	Original to Original	Grade/ Course	Original to Current	Original to Original
3	0.02	0.03	3	0.04	0.05
4	0.03	0.02	4	0.05	0.02
5	0.02	0.03	5	0.06	0.05
6	0.03	0.03	6	0.03	0.03
7	0.02	0.03	7	0.03	0.06
8	0.04	0.05	8	0.04	0.08
9	0.04	0.05	Algebra I	0.10	0.05
10	0.09	0.04	Geometry	0.07	0.06
			Algebra II	0.05	0.05

Table A.14.42 ELA/L Reading Claim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	46	0.9	30	0.86	0.85	0.01
4	64	0.88	42	0.83	0.83	0
5	64	0.9	42	0.85	0.86	-0.01
6	64	0.91	42	0.87	0.87	0
7	64	0.91	42	0.86	0.87	-0.01
8	64	0.9	42	0.85	0.86	-0.01
10	64	0.89	42	0.82	0.84	-0.02

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.43 ELA/L Writing Claim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	36	0.85	24	0.79	0.79	0
4	42	0.86	28	0.8	0.8	0
5	42	0.86	29	0.8	0.81	-0.01
6	45	0.87	30	0.82	0.82	0
7	45	0.88	30	0.83	0.83	0
8	45	0.89	30	0.85	0.84	0.01
10	45	0.88	30	0.84	0.83	0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.44 ELA/L Reading Information (RI) Subclaim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	17	0.74	11	0.68	0.65	0.03
4	26	0.76	16	0.62	0.66	-0.04
5	23	0.75	14	0.56	0.65	-0.09
6	24	0.76	16	0.67	0.68	-0.01
7	24	0.81	14	0.66	0.71	-0.05
8	21	0.78	15	0.71	0.72	-0.01
10	30	0.8	19	0.68	0.72	-0.04

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.45 ELA/L Reading Literature (RL) Subclaim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	19	0.8	11	0.71	0.7	0.01
4	26	0.73	17	0.66	0.64	0.02
5	26	0.79	17	0.74	0.71	0.03
6	26	0.84	18	0.76	0.78	-0.02
7	25	0.79	17	0.7	0.72	-0.02
8	26	0.79	16	0.69	0.7	-0.01
10	20	0.7	14	0.61	0.62	-0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.46 ELA/L Reading Vocabulary (RV) Subclaim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	10	0.68	8	0.61	0.63	-0.02
4	12	0.61	9	0.56	0.54	0.02
5	15	0.75	11	0.67	0.69	-0.02
6	14	0.72	8	0.58	0.56	-0.02
7	15	0.66	11	0.62	0.59	0.03
8	17	0.69	11	0.53	0.59	-0.06
10	14	0.6	10	0.47	0.52	-0.05

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.47 ELA/L Writing Knowledge and Conventions (WKL) Subclaim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	9	0.87	6	0.82	0.82	0
4	9	0.88	6	0.84	0.83	0.01
5	9	0.88	6	0.84	0.83	0.01
6	9	0.89	6	0.85	0.84	0.01
7	9	0.89	6	0.86	0.84	0.02
8	9	0.91	6	0.87	0.87	0
10	9	0.89	6	0.86	0.84	0.02

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.48 ELA/L Written Expression (WE) Subclaim Reliability

	Original		Current			
Grade	Points	Alpha	Points	Alpha	SB	Diff*
3	27	0.81	18	0.74	0.74	0
4	33	0.83	22	0.77	0.76	0.01
5	33	0.81	23	0.72	0.75	-0.03
6	36	0.86	24	0.81	0.8	0.01
7	36	0.88	24	0.85	0.83	0.02
8	36	0.9	24	0.86	0.86	0
10	36	0.88	24	0.85	0.83	0.02

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.49 Mathematics Subclaim A Reliability

	Original		Current			
Grade/Course	Points	Alpha	Points	Alpha	SB	Diff*
3	28	0.91	20	0.86	0.88	-0.02
4	31	0.9	21	0.86	0.86	0
5	30	0.9	20	0.86	0.86	0
6	26	0.88	20	0.83	0.85	-0.02
7	29	0.87	20	0.84	0.82	0.02
8	27	0.77	20	0.74	0.71	0.03
Algebra I	26	0.79	17	0.72	0.71	0.01
Geometry	30	0.84	18	0.79	0.76	0.03
Algebra II	25	0.74	16	0.66	0.65	0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.50 Mathematics Subclaim B Reliability

	Original		Current			
Grade/Course	Points	Alpha	Points	Alpha	SB	Diff*
3	12	0.76	10	0.69	0.73	-0.04
4	9	0.72	9	0.72	0.72	0
5	10	0.71	10	0.7	0.71	-0.01
6	14	0.77	10	0.67	0.71	-0.04
7	11	0.67	10	0.64	0.65	-0.01
8	13	0.53	10	0.49	0.46	0.03
Algebra I	17	0.73	9	0.64	0.59	0.05
Geometry	19	0.79	12	0.65	0.7	-0.05
Algebra II	20	0.7	12	0.55	0.58	-0.03

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.51 Mathematics Subclaim C Reliability

	Original		Current			
Grade/Course	Points	Alpha	Points	Alpha	SB	Diff*
3	14	0.62	10	0.48	0.54	-0.06
4	14	0.79	10	0.76	0.73	0.03
5	14	0.71	10	0.62	0.64	-0.02
6	14	0.78	10	0.71	0.72	-0.01
7	14	0.64	10	0.52	0.56	-0.04
8	14	0.59	10	0.54	0.51	0.03
Algebra I	14	0.75	10	0.7	0.68	0.02
Geometry	14	0.64	10	0.6	0.56	0.04
Algebra II	14	0.55	10	0.44	0.47	-0.03

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.14.52 Mathematics Subclaim D Reliability

	Original		Current			
Grade/Course	Pts.	Alpha	Pts.	Alpha	SB	Diff*
3	12	0.76	12	0.75	-	-
4	12	0.66	12	0.66	-	-
5	12	0.74	12	0.73	-	-
6	12	0.71	12	0.69	-	-
7	12	0.73	12	0.74	-	-
8	12	0.5	12	0.52	-	-
Algebra I	18	0.75	15	0.69	0.71	-0.02
Geometry	18	0.7	15	0.64	0.66	-0.02
Algebra II	18	0.59	15	0.56	0.55	0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Addendum

The addendum presents the results of analyses for the fall/winter block 2018 operational administration. These results are reported separately from the spring 2019 results since fall testing included additional states or agencies and consisted of a nonrepresentative subset of students testing only ELA/L grades 9, 10, and 11, as well as Algebra I, Geometry, and Algebra II. Both online and paper test forms were administered for each test.

To organize the addendum, tables are numbered sequentially according to the section represented by the tables. The reader can refer back to the corresponding section in the technical report for related information on the topic. For example, the first addendum table provides participation counts similar to those provided for Section 11; therefore it is numbered ADD.11.1. The second addendum table for Section 11 is numbered ADD.11.2, and so on.

Addendum 11: Student Characteristics

Table ADD.11.1 State Participation in ELA/L Fall 2018 Operational Tests, by Grade

English Language Arts-Literacy					
State	Category	Total	Grade 9	Grade 10	Grade 11
All States	N of Students	43,970	4,651	26,181	13,138
	N of CBT	43,806	4,643	26,130	13,033
	% of CBT	100	100	100	99
	N of PBT	164	8	51	105
	% of PBT	0	0	0	1
BIE	% of All Data	0	n/a	n/a	0
	N of Students	65	n/a	n/a	65
	N of CBT	14	n/a	n/a	14
	% of CBT	22	n/a	n/a	22
	N of PBT	51	n/a	n/a	51
	% of PBT	79	n/a	n/a	79
MD	% of All Data	51	n/a	46	4
	N of Students	22,212	n/a	20,318	1,894
	N of CBT	22,168	n/a	20,275	1,893
	% of CBT	100	n/a	100	100
	N of PBT	44	n/a	43	1
	% of PBT	0	n/a	0	0
NJ	% of All Data	32	10	13	9
	N of Students	13,979	4,264	5,628	4,087
	N of CBT	13,952	4,256	5,620	4,076
	% of CBT	100	100	100	100
	N of PBT	27	8	8	11
	% of PBT	0	0	0	0
NM	% of All Data	18	1	1	16
	N of Students	7,714	387	235	7,092
	N of CBT	7,672	387	235	7,050
	% of CBT	100	100	100	99
	N of PBT	42	0	0	42
	% of PBT	1	0	0	1

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; n/a=not applicable.

Table ADD.11.2 State Participation in Mathematics Fall 2018 Operational Tests, by Course

Mathematics State	Category	Total	A1	GO	A2
All States	N of Students	48,917	32,649	5,956	10,312
	N of CBT	48,671	32,576	5,916	10,179
	% of CBT	100	100	99	99
	N of PBT	246	73	40	133
	% of PBT	1	0	1	1
BIE	% of All Data	0	n/a	0	0
	N of Students	119	n/a	19	100
	N of CBT	26	n/a	7	19
	% of CBT	22	n/a	37	19
	N of PBT	93	n/a	12	81
	% of PBT	78	n/a	63	81
MD	% of All Data	42	38	0	4
	N of Students	20,342	18,461	53	1,828
	N of CBT	20,284	18,405	53	1,826
	% of CBT	100	100	100	100
	N of PBT	58	56	0	2
	% of PBT	0	0	0	0
NJ	% of All Data	44	28	8	8
	N of Students	21,594	13,683	4,124	3,787
	N of CBT	21,566	13,666	4,115	3,785
	% of CBT	100	100	100	100
	N of PBT	28	17	9	2
	% of PBT	0	0	0	0
NM	% of All Data	14	1	4	9
	N of Students	6,862	505	1,760	4,597
	N of CBT	6,795	505	1,741	4,549
	% of CBT	99	100	99	99
	N of PBT	67	0	19	48
	% of PBT	1	0	1	1

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; n/a=not applicable. A1=Algebra I, GO=Geometry, A2=Algebra II.

Table ADD.11.3 State Participation in Spanish Mathematics Fall 2018 Operational Tests, by Course

Mathematics State	Category	Total	A1	GO	A2
All States	N of Students	531	327	96	108
	N of CBT	512	308	96	108
	% of CBT	96	94	100	100
	N of PBT	19	19	n/a	n/a
	% of PBT	4	6	n/a	n/a
BIE	% of All Data	n/a	n/a	n/a	n/a
	N of Students	n/a	n/a	n/a	n/a
	N of CBT	n/a	n/a	n/a	n/a
	% of CBT	n/a	n/a	n/a	n/a
	N of PBT	n/a	n/a	n/a	n/a
	% of PBT	n/a	n/a	n/a	n/a
MD	% of All Data	24	24	n/a	n/a
	N of Students	128	128	n/a	n/a
	N of CBT	109	109	n/a	n/a
	% of CBT	85	85	n/a	n/a
	N of PBT	19	19	n/a	n/a
	% of PBT	15	15	n/a	n/a
NJ	% of All Data	62	37	12	12
	N of Students	330	198	66	66
	N of CBT	330	198	66	66
	% of CBT	100	100	100	100
	N of PBT	n/a	n/a	n/a	n/a
	% of PBT	n/a	n/a	n/a	n/a
NM	% of All Data	14	0	6	8
	N of Students	73	1	30	42
	N of CBT	73	1	30	42
	% of CBT	100	100	100	100
	N of PBT	n/a	n/a	n/a	n/a
	% of PBT	n/a	n/a	n/a	n/a

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT = paper-based test; n/a=not applicable. A1=Algebra I, GO=Geometry, A2=Algebra II.

Table ADD.11.4 All States Combined: Fall 2018 ELA/L Students by Grade and Gender

Grade	Mode	Valid Cases	Female		Male	
			N	%	N	%
9	All	4,651	2,352	50.6	2,299	49.4
	CBT	4,643	2,349	50.6	2,294	49.4
	PBT	8	n/r	n/r	n/r	n/r
10	All	26,181	11,459	43.8	14,722	56.2
	CBT	26,130	11,443	43.8	14,687	56.2
	PBT	51	n/r	n/r	35	68.6
11	All	13,138	5,591	42.6	7,547	57.4
	CBT	13,033	5,544	42.5	7,489	57.5
	PBT	105	47	44.8	58	55.2

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; n/r=not reported due to n<20.

Table ADD.11.5 All States Combined: Fall 2018 Mathematics Students by Course and Gender

Course	Mode	Valid Cases	Female		Male	
			N	%	N	%
A1	All	32,649	15,886	48.7	16,763	51.3
	CBT	32,576	15,856	48.7	16,720	51.3
	PBT	73	30	41.1	43	58.9
A2	All	10,312	5,227	50.7	5,085	49.3
	CBT	10,179	5,159	50.7	5,020	49.3
	PBT	133	68	51.1	65	48.9
GO	All	5,956	2,918	49.0	3,038	51.0
	CBT	5,916	2,898	49.0	3,018	51.0
	PBT	40	20	50.0	20	50.0

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; A1=Algebra I, GO=Geometry, A2=Algebra II.

Table ADD.11.6 All States Combined: Fall 2018 Spanish-Language Mathematics Students by Course and Gender

Course	Mode	Valid Cases	Female		Male	
			N	%	N	%
A1	All	327	172	52.6	155	47.4
	CBT	308	165	53.6	143	46.4
	PBT	19	n/r	n/r	n/r	n/r
A2	All	108	50	46.3	58	53.7
	CBT	108	50	46.3	58	53.7
GO	All	96	44	45.8	52	54.2
	CBT	96	44	45.8	52	54.2

Note: BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico; CBT=computer-based test; PBT=paper-based test; A1=Algebra I, GO=Geometry, A2=Algebra II. n/r=not reported due to n<20.

Table ADD.11.7 Demographic Information for Fall 2018 Grade 9 ELA/L, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	33.0	n/a	n/a	33.2	31.8
Student with Disabilities	18.6	n/a	n/a	19.7	6.2
English learner	1.4	n/a	n/a	0.8	7.8
Male	49.4	n/a	n/a	49.7	46.0
Female	50.6	n/a	n/a	50.3	54.0
American Indian/Alaska Native	1.2	n/a	n/a	n/r	12.1
Asian	6.6	n/a	n/a	7.0	n/r
Black/African American	18.6	n/a	n/a	20.2	n/r
Hispanic/Latino	24.2	n/a	n/a	21.4	55.3
White/Caucasian	46.0	n/a	n/a	48.0	24.0
Native Hawaiian/Pacific Islander	n/r	n/a	n/a	n/r	n/a
Two or More Races Reported	2.8	n/a	n/a	3.0	n/r
Unknown	n/r	n/a	n/a	n/a	n/r

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico. n/a = not applicable; and n/r = not reported due to n<20.

Table ADD.11.8 Demographic Information for Fall 2018 Grade 10 ELA/L, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	44.2	n/a	48.3	29.7	39.6
Student with Disabilities	24.5	n/a	26.1	19.5	n/r
English learner	13.8	n/a	17.0	2.4	8.5
Male	56.2	n/a	58.0	49.9	50.2
Female	43.8	n/a	42.0	50.1	49.8
American Indian/Alaska Native	0.4	n/a	0.3	n/r	19.1
Asian	3.8	n/a	2.5	8.3	n/r
Black/African American	40.9	n/a	47.0	20.5	n/r
Hispanic/Latino	22.3	n/a	22.5	20.5	47.7
White/Caucasian	29.7	n/a	24.6	48.3	30.2
Native Hawaiian/Pacific Islander	0.1	n/a	n/r	n/r	n/a
Two or More Races Reported	2.8	n/a	3.0	2.0	n/r
Unknown	n/a	n/a	n/a	n/a	n/a

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico n/a = not applicable; and n/r = not reported due to n<20.

Table ADD.11.9 Demographic Information for Fall 2018 Grade 11 ELA/L, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	47.2	95.4	40.0	31.3	57.9
Student with Disabilities	22.6	n/r	23.5	19.3	24.3
English learner	12.8	73.8	5.0	1.3	21.0
Male	57.4	60.0	62.5	51.9	59.2
Female	42.6	40.0	37.5	48.1	40.8
American Indian/Alaska Native	7.1	96.9	n/r	n/r	12.1
Asian	3.1	n/a	2.8	6.6	1.1
Black/African American	14.9	n/a	43.5	23.0	2.6
Hispanic/Latino	43.1	n/a	10.8	20.2	65.3
White/Caucasian	29.0	n/a	39.0	47.7	15.8
Native Hawaiian/Pacific Islander	0.2	n/a	n/r	n/r	n/r
Two or More Races Reported	1.9	n/r	3.7	2.2	1.2
Unknown	0.9	n/r	n/a	n/a	1.6

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico. n/a = not applicable; and n/r = not reported due to n<20.

Table ADD.11.10 Demographic Information for Fall 2018 Algebra I, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	43.3	n/a	49.4	35.0	47.7
Student with Disabilities	21.4	n/a	25.4	16.2	17.0
English learner	11.3	n/a	15.2	6.0	11.7
Male	51.3	n/a	53.0	49.2	48.1
Female	48.7	n/a	47.0	50.8	51.9
American Indian/Alaska Native	0.3	n/a	0.4	n/r	4.0
Asian	4.3	n/a	2.5	6.8	n/r
Black/African American	35.1	n/a	46.8	20.5	n/r
Hispanic/Latino	24.7	n/a	21.6	27.5	62.8
White/Caucasian	32.8	n/a	25.5	43.0	27.5
Native Hawaiian/Pacific Islander	0.2	n/a	0.1	0.2	n/r
Two or More Races Reported	2.5	n/a	3.0	1.8	n/a
Unknown	0.1	n/a	n/a	n/r	4.2

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico. n/a = not applicable; and n/r = not reported due to n<20.

Table ADD.11.11 Demographic Information for Fall 2018 Geometry, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	39.5	n/r	56.6	33.0	53.6
Student with Disabilities	18.4	n/r	n/r	18.2	19.0
English learner	8.0	n/r	n/r	3.4	18.0
Male	51.0	n/r	66.0	51.2	49.9
Female	49.0	n/r	n/r	48.8	50.1
American Indian/Alaska Native	4.0	n/r	n/a	n/r	12.3
Asian	5.1	n/a	n/r	7.0	n/r
Black/African American	14.7	n/a	n/r	19.9	2.8
Hispanic/Latino	35.9	n/a	n/r	24.0	64.9
White/Caucasian	37.5	n/a	75.5	46.7	15.2
Native Hawaiian/Pacific Islander	n/r	n/a	n/a	n/r	n/r
Two or More Races Reported	1.7	n/a	n/r	2.1	n/r
Unknown	0.8	n/a	n/a	n/r	2.8

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico. n/a = not applicable; and n/r = not reported due to n<20.

Table ADD.11.12 Demographic Information for Fall 2018 Algebra II, Overall and by State

Demographic	All States	BIE	MD	NJ	NM
Economically Disadvantaged	41.5	87.0	31.2	28.6	55.1
Student with Disabilities	13.5	n/r	10.3	13.9	14.7
English learner	10.5	69.0	2.0	3.6	18.3
Male	49.3	43.0	51.0	48.6	49.4
Female	50.7	57.0	49.0	51.4	50.6
American Indian/Alaska Native	8.1	97.0	n/r	n/r	15.7
Asian	5.0	n/a	4.4	10.4	0.9
Black/African American	11.6	n/a	21.1	18.4	2.5
Hispanic/Latino	36.4	n/a	9.3	21.0	60.7
White/Caucasian	35.5	n/a	59.5	48.2	16.3
Native Hawaiian/Pacific Islander	0.2	n/a	n/r	n/r	n/r
Two or More Races Reported	2.4	n/r	5.5	1.6	1.8
Unknown	0.8	n/r	n/a	n/a	1.8

Note: All States = data from all participating states combined; BIE=Bureau of Indian Education, MD=Maryland, NJ=New Jersey, and NM=New Mexico. n/a = not applicable; and n/r = not reported due to n<20.

Addendum 12: Scale Scores

Table ADD.12.1 Fall 2018 Subgroup Performance for ELA/L Scale Scores: Grade 9

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		4,651	744.52	35.20	650	850
Gender	Female	2,352	752.50	33.58	653	850
	Male	2,299	736.36	34.94	650	841
Ethnicity	American Indian/Alaska Native	56	751.79	31.41	661	808
	Asian	307	763.90	33.57	650	850
	Black/African American	865	731.91	32.70	650	850
	Hispanic/Latino	1,127	736.33	32.83	650	841
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	132	743.40	34.08	661	825
	White	2,140	750.73	34.97	650	850
Economic Status*	Not Economically Disadvantaged	3,114	750.32	34.38	650	850
	Economically Disadvantaged	1,537	732.77	33.89	650	850
English Learner Status	Non English Learner	4,588	745.00	35.00	650	850
	English Learner	63	709.32	31.55	650	799
Disabilities	Students without Disabilities	3,785	750.87	32.82	650	850
	Students with Disabilities	866	716.76	31.61	650	850
Reading Summative Score		4,651	48.17	14.39	10	90
Gender	Female	2,352	50.12	13.95	12	90
	Male	2,299	46.16	14.57	10	90
Ethnicity	American Indian/Alaska Native	56	51.27	12.33	15	75
	Asian	307	55.12	13.61	10	90
	Black/African American	865	43.64	13.32	10	84
	Hispanic/Latino	1,127	44.72	13.59	10	87
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	132	47.49	13.74	15	79
	White	2,140	50.68	14.44	10	90
Economic Status*	Not Economically Disadvantaged	3,114	50.42	14.13	10	90
	Economically Disadvantaged	1,537	43.60	13.83	10	87
English Learner Status	Non English Learner	4,588	48.36	14.32	10	90
	English Learner	63	34.43	13.03	10	81
Disabilities	Students without Disabilities	3,785	50.56	13.60	10	90
	Students with Disabilities	866	37.73	13.06	10	90

Group Type	Group	N	Mean	SD	Min	Max
Writing Summative Score		4,651	31.72	11.27	10	60
Gender	Female	2,352	35.01	9.80	10	60
	Male	2,299	28.36	11.68	10	60
Ethnicity	American Indian/Alaska Native	56	33.23	10.45	10	49
	Asian	307	37.43	9.48	10	60
	Black/African American	865	27.88	11.32	10	60
	Hispanic/Latino	1,127	29.72	11.14	10	53
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	132	31.94	11.14	10	60
	White	2,140	33.37	10.88	10	60
Economic Status*	Not Economically Disadvantaged	3,114	33.35	10.81	10	60
	Economically Disadvantaged	1,537	28.43	11.46	10	60
English Learner Status	Non English Learner	4,588	31.86	11.20	10	60
	English Learner	63	21.59	11.39	10	41
Disabilities	Students without Disabilities	3,785	33.70	10.25	10	60
	Students with Disabilities	866	23.07	11.45	10	53

*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.

Table ADD.12.2 Fall 2018 Subgroup Performance for ELA/L Scale Scores: Grade 10

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		26,181	717.92	46.86	650	850
Gender	Female	11,459	727.05	49.70	650	850
	Male	14,722	710.81	43.21	650	850
Ethnicity	American Indian/Alaska Native	108	720.44	34.22	650	799
	Asian	989	752.04	56.69	650	850
	Black/African American	10,715	702.16	34.52	650	850
	Hispanic/Latino	5,835	704.13	39.72	650	850
	Native Hawaiian/Pacific Islander	26	736.42	51.11	662	843
	Two or more races	724	726.30	46.46	650	850
	White	7,784	744.73	50.69	650	850
Economic Status*	Not Economically Disadvantaged	14,609	730.29	50.40	650	850
	Economically Disadvantaged	11,572	702.31	36.39	650	850
English Learner Status	Non English Learner	22,580	722.96	47.24	650	850
	English Learner	3,601	686.28	28.47	650	799
Disabilities	Students without Disabilities	19,761	724.08	47.77	650	850
	Students with Disabilities	6,419	698.95	38.10	650	850
Reading Summative Score		26,181	37.83	18.76	10	90
Gender	Female	11,459	40.10	19.62	10	90
	Male	14,722	36.05	17.85	10	90
Ethnicity	American Indian/Alaska Native	108	38.25	14.36	10	74
	Asian	989	50.63	23.19	10	90
	Black/African American	10,715	32.30	14.34	10	90
	Hispanic/Latino	5,835	31.61	15.81	10	90
	Native Hawaiian/Pacific Islander	26	45.04	22.76	10	90
	Two or more races	724	41.52	18.59	10	90
	White	7,784	48.08	20.24	10	90
Economic Status*	Not Economically Disadvantaged	14,609	42.61	20.09	10	90
	Economically Disadvantaged	11,572	31.78	14.87	10	90
English Learner Status	Non English Learner	22,580	39.99	18.83	10	90
	English Learner	3,601	24.25	11.03	10	79
Disabilities	Students without Disabilities	19,761	40.02	19.08	10	90
	Students with Disabilities	6,419	31.09	15.94	10	90
Writing Summative Score		26,181	25.88	12.88	10	60
Gender	Female	11,459	29.07	13.26	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	14,722	23.40	12.00	10	60
Ethnicity	American Indian/Alaska Native	108	27.08	10.46	10	48
	Asian	989	34.42	14.06	10	60
	Black/African American	10,715	21.63	10.47	10	60
	Hispanic/Latino	5,835	23.33	11.46	10	60
	Native Hawaiian/Pacific Islander	26	30.77	12.24	10	51
	Two or more races	724	27.62	12.87	10	60
	White	7,784	32.37	13.59	10	60
Economic Status*	Not Economically Disadvantaged	14,609	28.85	13.57	10	60
	Economically Disadvantaged	11,572	22.13	10.84	10	60
English Learner Status	Non English Learner	22,580	26.93	13.03	10	60
	English Learner	3,601	19.31	9.54	10	47
Disabilities	Students without Disabilities	19,761	27.60	12.97	10	60
	Students with Disabilities	6,419	20.59	11.03	10	60

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

Table ADD.12.3 Fall 2018 Subgroup Performance for ELA/L Scale Scores: Grade 11

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		13,138	719.36	38.49	650	850
Gender	Female	5,591	728.99	39.61	650	850
	Male	7,547	712.22	36.02	650	850
Ethnicity	American Indian/Alaska Native	930	715.01	31.74	650	827
	Asian	403	746.89	47.49	650	850
	Black/African American	1,951	712.87	36.34	650	839
	Hispanic/Latino	5,657	710.00	33.02	650	842
	Native Hawaiian/Pacific Islander	25	731.12	42.37	653	808
	Two or more races	249	732.58	42.36	650	842
	White	3,810	733.80	41.43	650	850
Economic Status*	Not Economically Disadvantaged	6,733	728.33	41.20	650	850
	Economically Disadvantaged	6,207	709.69	32.93	650	842
English Learner Status	Non English Learner	11,454	722.61	39.01	650	850
	English Learner	1,684	697.24	25.40	650	798
Disabilities	Students without Disabilities	10,056	725.34	38.66	650	850
	Students with Disabilities	2,968	699.03	30.42	650	829
Reading Summative Score		13,138	38.87	15.25	10	90
Gender	Female	5,591	41.83	15.65	10	90
	Male	7,547	36.68	14.56	10	90
Ethnicity	American Indian/Alaska Native	930	35.42	12.24	10	84
	Asian	403	48.87	18.17	10	90
	Black/African American	1,951	36.53	14.38	10	86
	Hispanic/Latino	5,657	35.27	13.08	10	89
	Native Hawaiian/Pacific Islander	25	44.64	17.59	13	79
	Two or more races	249	44.16	16.66	10	89
	White	3,810	44.83	16.53	10	90
Economic Status*	Not Economically Disadvantaged	6,733	42.48	16.28	10	90
	Economically Disadvantaged	6,207	34.96	13.04	10	89
English Learner Status	Non English Learner	11,454	40.24	15.44	10	90
	English Learner	1,684	29.56	9.69	10	74
Disabilities	Students without Disabilities	10,056	41.17	15.31	10	90
	Students with Disabilities	2,968	31.06	12.27	10	82
Writing Summative Score		13,138	22.94	13.02	10	60
Gender	Female	5,591	26.73	13.18	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	7,547	20.13	12.16	10	60
Ethnicity	American Indian/Alaska Native	930	24.20	11.69	10	56
	Asian	403	31.17	14.75	10	60
	Black/African American	1,951	20.82	12.45	10	52
	Hispanic/Latino	5,657	20.22	11.79	10	60
	Native Hawaiian/Pacific Islander	25	24.84	14.11	10	44
	Two or more races	249	26.31	14.03	10	56
	White	3,810	26.62	13.77	10	60
Economic Status*	Not Economically Disadvantaged	6,733	25.33	13.67	10	60
	Economically Disadvantaged	6,207	20.36	11.81	10	60
English Learner Status	Non English Learner	11,454	23.70	13.22	10	60
	English Learner	1,684	17.77	10.20	10	46
Disabilities	Students without Disabilities	10,056	24.64	13.22	10	60
	Students with Disabilities	2,968	17.15	10.52	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.

Table ADD.12.4 Subgroup Performance for Mathematics Scale Scores: Algebra I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		32,649	725.43	30.32	650	850
Gender	Female	15,886	727.24	30.36	650	850
	Male	16,763	723.72	30.17	650	850
Ethnicity	American Indian/Alaska Native	108	718.12	28.37	650	793
	Asian	1,411	746.75	37.01	650	850
	Black/African American	11,459	714.02	24.16	650	850
	Hispanic/Latino	8,060	719.67	27.27	650	850
	Native Hawaiian/Pacific Islander	61	734.98	31.31	650	807
	Two or more races	805	726.8	31	650	841
	White	10,720	738.99	30.57	650	850
Economic Status*	Not Economically Disadvantaged	18,497	731.89	31.76	650	850
	Economically Disadvantaged	14,149	717	26	650	850
English Learner Status	Non English Learner	28,962	728.04	30.17	650	850
	English Learner	3,687	705	22.73	650	844
Disabilities	Students without Disabilities	25,644	729.02	30.38	650	850
	Students with Disabilities	7,001	712.32	26.21	650	841
Language Form	Spanish	327	703.13	20.3	650	754

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

Table ADD.12.5 Subgroup Performance for Mathematics Scale Scores: Geometry

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		10,312	713.12	39.07	650	850
Gender	Female	5,227	714.46	37.73	650	850
	Male	5,085	711.74	40.35	650	850
Ethnicity	American Indian/Alaska Native	832	695.45	23.59	650	809
	Asian	514	761.25	49.22	650	850
	Black/African American	1,196	701.48	30.06	650	850
	Hispanic/Latino	3,755	698.83	27.46	650	850
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	243	718	40.67	650	833
	White	3,663	728.72	41.55	650	850
Economic Status*	Not Economically Disadvantaged	5,808	723.97	42.99	650	850
	Economically Disadvantaged	4,276	699.26	27.81	650	850
English Learner Status	Non English Learner	9,227	716.04	39.58	650	850
	English Learner	1,083	688.21	22.22	650	829
Disabilities	Students without Disabilities	8,814	716.46	39.49	650	850
	Students with Disabilities	1,397	693.06	29.91	650	840
Language Form	Spanish	108	679.4	18.26	650	725

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

Table ADD.12.6 Subgroup Performance for Mathematics Scale Scores: Algebra II

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		5,956	725.07	25.61	650	850
Gender	Female	2,918	725.8	25.06	650	850
	Male	3,038	724.38	26.12	650	845
Ethnicity	American Indian/Alaska Native	238	709.8	14.53	660	750
	Asian	306	751.33	32.25	680	845
	Black/African American	876	718.93	22.12	650	786
	Hispanic/Latino	2,139	715.67	21.18	650	811
	Native Hawaiian/Pacific Islander	n/r	n/r	n/r	n/r	n/r
	Two or more races	101	729.97	23.86	669	800
	White	2,234	734.05	24.78	654	850
Economic Status*	Not Economically Disadvantaged	3,585	730.5	26.77	650	850
	Economically Disadvantaged	2,353	716.89	21.32	650	800
English Learner Status	Non English Learner	5,482	726.7	25.58	650	850
	English Learner	474	706.31	17.1	650	785
Disabilities	Students without Disabilities	4,841	727.93	25.98	650	850
	Students with Disabilities	1,096	712.58	19.65	650	790
Language Form	Spanish	96	702.61	15.55	650	740

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.

Addendum 13: Reliability

Table ADD.13.1 shows the total group level reliability estimates and raw score SEM for the fall 2018 forms. Tables ADD.13.2 – ADD.13.7 show the subgroup reliability estimates and raw score SEM. A minimum sample size of 100 per core form was required for calculating the reliability estimates for subgroups; therefore, the subgroup totals may not equal the total group sample size. Tables ADD.13.8 – ADD.13.10 provide the claim and subclaim reliability and raw score SEM estimates for the fall 2018 forms. The paper-based tests did not have sufficient sample sizes for reliability analyses.

Table ADD.13.1 Summary of ELA/L Test Reliability Estimates for Fall 2018 Total Group

Grade Level	Number of Forms	Avg. Max. Possible Score	Avg. Raw Score SEM	Average Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
ELA09	2	109	5.79	0.93	177	0.87	4,318	0.93
ELA10	2	109	5.51	0.94	204	0.87	13,359	0.94
ELA11	2	109	5.30	0.93	175	0.79	9,920	0.93
ALG01	2	81	3.45	0.93	13,504	0.93	1,312	0.93
GEO01	2	81	3.28	0.92	779	0.91	4,216	0.92
ALG02	2	81	3.39	0.94	709	0.93	7,238	0.94

Table ADD.13.2 Summary of Test Reliability Estimates for Fall 2018 Subgroups: ELA/L Grade 9

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	109	5.79	0.93	177	0.87	4318	0.93
Gender							
Male	109	5.61	0.93	112	0.87	2,089	0.93
Female	109	5.99	0.92	2,229	0.92	2,229	0.92
Ethnicity							
White	109	5.91	0.93	2,007	0.93	2,007	0.93
Black/African American	109	5.76	0.92	762	0.92	762	0.92
Asian/Pacific Islander	109	5.77	0.93	300	0.93	300	0.93
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	109	5.81	0.92	1,048	0.92	1,048	0.92
Multiple	109	5.74	0.94	126	0.94	126	0.94
Special Instruction Needs							
Economically Disadvantaged	109	5.70	0.92	1,383	0.92	1,383	0.92
Not Economically Disadvantaged	109	5.91	0.93	2,935	0.93	2,935	0.93
English Learner	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Non-English Learner	109	5.80	0.93	177	0.87	4,258	0.93
Students with Disabilities	109	5.21	0.91	177	0.87	649	0.92
Students without Disabilities	109	5.92	0.92	3,669	0.92	3,669	0.92
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	109	4.65	0.87	177	0.87	177	0.87

n/r = not reported due to n<100.

Table ADD.13.3 Summary of Test Reliability Estimates for Fall 2018 Subgroups: ELA/L Grade 10

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	109	5.51	0.94	204	0.87	13359	0.94
Gender							
Male	109	5.33	0.93	131	0.86	7,039	0.94
Female	109	5.71	0.94	6,320	0.94	6,320	0.94
Ethnicity							
White	109	5.77	0.93	5,863	0.93	5,863	0.93
Black/African American	109	5.11	0.90	3,883	0.90	3,883	0.90
Asian/Pacific Islander	109	5.89	0.94	697	0.94	697	0.94
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	109	5.30	0.92	2,458	0.92	2,458	0.92
Multiple	109	5.62	0.94	376	0.94	376	0.94
Special Instruction Needs							
Economically Disadvantaged	109	5.22	0.91	4,439	0.91	4,439	0.91
Not Economically Disadvantaged	109	5.65	0.94	112	0.87	8,920	0.94
English Learner	109	4.50	0.82	1,042	0.82	1,042	0.82
Non-English Learner	109	5.58	0.94	197	0.87	12,317	0.94
Students with Disabilities	109	4.99	0.92	204	0.87	2,783	0.93
Students without Disabilities	109	5.65	0.94	10,575	0.94	10,575	0.94
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	109	4.64	0.84	190	0.84	190	0.84

n/r = not reported due to n<100.

Table ADD.13.4 Summary of Test Reliability Estimates for Fall 2018 Subgroups: ELA/L Grade 11

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	109	5.30	0.93	175	0.79	9,920	0.93
Gender							
Male	109	4.97	0.92	128	0.82	5,520	0.92
Female	109	5.70	0.93	4,400	0.93	4,400	0.93
Ethnicity							
White	109	5.81	0.93	3,217	0.93	3,217	0.93
Black/African American	109	5.41	0.92	1,099	0.92	1,099	0.92
Asian/Pacific Islander	109	5.67	0.95	349	0.95	349	0.95
American Indian/Alaska Native	109	5.07	0.91	509	0.91	509	0.91
Hispanic/Latino	109	4.85	0.90	4,464	0.90	4,464	0.90
Multiple	109	5.98	0.93	189	0.93	189	0.93
Special Instruction Needs							
Economically Disadvantaged	109	4.90	0.91	4,461	0.91	4,461	0.91
Not Economically Disadvantaged	109	5.64	0.93	5,326	0.93	5,326	0.93
English Learner	109	4.18	0.77	1,117	0.77	1,117	0.77
Non-English Learner	109	5.42	0.93	152	0.80	8,803	0.93
Students with Disabilities	109	4.39	0.88	175	0.79	2,103	0.89
Students without Disabilities	109	5.53	0.93	7,757	0.93	7,757	0.93
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	109	4.40	0.79	169	0.79	169	0.79

n/r = not reported due to n<100.

Table ADD.13.5 Summary of Test Reliability Estimates for Subgroups: Algebra I

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	81	3.45	0.93	13,504	0.93	1,312	0.93
Gender							
Male	81	3.40	0.93	6,714	0.93	673	0.94
Female	81	3.50	0.93	639	0.92	6,790	0.93
Ethnicity							
White	81	3.62	0.93	5,569	0.93	555	0.93
Black/African American	81	3.11	0.86	245	0.84	3,961	0.87
Asian/Pacific Islander	81	3.71	0.96	722	0.96	722	0.96
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	81	3.29	0.89	401	0.86	2,786	0.90
Multiple	81	3.47	0.93	367	0.93	367	0.93
Special Instruction Needs							
Economically Disadvantaged	81	3.55	0.93	8,755	0.93	814	0.93
Not Economically Disadvantaged	81	2.85	0.86	201	0.82	958	0.87
English Learner	81	3.48	0.93	1,111	0.93	12,546	0.93
Non-English Learner	81	3.09	0.89	392	0.85	2,574	0.90
Students with Disabilities	81	3.51	0.93	920	0.93	10,926	0.93
Students without Disabilities	81	3.55	0.93	8,755	0.93	814	0.93
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	81	3.60	0.93	1,199	0.93	1,199	0.93
Students Taking Translated Forms							
Spanish Language Form	81	2.61	0.56	126	0.56	126	0.56

n/r = not reported due to n<100.

Table ADD.13.6 Summary of Test Reliability Estimates for Subgroups: Geometry

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	81	3.28	0.92	779	0.91	4,216	0.92
Gender							
Male	81	3.24	0.93	428	0.92	2,088	0.93
Female	81	3.31	0.91	351	0.91	2,128	0.92
Ethnicity							
White	81	3.50	0.91	340	0.90	1,666	0.92
Black/African American	81	3.06	0.87	107	0.86	592	0.88
Asian/Pacific Islander	81	3.94	0.96	245	0.96	245	0.96
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	81	2.94	0.86	1,413	0.85	270	0.92
Multiple	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Special Instruction Needs							
Economically Disadvantaged	81	2.99	0.86	1,596	0.85	269	0.91
Not Economically Disadvantaged	81	3.42	0.93	509	0.91	2,611	0.93
English Learner	81	2.54	0.76	239	0.73	129	0.82
Non-English Learner	81	3.32	0.92	650	0.91	3,977	0.92
Students with Disabilities	81	2.80	0.83	223	0.81	665	0.84
Students without Disabilities	81	3.36	0.93	556	0.91	3,540	0.93
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	81	3.36	0.91	689	0.91	689	0.91
Students Taking Translated Forms							
Spanish Language Form	n/r	n/r	n/r	n/r	n/r	n/r	n/r

n/r = not reported due to n<100.

Table ADD.13.7 Summary of Test Reliability Estimates for Subgroups: Algebra II

	Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	81	3.39	0.94	709	0.93	7,238	0.94
Gender							
Male	81	3.33	0.95	407	0.93	3,448	0.95
Female	81	3.44	0.93	302	0.92	3,790	0.93
Ethnicity							
White	81	3.75	0.94	317	0.94	2,675	0.94
Black/African American	81	3.02	0.88	751	0.88	751	0.88
Asian/Pacific Islander	81	4.04	0.96	408	0.96	408	0.96
American Indian/Alaska Native	81	2.86	0.76	404	0.76	404	0.76
Hispanic/Latino	81	2.97	0.84	248	0.80	2,740	0.85
Multiple	81	3.66	0.94	177	0.94	177	0.94
Special Instruction Needs							
Economically Disadvantaged	81	2.97	0.85	2,891	0.85	261	0.88
Not Economically Disadvantaged	81	3.64	0.95	441	0.93	4,163	0.95
English Learner	81	2.62	0.74	155	0.69	546	0.75
Non-English Learner	81	3.45	0.94	554	0.93	6,691	0.94
Students with Disabilities	81	2.86	0.88	276	0.85	794	0.90
Students without Disabilities	81	3.46	0.94	431	0.93	6,370	0.94
Students Taking Accommodated Forms							
ASL	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-Caption	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	81	3.36	0.91	689	0.91	689	0.91
Students Taking Translated Forms							
Spanish Language Form	81	3.18	0.93	633	0.93	633	0.93

n/r = not reported due to n<100.

Table ADD.13.8 Average ELA/L Reliability Estimates for Fall 2018 Total Test and Subscores

	Reading: Total		Reading: Literature		Reading: Information		Reading: Vocabulary		Writing: Total		Writing Expression		Writing: Knowledge Language and Conventions	
Grade Level	Max Raw Score	Average Reliability	Max Raw Score	Average Reliability	Raw Score	Average Reliability	Range of Max Raw Score	Average Reliability	Raw Score	Average Reliability	Raw Score	Average Reliability	Raw Score	Average Reliability
9	64	0.89	24	0.77	24	0.75	16	0.64	45	0.88	36	0.88	9	0.88
10	64	0.88	24	0.74	24	0.74	16	0.68	45	0.90	36	0.91	9	0.92
11	64	0.89	24	0.78	28	0.77	12	0.57	45	0.88	36	0.88	9	0.89

Table ADD.13.9 Average Mathematics Reliability Estimates for Fall 2018 Total Test and Subscores

	Major Content		Additional & Supporting Content		Mathematics Reasoning		Modeling Practice	
Grade Level	Max Raw Score	Average Reliability	Max Raw Score	Average Reliability	Max Raw Score	Average Reliability	Max Raw Score	Average Reliability
A1	26	0.81	17	0.63	14	0.74	18	0.76
GO	30	0.82	19	0.69	14	0.75	18	0.71
A2	22	0.80	20	0.71	14	0.76	18	0.77

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II,

Tables ADD.13.10 and ADD.13.11 provide information about the accuracy and the consistency of two classifications made on the basis of the scores on the fall block 2018 English language arts/literacy and mathematics assessments, respectively. The columns labeled "Exact level" provide the classification of the student into one of five achievement levels. The columns labeled "Level 4 or higher vs. 3 or lower" provide the classification of the student 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).

Tables ADD.13.12 to ADD.13.17 provide more detailed information about the accuracy and the consistency of the classification of students into proficiency levels for each fall block 2018 assessment. Each cell in the 5-by-5 table shows the estimated proportion of students who would be classified into a particular combination of proficiency levels. The sum of the five bold values on the diagonal should equal the exact level of decision accuracy or consistency presented in Tables ADD.13.10or ADD.13.11 for the corresponding assessment. For "Level 4 and higher vs. 3 and lower" found in Tables ADD.13.10 or ADD.13.11, the sum of the shaded values in Tables ADD.13.12 to ADD.13.17 should equal the level of decision accuracy or consistency for the corresponding assessment in ADD.13.10 or ADD.13.11. Note that the sums based on values may not match exactly to the values due to truncation and rounding.

Table ADD.13.10 Reliability of Classification: Summary for ELA/L Fall 2018

	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
Level	Exact Level	Level 4 or higher vs. 3 or lower	Exact Level	Level 4 or higher vs. 3 or lower
9	0.77	0.68	0.92	0.89
10	0.75	0.67	0.95	0.92
11	0.76	0.68	0.94	0.92

Table ADD.13.11 Reliability of Classification: Summary for Mathematics Fall 2018

	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
Level	Exact Level	Level 4 or higher vs. 3 or lower	Exact Level	Level 4 or higher vs. 3 or lower
A1	0.78	0.70	0.95	0.93
GO	0.79	0.69	0.95	0.93
A2	0.80	0.73	0.96	0.94

Note: A1 = Algebra I, GO = Geometry, A2 = Algebra II.

Table ADD.13.12 Reliability of Classification: Grade 9 ELA/L Fall 2018

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.09**	0.01	0.00	0.00	0.00	0.11
	700-724	0.02	**0.12**	0.03	0.00	0.00	0.17
	725-749	0.00	0.03	**0.18**	0.04	0.00	0.26
	750-809	0.00	0.00	0.04	**0.30**	0.03	0.37
	810-850	0.00	0.00	0.00	0.02	**0.07**	0.09
Decision Consistency	650-699	**0.09**	0.02	0.00	0.00	0.00	0.11
	700-724	0.03	**0.10**	0.05	0.00	0.00	0.18
	725-749	0.00	0.04	**0.15**	0.06	0.00	0.25
	750-809	0.00	0.00	0.06	**0.27**	0.03	0.36
	810-850	0.00	0.00	0.00	0.03	**0.07**	0.10

Table ADD.13.13 Reliability of Classification: Grade 10 ELA/L Fall 2018

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.34**	0.05	0.00	0.00	0.00	0.39
	700-724	0.03	**0.14**	0.04	0.00	0.00	0.22
	725-749	0.00	0.04	**0.10**	0.03	0.00	0.17
	750-809	0.00	0.00	0.03	**0.11**	0.02	0.16
	810-850	0.00	0.00	0.00	0.01	**0.06**	0.07
Decision Consistency	650-699	**0.33**	0.06	0.01	0.00	0.00	0.40
	700-724	0.04	**0.11**	0.04	0.00	0.00	0.20
	725-749	0.00	0.05	**0.08**	0.03	0.00	0.16
	750-809	0.00	0.01	0.04	**0.10**	0.02	0.16
	810-850	0.00	0.00	0.00	0.02	**0.06**	0.08

Table ADD.13.14 Reliability of Classification: Grade 11 ELA/L Fall 2018

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.31**	0.04	0.00	0.00	0.00	0.35
	700-724	0.04	**0.16**	0.04	0.00	0.00	0.25
	725-749	0.00	0.04	**0.13**	0.03	0.00	0.19
	750-809	0.00	0.00	0.03	**0.13**	0.01	0.17
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.04
Decision Consistency	650-699	**0.30**	0.05	0.00	0.00	0.00	0.36
	700-724	0.05	**0.13**	0.05	0.00	0.00	0.24
	725-749	0.00	0.05	**0.10**	0.03	0.00	0.19
	750-809	0.00	0.00	0.04	**0.11**	0.01	0.17
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.05

Table ADD.13.15 Reliability of Classification: Algebra I Fall 2018

	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	0.21
	700-724	0.04	**0.25**	0.05	0.00	0.00	0.33
	725-749	0.00	0.04	**0.20**	0.03	0.00	0.27
	750-809	0.00	0.00	0.03	**0.15**	0.00	0.18
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01
Decision Consistency	650-699	**0.16**	0.05	0.00	0.00	0.00	0.22
	700-724	0.05	**0.21**	0.06	0.00	0.00	0.32
	725-749	0.00	0.06	**0.17**	0.03	0.00	0.26
	750-809	0.00	0.00	0.04	**0.14**	0.01	0.19
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01

Table ADD.13.16 Reliability of Classification: Geometry Fall 2018

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.09**	0.04	0.00	0.00	0.00	0.12
	700-724	0.03	**0.37**	0.04	0.00	0.00	0.43
	725-749	0.00	0.05	**0.19**	0.03	0.00	0.28
	750-809	0.00	0.00	0.02	**0.12**	0.01	0.15
	810-850	0.00	0.00	0.00	0.00	**0.02**	0.02
Decision Consistency	650-699	**0.08**	0.06	0.00	0.00	0.00	0.14
	700-724	0.03	**0.32**	0.05	0.00	0.00	0.41
	725-749	0.00	0.07	**0.17**	0.04	0.00	0.28
	750-809	0.00	0.00	0.03	**0.11**	0.01	0.15
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.02

Table ADD.13.17 Reliability of Classification: Algebra II Fall 2018

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.39**	0.05	0.00	0.00	0.00	0.45
	700-724	0.03	**0.18**	0.03	0.00	0.00	0.24
	725-749	0.00	0.04	**0.08**	0.02	0.00	0.14
	750-809	0.00	0.00	0.02	**0.12**	0.01	0.15
	810-850	0.00	0.00	0.00	0.00	**0.02**	0.02
Decision Consistency	650-699	**0.38**	0.07	0.00	0.00	0.00	0.45
	700-724	0.05	**0.15**	0.04	0.00	0.00	0.23
	725-749	0.00	0.05	**0.07**	0.02	0.00	0.14
	750-809	0.00	0.00	0.03	**0.11**	0.01	0.15
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.02

Addendum 14: Validity

The intercorrelations for the fall 2018 assessments are presented in Tables ADD.14.1 through ADD.14.3 for ELA/L grades 9, 10, and 11 and Tables ADD.14.4 though ADD.14.6 for the traditional mathematics courses (A1, GO, A2). Like the spring intercorrelations, the ELA/L all have moderate to high values with the writing subclaims being highly intercorrelated. The mathematics intercorrelations have moderate values. Tables ADD.14.7 through ADD.14.9 are the correlations between ELA/L and mathematics from the fall block.

Table ADD.14.1 Average Intercorrelations and Reliability between Grade 9 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.89	4,630	4,630	4,630	4,630	4,630	4,630
RL	0.92	0.77	4,630	4,630	4,630	4,630	4,630
RI	0.91	0.73	0.75	4,630	4,630	4,630	4,630
RV	0.85	0.68	0.67	0.64	4,630	4,630	4,630
WR	0.75	0.72	0.71	0.54	0.88	4,630	4,630
WE	0.74	0.72	0.70	0.53	1	0.88	4,630
WKL	0.74	0.72	0.70	0.53	0.98	0.96	0.88

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

Table ADD.14.2 Average Intercorrelations and Reliability between Grade 10 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.88	15,277	15,277	15,277	15,277	15,277	15,277
RL	0.92	0.74	15,277	15,277	15,277	15,277	15,277
RI	0.90	0.74	0.74	15,277	15,277	15,277	15,277
RV	0.87	0.72	0.68	0.68	15,277	15,277	15,277
WR	0.81	0.77	0.77	0.64	0.90	15,277	15,277
WE	0.81	0.76	0.77	0.63	1.00	0.91	15,277
WKL	0.81	0.76	0.77	0.64	0.98	0.97	0.92

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

Table ADD.14.3 Average Intercorrelations and Reliability between Grade 11 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.89	11,086	11,086	11,086	11,086	11,086	11,086
RL	0.92	0.78	11,086	11,086	11,086	11,086	11,086
RI	0.93	0.76	0.77	11,086	11,086	11,086	11,086
RV	0.79	0.64	0.63	0.57	11,086	11,086	11,086
WR	0.77	0.73	0.75	0.52	0.88	11,086	11,086
WE	0.77	0.73	0.75	0.52	1.00	0.88	11,086
WKL	0.76	0.72	0.74	0.52	0.98	0.97	0.89

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

Table ADD.14.4 Average Intercorrelations and Reliability between Algebra I Subclaims

	Mathematics			
	MC	ASC	MR	MP
MC	0.81	16,282	16,282	16,282
ASC	0.78	0.62	16,282	16,282
MR	0.75	0.69	0.74	16,282
MP	0.78	0.70	0.72	0.76

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

Table ADD.14.5 Average Intercorrelations and Reliability between Geometry Subclaims

	Mathematics			
	MC	ASC	MR	MP
MC	0.82	5,505	5,505	5,505
ASC	0.76	0.69	5,505	5,505
MR	0.72	0.67	0.75	5,505
MP	0.75	0.68	0.78	0.71

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

Table ADD.14.6 Average Intercorrelations and Reliability between Algebra II Subclaims

	Mathematics			
	MC	ASC	MR	MP
MC	0.80	9,138	9,138	9,138
ASC	0.76	0.71	9,138	9,138
MR	0.76	0.73	0.76	9,138
MP	0.79	0.77	0.79	0.77

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

Table ADD.14.7 Average Correlations between ELA/L and Mathematics for High School

ELA/L	Mathematics		
	A1	GO	A2
9	0.72	0.77	
	(1,051)	(366)	
10	0.57	0.67	0.76
	(6,419)	(935)	(698)
11	0.65	0.41	0.55
	(208)	(1,268)	(3,588)

Note: ELA/L = English language arts/literacy, A1 = Algebra I, GO = Geometry, A2 = Algebra II. The correlations are provided with the sample sizes, below in parentheses.

Table ADD.14.8 Average Correlations between Reading and Mathematics for High School

RD	Mathematics		
	A1	GO	A2
9	0.70	0.76	
	(1,051)	(366)	
10	0.55	0.66	0.75
	(6,419)	(935)	(698)
11	0.67	0.43	0.56
	(208)	(1,268)	(3,588)

Note: RD = Reading, A1 = Algebra I, GO = Geometry, A2 = Algebra II. The correlations are provided with the sample sizes, below in parentheses.

Table ADD.14.9 Average Correlations between Writing and Mathematics for High School

WR	Mathematics		
	A1	GO	A2
9	0.62	0.67	
	(1,051)	(366)	
10	0.46	0.58	0.69
	(6,419)	(935)	(698)
11	0.55	0.29	0.43
	(208)	(1,268)	(3,588)

Note: WR = Writing, A1 = Algebra I, GO = Geometry, A2 = Algebra II. The average correlations are provided with the sample sizes, below in parentheses.