New Meridian

Technical Report 2020–2021
Alternate Blueprint

February 28, 2022

Table of Contents

List of Tables

List of Figures

Executive Summary

The purpose of this report is to describe the technical qualities of the 2020–2021 operational administration of the English language arts/literacy (ELA/L) and mathematics assessments in grades 3 through 8 and high school. Due to the outbreak of the global COVID-19 pandemic, the spring 2020 administration was suspended in March 2020 and ultimately cancelled for all participating states. At that time, testing only occurred for a small number of students in grades 3 through 8 in Illinois, although other states had planned to administer tests in grades 3 through 8 as well as high school. For spring 2021, two participating states cancelled their administration while Illinois, Department of Defense Education Activity, and Bureau of Indian Education administered the assessments. Illinois provided the option to test either in spring 2021 or in fall 2021. The forms administered were the same between spring and fall. Due to the substantial difference in the testing window, fall testers are not included in the analyses in this report, but will be reported separately in a forthcoming document. Please note that due to the ongoing effects of the COVID-19 pandemic, the results contained in this document should be interpreted with caution.

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 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.

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 Common Core State Standards 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, American Psychological Association, and National Council on Measurement in Education, 2014).

<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 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 of the contract to the Council of Chief State School Officers, and also contracted with New Meridian to manage the intellectual property and provide item development, forms construction, and governance. Starting in August 2017, New Meridian oversaw item development, data review for field test items, and test construction activities.

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, Department of Defense Education Activity, the District of Columbia Illinois, New Jersey, and New Mexico. For the academic years 2019–2020 and 2020-2021, participating states and agencies included the Bureau of Indian Education, the District of Columbia, Department of Defense Education Activity, Illinois, and New Jersey, but not all participating states administered forms. Most testing in spring 2020 was cancelled due to COVID-19, with the exception of a small number of students in Illinois who tested prior to the closure of schools. Only Bureau of Indian Education, Department of Defense Education Activity, and Illinois administered forms in spring 2021. Illinois provided the option for students to test in fall 2021 instead of spring 2021. The results for those students will be reported in a forthcoming document. Please note that due to the ongoing effects of the COVID-19 pandemic, the results contained in this document should be interpreted with caution.

The purpose of this technical report is to describe the operational administration of the summative assessments in the 2020–2021 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 Common Core State Standards 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 test(CBT) or paper-based test (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; 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 assumed the responsibility for the management of the assessments in 2017, as well as the responsibility for 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, which makes major policy and operational decisions;
- the Technical Advisory Committee, which helps ensure all assessments will provide reliable results to inform valid instructional and accountability decisions;
- the State Lead Council, which 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 analyses, item calibrations based on item response theory IRT, scaling, and development of all conversion tables.

1.6 Overview of the Technical Report

This report begins by providing explanations of the test form construction process, test administration, and scoring of the test items. Subsequent sections of the report present descriptions of student characteristics, results of classical item analyses, 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 Model and Parameters
This section presents the information related to the IRT models used and the descriptive statistics of the item parameters. Note that all tests delivered in 2021 employed a pre-equated model, in which previously estimated item parameters are used to generate scoring tables.

Section 8 – Performance Level Setting
Performance levels and policy definitions, as well as the processes followed to establish performance level thresholds, 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 scale score reliability and 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. 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.

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
CBT	computer-based test
CCSS	Common Core State Standards
CDQ	customer data quality
COVID-19	coronavirus disease 2019
CSEM	conditional standard error of measurement
DIF	differential item functioning
DPL	digital production line
DPP	digital pre-press
EcDis	economically disadvantaged
EBSS	evidence-based standard setting
ELA/L	English language arts/literacy
EL	English learners
ELN	not an English learner
ELY	English learners
EOY	end-of-year
ePEN2	Electronic Performance Evaluation Network second generation
ESEA	Elementary and Secondary Education Act
FRL	free or reduced-price lunch
FT	field test
GO	Geometry
HOSS	highest obtainable scale score
IA	item analysis
IDEA	Individuals with Disabilities Education Act
IEA	Intelligent Essay Assessor
IEP	Individualized Education Program
INF	information curve
IP	intellectual property
IRF	item response file
IRT	item response theory
ISR	individual student report
K–12	kindergarten to grade 12
LEA	local education agency

Abbreviation/Acronym	Definition
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
NCEO	National Center for Educational Outcomes
NCLB	No Child Left Behind
NCME	National Council on Measurement in Education
NoEconDis	not economically disadvantaged
n/r	not reported
NTC	nontablet condition
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
PIRLS	Progress in International Reading Literacy Study
PISA	Programme of International Student Achievement
PLD	performance level descriptor
PLS	performance level setting
pre-equating	Scoring tables are built prior to administration with existing parameters
PV	product validation
QA	quality assurance
QTI	question and test interoperability
RD	Reading (ELA/L)
RI	Reading Information (ELA/L)
RL	Reading Literature (ELA/L)
RV	Reading Vocabulary (ELA/L)
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
SMD	standardized mean difference
SSMC	single select multiple choice

Abbreviation/Acronym	Definition
SWD	students with disabilities
SWDN	not student with disability
SWDY	students with disabilities
TCC	test characteristic curve
TIMSS	Trends in International Mathematics and Science Study
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)

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. TEIs 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 vehicle through 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 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 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 and paper-based tests 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 English Language Arts/Literacy

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 stimuli. Authentic texts are grade-appropriate texts that are not developed for the purposes of the assessment or to achieve a particular readability metric, but reflect the original language of the authors. 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 were 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 involves 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 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 the 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 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% 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, and 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 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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, LP, human reader/human signers, assistive technology, TTS, 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.

- constructed the online forms to match the blueprint and test construction specifications
- constructed the paper forms to match the blueprint and test construction specifications; and
- 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, and
- recommended accommodated forms

Test Form Verification Meeting to Review Test Construction Inputs

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

Accommodated Form Review Process

In addition to participating in many of the development activities including the text review and the bias and sensitivity review meetings, the AAF OWG reviewed the proposed accommodated forms at the test construction committee meeting for accessibility to make sure that the content can be accommodated for students with disabilities and English learners without changing the underlying measured construct.

Forms were identified to support the following accommodations:

Accommodated Base 1

- 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

Accommodated Base 2

- closed captioning
- TTS first form
- Spanish online
- Spanish TTS

Accommodated Base 3 (mathematics only)

- TTS 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 to the maximum extent possible, 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 LP and TTS versions. Once the operational form was approved, the form was sent to Pearson's subcontractor, Teneo, for transadaptation of the items. Transadaptation differs from translation in that it takes into consideration the grade-level appropriateness of the words, as well as the linguistic and cultural differences that exist between speakers of two different languages. Accounting for these differences allows the item to measure the achievement of Spanish-language speakers in the same way that the original version of the item does for native speakers of English. The Spanish Glossary provided guidance to the translator conducting the transadaptation in grade-level and culturally appropriate ways of transadapting the items. For the Spanish-language TTS form, the alternate text (used for description and/or text in art and graphics) was transadapted from the alternate text for the English language version of the TTS form. Phonetic markup, 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: 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 including horizontal linking and across-administration linking. Horizontal linking consisted of linking items, or common items, included in both forms in a single administration, which was the case for mathematics forms and some ELA/L forms. Across-administration linking, or year-to-year linking, consisted of common items included in two different administrations, which was used for all forms due to the pre-equated model. The placement of linking items across forms or administrations supports the development of comparable scores.

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

2.2.5 Field-Test Data Collection Overview

Field-test items were embedded in the spring operational mathematics forms, but the data were not analyzed. Field-test items for ELA/L operational forms were not administered.

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 tests [PBTs]) or online (i.e., computer-based tests [CBTs]). For the paper-based administration, students used paper-based answer documents (except in grade 3 where students responded directly into test booklets). Nearly all of the summative assessments administered during the 2020–2021 administration were online assessments (see Tables 11.1 through 11.3).

3.1.1 Secure Versus 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, and so on. For PBTs, secure materials include both used and unused test booklets and used scratch paper, while for CBTs, 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, and so on.

3.1.2 Scorable Versus Nonscorable Materials

PBTs have both scorable and nonscorable materials while CBTs 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 PBTs, such as blank (i.e., unused) test booklets, test administration manuals, scratch paper, mathematics reference sheets, and so on, are deemed nonscorable. For CBTs, 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 CBT 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 PBT 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 students taking the PBT 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 PBTs are included in each local education agency (LEA) or school's test materials shipment.

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

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

3.2 Accessibility Features and Accommodations

3.2.1 Participation Guidelines for Assessments

All students, including students with disabilities and English learners, are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for English learners in their first year in a U.S. school, and certain students with disabilities who have been identified by the Individualized Education Program (IEP) team to take their state's alternate assessment. Federal laws governing student participation in statewide assessments include the No Child Left Behind Act of 2001, the Individuals with Disabilities Education Act of 2004, Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008), and the Elementary and Secondary Education Act of 1965, as amended. All students can receive accessibility features on the summative assessments.

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

1. students with disabilities who have an IEP;
2. students with a Section 504 plan who have a physical or mental 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 CBTs, 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 <u>not</u> provided a test accommodation that <u>was</u> 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 is 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.

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 tasks in the English language arts/literacy content area. This analysis was conducted for prose constructed-response tasks administered online using some of the same artificial intelligence techniques

[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.

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 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, and so on. 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.
- The project manager documented the procedures, identified risks, and managed day-to-day administrative matters.
- A portfolio manager provided oversight for the entire scoring process.

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

4.2.1 Scorer Training

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

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

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

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

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

Qualification sets were used to confirm that scorers understood how to score student responses accurately. 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 four prototype sets per grade level or course.

The prototype training approach promoted consistency in scoring, as each subsequent abbreviated training set for the ELA/L task type or mathematics item grouping was based on the prototype. Once a prototype was chosen, full training materials were developed for that item, and at each grade level, scorers were trained to score a particular item type using the prototype training materials for that type.

Abbreviated training sets were prepared for all items not selected for prototype training sets. The abbreviated training sets included an anchor set and two practice sets so scorers could internalize the scoring standards for these new items, which were similar to prototype items they had previously scored.

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

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

Table 4.1 Training Materials Used During Scoring

Training Set Development	
Description	**Specification**
Anchor Set	
The anchor set is the primary reference for scorers as they internalize the rubric during training. All scorers have access to the anchor set when they are training and scoring, and are directed to refer to it regularly. 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 for Research Simulation and Literary Analysis Tasks, Written Expression for Narrative Writing Tasks, and Knowledge of Language and Conventions for all task types).
Practice Sets	
Practice sets are used to help 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 in 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 defined and documented qualifying standards in conjunction with participating states and agencies prior to scoring.	The qualification sets for mathematics prototype items include three sets of ten responses each (not annotated). 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.

Note. ELA/L = English language arts/literacy

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

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 as set forth in Table 4.2 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% 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 (IEA). 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% 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, that is, 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%

Note. ELA/L = English language arts/literacy

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

Calibration Sets

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

Inter-rater Agreement

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

Table 4.5 Inter-rater Agreement Expectations and Results

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

Note. ELA/L = English language arts/literacy

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

Pearson's ePEN2 scoring system included comprehensive inter-rater agreement reports that allowed supervisory personnel to monitor both individual and group performance. Based on reviews of these reports, scoring experts targeted individuals for increased backreading and feedback, and if necessary, retraining.

The perfect agreement rate for mathematics responses scored by two scorers ranged from 73% to 100% and the within one point rate ranged from 98% to 100%. For all ELA/L responses scored by two scorers, the perfect

agreement rate ranged from 73 percent to 100 percent and the within one point rate ranged from 98 percent to 100 percent.

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

4.3 Automated Scoring for Prose Constructed-Responses

Automated scoring performed by Pearson's 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% of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. Human scoring was applied to responses that were scored while IEA was being trained as well as to additional responses routed to human scoring when there was uncertainty about the automated scores.

For 10% of responses, a second "reliability" score was assigned. 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 following discussion 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 Intelligent Essay Assessor 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% reliability sample and other responses that may be routed accordingly).

Smart Routing

Smart routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores. Smart routing can be applied prompt by prompt to the extent needed to meet scoring quality criteria for automated scoring.

Quality Criteria for Evaluating Automated Scoring

The state leads approved specific quality criteria for evaluating automated scoring. The primary evaluation criteria for IEA were 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 and human should be within 0.1 of human-human.
- Kappa between IEA and human should be within 0.1 of human-human.
- Quadratic-weighted kappa between IEA and human should be within 0.1 of human-human.
- Exact agreement between IEA and human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA and human should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria, and are as follows:

- **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 Williamson et al.'s (2012) tolerances for subgroups with at least 50 responses.

Hierarchy of Assigned Scores for Reporting

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

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

4.3.2 Sampling Responses Used for Training IEA

For prompts trained using 2021 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% for each trait),

- exact agreement between human scores conditioned on score point (the goal was for this to be at least 50% for each trait),
- the number of responses at each score point (the goal was to have at least 40 responses at the highest score points in the training samples used by IEA), 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 was 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 Intelligent Essay Assessor Performance

The primary criteria for evaluating IEA performance are 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% 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 Intelligent Essay Assessor Performance

For many of the prompts trained in 2021, 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% of the human-human agreement, IEA can be deployed operationally.

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

conditioned on the human score). 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, for example, 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, for example, 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 for which it makes sense to be 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 Intelligent Essay Assessor 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, 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 ten comparison groups, as set forth in Table 4.6.

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
	Students with disabilities

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% 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 and human should be 0.40 or above.
- Quadratic-weighted kappa between IEA and human should be 0.70 or above.
- Exact agreement between IEA and human should be 65% or above.

These targets were not intended to be directly applied in decisions about whether to deploy IEA operationally. 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 PCR tasks by trait and grade level. PCR items are scored on two traits: (1) Reading Comprehension and Written Expression and (2) Knowledge of Language and Conventions for Research Simulation for Literary Analysis tasks and (1) Written Expression and (2) Knowledge of Language and Conventions for the Narrative task.

For 10% 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% or higher between two scorers. When IEA provided the first score of record, the second reliability score was a human score. For a subset of responses, the first and second score were both human scores.

Table 4.7 presents the average agreement across the PCRs for each grade level by trait. The number of prompts included in the analyses is listed for each grade level. The agreement indices (exact agreement, kappa, quadratic-weighted kappa, and Pearson correlation) were calculated separately by PCR for each trait (Reading Comprehension and Written Expression or Written Expression and Conventions). For each grade level, the agreement indices were averaged across the PCRs. Table 4.7 presents the average count and the average for the agreement indices.

The exact agreement for the PCR traits is above the criterion of a 65% agreement rate for all PCRs. The strength of agreement between raters is moderate to substantial agreement as defined by Landis and Koch (1977) for all PCRs. The quadratic-weighted kappa (QW kappa) distinguishes between differences in ratings that are close to each other versus larger differences. The weighted kappa is substantial to almost perfect agreement for all grades. The Pearson correlations (r) ranged from 0.71 to 0.96.

Table 4.7 PCR Average Agreement Indices by Test

			Written Expression				Conventions			
Test	**# of PCRs**	**Count**	**Exact**	**Kappa**	**QW Kappa**	***r***	**Exact**	**Kappa**	**QW Kappa**	***r***
ELA03	4	12343	75.58	0.56	0.71	0.71	75.28	0.58	0.76	0.76
ELA04	3	22761	77.67	0.62	0.78	0.79	77.13	0.63	0.80	0.80
ELA05	4	12618	75.48	0.60	0.79	0.80	76.10	0.62	0.81	0.81
ELA06	4	12624	80.13	0.67	0.84	0.84	80.30	0.68	0.84	0.85
ELA07	4	12246	75.15	0.63	0.85	0.85	75.60	0.64	0.84	0.85
ELA08	4	12334	79.15	0.69	0.88	0.88	79.68	0.70	0.87	0.87
ELA10	4	181	91.48	0.88	0.96	0.96	90.20	0.86	0.94	0.94
ELA11	4	20	91.15	0.83	0.88	0.89	84.93	0.71	0.82	0.83

Note. ELA03 – ELA11 = English language arts/literacy tests in grades 3 through 11.

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 English language arts/literacy (ELA/L) and mathematics assessments were pre-equated. The item statistics provided in this section were from prior operational administrations and reflect the statistics that were used in test construction and for score reporting for some states and agencies. Item analysis serves two purposes: to inform item exclusion decisions for item response theory (IRT) analysis and to provide item statistics for the item bank.

Item analysis included data from the following types of items: key-based selected-response items, rule-based machine-scored items, and hand-scored constructed-response items. For each item, the analysis produced item difficulty, item discrimination, and item response frequencies.

5.2 Data Screening Criteria

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

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% 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.

Distractor-Total Correlation

For selected-response items, this estimate describes the relationship between selecting an incorrect response (i.e., a distractor) for a specific item and performance on the total test. The item-total correlation is calculated for the distractors. Items with distractor-total correlations above .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% 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% for selected-response items and 15% 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 three-category item), this may indicate that there are problems with the item or the rubric so that students can receive either full credit or no credit at all, but not partial credit.

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

5.4 Summary of Classical Item Analysis Flagging Criteria

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

1. p-value above .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%) choosing a distractor rather than the keyed response,
6. high percentage of omits: above 5% for selected-response items and above 15% for constructed-response items,
7. high percentage that did not reach the item: above 5% for selected-response items and above 15% for constructed-response items, or
8. constructed-response items with a score value obtained by less than 3% of responses.

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

5.5 Classical Item Analysis Results

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

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

Some forms in the spring 2021 administration were based on previous administrations, with many of them being reused from the spring 2020 administration; therefore, the item analyses for these forms were reported in the associated 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	60	0.44	0.18	0.13	0.78	0.4
4	76	0.45	0.15	0.16	0.72	0.47
5	74	0.46	0.16	0.17	0.84	0.44
6	78	0.48	0.14	0.18	0.76	0.46
7	75	0.47	0.14	0.23	0.8	0.46
8	76	0.48	0.15	0.19	0.84	0.46
9	55	0.45	0.13	0.23	0.75	0.44
10	78	0.42	0.12	0.2	0.71	0.41
11	80	0.35	0.11	0.13	0.73	0.34

Note. ELA/L = English language arts/literacy; SD = standard deviation.

Table 5.2 Summary of Pre-Administration 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	91	0.58	0.2	0.18	0.93	0.56
4	87	0.52	0.19	0.19	0.95	0.52
5	85	0.48	0.17	0.13	0.83	0.5
6	74	0.42	0.21	0.08	0.94	0.4
7	86	0.42	0.19	0.07	0.85	0.35
8	76	0.33	0.22	0.05	0.82	0.27
A1	118	0.3	0.18	0.05	0.71	0.27
GO	129	0.33	0.22	0.05	0.91	0.29
A2	128	0.31	0.17	0.05	0.82	0.29
M1	34	0.35	0.17	0.02	0.7	0.4
M2	30	0.33	0.19	0.01	0.72	0.4

Note. SD = standard deviation; A1 = Algebra I; GO = Geometry; A2 = Algebra II; M1 = Integrated Mathematics I; M2 = Integrated Mathematics II.

Table 5.3 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	60	0.54	0.13	0.23	0.79	0.56
4	76	0.5	0.14	0.22	0.81	0.47
5	74	0.49	0.16	0.2	0.86	0.46
6	78	0.53	0.15	0.28	0.87	0.51
7	75	0.52	0.17	0.23	0.86	0.46
8	76	0.52	0.18	0.22	0.86	0.48
9	55	0.49	0.17	0.2	0.86	0.48
10	78	0.49	0.18	0.18	0.86	0.46
11	80	0.48	0.19	0.17	0.86	0.43

Note. ELA/L = English language arts/literacy; SD = standard deviation.

Table 5.4 Summary of Pre-Administration 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	91	0.5	0.13	0.18	0.82	0.51
4	87	0.51	0.13	0.28	0.78	0.53
5	85	0.51	0.13	0.21	0.76	0.52
6	74	0.54	0.14	0.24	0.78	0.55
7	86	0.5	0.16	0.18	0.81	0.47
8	76	0.46	0.15	0.19	0.8	0.46
A1	118	0.48	0.15	0.15	0.89	0.47
GO	129	0.51	0.15	0.19	0.83	0.49
A2	128	0.46	0.14	0.16	0.76	0.46
M1	34	0.5	0.18	0.15	0.8	0.54
M2	30	0.46	0.19	0.07	0.95	0.43

Note. SD = standard deviation; A1 = Algebra I; GO = Geometry; A2 = Algebra II; M1 = Integrated Mathematics I, M2 = Integrated Mathematics II.

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 English language arts/literacy (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.

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

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, 2013; Dorans & Schmitt, 1991; Zwick et al., 1997), 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-".

Note. DIF = differential item functioning.

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. DIF = differential item functioning; *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 (SWDs) or English learners (ELs). 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 = differential item functioning.

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 prose constructed-responses 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 through 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

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 versus Female	51					51	100				
White versus Black	51			1	2	50	98				
White versus Hispanic	51					51	100				
White versus Asian	51					51	100				
White versus AmerIndian	51					51	100				
White versus Pacific Islander	51			2	4	49	96				
White versus Multiracial	51					50	98	1	2		
NoEcnDis versus EcnDis	51					51	100				
ELN versus ELY	51			4	8	47	92				
SWDN versus SWDY	51			1	2	50	98				

Note. DIF = differential item functioning; ELA/L = English language arts/literacy; 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 Pre-Administration Differential Item Functioning for Mathematics Grade 3

DIF Comparison	Total N of Unique Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	% of Total	N	% of Total	N	% of Total	N	% of Total	N	% of Total
Male versus Female	89			1	1	87	98	1	1		
White versus Black	89	1	1	6	7	79	89	3	3		
White versus Hispanic	89			1	1	88	99				
White versus Asian	89					81	91	7	8	1	1
White versus AmerIndian	89			1	1	88	99				
White versus Pacific Islander	89			2	2	86	97	1	1		
White versus Multiracial	89					88	99	1	1		
NoEcnDis versus EcnDis	89					89	100				
ELN versus ELY	89					89	100				
SWDN versus SWDY	89			2	2	87	98				

Note. DIF = differential item functioning; 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: Item Response Theory Model and Parameters

7.1 Overview

Multiple operational core forms were administered for each grade in English language arts/literacy (ELA/L) and mathematics assessments. All tests in spring 2021 were pre-equated, meaning that scoring tables were constructed prior to the administration with existing parameters, whose values were estimated in in 2019 or earlier. This section describes the item response theory (IRT) model used in this assessment program and provides descriptive statistics of the item parameters.

7.2 Two-Parameter Logistic/Generalized Partial Credit Model

The operational items used pre-equated parameters in the context of the two-parameter logistic/generalized partial credit model, which is denoted as

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $p_{im}(\theta_j)$ is the probability of a student with 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 $M_i - 1$; and v indexes the response categories and is iterated from 0 to $M_i - 1$.

7.3 Summary Statistics and Distributions From IRT Analyses

Tables 7.1 through 7.4 present summary statistics for the IRT (*b*- and *a*-) parameter estimates, the standard errors of the parameter estimates, and the IRT model fit values (chi-square and adjusted fit) for ELA/L and mathematics assessments. The summary statistics for IRT parameter estimates include all the items administered in the spring administration except the items on the reused forms, if applicable, for which the summary results were reported in the technical reports of the source administrations.

The information is provided by content area (ELA/L and mathematics) for all items at each grade level or course. The summary statistics shown include the total number of items and score points, along with the mean, standard deviation (SD), minimum, and maximum.

7.3.1 IRT Summary Statistics for English Language Arts/Literacy

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

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

			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	136	60	0.54	1.05	-1.64	3.35	0.57	0.23	0.12	1.04
4	169	76	0.45	0.95	-1.41	2.95	0.46	0.21	0.13	0.99
5	168	74	0.52	1.00	-1.70	3.59	0.48	0.25	0.10	0.99
6	177	78	0.34	0.74	-1.09	1.89	0.51	0.23	0.18	1.16
7	171	75	0.28	0.79	-1.70	1.60	0.50	0.28	0.13	1.23
8	176	76	0.24	0.85	-1.39	2.83	0.52	0.29	0.18	1.24
10	177	78	0.69	0.84	-0.77	4.03	0.49	0.28	0.13	1.19
11	181	80	1.04	0.87	-1.09	4.21	0.45	0.25	0.10	1.10

Note. IRT = item response theory; ELA/L = English language arts/literacy; SD = standard deviation.

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

Grade	ALL	2014	2015	2016	2017	2018	2019
3	60	0	7	11	5	11	26
4	76	1	18	8	6	12	31
5	74	0	0	8	4	30	32
6	78	1	16	0	9	27	25
7	75	0	12	10	3	22	28
8	76	0	1	19	6	14	36
10	78	0	1	10	31	33	3
11	80	0	20	10	6	16	28

Note. IRT = item response theory; ELA/L = English language arts/literacy.

7.3.2 IRT Summary Statistics for Mathematics

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

Table 7.3 Pre-Equated IRT Parameter Estimates Summary for All Items for Mathematics by Grade/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	133	91	-0.36	0.99	-2.52	1.62	0.75	0.23	0.22	1.31
4	136	87	-0.08	0.99	-2.69	1.86	0.72	0.21	0.32	1.38
5	144	85	0.08	0.91	-2.06	2.45	0.67	0.23	0.17	1.50
6	138	74	0.31	1.00	-2.38	2.06	0.75	0.25	0.33	1.44
7	142	86	0.55	1.05	-1.78	2.78	0.65	0.26	0.19	1.22
8	135	76	0.97	1.22	-1.52	3.18	0.63	0.24	0.21	1.34
A1	228	118	1.32	1.10	-0.99	3.62	0.62	0.26	0.16	1.34
G0	236	129	1.01	1.17	-1.60	3.83	0.78	0.33	0.18	1.78
A2	229	128	1.28	1.03	-1.53	3.67	0.63	0.28	0.16	1.28
M1	62	34	1.10	1.09	-0.95	4.02	0.61	0.31	0.11	1.61
M2	55	30	1.46	1.23	-0.97	3.96	0.56	0.27	0.06	1.41
M3	55	29	1.52	1.44	-1.02	4.30	0.52	0.26	0.17	1.24

Note. SD = standard deviation; A1 = Algebra I; GO = Geometry; A2 = Algebra II; M1 = Integrated Mathematics I; M2 = Integrated Mathematics II; M3 = Integrated Mathematics III.

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

Grade	ALL	2014	2015	2016	2017	2018	2019
3	91	0	13	8	13	8	49
4	87	1	9	12	16	9	40
5	85	0	8	5	9	9	54
6	74	0	8	3	10	8	45
7	86	1	13	12	13	8	39
8	76	0	9	7	6	5	49
A1	118	1	19	31	23	16	28
G0	129	1	27	38	28	12	23
A2	128	0	16	24	22	25	41
M1	34	1	23	10	0	0	0
M2	30	0	24	6	0	0	0
M3	29	0	22	7	0	0	0

Note: SD = standard deviation; 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 as follows:

- 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

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

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 would 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 component and then the end-of-year 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 that order. The panelists entered their item judgments for each round by completing an online item judgment survey. Educator judgments were summed across items to create an estimated total score on the reference form for each performance level threshold. Feedback data relative to panelist agreement, student performance on the items, and student performance on the test as a whole were provided in between each of the three rounds of judgment. Panelists were shown the

pre-policy reasonable ranges prior to making their round 1 judgments and again as feedback data following each round of judgment.

A dry run of the PLS meeting process was held for grade 11 ELA/L and Algebra II in order to evaluate the implementation of the PLS method with the innovative characteristics of the summative assessments. These content areas were selected because they combined all the various aspects of the assessments, including the various types of items, scoring rules, and performance level decisions. The dry-run PLS meetings provided the opportunity to implement and evaluate multiple aspects of the operational plan for the actual PLS meeting, including pre-work, meeting materials, data analysis and feedback, and staff and panelist functions. The results of the dry-run PLS meeting were used to implement improvements in the process for the operational PLS meetings. Additional information about the methods and results of the dry-run PLS meeting is available in the 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 12 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 two 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 Pearson Assessment Banking for Building and Interoperability (ABBI) system 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 now being created within the ABBI system, which employs templates to control the consistency of the underlying scoring logic and question and test interoperability 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 provides 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 on forms evaluating several item characteristics (e.g., content accuracy, completeness, style guide conformity, tools function). Revisions were incorporated into the forms before final review and approval. Section 2.2 provides more details on the form development process.

The forms quality assurance was performed by Pearson's Assessment and Information Quality (AIQ) organization. AIQ completed a comprehensive review of all *online* forms for the administration cycle. This

group is part of Pearson's larger Organizational Quality Group and operates exclusively to validate form operability. The group validates that the functionality of every online form is working to specifications. The overall functionality and maneuverability of each form is checked, and the behavior of each item within the form is verified. (Quality processes for paper forms are described in Section 9.3.)

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

Pearson also performed quality control tests to verify that a standard set of responses was outputted to the XML as expected after the final version of the form was approved. These responses were based on the keys provided in the test map or a standard open-ended 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 for digital print methodology, or plateroom for offset print methodology. Both the digital pre-pres and plateroom functions

verify content, file naming, imposition, pagination, numbering stream, registration of technical components, color mapping, workflow, and file integrity. Records of these checks are created and saved.

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

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

Pearson's Quality Assurance (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 QA 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 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 include 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 the following:

- incorrect or double gridding,
- incorrect dates (including birth year),
- mismatches between pre-ID label and gridded information, and
- 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); and
- 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; and
- 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 was responsible for the psychometric analyses of the operational administration and implemented measures to ensure the quality of work. The psychometric analyses were all conducted according to well-defined specifications. Data cleaning rules were clearly articulated and applied consistently throughout the process. Results from all analyses underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were used by members of the team for each statistical procedure.

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

Data Screening

Data screening is an important first step to ensure quality data input for meaningful analysis. The Pearson Customer Data Quality team validated all student data files used in the operational psychometric analyses. The data validation for the student data files and item response files included the following steps:

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

Classical Item Analysis

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

Conversion Tables

Conversion tables must be accurate because they are used to generate reported scores for students. Comprehensive records were meticulously maintained on item-level decisions, and thorough checks were made to ensure that the correct items were included in the final score. Pre-equated conversion tables were developed independently by two psychometricians and completely matched. A reasonableness check was also conducted by psychometricians for each content and grade level to make sure the results were in alignment with observations during the analyses prior to conversion table creation. Refer to Section 12.3 for the procedure to create conversion tables.

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 Common Core State Standards that are assessed in more than one evidence statement when selecting the items for the spring 2021 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 Council of Chief State School Officers 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 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

Note. ELA/L = English language arts/literacy; 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 3–8 and 10-11 English language arts/literacy (ELA/L) and Integrated Mathematics I and II, and 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 or a paper-based test 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

Over a million forms were administered in the Bureau of Indian Education, the Department of Defense Education Activity, and Illinois during the 2020–2021 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 8 and grades 10 and 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 and II). The student counts in Integrated Mathematics were not large enough for most analyses in this report. A small subset of students tested in ELA/L grades 3 through 8, and mathematics grades 3 through 8 during the fall 2021. Student characteristics for this group will be presented in a forthcoming addendum. The majority of students tested during the spring administration when all grades and content areas were administered mostly online with small numbers of paper testers.

11.2 Rules for Inclusion of Students in Analyses

Criteria for inclusion of students were implemented prior to all operational analyses. These rules were established by Pearson psychometricians in consultation with 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:

- valid form numbers were observed for each unit for online assessments or for the full form for paper assessments,
- student records were not flagged as “void” (i.e., do not score), and
- the student attempted at least 25% 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, computer-based test (CBT) or paper-based test (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 were greater than 99%, except for grade 11, which had a low overall count. For all mathematics students, the percentages of students testing online was greater than 98%. The percentages of students taking Spanish-language mathematics online forms was greater than or equal to 99%. 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

Grade	No. of Valid Cases	CBT		PBT	
		N	%	N	%
3	96,928	96,041	99.1	887	0.9
4	99,006	98,343	99.3	663	0.7
5	99,632	99,004	99.4	628	0.6
6	98,590	98,352	99.8	238	0.2
7	96,950	96,723	99.8	227	0.2
8	96,028	95,786	99.7	242	0.3
10	2,767	2,765	99.9	2	0.1
11	413	371	89.8	42	10.2
Grand Total	590,314	587,385	99.5	2,929	0.5

Note: Includes students taking accommodated forms of ELA/L. ELA/L = English language arts/literacy; 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	96,011	95,149	99.1	862	0.9
4	97,740	97,088	99.3	652	0.7
5	98,306	97,690	99.4	616	0.6
6	96,924	96,684	99.8	240	0.2
7	91,315	91,100	99.8	215	0.2
8	92,946	92,711	99.7	235	0.3
A1	3,424	3,380	98.7	44	1.3
GO	2,922	2,920	99.9	2	0.1
A2	2,726	2,725	100	1	0
M1	17	17	100	n/a	n/a
M2	1	1	100	n/a	n/a
Grand total	575,018	572,151	99.5	2,867	0.5

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; n/a = not applicable.

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	1,729	1,726	99.8	3	0.2
4	1,504	1,496	99.5	8	0.5
5	1,305	1,303	99.8	2	0.2
6	961	957	99.6	4	0.4
7	367	364	99.2	3	0.8
8	290	287	99	3	1
Grand total	6,156	6,133	99.6	23	0.4

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

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, 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} or PearsonAccess 5, depending on which platform was used by the respective state, by means of a student data upload. The demographic data was verified by the states and districts prior to score reporting. Not all demographics were provided for all students. Students missing information on one or more demographic variables were omitted from the corresponding subgroup analyses.

Tables A.11.7 through A.11.14 provide demographic information for students with valid ELA/L scores, and Tables A.11.15 through A.11.25 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. A summative scale score is reported for each mathematics assessment. A summative scale score and separate claim scores for Reading and Writing are reported for each English language arts/literacy (ELA/L) assessment.

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

12.1 Operational Test Content (Claims and Subclaims)

A claim is a statement about student performance based on how students respond to test questions. The tests are designed to elicit evidence from students that supports valid and reliable claims about the extent to which they are college- and career-ready or on track toward that goal and are making expected academic gains based on the Common Core State Standards (CCSS).

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

12.1.1 English Language Arts/Literacy

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

[3] 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 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. ELA/L = English language arts/literacy.

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 PCR 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

Note. * ELA/L assessments consist of the Research Simulation Task and either the Literary Analysis Task or the Narrative Writing Task. ELA/L = English language arts/literacy.

12.1.2 Mathematics

Table 12.3 [4] includes the numbers of items and points associated with subclaim scores for mathematics grade 3, as an example of the composition of the mathematics tests.

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 [5] with Level 1 indicating the lowest level of performance and Level 5 indicating the highest level of performance. Threshold or cut scores associated with performance levels were initially expressed as raw scores on the performance level setting (PLS) forms approved by the Governing Board. A scale score task force was assembled, which made recommendations about how threshold levels would be represented on the reporting scale.

[4] Table A.12.9 in Appendix 12.1 is identical to Table 12.3.
[5] Section 8 provides an overview of the performance level setting process, and detailed information can be found in the Performance Level Setting Technical Report.

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 cuts. 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 PLS 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 item response theory (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 ($ScaleScore_{2015}$):

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

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

Note. 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:

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

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

where $a_i\left(\theta_j-b_i+d_{i0}\right)\equiv 0$; $s_i\left(\theta_j\right)$ is the expected item score for item i on theta, θ_j; $p_{im}\left(\theta_j\right)$ 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:

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

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:

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

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

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

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:

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

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

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

Appendix 12.3 contains TCC, CSEM, and INF curves for all ELA/L grades and all mathematics grades/courses. The curves are based on IRT parameters from a prior operational or field-test administration.

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

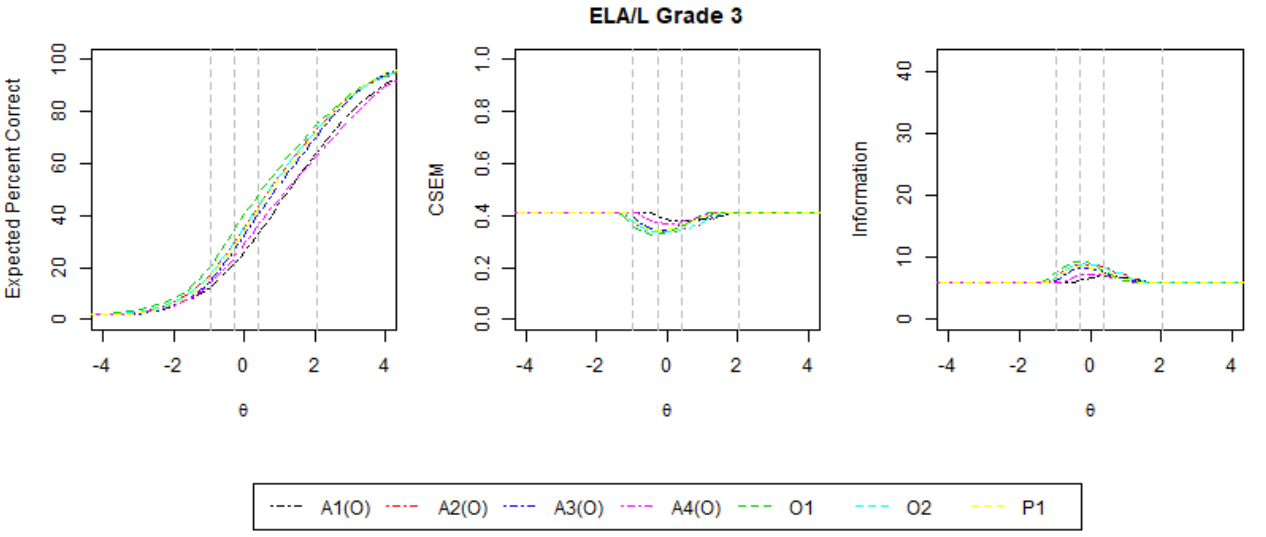

Figure 12.1 Test Characteristic Curves, Conditional Standard Error of Measurement Curves, and Information Curves for ELA/L Grade 3

12.4 Score Distributions

12.4.1 Score Distributions for English Language Arts/Literacy

Figures 12.2 through 12.4 graphically represent the distributions of scale scores for grades 3 through 8 and 10 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) or slightly below, except for grade 11, which was centered closer to 700.

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, except for grade 11, which was skewed right.

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 10 in Writing across all ELA/L grades. Students with zero raw score points on the written portion of the assessment are automatically assigned the LOSS value of a scale. Writing items are embedded exclusively in PCR tasks, which tended to be difficult. The Written Expression trait also tended to be the most difficult of the PCR traits.

Across the ELA/L grades, there are relatively few students between 11 and about 20, depending on the grade. [7] As noted in Section 12.2.2, the scale score task force selected 10 as the LOSS. This value was selected to be consistent with the Reading LOSS and reduce truncation at the lower ends of the scale. However, the scale is defined by the theta values associated with the Level 2 and Level 4 performance levels. All other scale score values are identified through a theta-to-scale score linear transformation applying the scaling constants (Table 12.4). For Writing, the lowest theta estimate associated with raw scores ranging from one to two are linearly transformed to scale score values generally between 15 and 20, meaning that there may be multiple scale scores between 11 and 20 that are not assigned to a raw score. In contrast, the Reading lowest theta estimates associated with raw scores ranging from one to two are linearly transformed to scale score values closer to the LOSS. The gap in the proportion of students at the scale scores between the LOSS value of 10 and the scale score values around 17 to 19 is an artifact of scale score task force selecting the LOSS value of 10.

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

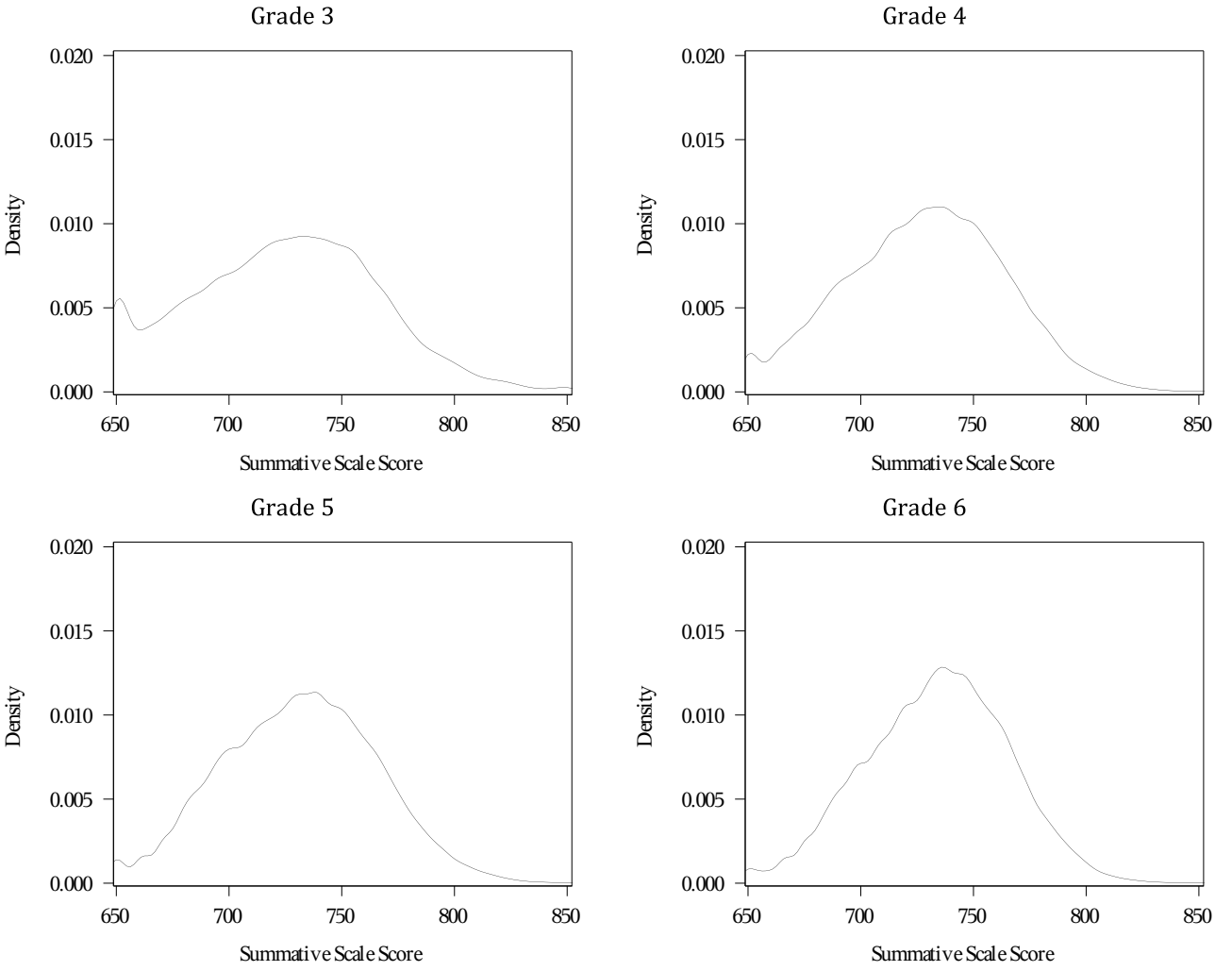

Figure 12.2 Distributions of ELA/L Scale Scores: Grades 3-8, and 10-11

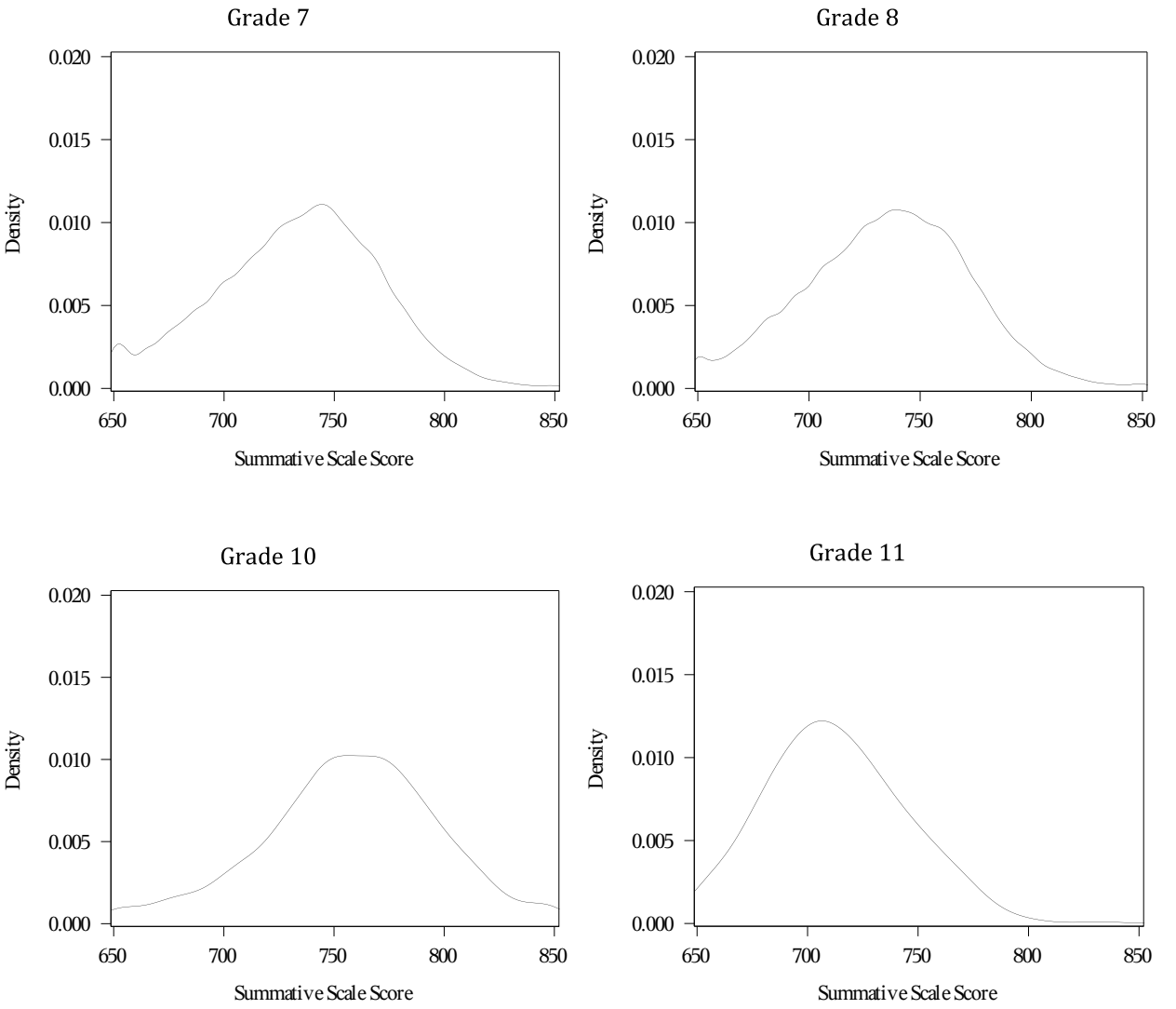

Figure 12.2 (continued) Distributions of ELA/L Scale Scores: Grades 3-8, and 10-11

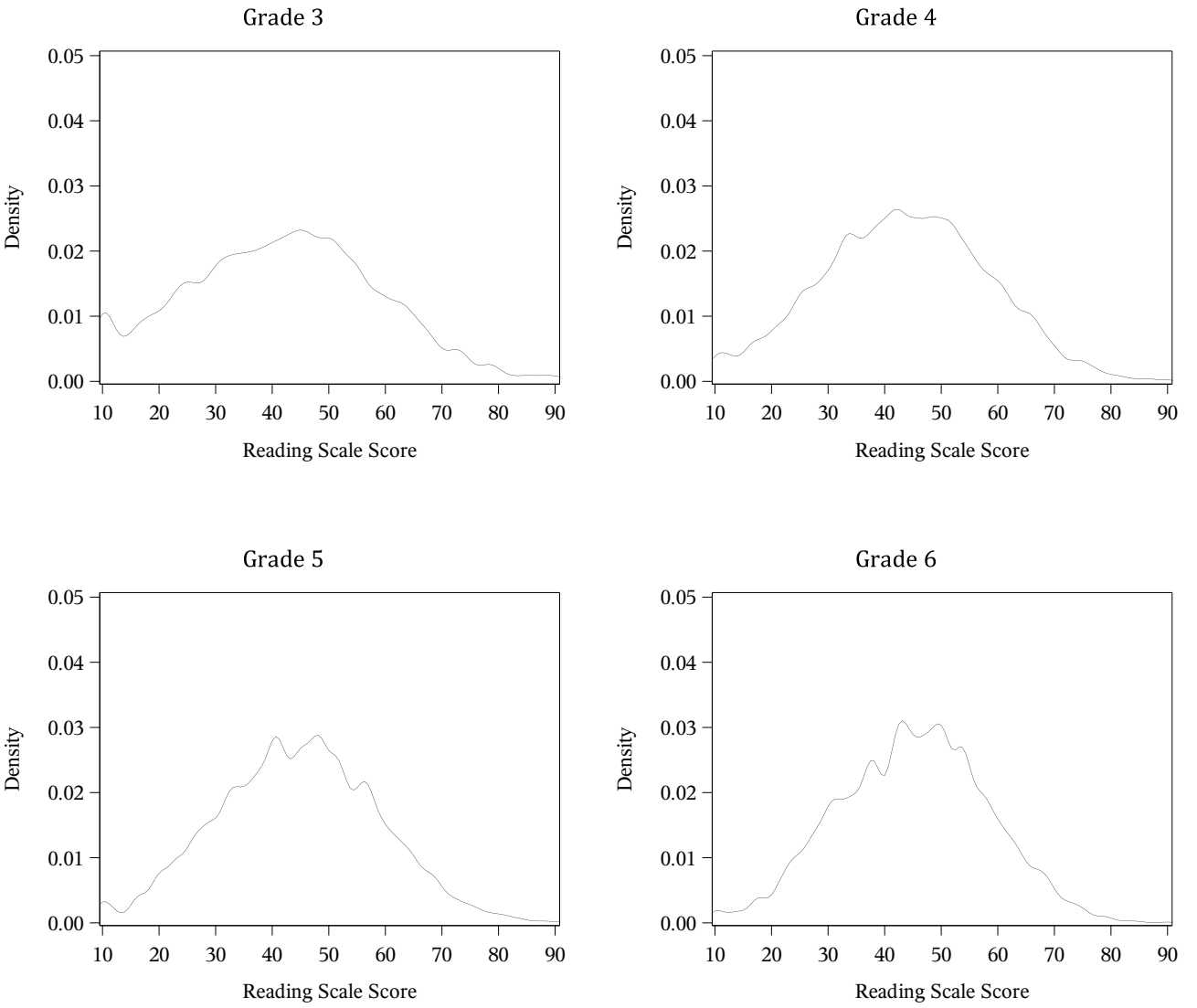

Figure 12.3 Distributions of Reading Scale Scores: Grades 3-8, and 10-11

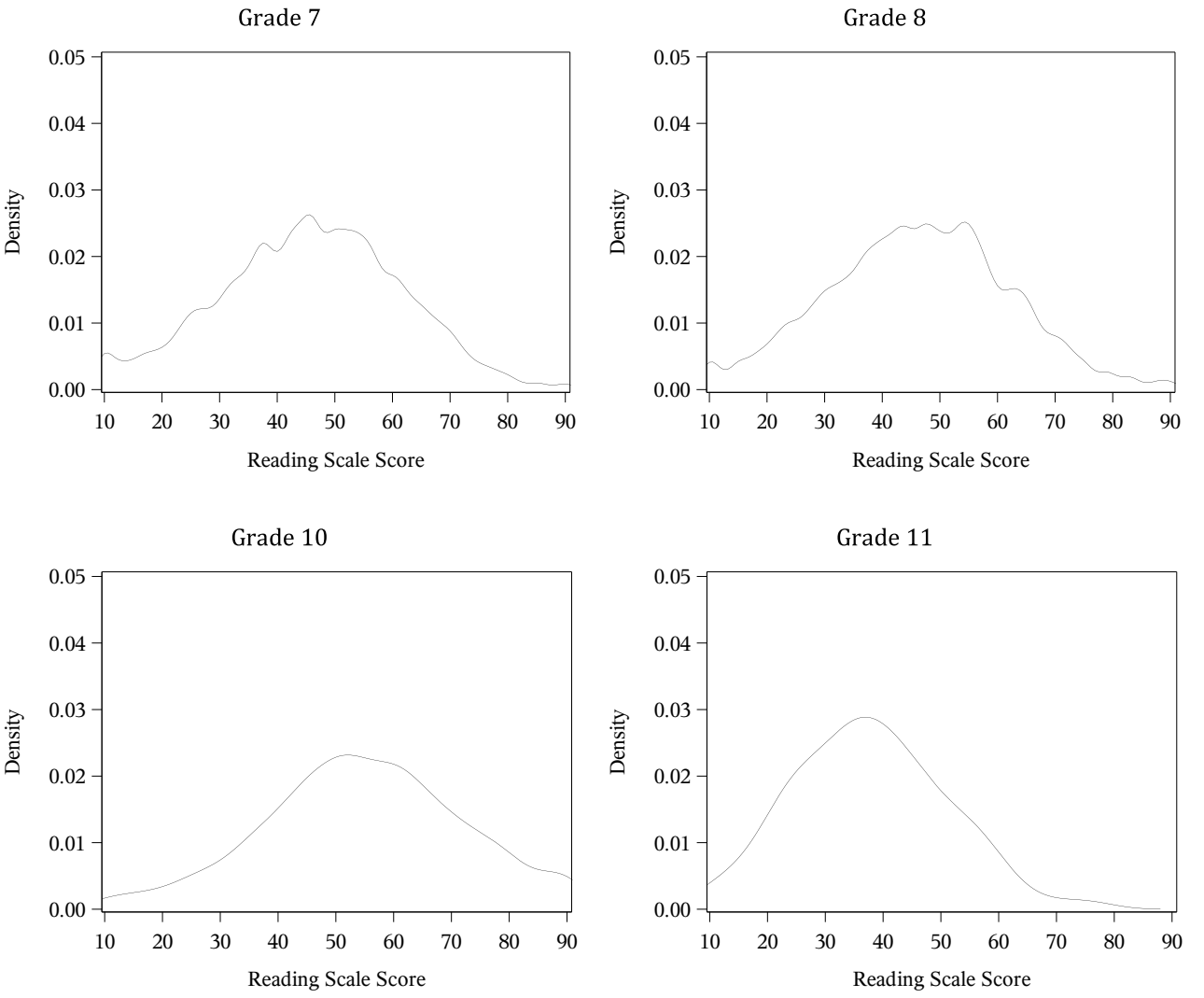

Figure 12.3 (continued) Distributions of Reading Scale Scores: Grades 3-8, and 10-11

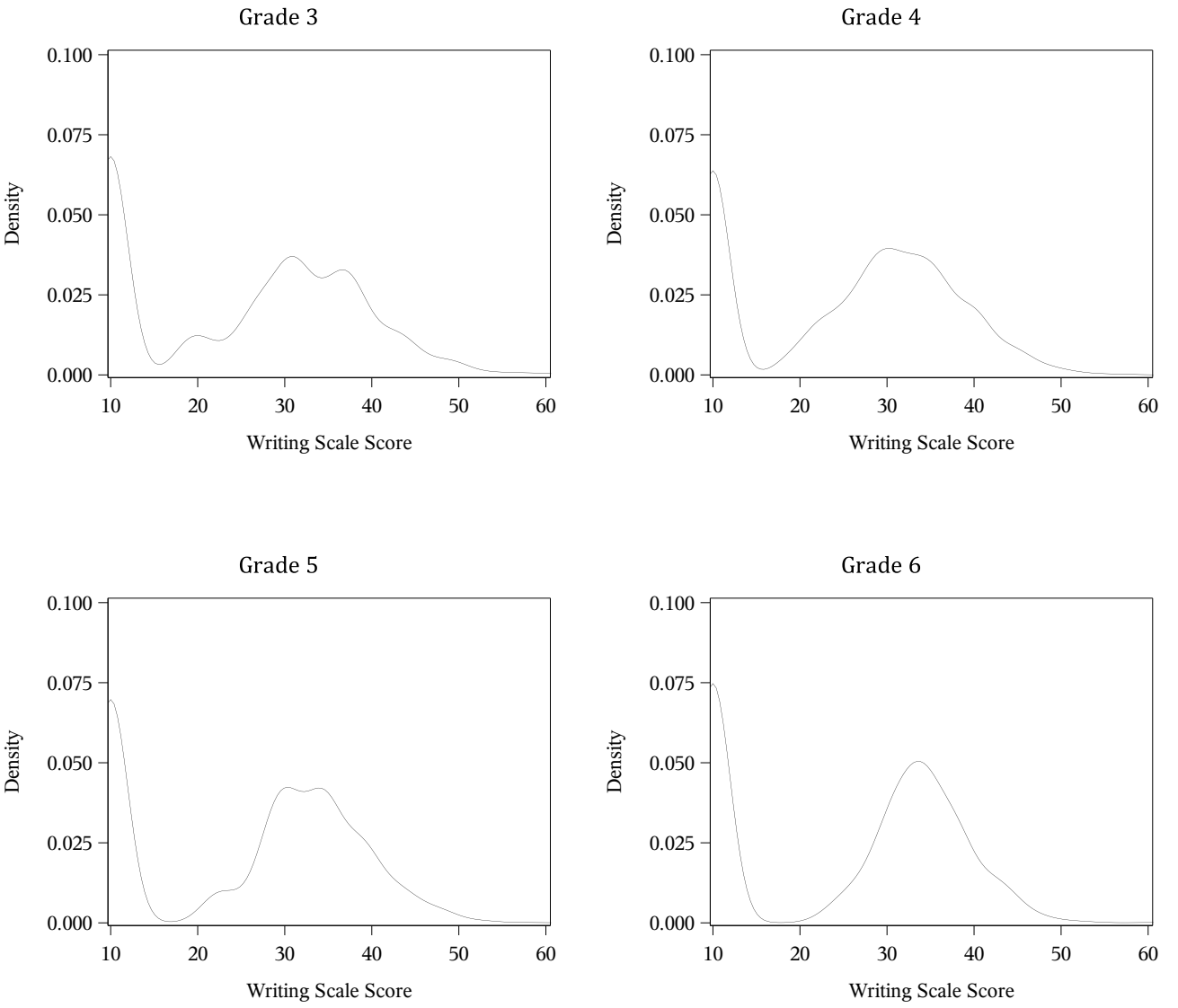

Figure 12.4 Distributions of Writing Scale Scores: Grades 3-8, and 10-11

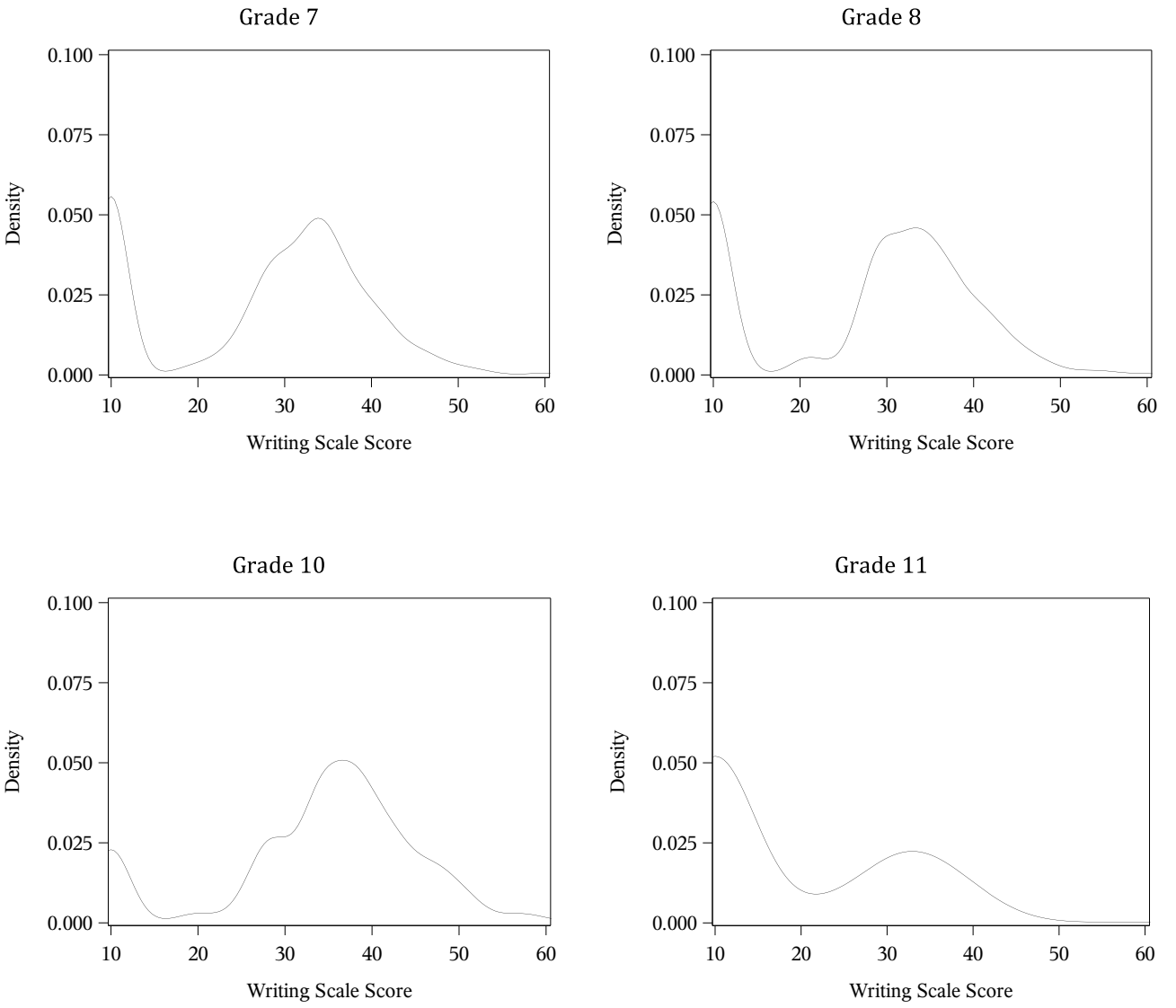

Figure 12.4 (continued) Distributions of Writing Scale Scores: Grades 3-8, and 10-11

12.4.2 Scale Score Cumulative Frequencies for English Language Arts/Literacy

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 English Language Arts/Literacy Groups

Subgroup statistics for ELA/L full summative, Reading, and Writing scale scores are presented in Tables 12.5 and 12.6 [8] for ELA/L grades 3 and 10, 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. [9]

[8] Due to omitted demographic values, subgroup sample sizes may not sum to the total sample size.
[9] Table A.12.44 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		**96,928**	**724.36**	**41.06**	**650**	**850**
Gender	Female	47,502	729.22	41.61	650	850
	Male	49,342	719.66	39.96	650	850
Ethnicity	American Indian/Alaska Native	349	708.90	42.14	650	829
	Asian	5,146	747.36	39.10	650	850
	Black/African American	11,883	701.50	36.74	650	850
	Hispanic/Latino	21,157	709.13	38.57	650	850
	Native Hawaiian or Pacific Islander	180	740.79	37.88	650	850
	Two or more races	4,773	729.37	40.86	650	850
	White	52,318	733.43	38.77	650	850
Economic status*	Not economically disadvantaged	50,287	737.58	38.44	650	850
	Economically disadvantaged	40,130	705.88	36.94	650	850
English learner status	Non-English learner	75,834	728.03	40.47	650	850
	English learner	15,238	701.17	35.39	650	850
Disabilities	Students without disabilities	79,922	729.48	39.88	650	850
	Students with disabilities	15,824	698.60	37.17	650	850
Reading summative score		**96,928**	**41.75**	**16.81**	**10**	**90**
Gender	Female	47,502	43.32	16.88	10	90
	Male	49,342	40.23	16.59	10	90
Ethnicity	American Indian/Alaska Native	349	35.49	16.86	10	90
	Asian	5,146	50.87	16.14	10	90
	Black/African American	11,883	32.92	15.23	10	90
	Hispanic/Latino	21,157	35.65	15.78	10	90
	Native Hawaiian or Pacific Islander	180	47.10	15.09	10	87
	Two or more races	4,773	43.71	16.77	10	90
	White	52,318	45.32	15.95	10	90
Economic status*	Not economically disadvantaged	50,287	47.16	15.87	10	90
	Economically disadvantaged	40,130	34.29	15.03	10	90
English learner status	Non-English learner	75,834	43.33	16.60	10	90
	English learner	15,238	32.03	14.13	10	90
Disabilities	Students without disabilities	79,922	43.79	16.31	10	90
	Students with disabilities	15,824	31.47	15.49	10	90
Writing Summative Score		**96,928**	**25.28**	**12.53**	**10**	**60**
Gender	Female	47,502	26.98	12.59	10	60
	Male	49,342	23.63	12.26	10	60
Ethnicity	American Indian/Alaska Native	349	21.45	12.53	10	53
	Asian	5,146	31.45	11.80	10	60
	Black/African American	11,883	18.93	11.05	10	60
	Hispanic/Latino	21,157	21.37	11.81	10	60
	Native Hawaiian or Pacific Islander	180	30.56	12.13	10	60
	Two or more races	4,773	26.66	12.50	10	60
	White	52,318	27.68	12.20	10	60
Economic status*	Not economically disadvantaged	50,287	28.56	12.14	10	60
	Economically disadvantaged	40,130	20.52	11.51	10	60
English learner status	Non-English learner	75,834	26.04	12.51	10	60

Group Type	Group	N	Mean	SD	Min	Max
	English learner	15,238	19.91	11.20	10	60
Disabilities	Students without disabilities	79,922	26.59	12.40	10	60
	Students with disabilities	15,824	18.67	11.07	10	60

Note. ELA/L = English/language arts/literacy; SD = standard deviation. *Economic status was based on participation in National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Grade 10 subgroup statistics for ELA/L, Reading, and Writing scale scores are presented in Table 12.6. [10] Mean scores were very similar to what was observed for grades 3 through 8. Corresponding tables for grades 10 and 11 are presented in Appendix 12.5.

[10] Table A.12.50 in Appendix 12.5 is identical to Table 12.6.

Table 12.6 Subgroup Performance for ELA/L Scale Scores: Grade 10

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		**2,767**	**757.37**	**40.11**	**650**	**850**
Gender	Female	1,347	764.34	38.37	650	850
	Male	1,376	749.99	40.68	650	850
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	757.54	41.64	650	850
	Black/African American	263	740.13	40.24	650	850
	Hispanic/Latino	605	750.58	39.38	650	850
	Native Hawaiian or Pacific Islander	54	752.93	39.47	650	838
	Two or more races	371	760.51	40.15	650	850
	White	1,139	763.81	38.70	650	850
Economic status*	Not economically disadvantaged	143	714.02	38.25	650	795
	Economically disadvantaged	2,358	761.89	38.25	650	850
English learner status	Non-English learner	365	726.06	38.58	650	850
	English learner	2,767	757.37	40.11	650	850
Disabilities	Students without disabilities	1,347	764.34	38.37	650	850
	Students with disabilities	1,376	749.99	40.68	650	850
Reading summative score		**2,767**	**54.81**	**17.24**	**10**	**90**
Gender	Female	1,347	56.74	16.79	10	90
	Male	1,376	52.74	17.51	10	90
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	53.92	17.75	10	90
	Black/African American	263	47.46	16.37	10	90
	Hispanic/Latino	605	52.37	16.89	10	90
	Native Hawaiian or Pacific Islander	54	50.98	15.66	11	90
	Two or more races	371	55.54	17.12	10	90
	White	1,139	57.85	16.93	10	90
Economic status*	Not economically disadvantaged	143	36.67	16.20	10	85
	Economically disadvantaged	2,358	56.57	16.64	10	90
English learner status	Non-English learner	365	42.73	16.48	10	90
	English learner	2,767	54.81	17.24	10	90
Disabilities	Students without disabilities	1,347	56.74	16.79	10	90
	Students with disabilities	1,376	52.74	17.51	10	90
Writing Summative Score		**2,767**	**33.95**	**11.48**	**10**	**60**
Gender	Female	1,347	36.37	10.40	10	60
	Male	1,376	31.41	12.01	10	60
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	34.68	11.45	10	60
	Black/African American	263	29.80	12.09	10	55
	Hispanic/Latino	605	32.19	11.42	10	60
	Native Hawaiian or Pacific Islander	54	34.28	11.50	10	57
	Two or more races	371	35.11	11.21	10	60
	White	1,139	35.15	11.29	10	60
Economic status*	Not economically disadvantaged	143	24.06	11.73	10	42
	Economically disadvantaged	2,358	35.18	10.81	10	60
English learner status	Non-English learner	365	25.35	12.23	10	57
	English learner	2,767	33.95	11.48	10	60

Group Type	Group	N	Mean	SD	Min	Max
Disabilities	Students without disabilities	1347	36.37	10.40	10	60
	Students with disabilities	1376	31.41	12.01	10	60

Note. ELA/L = English/language arts/literacy; SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

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, and Algebra II. Scale score distributions generally peaked between approximately 700 and the 750 Level 4 performance level cut score for Algebra I and Geometry. Integrated Mathematics results are omitted from this section due to low sample size.

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.

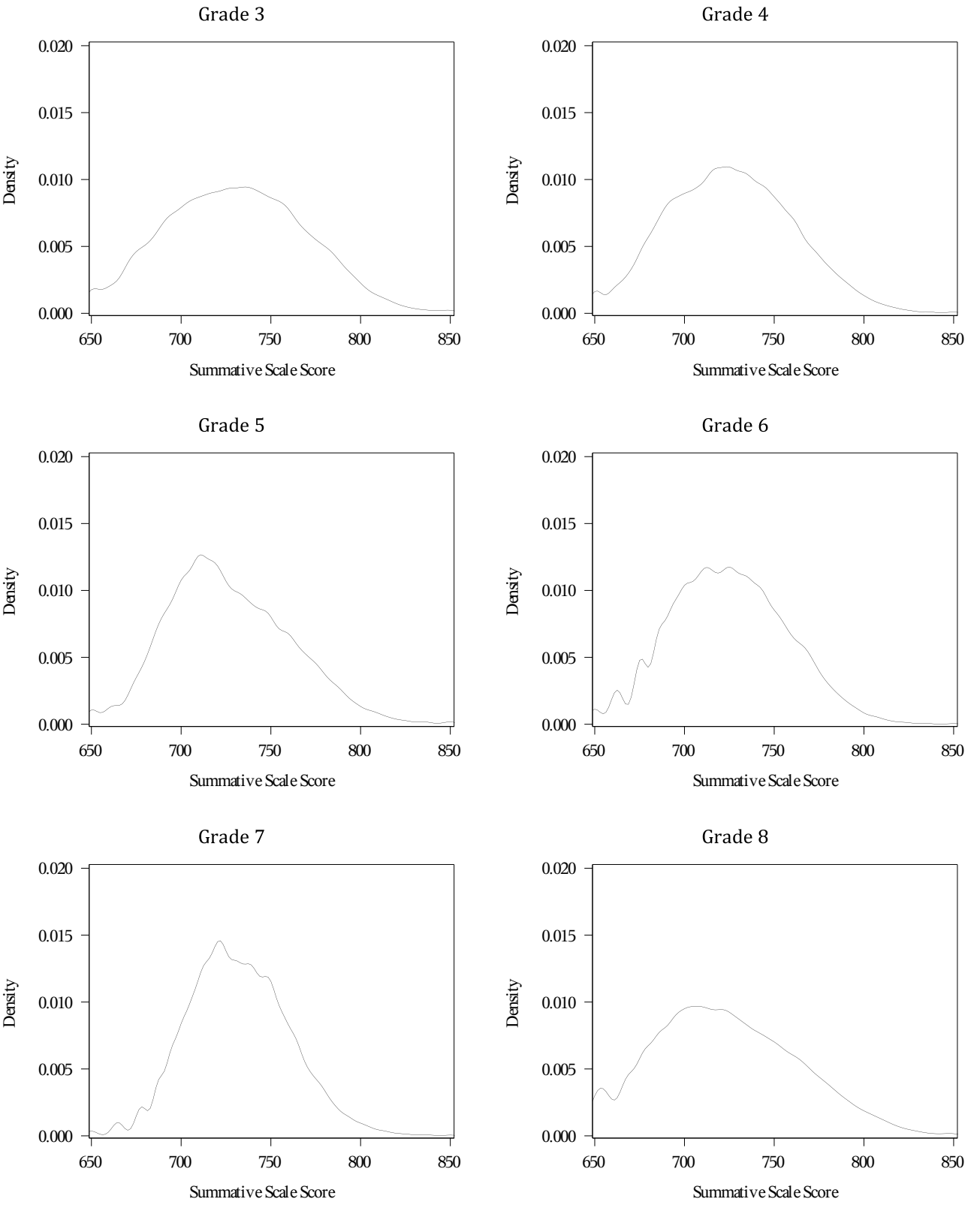

Figure 12.5 Distributions of Mathematics Scale Scores: Grades 3–8

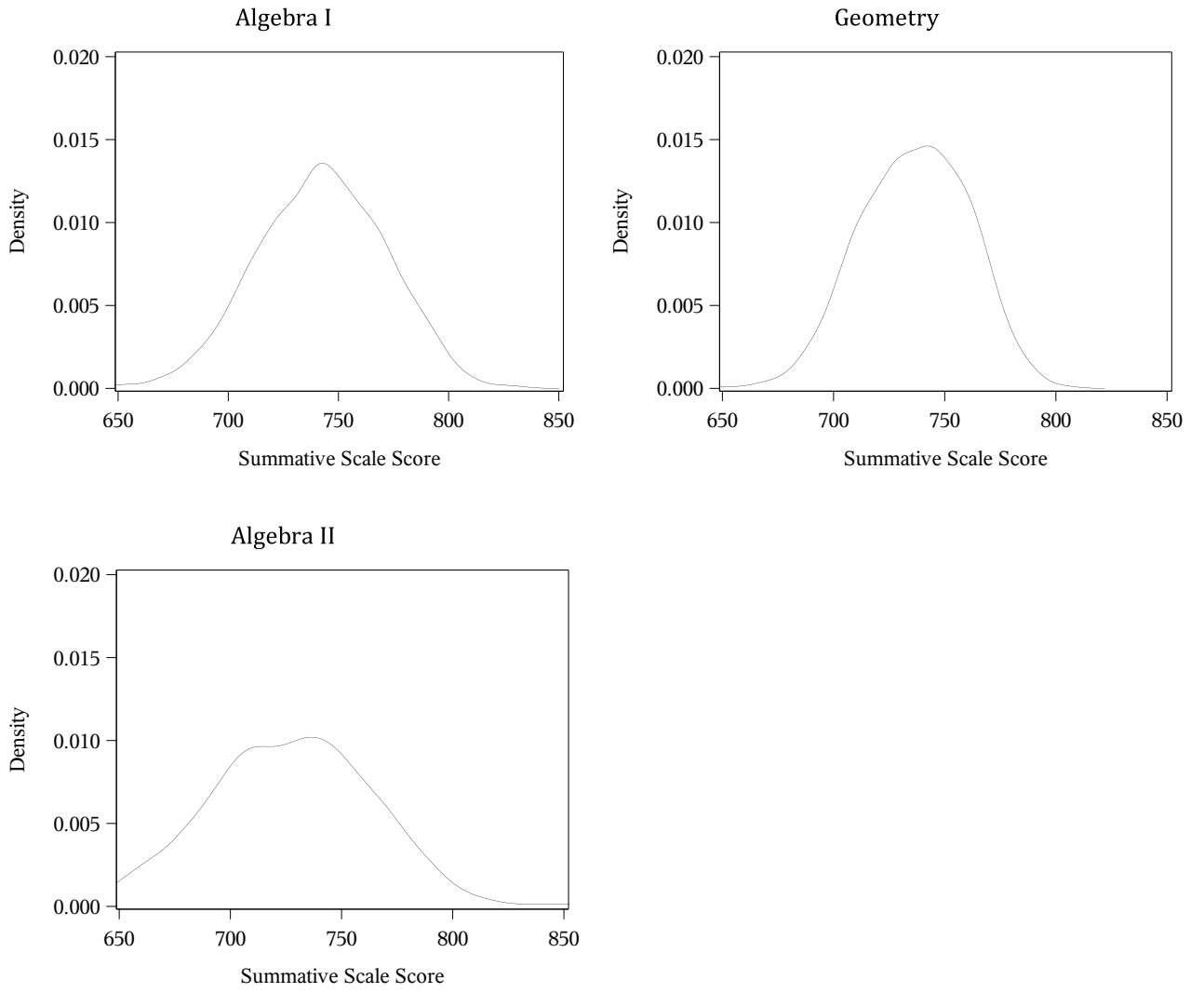

Figure 12.6 Distributions of Mathematics Scale Scores: High School

12.4.6 Summary Scale Score Statistics for Mathematics Groups

Subgroup statistics for mathematics scale scores are presented in Tables 12.7 and 12.8 [11] for grade 3 and Algebra I, respectively. Grade 3 subgroup statistics are presented in Table 12.7. [12] Students using the Spanish language form tended to have lower mean scores. Corresponding tables for all grades/courses are presented in Appendix 12.5.

[11] Due to omitted demographic values, subgroup sample sizes in these tables may not sum to total sample size.
[12] Table A.12.52 in Appendix 12.5 is identical to Table 12.7.

Table 12.7 Subgroup Performance for Mathematics Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		**96,011**	**730.93**	**38.66**	**650**	**850**
Gender	Female	47,029	729.42	37.71	650	850
	Male	48,909	732.36	39.50	650	850
Ethnicity	American Indian/Alaska Native	339	714.08	38.53	650	830
	Asian	5,130	760.21	39.24	650	850
	Black/African American	11,609	703.07	32.33	650	850
	Hispanic/Latino	20,914	714.39	33.84	650	850
	Native Hawaiian or Pacific Islander	179	738.60	34.18	670	814
	Two or more races	4,739	734.44	38.81	650	850
	White	52,020	741.11	35.37	650	850
Economic status*	Not economically disadvantaged	49,969	745.31	36.08	650	850
	Economically disadvantaged	39,577	711.07	32.85	650	850
English learner status	Non-English learner	75,112	734.10	38.50	650	850
	English learner	15,095	710.97	32.90	650	850
Disabilities	Students without disabilities	79,199	734.96	37.82	650	850
	Students with disabilities	15,632	710.76	36.67	650	850
Language form	Spanish	1,729	700.02	26.93	650	810

Note. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table 12.8 Subgroup Performance for Mathematics Scale Scores: Algebra I

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		**3,424**	**742.39**	**29.14**	**650**	**833**
Gender	Female	1,626	741.62	28.33	650	829
	Male	1,756	743.20	29.90	650	833
Ethnicity	American Indian/Alaska Native	25	716.00	30.81	668	793
	Asian	252	748.23	30.70	650	832
	Black/African American	321	729.98	24.43	668	797
	Hispanic/Latino	679	737.02	27.88	650	833
	Native Hawaiian or Pacific Islander	65	733.88	28.28	659	787
	Two or more races	493	744.62	28.55	674	828
	White	1,376	749.01	27.93	650	829
Economic status*	Not economically disadvantaged					
	Economically disadvantaged					
English learner status	Non-English learner					
	English learner	244	726.47	30.09	650	829
Disabilities	Students without disabilities	2,956	745.85	27.56	650	832
	Students with disabilities	427	718.73	28.97	650	833

Note. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

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 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 accounts 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 following formula:

$$\alpha = \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right], \tag{13-1}$$

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

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

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

The formula for the standard error of measurement is

$$\sigma_E = \sigma_X\sqrt{1-\rho_{XX'}}, \tag{13-3}$$

where σ_X is the standard deviation of the test raw score and $\rho_{xx'}$ is the reliability estimated by substitution of appropriate statistics for the parameters 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 (Cortina, 1993, Lord & Novick, 1968; Tavakol & Dennick, 2011). As test length decreases and samples become smaller and more homogeneous, lower estimates of alpha are obtained (Pike & Hudson, 1998; Tavakol & Dennick, 2011). 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 students would be to obtain similar scores upon repeated testing occasions, if the they do not change in their level of the knowledge or skills measured by the test. Because the scale scores are computed from a total score and do not have an item-level component, a stratified alpha coefficient cannot be computed for scale scores. Instead, Kolen et al.'s (1996) method for scale score reliability was used.

The general formula for a reliability coefficient,

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

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 tables provide the average reliability, which is estimated by averaging the internal consistency estimates computed for all the individual forms of the test and the raw score SEMs. In addition, the number of forms, the sample size of the minimum reliability, sample size of the maximum reliability, and the average maximum possible score for each set of tests are provided. Estimates were calculated only for groups of 100 or more students administered a specific test form.

English Language Arts/Literacy

The average reliability estimates for grades 3 through 8 and 10 through 11 ELA/L range from a low of .79 to a high of .89; note that grade 11 had a low sample size. The average raw score SEM is consistently between about 6 percent and 8 percent of the maximum possible score.

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

Grade Level	Number of Forms	Avg. Max Possible Score	Avg. Raw Score SEM	Average Reliability	Minimum Reliability		Maximum Reliability	
					N	Alpha	N	Alpha
3	4	54	3.73	0.86	1739	0.77	44314	0.87
4	4	70	4.41	0.86	1901	0.75	44291	0.87
5	4	70	4.37	0.87	1951	0.72	45896	0.88
6	4	72	4.52	0.88	1823	0.76	238	0.89
7	5	72	4.71	0.89	1890	0.84	127	0.92
8	4	72	4.82	0.88	242	0.85	130	0.93
10	2	70	5.51	0.82	1888	0.8	868	0.87
11	2	72	4.04	0.79	211	0.78	131	0.8

Mathematics

The average reliability estimates for mathematics assessments range from .86 to .92. The raw score SEM consistently ranges from about 5% to 7% of the maximum score. Integrated Mathematics is omitted from this section due to low sample sizes.

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	6	52	3.03	0.92	654	0.9	33023	0.93
4	6	52	3.07	0.92	572	0.85	34974	0.92
5	6	52	3.26	0.91	557	0.83	35732	0.91
6	6	52	2.94	0.92	395	0.83	9475	0.93
7	6	52	3.23	0.91	274	0.82	9704	0.93
8	6	52	2.74	0.91	333	0.71	8995	0.92
A1	2	55	2.79	0.87	1932	0.86	1179	0.87
GO	2	55	2.96	0.87	1626	0.87	883	0.87
A2	2	55	3.05	0.86	1601	0.86	974	0.87

Note. A1=Algebra I, GO=Geometry, A2=Algebra II.

13.3.2 Scale Score Reliability Results

Tables 13.3 and 13.4 summarize scale score reliability estimates for the total testing group for ELA/L and mathematics for spring 2021. 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. Scale score reliability requires an ability distribution, which is not reasonable to assume for Integrated Mathematics due to the low sample sizes.

English Language Arts/Literacy

Reliability estimates for ELA/L are presented in Table 13.3. Average reliabilities range from .86 to .86. The average SEM ranges from 10.85 to 14.97.

Table 13.3 Summary of ELA/L Test 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	7	14.97	0.86	0.84	0.88
4	7	12.18	0.87	0.87	0.88
5	7	11.84	0.87	0.85	0.88
6	6	10.85	0.88	0.87	0.89
7	6	12.04	0.89	0.88	0.9
8	6	11.94	0.89	0.89	0.9
10	5	14.75	0.89	0.88	0.9
11	6	14.56	0.86	0.84	0.87

Mathematics

The scale score reliability estimates for the mathematics assessments are presented in Table 13.4. Average scale score reliability estimates for the grades 3 through 8 mathematics assessments range from .86 to .91. For the high school assessments, these quantities range from .85 to .87. For grades 3 through 8, the average scale score SEM ranges from 9.33 to 13.26. For high school tests, the average scale score SEM ranges from 10.4 to 15.4.

Table 13.4 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	6	10.14	0.91	0.91	0.92
4	6	9.81	0.91	0.9	0.91
5	6	9.98	0.9	0.89	0.9
6	6	9.47	0.9	0.9	0.9
7	6	9.33	0.89	0.88	0.89
8	6	13.26	0.86	0.85	0.87
A1	7	12.79	0.86	0.83	0.87
GO	7	10.4	0.87	0.82	0.88
A2	7	15.4	0.85	0.84	0.86

Note. A1=Algebra I, GO=Geometry, A2=Algebra II.

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 differential item functioning analysis. Estimates were calculated only for groups of 100 or more students administered a specific test form.

Tables 13.5 and 13.6 summarize test reliability for groups of interest for ELA/L grade 3 and mathematics grade 3, respectively. Corresponding information is provided in Appendix 13.1 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 .02. The SEMs for females were slightly higher than for males for all ELA/L assessments.

Mathematics

As with the ELA/L test components, the average reliability estimates and SEMs for males and females reflect the corresponding reliabilities for the total group. For most tests, the reliabilities between males and females are equal or within .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 .84 and .88. Average SEMs were generally slightly higher for White and Asian/Pacific Islander students than for Black/African American and Hispanic/Latino students.

Mathematics

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

13.4.3 Reliability Results for Special Education Needs

English Language Arts/Literacy

The average reliabilities for five groups of students (economically disadvantaged, not economically disadvantaged, non-English learner, students with disabilities, and students without disabilities) are generally equal to or .01 to .02 less than the average reliability for the total group of students. Average reliabilities for English learner students are lower, ranging from .76 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

Reliability information for accommodated forms is sparse due to small sample sizes or because the form was not administered. Reliabilities for test-to-speech forms tended to be lower than the overall reliabilities, while those for closed-caption forms tended to be higher.

Mathematics

The text-to-speech forms had sufficient sample sizes for reliability and SEM estimation across grades/subjects, except for high school 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

There were sufficient numbers of students taking the Spanish-language form for reliability and SEM estimation for grades 3 through 8. The average reliability ranged from .82 to .86. The SEMs are generally lower for the students administered the Spanish-language forms.

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

	Max. Raw Score	Avg. SEM	Average Reliability	Minimum Reliability N	Minimum Reliability Alpha	Maximum Reliability N	Maximum Reliability Alpha
Total Group	54	3.73	0.86	1739	0.77	44314	0.87
Gender							
Male	54	3.6	0.86	1142	0.76	506	0.87
Female	54	3.85	0.86	595	0.79	21802	0.86
Ethnicity							
White	54	3.9	0.84	918	0.79	141	0.85
Black/African American	54	3.28	0.85	250	0.77	5432	0.86
Asian/Pacific Islander	54	4.25	0.82	2377	0.81	2407	0.82
American Indian/Alaska Native	53	3.31	0.89	144	0.88	152	0.9
Hispanic/Latino	54	3.44	0.86	392	0.73	9686	0.87
Multiple	53	4.03	0.84	1927	0.84	1931	0.84
Special Instruction Needs							
Economically Disadvantaged	54	3.36	0.85	1080	0.74	19208	0.86
Not Economically Disadvantaged	54	3.76	0.85	594	0.81	24602	0.86
English Learner	54	3.33	0.83	353	0.65	6967	0.85
Non-English Learner	54	3.65	0.87	1319	0.79	36834	0.87
Students with Disabilities	54	3.15	0.86	1684	0.76	6320	0.87
Students without Disabilities	53	3.85	0.85	301	0.84	37501	0.85
Students Taking Accommodated Forms							
ASL							
Closed-Caption							
Screen Reader							
Text-to-Speech	54	2.66	0.75	1487	0.75	1487	0.75

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

Table 13.6 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	3.03	0.92	654	0.9	33023	0.93
Gender							
Male	52	3.02	0.93	401	0.9	16735	0.93
Female	52	3.04	0.92	344	0.89	5097	0.92
Ethnicity							
White	52	3.13	0.91	362	0.9	3724	0.93
Black/African American	52	2.72	0.9	102	0.81	2907	0.91
Asian/Pacific Islander	52	3.17	0.92	1879	0.92	530	0.93
American Indian/Alaska Native							
Hispanic/Latino	52	2.86	0.9	368	0.83	5324	0.92
Multiple	52	3.06	0.93	1865	0.92	345	0.93
Special instruction needs							
Economically disadvantaged	52	2.82	0.9	506	0.83	11048	0.91
Not economically disadvantaged	52	3.16	0.91	285	0.91	4027	0.93
English learner	52	2.81	0.9	261	0.83	3383	0.91
Non-English learner	52	3.07	0.92	481	0.89	6675	0.93
Students with disabilities	52	2.81	0.91	2977	0.88	3805	0.93
Students without disabilities	52	3.07	0.92	161	0.88	7830	0.93
Students taking accommodated forms							
American Sign Language							
Closed-caption							
Screen reader							
Text-to-speech	52	2.88	0.93	9438	0.92	9841	0.93
Students taking translated forms							
Spanish language form	52	2.59	0.86	1505	0.86	1505	0.86

Note. ELA/L = English language arts/literacy; 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 13.7 for a summary of the ELA/L claims and subclaims.

Table 13.7 Descriptions of ELA/L Claims and Subclaims

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

Reliability indices were calculated for each major claim and subclaim. Table 13.8 presents the average reliability estimates for all forms of the test at the specified grade 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. Reliabilities from grade 11 tended to be lower than the other grades, so they are omitted from the descriptions in the following paragraphs. However, they can be found in Table 13.8.

The average reliabilities for the Reading claim for grades 3 through 8 and 10 range from .8 to .85. They are based on maximum scores of 38 to 44 points per form, except for grade 3 (28 to 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 .76 to .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 vary from .61 to .77. The maximum number of points per form ranges from 11 to 18. The average reliabilities of the Reading Information subclaim scores vary from .54 to .72, with 7 to 24 points per form. The average reliabilities of the Reading Vocabulary subclaim scores vary from .47 to .62. 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 to 24 points for grades 4 and 5. Grades 6 through 11 are based on 24 points for all forms. The average reliabilities range from .66 to

.81. The Writing Knowledge of Language and Conventions subclaims are all based on six points. The reliabilities range from .73 to .84.

Table 13.8 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	0.85	11-12	0.71	7-11	0.61	8-10	0.61	24-24	0.76	18-18	0.67	6-6	0.79
4	40-44	0.83	18-18	0.76	12-14	0.54	8-14	0.57	27-30	0.76	21-24	0.69	6-6	0.77
5	40-44	0.84	16-18	0.66	14-16	0.62	10-14	0.62	27-30	0.8	21-24	0.74	6-6	0.79
6	40-44	0.84	14-18	0.77	14-16	0.66	8-14	0.51	30-30	0.78	24-24	0.74	6-6	0.8
7	40-44	0.85	16-18	0.68	14-16	0.72	8-14	0.54	30-30	0.82	24-24	0.81	6-6	0.84
8	40-44	0.85	16-18	0.7	14-14	0.66	10-14	0.57	30-30	0.82	24-24	0.8	6-6	0.82
10	38-44	0.8	12-18	0.61	14-22	0.64	8-12	0.47	30-30	0.77	24-24	0.69	6-6	0.73
11	40-44	0.7	12-16	0.51	14-24	0.43	8-12	0.36	30-30	0.77	24-24	0.66	6-6	0.68

Note. ELA/L = English language arts/literacy.

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 13.9 presents the average reliability estimates for mathematics subclaims.

Subclaims with greater numbers of points tend to have greater reliability estimates. The Major Content subclaim has the largest number of points for each assessment and, accordingly, has higher average reliabilities than the other three subclaims. For grades 3 through 8, Algebra I, Geometry, and Algebra II, the median of the average reliabilities for the Major Content range from .64 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 ranges from .54 to .68. The maximum number of points per form for this subclaim ranges from 9 to 12.

The average reliabilities for Mathematics Reasoning range from .5 to .7 for grades 3 through 8, Algebra I, Geometry, and Algebra II. 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 range from .57 to .73. The number of points is 12 for grades 3 through 8 and 15 for all high school courses.

The Integrated Mathematics assessments do not have sufficient sample sizes for reliability analyses.

Table 13.9 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	0.68	10-10	0.65	12-12	0.73
4	21-21	0.85	9-9	0.64	10-10	0.66	12-12	0.61
5	20-20	0.81	10-10	0.65	10-10	0.56	12-12	0.69
6	20-20	0.82	10-10	0.62	10-10	0.7	12-12	0.7
7	20-20	0.83	10-10	0.58	10-10	0.6	12-12	0.66
8	20-20	0.79	10-10	0.54	10-10	0.7	12-12	0.69
A1	17-17	0.68	9-9	0.53	10-10	0.57	15-15	0.62
GO	18-18	0.75	12-12	0.55	10-10	0.56	15-15	0.57
A2	16-18	0.64	12-12	0.62	10-10	0.5	15-15	0.62

Note. * Cronbach alpha below .50, further investigation summarized at the end of Section 13.6. A1 = Algebra I, GO = Geometry, A2 = Algebra II.

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 versus 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 .64 to .71; the proportion who would be consistently classified on two different test forms ranges from .53 to .61. For classifying each student as being at Level 4 or higher versus being at Level 3 or lower, the proportion accurately classified ranges from .88 to .92; the proportion who would be consistently classified this way on two different test forms ranges from .83 to .89.

Table 13.10 Reliability of Classification: Summary for ELA/L

Level	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
	Exact Level	Level 4 or Higher versus 3 or Lower	Exact Level	Level 4 or Higher versus 3 or Lower
3	0.68	0.90	0.59	0.85
4	0.67	0.89	0.56	0.85
5	0.70	0.90	0.60	0.86
6	0.71	0.90	0.61	0.86
7	0.68	0.90	0.58	0.86
8	0.69	0.90	0.59	0.86
10	0.64	0.88	0.53	0.83
11	0.65	0.92	0.55	0.89

Note. ELA/L = English language arts/literacy.

Table 13.11 provides more detailed information about the accuracy and the consistency of the classification of students into performance levels for ELA/L grade 3. Each cell in the five-by-five table shows the estimated proportion of students who would be classified into a particular combination of performance levels. The sum of the five bold values on the diagonal is approximately equal to the level of decision accuracy or consistency presented in Table 13.10. For "Level 4 and Higher versus 3 and Lower" found in Table 13.10, the sum of the shaded values in Table 13.11 is approximately equal to the level of decision accuracy or consistency presented in Table 13.10. Note that the sums based on values in Table 13.11 may not match exactly to the values in Table 13.10 due to truncation and rounding.

Detailed information for all ELA/L spring results are provided in Tables A.13.18 through A.13.25. The structure of these tables is the same as that of Table 13.11 and the values in the tables should be interpreted in the same manner. Table 13.11 includes the same information as Table A.13.18.

Table 13.11 Reliability of Classification: Grade 3 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision Accuracy	650-699	**0.23**	0.04	0.00	0.00	0.00	0.27
	700-724	0.04	**0.11**	0.05	0.00	0.00	0.21
	725-749	0.00	0.05	**0.11**	0.05	0.00	0.22
	750-809	0.00	0.00	0.05	**0.22**	0.02	0.30
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision Consistency	650-699	**0.22**	0.06	0.01	0.00	0.00	0.29
	700-724	0.05	**0.08**	0.05	0.01	0.00	0.20
	725-749	0.01	0.05	**0.08**	0.06	0.00	0.20
	750-809	0.00	0.02	0.07	**0.20**	0.02	0.30
	810-850	0.00	0.00	0.00	0.01	**0.00**	0.01

Note. ELA/L = English language arts/literacy.

13.7.2 Mathematics

Table 13.12 provides information about the accuracy and the consistency of two types of classifications made on the basis of the summative scale scores on the mathematics assessments. For the grades 3 through 8 mathematics tests, the table shows that for classifying each student into one of the five performance levels,

the proportion accurately classified ranges from .73 to .76; the proportion who would be consistently classified on two different test forms ranges from .59 to .66. 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 .69 to .74; the proportion who would be consistently classified on two different test forms ranges from .59 to .67.

For classifying each student as being at Level 4 or higher versus being at Level 3 or lower, for the grades 3 through 8 mathematics tests, the proportion accurately classified ranges from .92 to .93; the proportion who would be consistently classified on two different test forms is .89 to .90 for grades 3 and 8. For high school mathematics courses, the proportion accurately classified as being at Level 4 or higher versus being at Level 3 or lower ranges from .89 to .90; the proportion who would be consistently classified on two different test forms ranges from .84 to .86.

Appendix 13 Tables A.13.26 through A.13.34 provide more detailed information about the accuracy and the consistency of the classification of students into performance levels for mathematics. Each cell in the five-by-five table shows the estimated proportion of students who would be classified into a particular combination of performance levels.

Table 13.12 Reliability of Classification: Summary for Mathematics

	Decision Accuracy: Proportion Accurately Classified		Decision Consistency: Proportion Consistently Classified	
Level	Exact Level	Level 4 or Higher versus 3 or Lower	Exact Level	Level 4 or Higher versus 3 or Lower
3	0.73	0.92	0.63	0.89
4	0.75	0.93	0.66	0.9
5	0.74	0.93	0.64	0.9
6	0.76	0.93	0.67	0.9
7	0.76	0.92	0.66	0.89
8	0.73	0.92	0.63	0.89
A1	0.74	0.89	0.65	0.84
GO	0.74	0.89	0.64	0.85
A2	0.69	0.9	0.59	0.86

Note. A1 = Algebra I, GO = Geometry, A2 = Algebra II.

13.8 Inter-rater Agreement

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

Table 13.13 Inter-rater Agreement Expectations and Results

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

Note. A 0 or 1 score compared to a blank score will have a disagreement greater than 1 point. ELA/L= English language arts/literacy.

Pearson's ePEN2 scoring system included comprehensive inter-rater agreement reports that allowed supervisory personnel to monitor both individual and group performance. Based on reviews of these reports, scoring experts targeted individuals for increased backreading and feedback and, if necessary, retraining. Table 13.13 shows that the actual percentages for perfect reader agreement were higher than the inter-rater agreement expectations, and the percentages for within one point were very close. Refer to Section 4 for more information on 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 2020–2021 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 on Educational Outcomes 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 [CCSS]) 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 (PLDs). At the policy level, the PLDs include policy claims about the educational achievement of students who attain a particular performance level, and a broad description of the

grade-level knowledge, skills, and practices students performing at a particular achievement level are able to demonstrate. Those policy-level descriptors are the foundation for the subject- and grade-specific PLD, 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% 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, Trends in International Mathematics and Science Study, Programme of International Student Assessment, and Progress in International Reading Literacy Study tests were conducted (McClarty et al., 2015).

Gathering construct validity evidence for the assessments is embedded in the process by which the 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 Individualized Education Program, 504 Plan, or if required by the participating state or agency, an English Learner 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 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, PLDs, 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 is provided to summarize the results for the spring 2021 administration. Tables 14.1 through 14.8 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 among the subclaims are sufficiently high to provide evidence that the ELA/L tests are unidimensional. The moderate intercorrelations among the subclaims and claims suggest the claims may be sufficient for individual student reporting.

The intercorrelations and reliability estimates for mathematics are provided in Tables 14.9 through 14.17. 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 among the ASC, MR, and MP subclaims are usually slightly lower. The mathematics intercorrelations are sufficiently high to suggest that the mathematics tests are likely to be unidimensional with some minor secondary dimensions.

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

	RD	RL	RI	RV	WR	WE	WKL
RD	0.85	96,916	96,916	96,916	96,916	96,916	96,916
RL	0.91	0.71	96,916	96,916	96,916	96,916	96,916
RI	0.85	0.65	0.61	96,916	96,916	96,916	96,916
RV	0.86	0.66	0.60	0.61	96,916	96,916	96,916
WR	0.68	0.61	0.65	0.51	0.76	96,916	96,916
WE	0.66	0.59	0.64	0.49	0.98	0.67	96,916
WKL	0.62	0.56	0.57	0.48	0.87	0.77	0.79

Note. ELA/L = English language arts/literacy, 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	0.83	98,994	98,994	98,994	98,994	98,994	98,994
RL	0.93	0.76	98,994	98,994	98,994	98,994	98,994
RI	0.81	0.63	0.54	98,994	98,994	98,994	98,994
RV	0.82	0.64	0.54	0.57	98,994	98,994	98,994
WR	0.67	0.62	0.63	0.48	0.76	98,994	98,994
WE	0.65	0.60	0.62	0.47	0.99	0.69	98,994
WKL	0.63	0.58	0.58	0.46	0.91	0.83	0.77

Note. ELA/L = English language arts/literacy, 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	0.84	99,628	99,628	99,628	99,628	99,628	99,628
RL	0.89	0.66	99,628	99,628	99,628	99,628	99,628
RI	0.84	0.62	0.62	99,628	99,628	99,628	99,628
RV	0.85	0.64	0.60	0.62	99,628	99,628	99,628
WR	0.68	0.61	0.64	0.50	0.8	99,628	99,628
WE	0.67	0.60	0.64	0.49	0.99	0.74	99,628
WKL	0.65	0.59	0.61	0.49	0.94	0.89	0.79

Note. ELA/L = English language arts/literacy, 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	0.84	98,577	98,577	98,577	98,577	98,577	98,577
RL	0.92	0.77	98,577	98,577	98,577	98,577	98,577
RI	0.85	0.65	0.66	98,577	98,577	98,577	98,577
RV	0.80	0.63	0.57	0.51	98,577	98,577	98,577
WR	0.69	0.61	0.67	0.49	0.78	98,577	98,577
WE	0.68	0.60	0.66	0.48	0.99	0.74	98,577
WKL	0.67	0.59	0.64	0.48	0.94	0.90	0.8

Note. ELA/L = English language arts/literacy, 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	96,935	96,935	96,935	96,935	96,935	96,935
RL	0.90	0.68	96,935	96,935	96,935	96,935	96,935
RI	0.89	0.69	0.72	96,935	96,935	96,935	96,935
RV	0.80	0.59	0.61	0.54	96,935	96,935	96,935
WR	0.71	0.62	0.72	0.49	0.82	96,935	96,935
WE	0.70	0.61	0.72	0.48	1.00	0.81	96,935
WKL	0.70	0.60	0.70	0.48	0.96	0.93	0.84

Note. ELA/L = English language arts/literacy, 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	96,018	96,018	96,018	96,018	96,018	96,018
RL	0.90	0.7	96,018	96,018	96,018	96,018	96,018
RI	0.86	0.65	0.66	96,018	96,018	96,018	96,018
RV	0.82	0.61	0.59	0.57	96,018	96,018	96,018
WR	0.70	0.61	0.70	0.49	0.82	96,018	96,018
WE	0.69	0.61	0.69	0.48	1.00	0.8	96,018
WKL	0.69	0.61	0.68	0.49	0.97	0.95	0.82

Note. ELA/L = English language arts/literacy, 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 10 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.8	2,763	2,763	2,763	2,763	2,763	2,763
RL	0.86	0.61	2,763	2,763	2,763	2,763	2,763
RI	0.86	0.58	0.64	2,763	2,763	2,763	2,763
RV	0.75	0.51	0.51	0.47	2,763	2,763	2,763
WR	0.67	0.54	0.67	0.43	0.77	2,763	2,763
WE	0.66	0.53	0.67	0.43	1.00	0.69	2,763
WKL	0.65	0.53	0.65	0.42	0.96	0.93	0.73

Note. ELA/L = English language arts/literacy, 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 11 ELA/L Subclaims

	RD	RL	RI	RV	WR	WE	WKL
RD	0.7	413	413	413	413	413	413
RL	0.78	0.51	413	413	413	413	413
RI	0.78	0.39	0.43	413	413	413	413
RV	0.73	0.38	0.43	0.36	413	413	413
WR	0.58	0.49	0.56	0.33	0.77	413	413
WE	0.58	0.48	0.57	0.33	0.95	0.66	413
WKL	0.56	0.47	0.53	0.32	0.91	0.88	0.68

Note. ELA/L = English language arts/literacy, 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 3 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.86	95,980	95,980	95,980
ASC	0.74	0.68	95,980	95,980
MR	0.73	0.60	0.65	95,980
MP	0.73	0.57	0.62	0.73

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

Table 14.10 Average Intercorrelations and Reliability between Grade 4 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.85	97,714	97,714	97,714
ASC	0.70	0.64	97,714	97,714
MR	0.71	0.62	0.66	97,714
MP	0.71	0.60	0.69	0.61

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 5 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.81	98,283	98,283	98,283
ASC	0.70	0.65	98,283	98,283
MR	0.69	0.61	0.56	98,283
MP	0.76	0.67	0.66	0.69

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 6 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.82	96,905	96,905	96,905
ASC	0.72	0.62	96,905	96,905
MR	0.77	0.67	0.7	96,905
MP	0.75	0.64	0.71	0.7

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 7 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.83	91,306	91,306	91,306
ASC	0.72	0.58	91,306	91,306
MR	0.74	0.64	0.6	91,306
MP	0.77	0.65	0.69	0.66

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 8 Mathematics Subclaims

	MC	ASC	MR	MP
MC	0.79	92,936	92,936	92,936
ASC	0.66	0.54	92,936	92,936
MR	0.77	0.62	0.7	92,936
MP	0.67	0.45	0.68	0.69

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

Table 14.15 Average Intercorrelations and Reliability between Algebra I Subclaims

	MC	ASC	MR	MP
MC	0.68	3,424	3,424	3,424
ASC	0.60	0.53	3,424	3,424
MR	0.55	0.60	0.57	3,424
MP	0.60	0.55	0.55	0.62

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

Table 14.16 Average Intercorrelations and Reliability between Geometry Subclaims

	MC	ASC	MR	MP
MC	0.75	2,921	2,921	2,921
ASC	0.68	0.55	2,921	2,921
MR	0.60	0.58	0.56	2,921
MP	0.64	0.56	0.52	0.57

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

Table 14.17 Average Intercorrelations and Reliability between Algebra II Subclaims

	MC	ASC	MR	MP
MC	0.64	2,726	2,726	2,726
ASC	0.62	0.62	2,726	2,726
MR	0.57	0.59	0.5	2,726
MP	0.64	0.60	0.57	0.62

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 among 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 14.17. [13] The average reliabilities for ELA/L and mathematics summative assessments range from .79 up to .92. Tables 13.5 through 13.14 summarize test reliability for groups of interest for ELA/L grades 3 through 9 and 10 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

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

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

$$\alpha = \frac{l}{l-1}\frac{\sum_{i \neq i'} \sigma_{ii'}}{\sigma_X^2}, \tag{14-1}$$

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.

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

LID manifests itself as a residual correlation that is nonzero and large. For $Q3$, LID can be either positive or negative. Positive (negative) LID indicates that performance is higher (lower) than expectation. The residual $Q3$ correlation matrix can be inspected to determine if there are any blocks of locally dependent items (e.g., perhaps blocks of items belonging to the same reading passage). For $Q3$, the null hypothesis is that local independence holds. 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 computer-based test (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 14.18.

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, and 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 14.19 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% 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 14.19 centered around the expectation of -1/n-1.

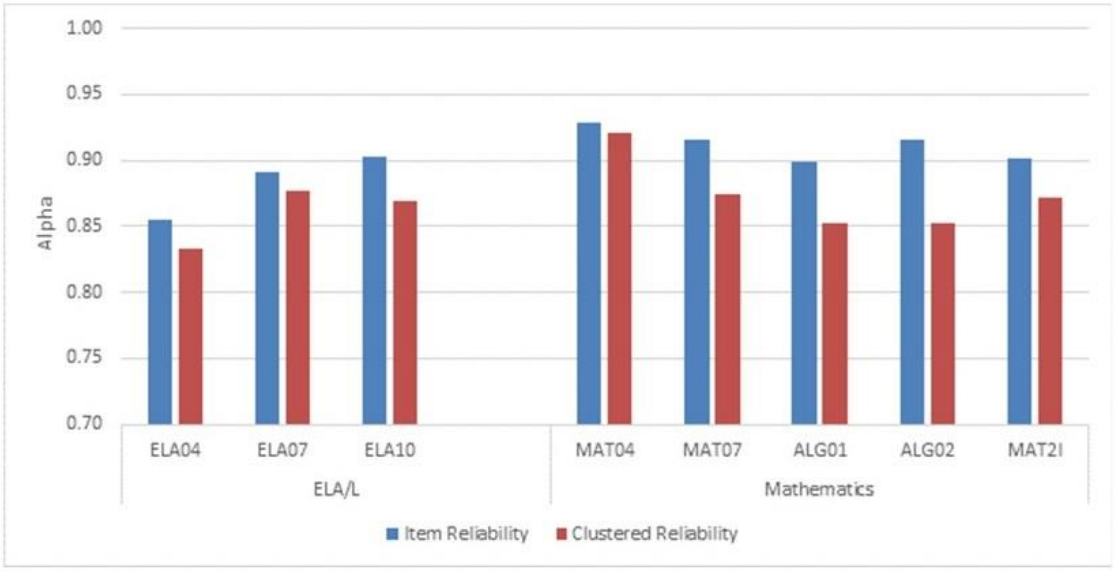

Figure 14.1 Comparison of Internal Consistency by Item and Cluster (Testlet)

Table 14.18 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. ELA/L = English language arts/literacy, A1 = Algebra I, A2 = Algebra II, M2 = Integrated Mathematics II.

Table 14.19 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

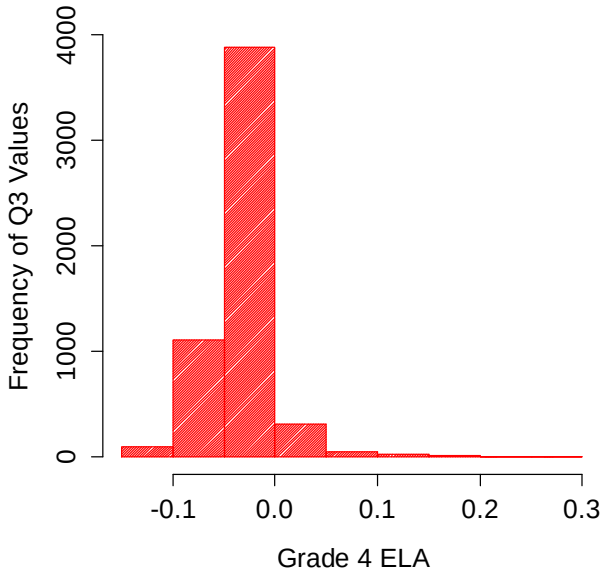

Figure 14.2 Distribution of Q3 Values for Grade 4 ELA/L (Spring 2015)

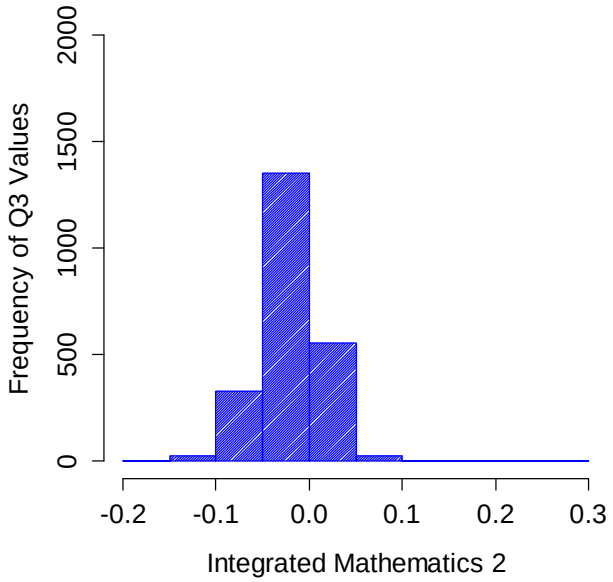

Figure 14.3 Distribution of Q3 Values for Integrated Mathematics II (Spring 2015)

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 14.20 through 14.25 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 2021. ELA/L, Reading (RD), and Writing (WR) are moderately correlated with mathematics; the correlations range from .68 up to .71 for grades 3 through 8. These correlations suggest that while there is a relationship between ELA/L and mathematics, they 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 14.26 through 14.28. 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. Blank cells indicate pairings with sample sizes less than 100. Across grades 8 through 11, ELA/L, Reading (RD), and Writing (WR) scores have correlations with high school mathematics tests that range from .38 to .55. Correlations between high school mathematics scores and corresponding ELA/L scores demonstrate low to moderate correlations.

Table 14.20 Correlations between ELA/L and Mathematics for Grade 3

	ELA/L	RD	WR	MA
ELA/L		94,728	94,728	94,728
RD	0.95		94,728	94,728
WR	0.86	0.70		94,728
MA	0.74	0.73	0.63	

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

Table 14.21 Correlations between ELA/L and Mathematics for Grade 4

	ELA/L	RD	WR	MA
ELA/L		96,567	96,567	96,567
RD	0.96		96,567	96,567
WR	0.83	0.68		96,567
MA	0.74	0.73	0.62	

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

Table 14.22 Correlations between ELA/L and Mathematics for Grade 5

	ELA/L	RD	WR	MA
ELA/L		97,031	97,031	97,031
RD	0.95		97,031	97,031
WR	0.84	0.68		97,031
MA	0.72	0.72	0.60	

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

Table 14.23 Correlations between ELA/L and Mathematics for Grade 6

	ELA/L	RD	WR	MA
ELA/L		95,509	95,509	95,509
RD	0.96		95,509	95,509
WR	0.82	0.69		95,509
MA	0.75	0.74	0.63	

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

Table 14.24 Correlations between ELA/L and Mathematics for Grade 7

	ELA/L	RD	WR	MA
ELA/L		90,033	90,033	90,033
RD	0.95		90,033	90,033
WR	0.87	0.71		90,033
MA	0.75	0.75	0.62	

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

Table 14.25 Correlations between ELA/L and Mathematics for Grade 8

	ELA/L	RD	WR	MA
ELA/L		91,432	91,432	91,432
RD	0.95		91,432	91,432
WR	0.86	0.71		91,432
MA	0.73	0.72	0.61	

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

Table 14.26 Correlations between ELA/L and Mathematics for High School

	Mathematics Courses		
ELA/L	A1	GO	A2
8	0.39		
	1,132		
9		0.55	0.50
		1,648	871
10			0.47
			194
11	0.39		
	1,132		

Note. ELA/L = English language arts/literacy, A1 = Algebra I, GO = Geometry, A2 = Algebra II.

Table 14.27 Correlations between ELA/L Reading and Mathematics for High School

	Mathematics		
RD	A1	GO	A2
8	0.43		
	1,132		
9		0.52	0.51
		1,648	871
10			0.45
			194
11	0.43		
	1,132		

Note. ELA/L = English language arts/literacy, RD = Reading, A1 = Algebra I, GO = Geometry, A2 = Algebra II.

Table 14.28 Correlations between ELA/L Writing and Mathematics for High School

	Mathematics		
WR	A1	GO	A2
8	0.25	0.52	
	1,132	55	
9		0.47	0.38
		1,648	871
10			0.40
			194
11	0.25		
	1,132		

Note: WR = Writing, A1 = Algebra I, GO = Geometry, A2 = Algebra II.

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 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 demands, 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% of grade 3 and 4 items, 95.5% of grade 6 items, and 94.6% of grade 7 items were fully aligned. For mathematics, 92.0% of grade 3, 91.1% of grade 4 items, 83.1% of grade 6 items, and 94.0% 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%), and grade 7 had the highest (96.3%). Overall the alignment study suggests that the identified CCSS capture the knowledge and skills required in the items.

In addition to the alignment study, HumRRO also evaluated the CCSSO criteria for content and depth for ELA/L and mathematics grades 3, 4, 6, and 7, 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 (McClarty et al., 2013) was used to establish the performance levels for its assessments. In evidence-based standard setting approach, 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, Trends in International Mathematics and Science Study, Progress in International Reading Literacy Study, National Assessment of Educational Progress, ACT, SAT, the Michigan Merit Exam, and the Virginia End-of-Course exams. For each external assessment, the best-matched performance level was determined and the percentage of students reaching that level across the nation and in the participating states and agencies was determined. Across all grades and subjects, the data indicated approximately 25% to 50% 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% to 86.5%. 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% to 79.7%. 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 few 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., 2018a) 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 was 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 (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 14.29 and 14.30. 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 14.29 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

Note. ELA/L = English language arts/literacy.

Table 14.30 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 14.31 for ELA/L and Table 14.32 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 14.31 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

Note. ELA/L = English language arts/literacy.

Table 14.32 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 through 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 .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.

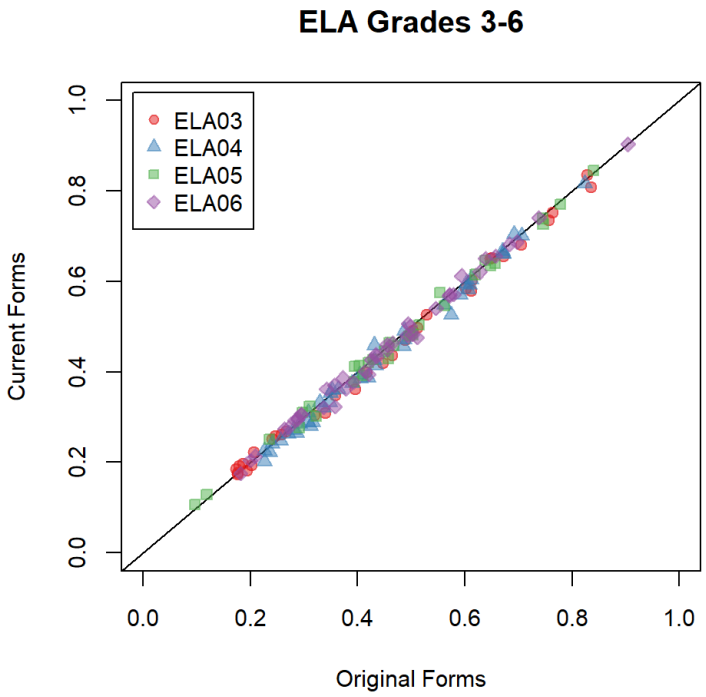

Figure 14.4 ELA/L Grades 3-6 P-Values

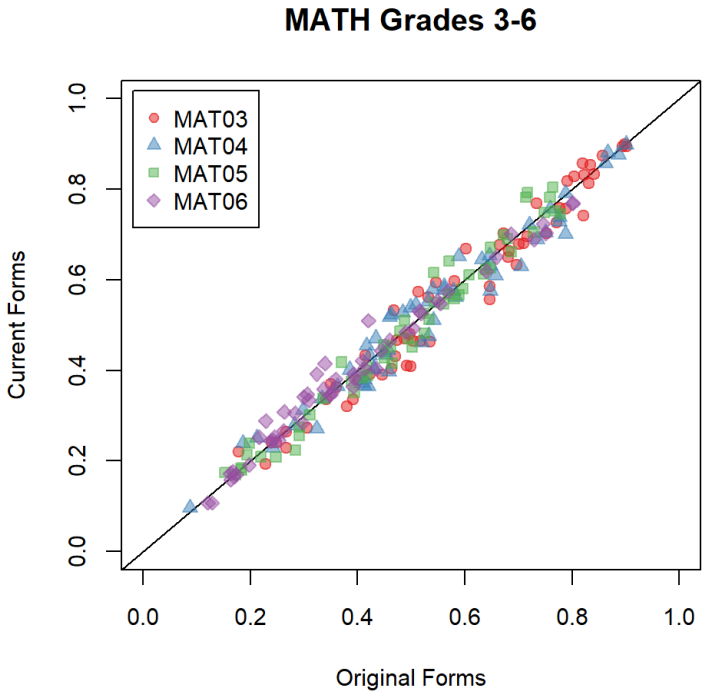

Figure 14.5 Mathematics Grades 3-6 P-Values

The scatterplots for all grades and courses are presented in Figures A.14.1 through 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 through 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 zero, 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 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 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 through A.14.16, while results for the claim and subclaim levels are available in and A.14.42 through 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 through 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 to 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 fewer 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 through 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. Finally, a follow-up study was conducted on grade 10 without using a prior score in the matching process, due to the potential aforementioned challenges. The results showed stronger similarity between the original and current forms than what is presented in this report.

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 CCSS and the National Council of Teachers of Mathematics 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 (Minchen et al., 2018b; Steedle & LaSalle, 2016) involved evaluating the effect of the tool in the context of the operational assessments. The third study was conducted in grade 3 and the fourth study was conducted in grades 4 and 5. To evaluate the drawing tool in context, a set of items was studied by field testing 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. [14]

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. 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.

[14] 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, many pieces of evidence that demonstrate and support the validity of this assessment program were included. Evidence has grown throughout the duration of the program, as additional studies have been conducted and added to this section. Included here is 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 few 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 DIF analyses.
- Section 11 presents information regarding student characteristics for the spring administration of the ELA/L and mathematics administration.
- Section 12 provides detailed information concerning the scores that were reported and the cut scores for ELA/L and mathematics.
- Section 13 provides information on the test reliability (total test score and for subclaims) and includes information on the interrater reliability/agreement.

Section 15: Student Growth Measures

Student growth percentiles (SGPs) are normative measures of annual progress. Normative measures are useful in answering questions like "How does my academic progress compare with the academic progress of my peers?" In contrast to criterion-referenced measures of growth, which describe academic growth toward a particular goal, norm-referenced measures of growth describe students' growth relative to that of students who performed similarly in the past (Betebenner, 2009).

SGPs measure individual student progress by tracking student scores from one year to the next. SGPs compare a student's performance to that of his or her academic peers both within the state and across the consortium. Academic peers are defined as students in the norm group who took the same assessment as the student in prior years and achieved a similar score.

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

The SGP describes a student's location in the distribution of current test scores for all students who performed similarly in the past. SGPs indicate the percentage of academic peers above whom the student scored. With a range of 1 to 99, higher numbers represent higher growth and lower numbers represent lower growth. For example, a SGP of 60 on grade 7 English language arts/literacy (ELA/L) means that the student scored better than 60% of the students in the state or consortium who took grade 7 ELA/L in spring 2019 *and* who had achieved a similar score as this student on the grade 6 ELA/L assessment in spring 2018 and the grade 5 ELA/L assessment in spring 2017. [15] A SGP of 50 represents typical (median) student growth for the state or consortium. Because students are only compared with other students who performed similarly in the past, all students, regardless of starting point, can demonstrate high or low growth.

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

15.1 Norm Groups

The norm groups consisted of students with the same prior scores based on grade or content area progressions (academic peers). SGPs were based on up to two years of prior test scores from spring 2018 and spring 2019 administrations. States administering traditional mathematics assessments in fall 2018 or fall 2019 may also have SGPs based on these prior scores. Tables 15.1 through 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,

[15] 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.

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

Note. ELA/L = English language arts/literacy, *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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 15.9 State-specific SGP Progressions

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

Note. SGP = student growth percentiles, ELA/L = English language arts/literacy.

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 $t \geq 2$), 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.

15.3 Student Growth Percentile Results/Model Fit for Total Group

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

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

In general, SGPs can be divided into three categories: below 30 indicating that a student is not meeting a year's worth of growth, an SGP of 30 to 70 indicating that a student did achieve a year's worth of growth, and

an SGP over 70 indicating that the student surpassed a year's worth of growth. It is important to note that definitions such as these are not inherent to the SGP method, but rather require expert judgment (Betenbenner, 2009). The observed standard errors, ranging from 12.99 to 16.10, support these interpretations (Betenbenner et al., 2016).

Table 15.10 Summary of ELA/L SGP Estimates for Total Group

Grade	Sample Size	Average SGP	Average Standard Error	Median SGP
5	90,323	49.94	13.27	50
6	89,888	50.03	13.87	50
7	88,706	50.01	13.88	50
8	91,137	50.33	13.91	50
10	1,597	49.76	14.64	50
11	90,323	49.94	13.27	50

Note. ELA/L = English language arts/literacy; SGP = student growth percentile.

Table 15.11 Summary of Mathematics SGP Estimates for Total Group

Grade	Sample Size	Average SGP	Average Standard Error	Median SGP
4	--	--	--	--
5	90,507	50.10	12.99	50
6	90,218	50.20	14.95	50
7	87,165	50.03	15.31	50
8	88,595	50.01	15.88	50
A1	--	--	--	--
GO	--	--	--	--
A2	1,337	49.56	16.10	50

Note. "--" indicates insufficient sample for SGP calculation for these tests. ELA/L = English language arts/literacy; SGP = student growth percentile; A1 = Algebra I; GO = Geometry; A2 = Algebra II.

15.4 Student Growth Percentile Results for Subgroups of Interest

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

15.4.1 SGP Results for Gender

English Language Arts/Literacy

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

Mathematics

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

15.4.2 SGP Results for Ethnicity

English Language Arts/Literacy

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

Mathematics

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

15.4.3 SGP Results for Special Instructional Needs

English Language Arts/Literacy

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

Mathematics

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

Table 15.12 Summary of SGP Estimates for Subgroups: Grade 5 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,491	47.37	13.53	46
Female	43,832	52.67	12.99	54
Ethnicity				
White	50,369	53.25	13.04	54
African American	11,404	38.24	13.70	34
Asian/Pacific Islander	4,571	60.74	12.42	65
American Indian/Alaska Native	151	45.05	13.72	44
Hispanic	20,082	45.85	13.78	44
Multiple	3,683	50.50	13.28	51
Special instruction needs				
Economically disadvantaged	40,139	43.63	13.76	41
Not-economically disadvantaged	50,184	54.99	12.87	57
English learner	10,139	43.31	14.67	40
Non-English learner	80,184	50.78	13.09	51
Students with disabilities	15,804	43.17	14.64	40
Students without disabilities	74,519	51.38	12.98	52

Note. SGP = student growth percentile.

15.4.4 SGP Results for Students Taking Spanish Forms

Mathematics

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

Table 15.13 Summary of SGP Estimates for Subgroups: Grade 5 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,555	50.23	12.95	50
Female	43,910	49.95	13.04	50
Ethnicity				
White	51,038	53.86	12.30	55
African American	11,317	37.32	15.09	33
Asian/Pacific Islander	4,764	61.15	11.12	66
American Indian/Alaska Native	137	49.40	13.71	46
Hispanic	19,147	44.45	14.07	42
Multiple	3,990	52.18	12.82	53
Special instruction needs				
Economically disadvantaged	38,499	42.39	14.49	39
Not-economically disadvantaged	52,008	55.81	11.88	58
English learner	9,259	44.62	15.61	42
Non-English learner	81,248	50.72	12.69	51
Students with disabilities	15,811	47.95	15.12	47
Students without disabilities	74,696	50.55	12.54	51
Spanish language form	1,206	37.51	15.48	32

Note. SGP = student growth percentile.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* 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.http://images.pearsonassessments.com/images/tmrs/bulletin21_evidence_based_standard_setting.pdf

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

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

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

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

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

Brandt, R., Bercovitz, E., & Zimmerman, L. (2015b). *Drawing response interaction usability study for PARCC, November 16–19, 2015.* Pearson.

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

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

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

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

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

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

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

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

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

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

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

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* 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). Lawrence 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* (Version 1.0) [Computer software]. University of Iowa..

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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* [Computer software]. 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*(1–2), 149–163.

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

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

Schultz, S. R., 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). 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). Human Resources Research Organization.

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

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

Steedle, J., Quesen, S., & Boyd, A. (2017). *Longitudinal study of external validity of the PARCC performance levels: Phase I report*. Pearson.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

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

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

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

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

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 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*(2), 125–145.

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

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

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

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 versus Female	51					51	100				
White versus Black	51			1	2	50	98				
White versus Hispanic	51					51	100				
White versus Asian	51					51	100				
White versus AmerIndian	51					51	100				
White versus Pacific Islander	51			2	4	49	96				
White versus Multiracial	51					50	98	1	2		
NoEcnDis versus EcnDis	51					51	100				
ELN versus ELY	51			4	8	47	92				
SWDN versus SWDY	51			1	2	50	98				

Note. ELA/L = English language arts/literacy, 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 versus Female	69			7	10	60	87	2	3		
White versus Black	69			4	6	65	94				
White versus Hispanic	69	1	1	3	4	65	94				
White versus Asian	69			1	1	66	96	2	3		
White versus AmerIndian	69			2	3	67	97				
White versus Pacific Islander	69					69	100				
White versus Multiracial	69			1	1	68	99				
NoEcnDis versus EcnDis	69	1	1	3	4	65	94				
ELN versus ELY	69	2	3	8	12	59	86				
SWDN versus SWDY	69			3	4	66	96				

Note. ELA/L = English language arts/literacy, 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 versus Female	67			3	4	60	90	4	6		
White versus Black	67			3	4	64	96				
White versus Hispanic	67	1	1	4	6	62	93				
White versus Asian	67					66	99	1	1		
White versus AmerIndian	67	3	4	1	1	63	94				
White versus Pacific Islander	67			2	3	65	97				
White versus Multiracial	67					67	100				
NoEcnDis versus EcnDis	67					67	100				
ELN versus ELY	67	2	3	7	10	58	87				
SWDN versus SWDY	67	1	1	1	1	65	97				

Note. ELA/L = English language arts/literacy, 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 versus Female	70	3	4	3	4	64	91				
White versus Black	70	1	1	2	3	67	96				
White versus Hispanic	70	1	1	4	6	65	93				
White versus Asian	70			1	1	67	96	1	1	1	1
White versus AmerIndian	70	2	3	6	9	60	86	2	3		
White versus Pacific Islander	70			1	1	69	99				
White versus Multiracial	70					70	100				
NoEcnDis versus EcnDis	70					70	100				
ELN versus ELY	70	2	3	6	9	62	89				
SWDN versus SWDY	70			2	3	68	97				

Note. ELA/L = English language arts/literacy, 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 versus Female	68			7	10	61	90				
White versus Black	68			1	1	67	99				
White versus Hispanic	68			3	4	65	96				
White versus Asian	68					67	99			1	1
White versus AmerIndian	68			3	4	65	96				
White versus Pacific Islander	68			1	1	67	99				
White versus Multiracial	68					68	100				
NoEcnDis versus EcnDis	68					68	100				
ELN versus ELY	68	4	6	7	10	57	84				
SWDN versus SWDY	68					68	100				

Note. ELA/L = English language arts/literacy, 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 versus Female	67	2	3	5	7	58	87	2	3		
White versus Black	67	1	1	4	6	62	93				
White versus Hispanic	67			3	4	64	96				
White versus Asian	67					65	97	1	1	1	1
White versus AmerIndian	67	1	1	2	3	63	94	1	1		
White versus Pacific Islander	67			1	1	66	99				
White versus Multiracial	67					67	100				
NoEcnDis versus EcnDis	67			2	3	65	97				
ELN versus ELY	67	5	7	6	9	56	84				
SWDN versus SWDY	67			2	3	65	97				

Note. ELA/L = English language arts/literacy, 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 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 versus Female	52	1	2	2	4	48	92	1	2		
White versus Black	52			1	2	51	98				
White versus Hispanic	52	1	2	1	2	50	96				
White versus Asian	52					51	98	1	2		
White versus AmerIndian	52	1	2			51	98				
White versus Pacific Islander	52					52	100				
White versus Multiracial	52			1	2	51	98				
NoEcnDis versus EcnDis	52			1	2	51	98				
ELN versus ELY	52	3	6	4	8	44	85	1	2		
SWDN versus SWDY	52					52	100				

Note. ELA/L = English language arts/literacy, 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 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 versus Female	62			5	8	56	90	1	2		
White versus Black	62	1	2	3	5	57	92	1	2		
White versus Hispanic	62			2	3	59	95	1	2		
White versus Asian	62			2	3	59	95	1	2		
White versus AmerIndian	62	5	8	6	10	51	82				
White versus Pacific Islander	62			1	2	61	98				
White versus Multiracial	62			1	2	61	98				
NoEcnDis versus EcnDis	62			1	2	61	98				
ELN versus ELY	62	2	3	2	3	58	94				
SWDN versus SWDY	62			1	2	61	98				

Note. ELA/L = English language arts/literacy, 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 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 versus Female	89			1	1	87	98	1	1		
White versus Black	89	1	1	6	7	79	89	3	3		
White versus Hispanic	89			1	1	88	99				
White versus Asian	89					81	91	7	8	1	1
White versus AmerIndian	89			1	1	88	99				
White versus Pacific Islander	89			2	2	86	97	1	1		
White versus Multiracial	89					88	99	1	1		
NoEcnDis versus EcnDis	89					89	100				
ELN versus ELY	89					89	100				
SWDN versus SWDY	89			2	2	87	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.6.10 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 versus Female	84	2	2	2	2	79	94	1	1		
White versus Black	84			3	4	80	95	1	1		
White versus Hispanic	84					83	99	1	1		
White versus Asian	84					82	98	2	2		
White versus AmerIndian	84	1	1	2	2	79	94	2	2		
White versus Pacific Islander	84			1	1	82	98	1	1		
White versus Multiracial	84					84	100				
NoEcnDis versus EcnDis	84					84	100				
ELN versus ELY	84			3	4	81	96				
SWDN versus SWDY	84			1	1	83	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.6.11 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 versus Female	82			3	4	79	96				
White versus Black	82			2	2	80	98				
White versus Hispanic	82					82	100				
White versus Asian	82					80	98	2	2		
White versus AmerIndian	82			5	6	76	93			1	1
White versus Pacific Islander	82			1	1	81	99				
White versus Multiracial	82					82	100				
NoEcnDis versus EcnDis	82					82	100				
ELN versus ELY	82	1	1	3	4	78	95				
SWDN versus SWDY	82			1	1	79	96	1	1	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.6.12 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 versus Female	74			2	3	72	97				
White versus Black	74			2	3	71	96	1	1		
White versus Hispanic	74					74	100				
White versus Asian	74			1	1	65	88	7	9	1	1
White versus AmerIndian	74			3	4	68	92	3	4		
White versus Pacific Islander	74					74	100				
White versus Multiracial	74					74	100				
NoEcnDis versus EcnDis	74					74	100				
ELN versus ELY	74	1	1	2	3	71	96				
SWDN versus SWDY	74			2	3	71	96	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.6.13 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 versus Female	82			3	4	76	93	3	4		
White versus Black	82			2	2	80	98				
White versus Hispanic	82			1	1	81	99				
White versus Asian	82			1	1	77	94	4	5		
White versus AmerIndian	82			2	2	80	98				
White versus Pacific Islander	82			1	1	81	99				
White versus Multiracial	82					82	100				
NoEcnDis versus EcnDis	82					82	100				
ELN versus ELY	82			3	4	78	95	1	1		
SWDN versus SWDY	82					82	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.6.14 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 versus Female	76			1	1	75	99				
White versus Black	76			3	4	73	96				
White versus Hispanic	76					76	100				
White versus Asian	76					66	87	7	9	3	4
White versus AmerIndian	76	1	1	5	7	70	92				
White versus Pacific Islander	76					76	100				
White versus Multiracial	76			3	4	73	96				
NoEcnDis versus EcnDis	76			1	1	75	99				
ELN versus ELY	76	1	1	5	7	69	91	1	1		
SWDN versus SWDY	76			4	5	71	93			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.6.15 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 versus Female	89			3	3	86	97				
White versus Black	89			2	2	85	96	2	2		
White versus Hispanic	89					89	100				
White versus Asian	89					73	82	15	17	1	1
White versus AmerIndian	89			5	6	84	94				
White versus Pacific Islander	89			1	1	88	99				
White versus Multiracial	89					88	99	1	1		
NoEcnDis versus EcnDis	89					89	100				
ELN versus ELY	89	2	2	6	7	79	89	2	2		
SWDN versus SWDY	89					88	99	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.6.16 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 versus Female	97			2	2	95	98				
White versus Black	97	1	1	3	3	93	96				
White versus Hispanic	97			4	4	93	96				
White versus Asian	97					90	93	6	6	1	1
White versus AmerIndian	97	1	1	6	6	89	92	1	1		
White versus Pacific Islander	97			1	1	96	99				
White versus Multiracial	97					97	100				
NoEcnDis versus EcnDis	97			1	1	96	99				
NoEcnDis versus EcnDis	97	3	3	4	4	83	86	7	7		
SWDN versus SWDY	97	2	2	2	2	93	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.6.17 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 versus Female	101			3	3	95	94	3	3		
White versus Black	101			4	4	97	96				
White versus Hispanic	101			2	2	98	97	1	1		
White versus Asian	101			1	1	94	93	6	6		
White versus AmerIndian	101			4	4	96	95	1	1		
White versus Pacific Islander	101					101	100				
White versus Multiracial	101			1	1	100	99				
NoEcnDis versus EcnDis	101			1	1	100	99				
ELN versus ELY	101	4	4	4	4	89	88	4	4		
SWDN versus SWDY	101			5	5	96	95				

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 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 versus Female	3					3	100				
White versus Black	3					3	100				
White versus Hispanic	3					3	100				
White versus Asian	3					3	100				
White versus AmerIndian	3					3	100				
White versus Pacific Islander	3					3	100				
White versus Multiracial	3					3	100				
NoEcnDis versus EcnDis	3					3	100				
NoEcnDis versus EcnDis	3					3	100				
SWDN versus SWDY	3					3	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 Integrated Mathematics 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 versus Female	5					5	100				
White versus Black	5					5	100				
White versus Hispanic	5					5	100				
White versus Asian	5					5	100				
White versus AmerIndian	5					5	100				
White versus Pacific Islander	5					5	100				
White versus Multiracial	5					5	100				
NoEcnDis versus EcnDis	5					5	100				
NoEcnDis versus EcnDis	5					5	100				
SWDN versus SWDY	5					5	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.

Appendix 7.1: Pre-Equated IRT Results for Spring 2021 English Language Arts/Literacy (ELA/L)

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

				***b* Estimates Summary**				***a* Estimates Summary**			
Grade	**Item Grouping**	**No. of Score Points**	**No. of Items**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
E03	All Items	136	60	0.54	1.05	-1.64	3.35	0.57	0.23	0.12	1.04
	Reading	88	44	0.17	0.99	-1.64	3.35	0.47	0.17	0.12	0.84
	Writing	48	16	1.54	0.29	0.99	2.14	0.84	0.12	0.59	1.04
E04	All Items	169	76	0.45	0.95	-1.41	2.95	0.46	0.21	0.13	0.99
	Reading	128	64	0.31	0.95	-1.41	2.95	0.39	0.14	0.13	0.82
	Writing	41	12	1.20	0.43	0.67	1.84	0.85	0.08	0.69	0.99
E05	All Items	168	74	0.52	1.00	-1.70	3.59	0.48	0.25	0.10	0.99
	Reading	120	60	0.35	1.00	-1.70	3.59	0.38	0.16	0.10	0.77
	Writing	48	14	1.24	0.55	0.54	2.12	0.89	0.08	0.76	0.99
E06	All Items	177	78	0.34	0.74	-1.09	1.89	0.51	0.23	0.18	1.16
	Reading	128	64	0.14	0.64	-1.09	1.70	0.42	0.14	0.18	0.79
	Writing	49	14	1.26	0.41	0.67	1.89	0.88	0.16	0.61	1.16
E07	All Items	171	75	0.28	0.79	-1.70	1.60	0.50	0.28	0.13	1.23
	Reading	122	61	0.15	0.80	-1.70	1.60	0.38	0.14	0.13	0.78
	Writing	49	14	0.82	0.40	0.29	1.54	1.00	0.15	0.67	1.23
E08	All Items	176	76	0.24	0.85	-1.39	2.83	0.52	0.29	0.18	1.24
	Reading	120	60	0.13	0.90	-1.39	2.83	0.39	0.15	0.18	0.81
	Writing	56	16	0.66	0.48	-0.18	1.55	0.98	0.18	0.64	1.24
E10	All Items	177	78	0.69	0.84	-0.77	4.03	0.49	0.28	0.13	1.19
	Reading	128	64	0.65	0.92	-0.77	4.03	0.38	0.14	0.13	0.73
	Writing	49	14	0.89	0.31	0.41	1.35	1.01	0.13	0.77	1.19
E11	All Items	181	80	1.04	0.87	-1.09	4.21	0.45	0.25	0.10	1.10
	Reading	132	66	1.02	0.95	-1.09	4.21	0.36	0.15	0.10	0.84
	Writing	49	14	1.11	0.29	0.61	1.53	0.87	0.17	0.56	1.10

Note. E03 through E08 = English language arts/literacy (ELA/L) grades 3 through 8.

Appendix 7.2: Pre-Equated IRT Results for Spring 2019 Mathematics

Table A.7.2 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	133	91	-0.36	0.99	-2.52	1.62	0.75	0.23	0.22	1.31
	SSMC	32	32	-0.80	1.02	-2.50	1.07	0.68	0.21	0.22	0.95
	CR	101	59	-0.12	0.90	-2.52	1.62	0.78	0.23	0.39	1.31
	Type I	86	78	-0.56	0.91	-2.52	1.22	0.78	0.23	0.22	1.31
	Type II	20	6	0.93	0.56	0.24	1.62	0.48	0.07	0.39	0.58
	Type III	27	7	0.77	0.35	0.34	1.28	0.67	0.15	0.50	0.85
M04	All Items	136	87	-0.08	0.99	-2.69	1.86	0.72	0.21	0.32	1.38
	SSMC	23	23	-0.82	1.09	-2.69	1.86	0.66	0.21	0.34	1.09
	CR	113	64	0.19	0.81	-2.08	1.66	0.75	0.21	0.32	1.38
	Type I	86	73	-0.27	0.96	-2.69	1.86	0.74	0.21	0.34	1.38
	Type II	23	7	0.91	0.50	-0.17	1.40	0.67	0.16	0.40	0.92
	Type III	27	7	0.93	0.12	0.80	1.09	0.61	0.20	0.32	0.92
M05	All Items	144	85	0.08	0.91	-2.06	2.45	0.67	0.23	0.17	1.50
	SSMC	26	26	-0.45	0.68	-2.06	1.16	0.68	0.30	0.18	1.50
	CR	118	59	0.32	0.90	-2.03	2.45	0.67	0.21	0.17	1.18
	Type I	78	67	-0.11	0.90	-2.06	2.45	0.70	0.24	0.18	1.50
	Type II	30	9	0.73	0.55	-0.16	1.62	0.51	0.19	0.17	0.73
	Type III	36	9	0.89	0.42	0.02	1.38	0.58	0.15	0.44	0.91
M06	All Items	138	74	0.31	1.00	-2.38	2.06	0.75	0.25	0.33	1.44
	SSMC	19	19	-0.55	0.82	-1.98	1.35	0.65	0.23	0.36	1.44
	CR	119	55	0.61	0.87	-2.38	2.06	0.78	0.25	0.33	1.33
	Type I	78	58	0.14	1.02	-2.38	2.06	0.79	0.26	0.33	1.44
	Type II	27	8	0.69	0.60	-0.23	1.55	0.63	0.11	0.52	0.83
	Type III	33	8	1.15	0.52	0.49	1.73	0.57	0.14	0.37	0.79
M07	All Items	142	86	0.55	1.05	-1.78	2.78	0.65	0.26	0.19	1.22
	SSMC	37	37	0.01	1.09	-1.74	2.38	0.55	0.24	0.19	1.17
	CR	105	49	0.95	0.83	-1.78	2.78	0.72	0.25	0.31	1.22
	Type I	83	70	0.40	1.09	-1.78	2.66	0.67	0.28	0.19	1.22
	Type II	26	8	1.22	0.66	0.75	2.78	0.55	0.15	0.32	0.80
	Type III	33	8	1.16	0.27	0.70	1.64	0.57	0.09	0.38	0.69

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	135	76	0.97	1.22	-1.52	3.18	0.63	0.24	0.21	1.34
	SSMC	26	26	0.07	1.10	-1.52	3.18	0.46	0.19	0.21	0.83
	CR	109	50	1.44	0.99	-1.40	2.91	0.71	0.23	0.29	1.34
	Type I	78	61	0.76	1.24	-1.52	3.18	0.62	0.27	0.21	1.34
	Type II	24	7	1.68	0.57	1.00	2.66	0.64	0.17	0.45	0.91
	Type III	33	8	1.94	0.59	0.70	2.54	0.61	0.11	0.45	0.82
A1	All Items	228	118	1.32	1.10	-0.99	3.62	0.62	0.26	0.16	1.34
	SSMC	30	30	0.87	1.28	-0.99	3.62	0.47	0.21	0.16	0.85
	CR	146	62	1.50	1.04	-0.96	3.61	0.70	0.25	0.28	1.34
	Type I	98	73	1.10	1.20	-0.99	3.62	0.62	0.28	0.16	1.34
	Type II	33	10	2.13	0.59	1.55	3.61	0.69	0.20	0.29	0.91
	Type III	45	9	1.96	0.32	1.50	2.60	0.58	0.13	0.41	0.76
G1	All Items	236	129	1.01	1.17	-1.60	3.83	0.78	0.33	0.18	1.78
	SSMC	33	33	-0.08	1.29	-1.60	3.83	0.65	0.29	0.26	1.41
	CR	152	69	1.49	0.80	-0.66	3.50	0.83	0.32	0.19	1.68
	Type I	103	83	0.81	1.28	-1.60	3.83	0.77	0.34	0.19	1.68
	Type II	31	9	1.81	0.56	0.96	2.79	0.83	0.13	0.66	1.04
	Type III	51	10	1.66	0.40	1.05	2.14	0.76	0.28	0.36	1.18
A2	All Items	229	128	1.28	1.03	-1.53	3.67	0.63	0.28	0.16	1.28
	SSMC	31	31	0.77	1.23	-1.53	3.09	0.46	0.15	0.19	0.89
	CR	152	70	1.50	0.85	-0.63	3.67	0.71	0.29	0.31	1.28
	Type I	106	82	1.12	1.05	-1.53	3.67	0.64	0.29	0.19	1.28
	Type II	32	10	1.81	0.68	0.50	2.73	0.65	0.19	0.40	0.96
	Type III	45	9	2.05	0.52	1.06	2.79	0.62	0.24	0.38	0.99
M1	All Items	62	34	1.10	1.09	-0.95	4.02	0.61	0.31	0.11	1.61
	SSMC	13	13	0.90	1.16	-0.06	4.02	0.45	0.20	0.11	0.77
	CR	49	21	1.23	1.05	-0.95	2.85	0.71	0.33	0.18	1.61
	Type I	37	28	0.86	1.02	-0.95	4.02	0.60	0.34	0.11	1.61
	Type II	10	3	1.80	0.48	1.31	2.26	0.60	0.04	0.57	0.65
	Type III	15	3	2.66	0.20	2.46	2.85	0.75	0.10	0.68	0.87

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	55	30	1.46	1.23	-0.97	3.96	0.56	0.27	0.06	1.41
	SSMC	13	13	1.06	1.20	-0.97	3.75	0.42	0.18	0.06	0.71
	CR	42	17	1.77	1.20	-0.50	3.96	0.67	0.28	0.31	1.41
	Type I	30	24	1.27	1.10	-0.97	3.75	0.57	0.29	0.06	1.41
	Type II	10	3	2.78	1.03	2.04	3.96	0.53	0.14	0.41	0.68
	Type III	15	3	1.70	1.91	-0.50	2.98	0.50	0.17	0.31	0.64

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.

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

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

		English Language Arts-Literacy								
State	**Category**	**Total**	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Grade 10**	**Grade 11**
All States	N of Students	590,314	96,928	99,006	99,632	98,590	96,950	96,028	2,767	413
	N of CBT	587,385	96,041	98,343	99,004	98,352	96,723	95,786	2,765	371
	% of CBT	99.5	99.1	99.3	99.4	99.8	99.8	99.7	99.9	89.8
	N of PBT	2,929	887	663	628	238	227	242	n/r	42
	% of PBT	0.5	0.9	0.7	0.6	0.2	0.2	0.3	n/r	10.2
BIE	% of All Data	0.7	0.1	0.1	0.1	0.1	0.1	0.1	n/a	0.1
	N of Students	5,116	848	723	772	813	786	761	n/a	413
	N of CBT	5,007	821	705	769	804	778	759	n/a	371
	% of CBT	97.9	96.8	97.5	99.6	98.9	99.0	99.7	n/a	89.8
	N of PBT	109	27	n/r	n/r	n/r	n/r	n/r	n/a	42
	% of PBT	2.1	3.2	n/r	n/r	n/r	n/r	n/r	n/a	10.2
DD	% of All Data	5.3	0.9	0.9	0.8	0.8	0.7	0.7	0.5	n/a
	N of Students	30,645	5,306	5,303	4,776	4,578	4,037	3,878	2,767	n/a
	N of CBT	30,573	5,281	5,289	4,765	4,569	4,027	3,877	2,765	n/a
	% of CBT	99.8	99.5	99.7	99.8	99.8	99.8	100.0	99.9	n/a
	N of PBT	72	25	n/r	n/r	n/r	n/r	n/r	n/r	n/a
	% of PBT	0.2	0.5	n/r	n/r	n/r	n/r	n/r	n/r	n/a

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

		English Language Arts-Literacy								
State	**Category**	**Total**	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Grade 10**	**Grade 11**
IL	% of All Data	94.0	15.4	15.8	15.9	15.8	15.6	15.5	n/a	n/a
	N of Students	554,553	90,774	92,980	94,084	93,199	92,127	91,389	n/a	n/a
	N of CBT	551,805	89,939	92,349	93,470	92,979	91,918	91,150	n/a	n/a
	% of CBT	99.5	99.1	99.3	99.3	99.8	99.8	99.7	n/a	n/a
	N of PBT	2,748	835	631	614	220	209	239	n/a	n/a
	% of PBT	0.5	0.9	0.7	0.7	0.2	0.2	0.3	n/a	n/a

Note. BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity; CBT=computer-based test; PBT=paper-based test; n/a=not applicable; and n/r=not reported due to n<20 or missing demographic information.

Table A.11.2 All Mathematics Test Takers, 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**
All States	N of Students	582,332	96,011	97,740	98,306	96,924	91,315	92,946	3,424	2,922	2,726	17	1
	N of CBT	579,465	95,149	97,088	97,690	96,684	91,100	92,711	3,380	2,920	2,725	n/r	n/r
	% of CBT	99.5	99	99	99	100	100	100	99	100	100	n/r	n/r
	N of PBT	2,867	862	652	616	240	215	235	44	n/r	n/r	n/r	n/r
	% of PBT	0.5	1	1	1	0	0	0	1	n/r	n/r	n/r	n/r
BIE	% of All Data	0.6	0	0	0	0	0	0	0	0	0	0	0
	N of Students	4,896	807	699	727	748	776	742	103	56	220	17	1
	N of CBT	4,788	782	681	723	739	769	740	61	56	219	n/r	n/r
	% of CBT	97.8	97	97	99	99	99	100	59	100	100	n/r	n/r
	N of PBT	108	25	n/r	n/r	n/r	n/r	n/r	42	n/r	n/r	n/r	n/r
	% of PBT	2.2	3	n/r	n/r	n/r	n/r	n/r	41	n/r	n/r	n/r	n/r
DD	% of All Data	5.4	1	1	1	1	n/a	1	1	1	0	n/a	n/a
	N of Students	31,121	5,305	5,285	4,748	4,469	n/a	2,621	3,321	2,866	2,506	n/a	n/a
	N of CBT	31,050	5,279	5,270	4,736	4,459	n/a	2,617	3,319	2,864	2,506	n/a	n/a
	% of CBT	99.8	100	100	100	100	n/a	100	100	100	100	n/a	n/a
	N of PBT	71	26	n/r	n/r	n/r	n/a	n/r	n/r	n/r	n/r	n/a	n/a
	% of PBT	0.2	1	n/r	n/r	n/r	n/a	n/r	n/r	n/r	n/r	n/a	n/a

Table A.11.2 All Mathematics Test Takers, 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**
IL	% of All Data	93.7	15	16	16	16	16	15	n/a	n/a	n/a	n/a	n/a
	N of Students	546,315	89,899	91,756	92,831	91,707	90,539	89,583	n/a	n/a	n/a	n/a	n/a
	N of CBT	543,627	89,088	91,137	92,231	91,486	90,331	89,354	n/a	n/a	n/a	n/a	n/a
	% of CBT	99.5	99	99	99	100	100	100	n/a	n/a	n/a	n/a	n/a
	N of PBT	2,688	811	619	600	221	208	229	n/a	n/a	n/a	n/a	n/a
	% of PBT	0.5	1	1	1	0	0	0	n/a	n/a	n/a	n/a	n/a

Note. BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity;
A1=Algebra I, GO=Geometry, A2 = Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II.
CBT=computer-based test; PBT=paper-based test;
n/a=not applicable; and n/r=not reported due to n<20 or missing demographic information.

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

		Mathematics						
State	**Category**	**Total**	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
All States	N of Students	6,156	1,729	1,504	1,305	961	367	290
	N of CBT	6,133	1,726	1,496	1,303	957	364	287
	% of CBT	99.6	100	100	100	100	99	99
	N of PBT	23	n/r	n/r	n/r	n/r	n/r	n/r
	% of PBT	0.4	n/r	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 of Students	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
	% of CBT	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
	% of PBT	n/a	n/a	n/a	n/a	n/a	n/a	n/a
DD	% of All Data	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 of CBT	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 of PBT	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

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

		Mathematics						
State	**Category**	**Total**	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
IL	% of All Data	100.0	28	24	21	16	6	5
	N of Students	6,156	1,729	1,504	1,305	961	367	290
	N of CBT	6,133	1,726	1,496	1,303	957	364	287
	% of CBT	99.6	100	100	100	100	99	99
	N of PBT	23	n/r	n/r	n/r	n/r	n/r	n/r
	% of PBT	0.4	n/r	n/r	n/r	n/r	n/r	n/r

Note. BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity;
CBT=computer-based test; PBT = paper-based test;
n/a=not applicable; and n/r=not reported due to n<20 or missing demographic information.
* No students in BIE tested in mathematics using Spanish-language forms.

Table A.11.4 All States Combined: ELA/L Test Takers, by Grade, Mode, and Gender

Grade	Mode	Valid Cases	Female		Male	
			N	%	N	%
3	**All**	96,844	47,502	49.1	49,342	50.9
	CBT	95,958	47,122	49.1	48,836	50.9
	PBT	886	380	42.9	506	57.1
4	**All**	98,916	48,440	49.0	50,476	51.0
	CBT	98,253	48,140	49.0	50,113	51.0
	PBT	663	300	45.2	363	54.8
5	**All**	99,542	48,401	48.6	51,141	51.4
	CBT	98,914	48,115	48.6	50,799	51.4
	PBT	628	286	45.5	342	54.5
6	**All**	98,498	48,053	48.8	50,445	51.2
	CBT	98,260	47,949	48.8	50,311	51.2
	PBT	238	104	43.7	134	56.3
7	**All**	96,833	47,049	48.6	49,784	51.4
	CBT	96,606	46,955	48.6	49,651	51.4
	PBT	227	94	41.4	133	58.6
8	**All**	95,934	46,187	48.1	49,747	51.9
	CBT	95,692	46,086	48.2	49,606	51.8
	PBT	242	101	41.7	141	58.3
10	**All**	2,723	1,347	49.5	1,376	50.5
	CBT	2,721	1,347	49.5	1,374	50.5
	PBT	n/r	n/r	n/r	n/r	n/r
11	**All**	404	211	52.2	193	47.8
	CBT	362	185	51.1	177	48.9
	PBT	42	26	61.9	n/r	n/r

Note. ELA/L = English language arts/literacy, CBT=computer-based test; PBT=paper-based test.

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

Grade	Mode	Valid Cases	Female		Male	
			N	%	N	%
3	**All**	95,938	47,029	49.0	48,909	51.0
	CBT	95,077	46,654	49.1	48,423	50.9
	PBT	861	375	43.6	486	56.4
4	**All**	97,671	47,855	49.0	49,816	51.0
	CBT	97,019	47,564	49.0	49,455	51.0
	PBT	652	291	44.6	361	55.4
5	**All**	98,238	47,761	48.6	50,477	51.4
	CBT	97,622	47,477	48.6	50,145	51.4
	PBT	616	284	46.1	332	53.9
6	**All**	96,859	47,233	48.8	49,626	51.2
	CBT	96,619	47,127	48.8	49,492	51.2
	PBT	240	106	44.2	134	55.8
7	**All**	91,298	44,279	48.5	47,019	51.5
	CBT	91,083	44,187	48.5	46,896	51.5
	PBT	215	92	42.8	123	57.2
8	**All**	92,894	44,707	48.1	48,187	51.9
	CBT	92,659	44,607	48.1	48,052	51.9
	PBT	235	100	42.6	135	57.4
A1	**All**	3,382	1,626	48.1	1,756	51.9
	CBT	3,338	1,600	47.9	1,738	52.1
	PBT	44	26	59.1	n/r	n/r
GO	**All**	2,898	1,388	47.9	1,510	52.1
	CBT	2,896	1,388	47.9	1,508	52.1
	PBT	n/r	n/r	n/r	n/r	n/r
A2	**All**	2,697	1,354	50.2	1,343	49.8
	CBT	2,696	1,354	50.2	1,342	49.8
	PBT	n/r	n/r	n/r	n/r	n/r
M1	**All**	n/r	n/r	n/r	n/r	n/r
	CBT	n/r	n/r	n/r	n/r	n/r
M2	**All**	n/r	n/r	n/r	n/r	n/r
	CBT	n/r	n/r	n/r	n/r	n/r

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

			Female		Male	
Grade	**Mode**	**Valid Cases**	**N**	**%**	**N**	**%**

Note. A1=Algebra I, GO=Geometry, A2=Algebra II, M1=Integrated Mathematics I, M2=Integrated Mathematics II.
n/a=not applicable. and n/r=not reported due to n<20 or missing demographic information.

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

			Female		Male	
Grade	**Mode**	**Valid Cases**	**N**	**%**	**N**	**%**
	All	1,729	886	51.2	843	48.8
3	CBT	1,726	885	51.3	841	48.7
	PBT	n/r	n/r	n/r	n/r	n/r
	All	1,504	750	49.9	754	50.1
4	CBT	1,496	747	49.9	749	50.1
	PBT	n/r	n/r	n/r	n/r	n/r
	All	1,305	637	48.8	668	51.2
5	CBT	1,303	637	48.9	666	51.1
	PBT	n/r	n/r	n/r	n/r	n/r
	All	961	487	50.7	474	49.3
6	CBT	957	487	50.9	470	49.1
	PBT	n/r	n/r	n/r	n/r	n/r
	All	367	183	49.9	184	50.1
7	CBT	364	181	49.7	183	50.3
	PBT	n/r	n/r	n/r	n/r	n/r
	All	290	136	46.9	154	53.1
8	CBT	287	135	47.0	152	53.0
	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 M athematics II, CBT=computer-based test; PBT=paper-based test, n/a=not applicable. and n/r=not reported due to n<20 or missing demographic information.
* No students in BIE tested in mathematics using Spanish-language forms.

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

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.4	n/r	44.2	n/r
Student with disabilities	16.3	18.6	16.3	15.7
English learner	15.7	3.1	16.0	13.0
Male	50.9	49.1	50.9	50.7
Female	49.0	48.9	49.1	48.0
American Indian/Alaska Native	0.4	15.6	0.2	n/r
Asian	5.3	n/r	5.3	6.0
Black/African American	12.3	n/r	12.5	9.5
Hispanic/Latino	21.8	n/r	22.0	21.8
White/Caucasian	54.0	n/r	55.1	42.5
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.5
Two or more races reported	4.9	n/r	4.4	14.7
Unknown	1.2	84.2	0.2	3.8

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.4	n/r	44.1	n/r
Student with disabilities	17.0	23.4	17.0	16.3
English learner	14.7	n/r	14.9	12.6
Male	51.0	48.1	51.1	49.6
Female	48.9	50.1	48.9	49.0
American Indian/Alaska Native	0.3	15.4	0.2	n/r
Asian	5.2	n/r	5.2	5.0
Black/African American	12.3	n/r	12.5	10.0
Hispanic/Latino	21.8	n/r	21.9	22.8
White/Caucasian	54.4	n/r	55.5	42.0
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.7
Two or more races reported	4.8	n/r	4.3	14.4
Unknown	1.0	84.6	0.2	4.0

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.5	n/r	44.0	n/r
Student with disabilities	17.3	21.8	17.3	15.3
English learner	11.4	4.3	11.5	10.8
Male	51.3	48.1	51.5	49.3
Female	48.6	50.1	48.5	49.1
American Indian/Alaska Native	0.3	17.6	0.2	n/r
Asian	5.2	n/r	5.2	6.1
Black/African American	12.3	n/r	12.6	9.3
Hispanic/Latino	22.0	n/r	22.1	22.4
White/Caucasian	54.5	n/r	55.6	41.9
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.5
Two or more races reported	4.6	n/r	4.1	14.7
Unknown	1.0	82.1	0.2	3.9

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.0	n/r	43.3	n/r
Student with disabilities	16.9	18.6	17.0	15.0
English learner	9.4	n/r	9.4	10.2
Male	51.2	52.3	51.2	50.3
Female	48.7	46.6	48.8	47.9
American Indian/Alaska Native	0.3	16.2	0.2	n/r
Asian	5.1	n/r	5.1	5.3
Black/African American	12.3	n/r	12.5	10.1
Hispanic/Latino	21.9	n/r	22.0	23.5
White/Caucasian	54.6	n/r	55.7	41.0
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.9
Two or more races reported	4.5	n/r	4.1	13.6
Unknown	1.1	83.5	0.2	4.3

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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

Demographic	All States (%)	DC (%)	IL (%)	DD (%)
Economically disadvantaged	40.9	n/r	43.0	n/r
Student with disabilities	16.8	20.9	16.9	13.6
English learner	8.6	4.7	8.7	8.1
Male	51.4	50.6	51.4	49.3
Female	48.5	47.1	48.6	48.2
American Indian/Alaska Native	0.4	16.3	0.2	n/r
Asian	5.0	n/r	5.0	6.4
Black/African American	12.6	n/r	12.9	9.7
Hispanic/Latino	21.4	n/r	21.6	22.6
White/Caucasian	55.1	n/r	56.2	41.3
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.5
Two or more races reported	4.2	n/r	3.8	13.6
Unknown	1.1	83.3	0.2	4.5

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	40.9	n/r	43.0	n/r
Student with disabilities	17.2	20.0	17.3	13.3
English learner	7.4	5.9	7.4	8.5
Male	51.8	49.7	51.9	49.8
Female	48.1	48.5	48.1	48.1
American Indian/Alaska Native	0.3	13.9	0.2	n/r
Asian	4.7	n/r	4.6	6.3
Black/African American	13.0	n/r	13.3	9.4
Hispanic/Latino	21.7	n/r	21.8	22.2
White/Caucasian	55.0	n/r	56.0	41.9
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.9
Two or more races reported	4.1	n/r	3.7	13.5
Unknown	1.1	85.5	0.2	4.6

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.13 Demographic Information for Grade 10 ELA/L, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/a	n/a	n/r
Student with disabilities	13.2	n/a	n/a	13.2
English learner	5.2	n/a	n/a	5.2
Male	49.7	n/a	n/a	49.7
Female	48.7	n/a	n/a	48.7
American Indian/Alaska Native	n/r	n/a	n/a	n/r
Asian	7.6	n/a	n/a	7.6
Black/African American	9.5	n/a	n/a	9.5
Hispanic/Latino	21.9	n/a	n/a	21.9
White/Caucasian	41.2	n/a	n/a	41.2
Native Hawaiian/Pacific Islander	2.0	n/a	n/a	2.0
Two or more races reported	13.4	n/a	n/a	13.4
Unknown	4.1	n/a	n/a	4.1

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.14 Demographic Information for Grade 11 ELA/L, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/a
Student with disabilities	16.5	16.5	n/a	n/a
English learner	7.0	7.0	n/a	n/a
Male	46.7	46.7	n/a	n/a
Female	51.1	51.1	n/a	n/a
American Indian/Alaska Native	14.8	14.8	n/a	n/a
Asian	n/r	n/r	n/a	n/a
Black/African American	n/r	n/r	n/a	n/a
Hispanic/Latino	n/r	n/r	n/a	n/a
White/Caucasian	n/r	n/r	n/a	n/a
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a
Two or more races reported	n/r	n/r	n/a	n/a
Unknown	83.5	83.5	n/a	n/a

Note. ELA/L = English language arts/literacy, All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.15 Demographic Information for Grade 3 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.2	n/r	44.0	n/r
Student with disabilities	16.3	18.7	16.3	15.6
English learner	15.7	n/r	16.0	13.5
Male	50.9	49.8	51.0	50.8
Female	49.0	48.2	49.0	48.1
American Indian/Alaska Native	0.4	15.5	0.2	n/r
Asian	5.3	n/r	5.4	6.0
Black/African American	12.1	n/r	12.4	9.5
Hispanic/Latino	21.8	n/r	22.0	21.9
White/Caucasian	54.2	n/r	55.4	42.6
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.5
Two or more races reported	4.9	n/r	4.4	14.6
Unknown	1.1	84.3	0.2	3.6

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.16 Demographic Information for Grade 4 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.2	n/r	43.8	n/r
Student with disabilities	17.0	23.2	17.0	16.3
English learner	14.7	n/r	14.9	12.9
Male	51.0	48.2	51.1	49.8
Female	49.0	50.1	48.9	49.1
American Indian/Alaska Native	0.3	16.5	0.2	n/r
Asian	5.2	n/r	5.3	5.1
Black/African American	12.1	n/r	12.3	10.0
Hispanic/Latino	21.8	n/r	21.9	22.8
White/Caucasian	54.6	n/r	55.8	42.1
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.7
Two or more races reported	4.8	n/r	4.3	14.4
Unknown	1.0	83.5	0.2	3.7

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.17 Demographic Information for Grade 5 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.3	n/r	43.7	n/r
Student with disabilities	17.2	21.5	17.2	15.2
English learner	11.3	3.4	11.4	11.1
Male	51.3	47.3	51.5	49.5
Female	48.6	50.6	48.5	49.4
American Indian/Alaska Native	0.3	18.4	0.2	n/r
Asian	5.2	n/r	5.2	6.1
Black/African American	12.1	n/r	12.3	9.3
Hispanic/Latino	21.8	n/r	22.0	22.4
White/Caucasian	54.9	n/r	55.9	42.3
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	1.5
Two or more races reported	4.6	n/r	4.1	14.8
Unknown	1.0	81.3	0.2	3.4

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.18 Demographic Information for Grade 6 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	40.8	n/r	43.1	n/r
Student with disabilities	16.9	19.0	16.9	15.1
English learner	9.3	n/r	9.3	10.4
Male	51.2	51.3	51.2	50.7
Female	48.7	47.7	48.8	48.0
American Indian/Alaska Native	0.3	17.8	0.2	n/r
Asian	5.1	n/r	5.1	5.2
Black/African American	12.1	n/r	12.3	10.0
Hispanic/Latino	21.8	n/r	21.9	23.6
White/Caucasian	55.0	n/r	56.1	41.4
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	2.0
Two or more races reported	4.5	n/r	4.1	13.6
Unknown	1.0	82.0	0.2	3.9

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.19 Demographic Information for Grade 7 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	42.4	n/r	42.7	n/a
Student with disabilities	16.9	20.6	16.9	n/a
English learner	8.5	4.6	8.6	n/a
Male	51.5	49.9	51.5	n/a
Female	48.5	47.9	48.5	n/a
American Indian/Alaska Native	0.3	17.0	0.2	n/a
Asian	5.0	n/r	5.0	n/a
Black/African American	12.5	n/r	12.6	n/a
Hispanic/Latino	21.3	n/r	21.4	n/a
White/Caucasian	56.1	n/r	56.6	n/a
Native Hawaiian/Pacific Islander	0.1	n/r	0.1	n/a
Two or more races reported	3.8	n/r	3.8	n/a
Unknown	0.9	82.6	0.2	n/a

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.20 Demographic Information for Grade 8 Mathematics, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	41.2	n/r	42.8	n/r
Student with disabilities	17.3	19.9	17.2	17.4
English learner	7.3	5.9	7.2	11.0
Male	51.8	49.6	51.9	50.6
Female	48.1	48.4	48.1	48.0
American Indian/Alaska Native	0.3	14.4	0.2	n/r
Asian	4.7	n/r	4.7	5.5
Black/African American	12.8	n/r	13.0	11.1
Hispanic/Latino	21.6	n/r	21.7	24.4
White/Caucasian	55.5	n/r	56.4	39.3
Native Hawaiian/Pacific Islander	0.2	n/r	0.1	2.3
Two or more races reported	4.0	n/r	3.7	13.2
Unknown	1.0	85.2	0.2	4.1

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.21 Demographic Information for Algebra I, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/r
Student with disabilities	12.5	23.3	n/a	12.1
English learner	7.1	n/r	n/a	7.1
Male	51.3	42.7	n/a	51.6
Female	47.5	56.3	n/a	47.2
American Indian/Alaska Native	0.7	n/r	n/a	n/r
Asian	7.4	n/r	n/a	7.6
Black/African American	9.4	n/r	n/a	9.7
Hispanic/Latino	19.8	n/r	n/a	20.4
White/Caucasian	40.2	n/r	n/a	41.4
Native Hawaiian/Pacific Islander	1.9	n/r	n/a	2.0
Two or more races reported	14.4	n/r	n/a	14.8
Unknown	6.2	88.3	n/a	3.7

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.22 Demographic Information for Geometry, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/r
Student with disabilities	12.7	n/r	n/a	12.2
English learner	5.7	n/r	n/a	5.7
Male	51.7	55.4	n/a	51.6
Female	47.5	39.3	n/a	47.7
American Indian/Alaska Native	n/r	n/r	n/a	n/r
Asian	7.9	n/r	n/a	8.0
Black/African American	9.3	n/r	n/a	9.5
Hispanic/Latino	20.9	n/r	n/a	21.3
White/Caucasian	41.1	n/r	n/a	41.9
Native Hawaiian/Pacific Islander	1.9	n/r	n/a	1.9
Two or more races reported	13.4	n/r	n/a	13.7
Unknown	5.1	91.1	n/a	3.4

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.23 Demographic Information for Algebra II, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/r
Student with disabilities	10.5	9.5	n/a	10.5
English learner	4.6	n/r	n/a	4.5
Male	49.3	47.3	n/a	49.4
Female	49.7	51.4	n/a	49.5
American Indian/Alaska Native	2.1	21.4	n/a	n/r
Asian	8.3	n/r	n/a	8.9
Black/African American	8.3	n/r	n/a	9.1
Hispanic/Latino	19.3	n/r	n/a	20.8
White/Caucasian	38.8	n/r	n/a	42.2
Native Hawaiian/Pacific Islander	1.7	n/r	n/a	1.8
Two or more races reported	12.6	n/r	n/a	13.6
Unknown	9.0	75.5	n/a	3.2

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.24 Demographic Information for Integrated Mathematics I, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/a
Student with disabilities	n/r	n/r	n/a	n/a
English learner	n/r	n/r	n/a	n/a
Male	n/r	n/r	n/a	n/a
Female	n/r	n/r	n/a	n/a
American Indian/Alaska Native	n/r	n/r	n/a	n/a
Asian	n/r	n/r	n/a	n/a
Black/African American	n/r	n/r	n/a	n/a
Hispanic/Latino	n/r	n/r	n/a	n/a
White/Caucasian	n/r	n/r	n/a	n/a
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a
Two or more races reported	n/r	n/r	n/a	n/a
Unknown	n/r	n/r	n/a	n/a

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

Table A.11.25 Demographic Information for Integrated Mathematics II, Overall and by State

Demographic	All States (%)	BIE (%)	IL (%)	DD (%)
Economically disadvantaged	n/r	n/r	n/a	n/a
Student with disabilities	n/r	n/r	n/a	n/a
English learner	n/r	n/r	n/a	n/a
Male	n/r	n/r	n/a	n/a
Female	n/r	n/r	n/a	n/a
American Indian/Alaska Native	n/r	n/r	n/a	n/a
Asian	n/r	n/r	n/a	n/a
Black/African American	n/r	n/r	n/a	n/a
Hispanic/Latino	n/r	n/r	n/a	n/a
White/Caucasian	n/r	n/r	n/a	n/a
Native Hawaiian/Pacific Islander	n/r	n/r	n/a	n/a
Two or more races reported	n/r	n/r	n/a	n/a
Unknown	n/r	n/r	n/a	n/a

Note. All States=data from all participating states combined; BIE=Bureau of Indian Education, IL=Illinois, and DD=Department of Defense Education Activity. n/a=not applicable; n/r=not reported due to n<20 or missing demographic information.

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. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

Table A.12.7 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

Note. ELA/L = English language arts/literacy.

Table A.12.8 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

Note. ELA/L = English language arts/literacy.

Table A.12.9 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.12.10 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.12.11 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.12.12 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.12.13 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.12.14 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.12.15 Form Composition for Algebra I

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	12	17
	Additional & Supporting Content	8-9	9-11
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
	Integrated (Ψ*)	1-2	2-4
Total		28	55

Table A.12.16 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.12.17 Form Composition for Algebra II

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	13-14	16-18
	Additional & Supporting Content	9	12
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
	Integrated (Ψ*)	0-2	0-2
Total		29	55

Table A.12.18 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.12.19 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

Appendix 12.2: Threshold Scores and Scaling Constants

Table A.12.20 Threshold Scores and Scaling Constants for ELA/L Grades 3 to 8

Assessment	Threshold Cut	Theta	Scale Score	A	B
Grade 3 ELA/L	Level 2 Cut	-0.9648	700	36.7227	735.4297
	Level 3 Cut	-0.2840	725		
	Level 4 Cut	0.3968	750		
	Level 5 Cut	2.0360	810		
Grade 4 ELA/L	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/L	Level 2 Cut	-1.3411	700	29.4580	739.5050
	Level 3 Cut	-0.4924	725		
	Level 4 Cut	0.3563	750		
	Level 5 Cut	2.0224	799		
Grade 6 ELA/L	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/L	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/L	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		

Note. ELA/L = English language arts/literacy.

Table A.12.21 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	725		
	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	725		
	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	725		
	Level 4 Cut	0.6323	750		
	Level 5 Cut	2.1896	801		

Table A.12.22 Threshold Scores and Scaling Constants for High School ELA/L

Assessment	Threshold Cut	Theta	Scale Score	A	B
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	725		
	Level 4 Cut	0.3298	750		
	Level 5 Cut	1.5206	792		

Note. ELA/L = English language arts/literacy.

Table A.12.23 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	725		
	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	725		
	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	725		
	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	725		
	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		

Table A.12.24 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 10 ELA/L	17.2512	45.3690	8.6256	32.6845
Grade 11 ELA/L	13.9712	45.3920	6.9856	32.6961

Note. ELA/L = English language arts/literacy.

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

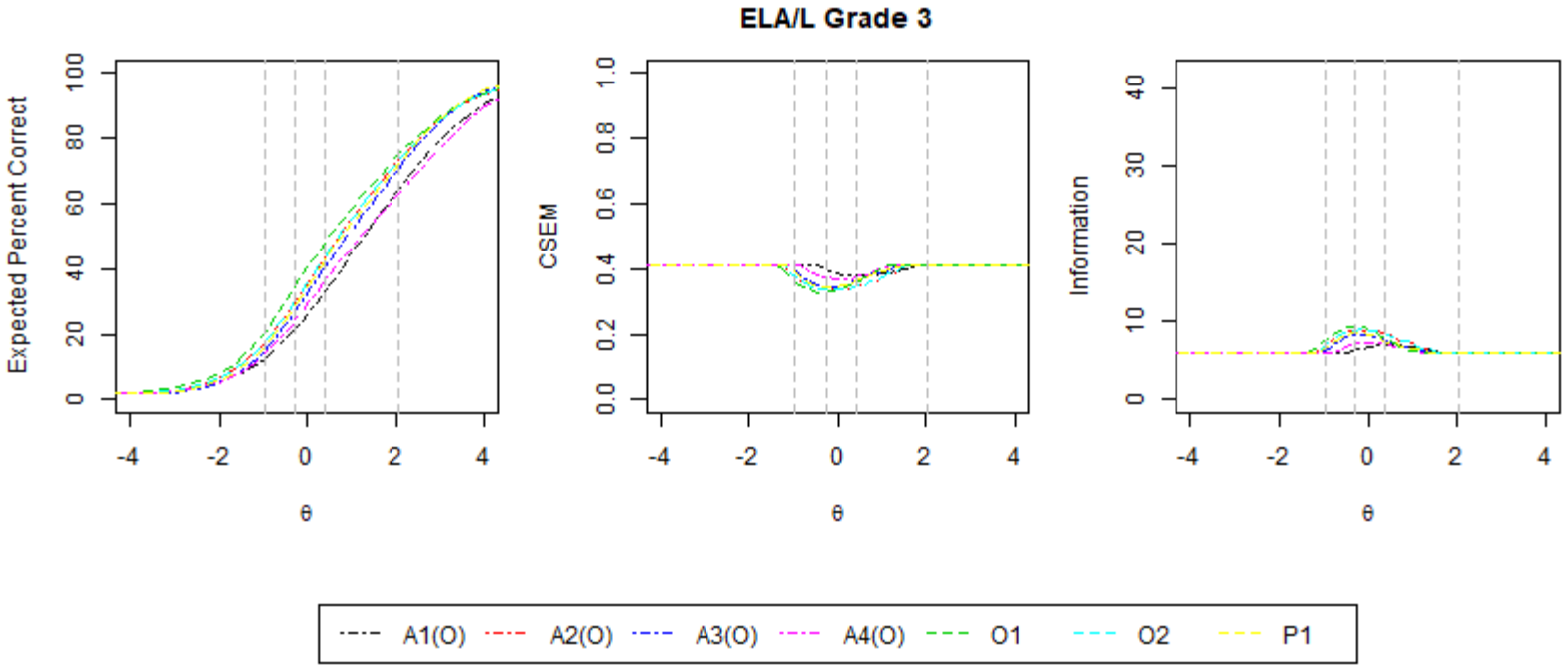

Figure A.12.1 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 3

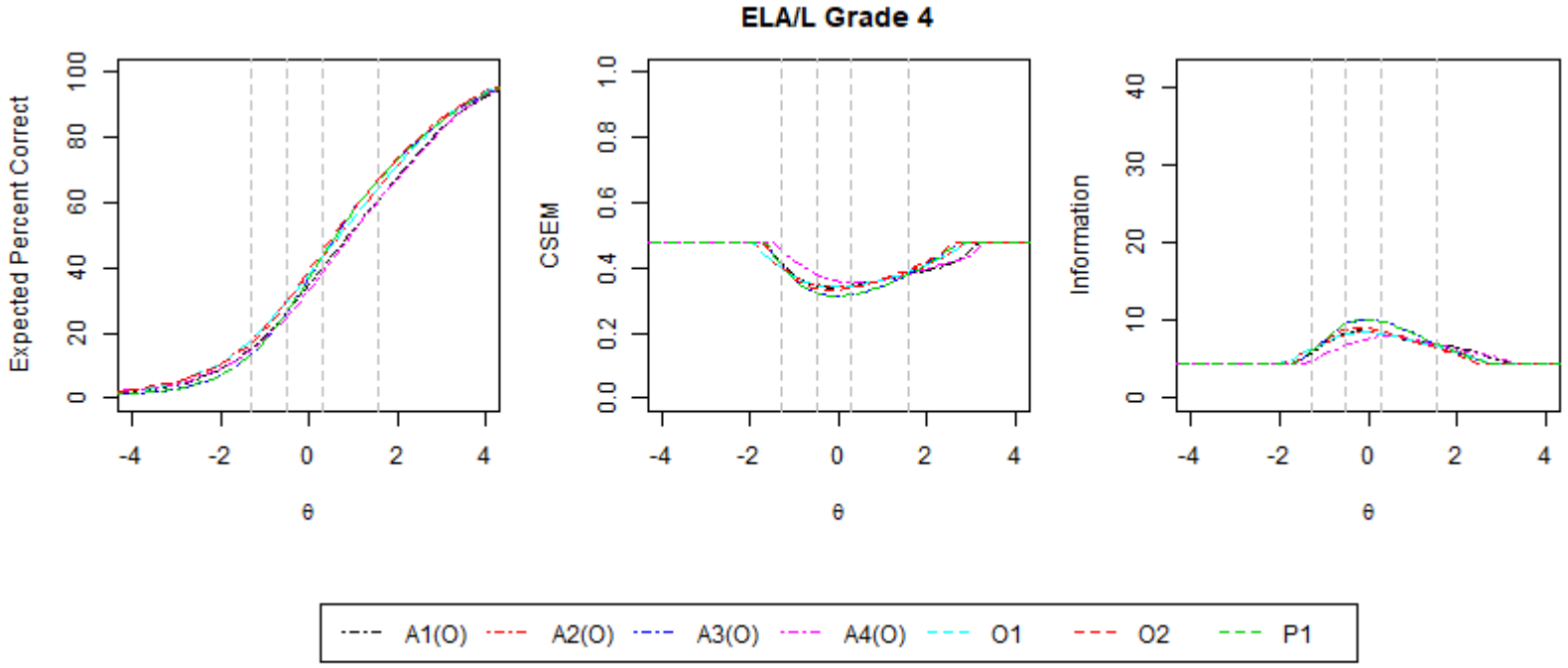

Figure A.12.2 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 4

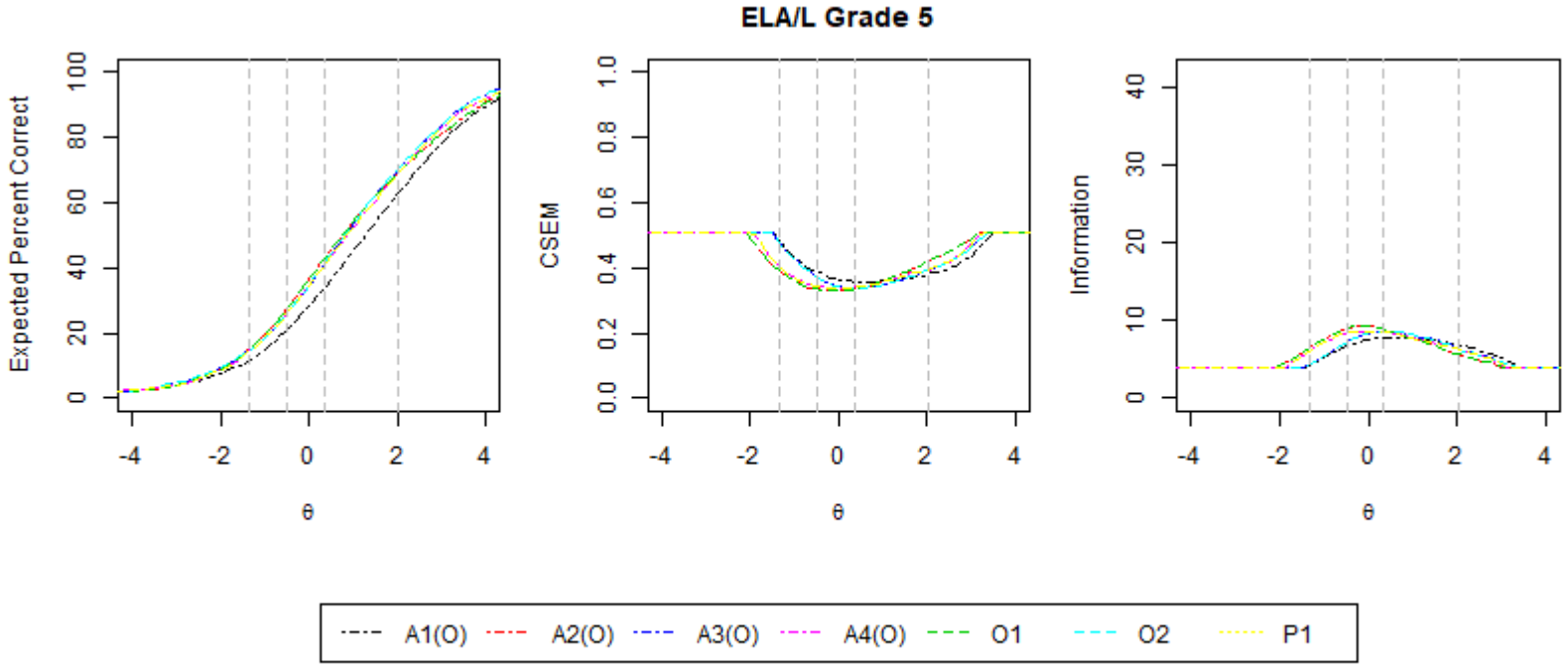

Figure A.12.3 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 5

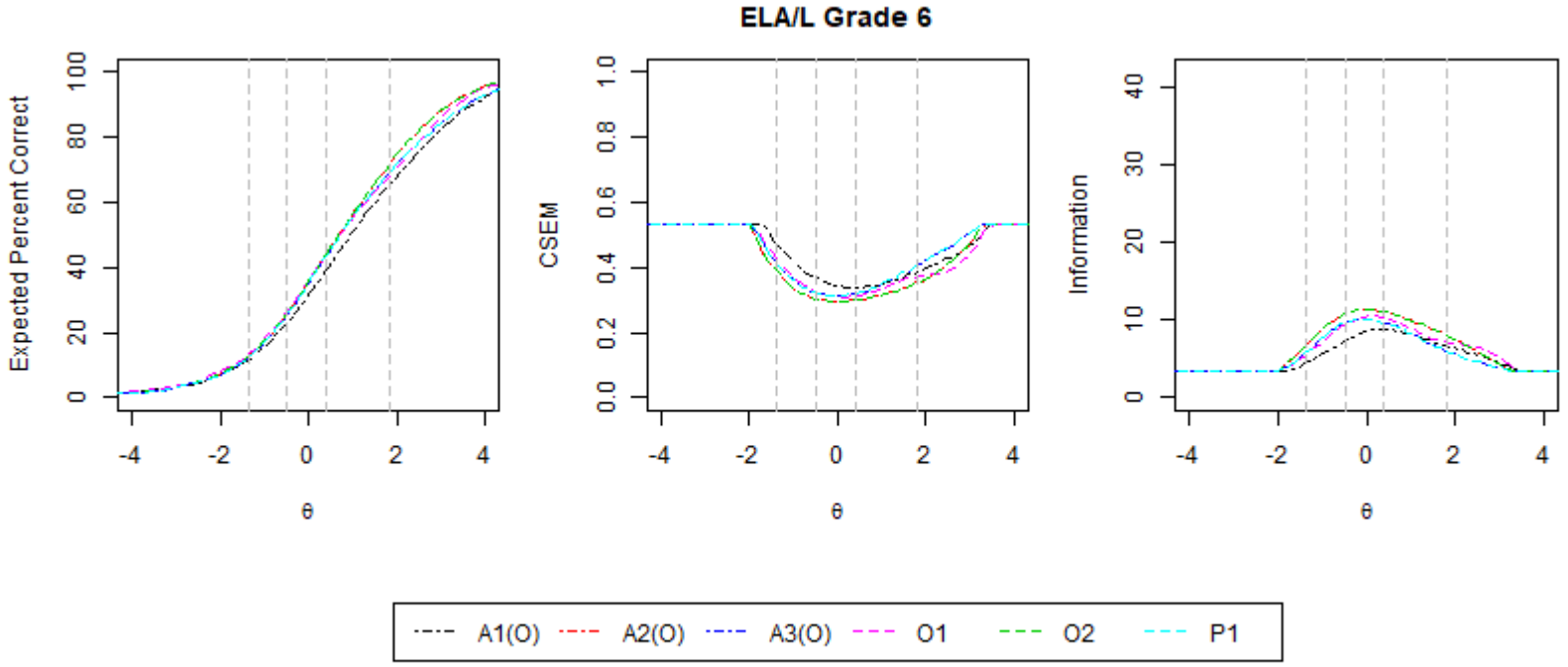

Figure A.12.4 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 6

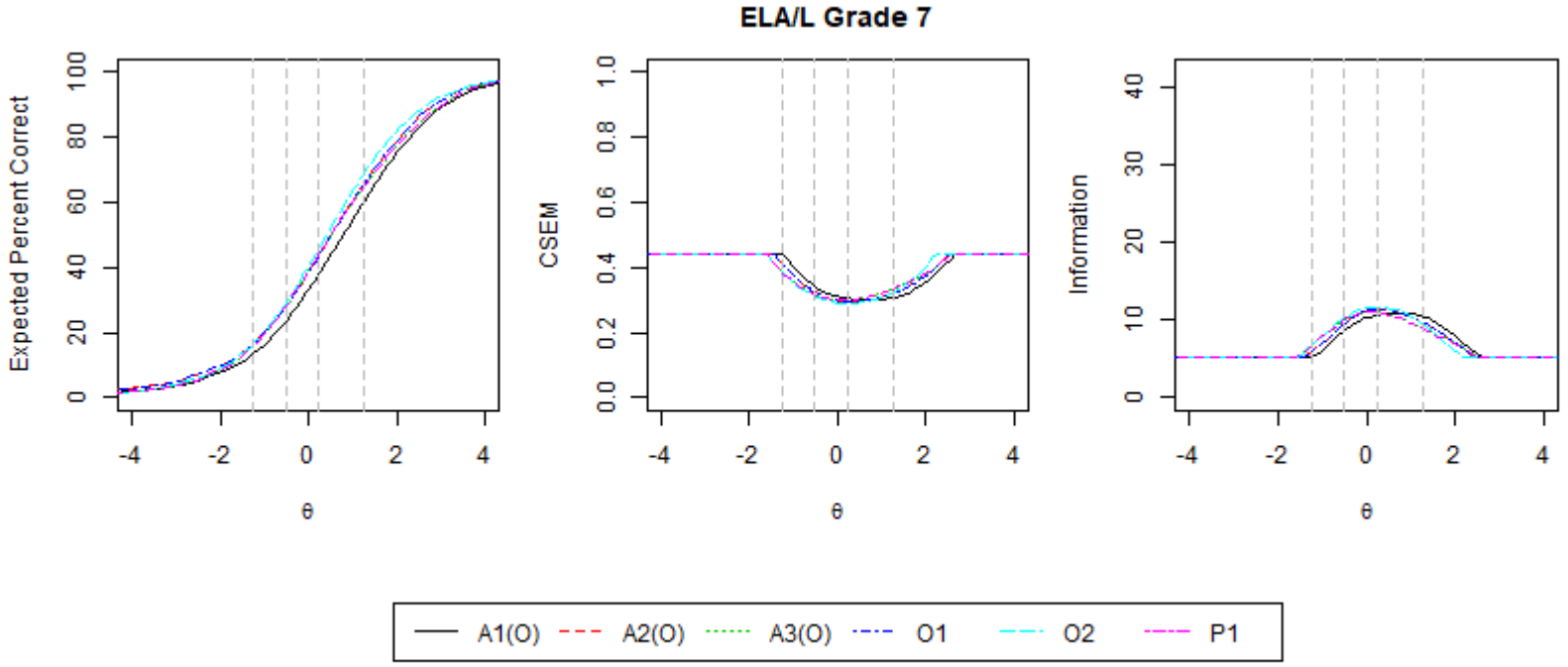

Figure A.12.5 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 7

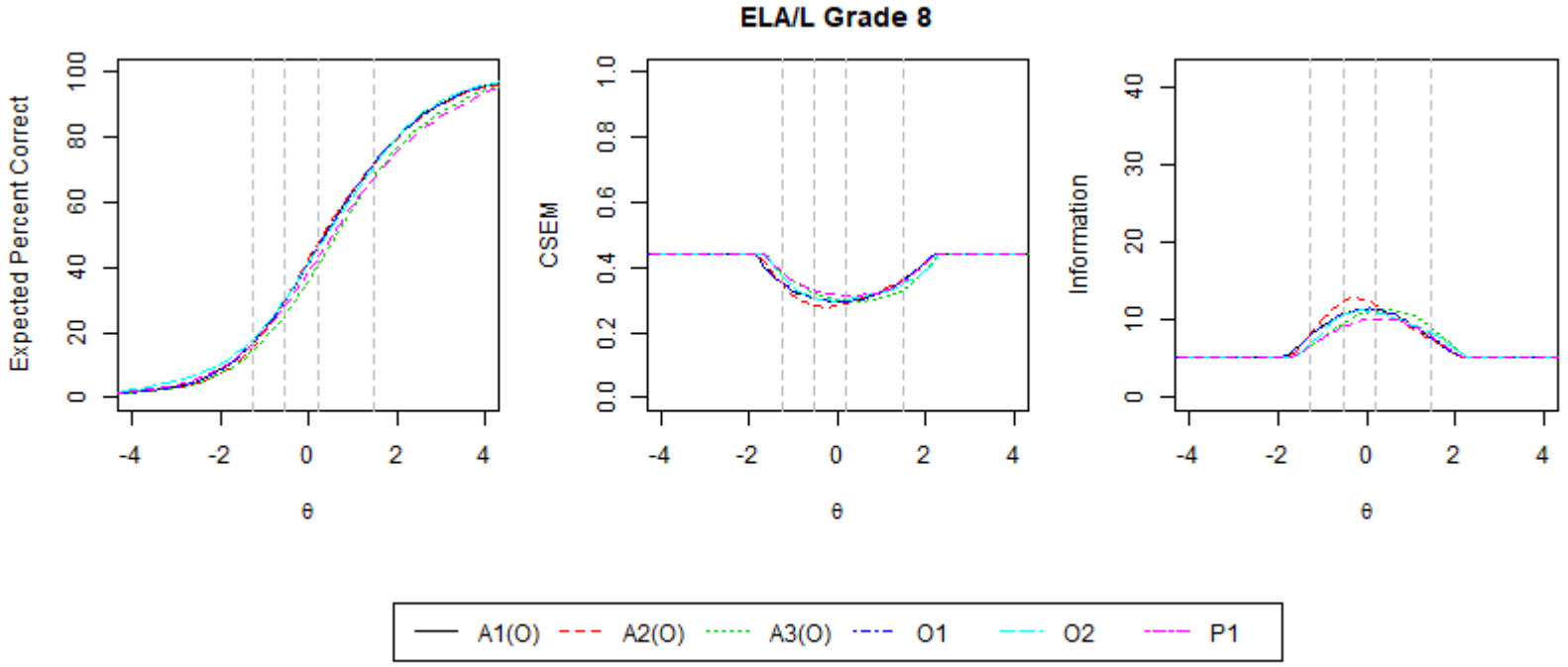

Figure A.12.6 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 8

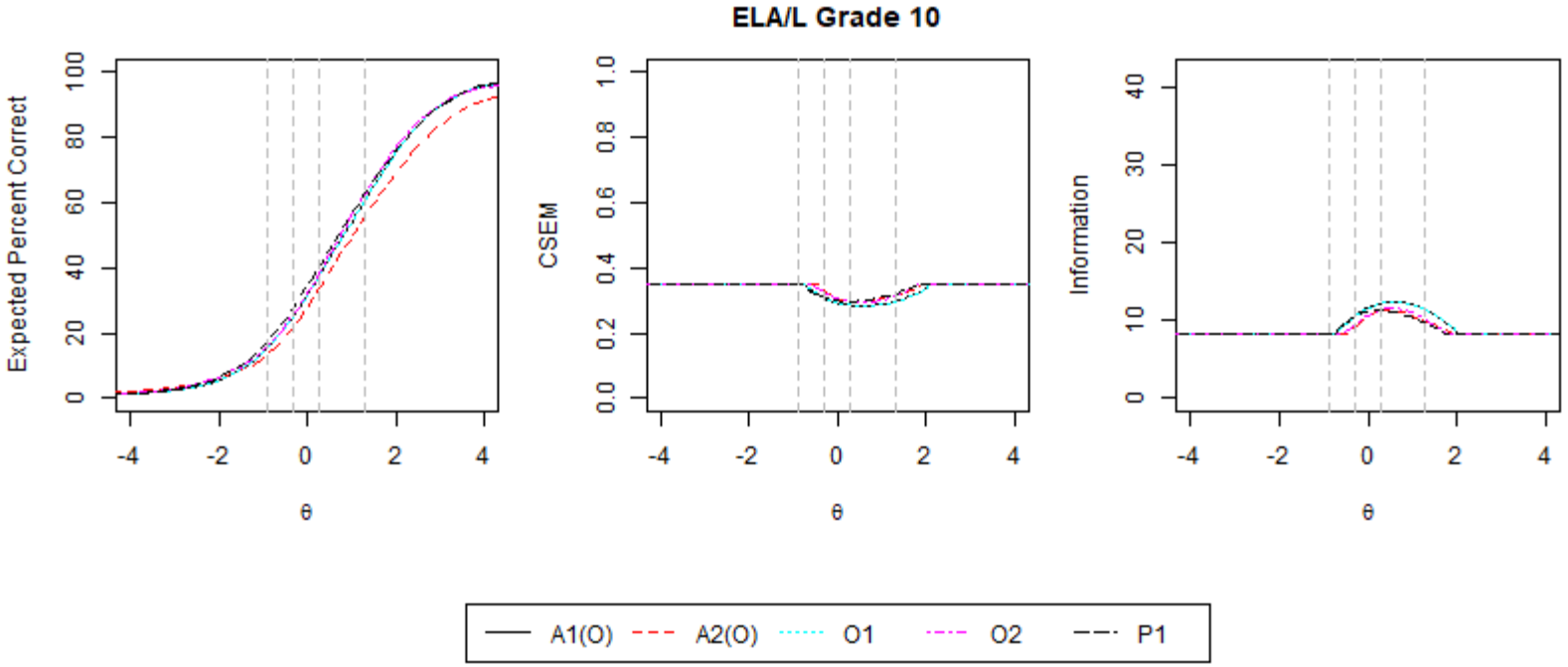

Figure A.12.7 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 10

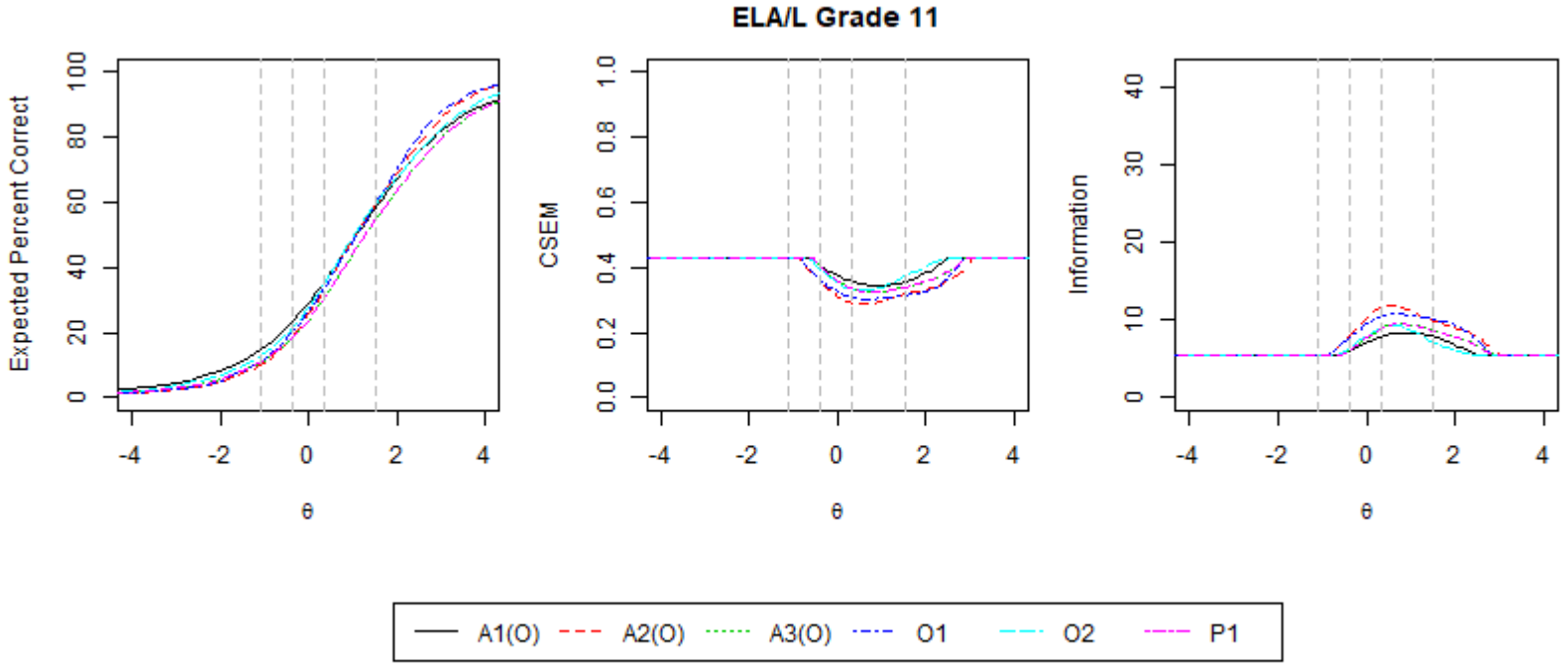

Figure A.12.8 Pre-Equated IRT Test Characteristic Curves, Information Curves, and CSEM Curves ELA/L Grade 11

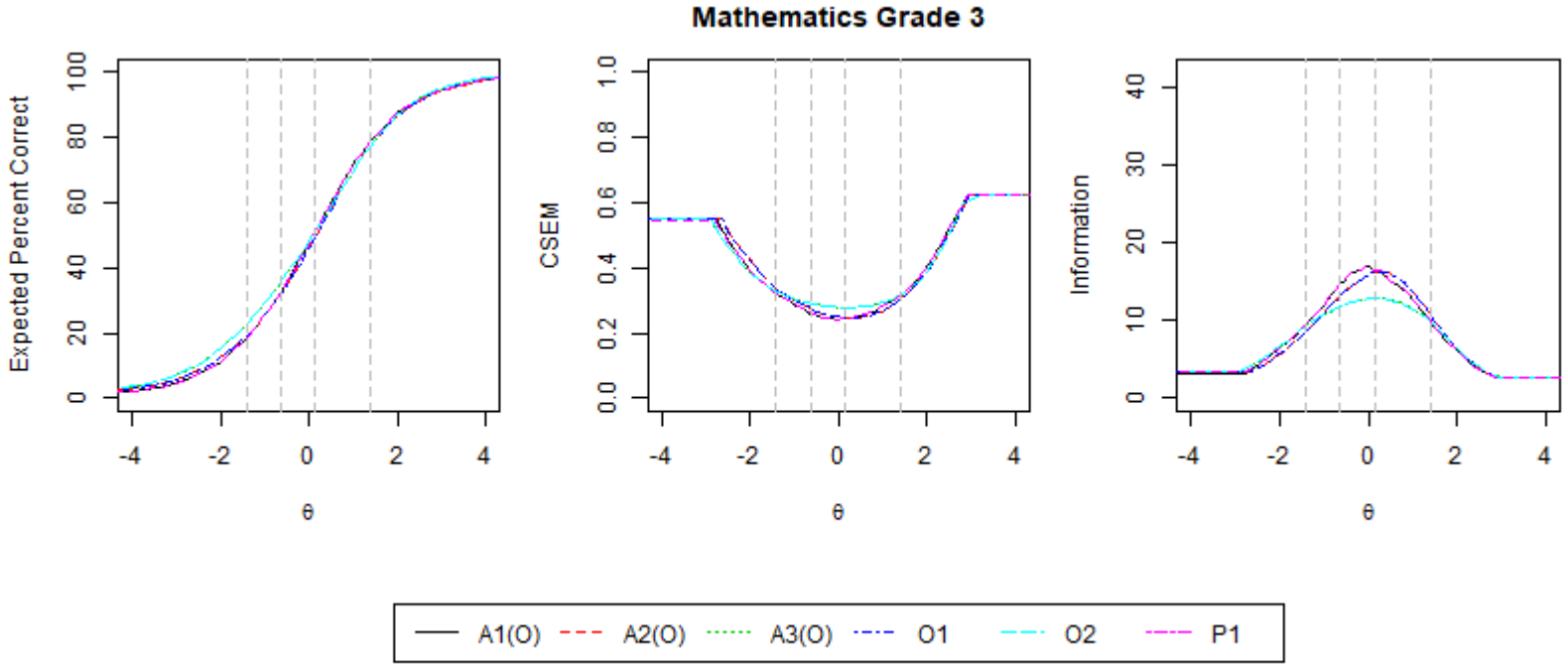

Figure A.12.9 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 3

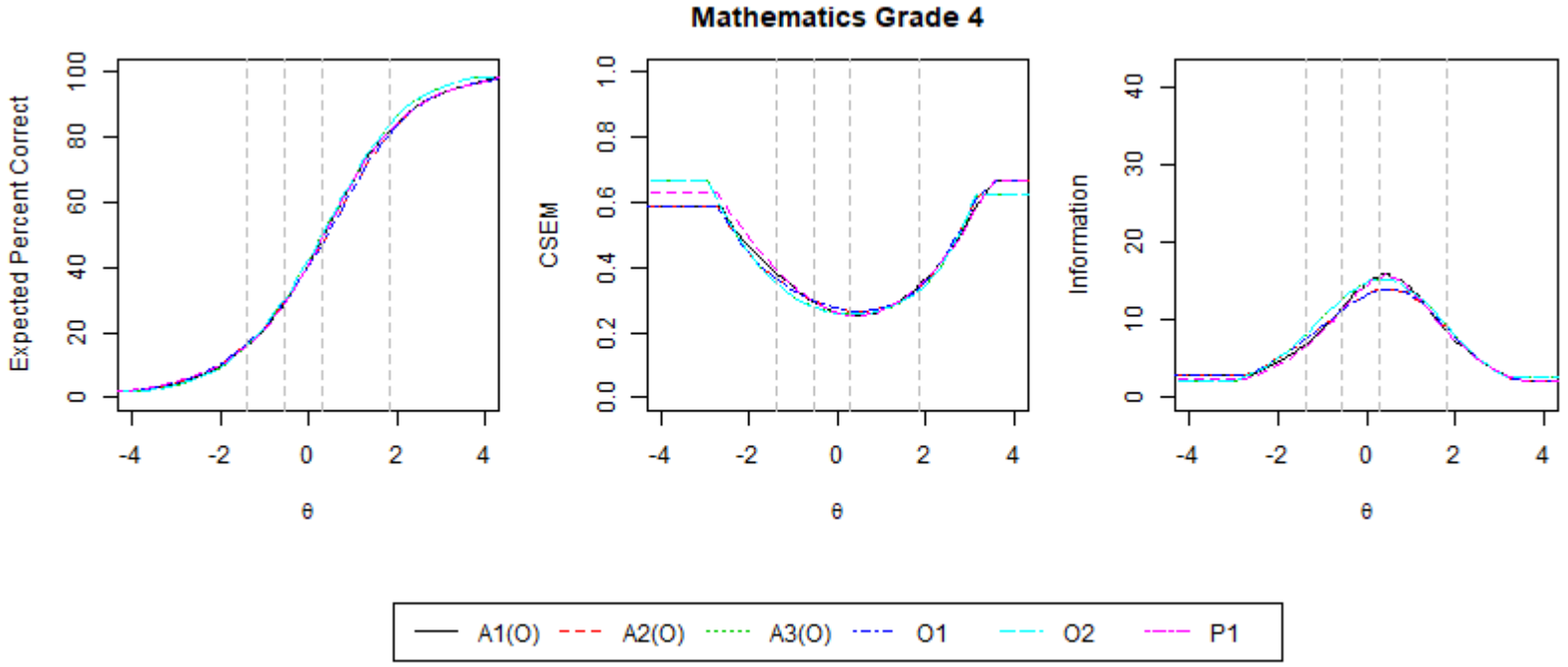

Figure A.12.10 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 4

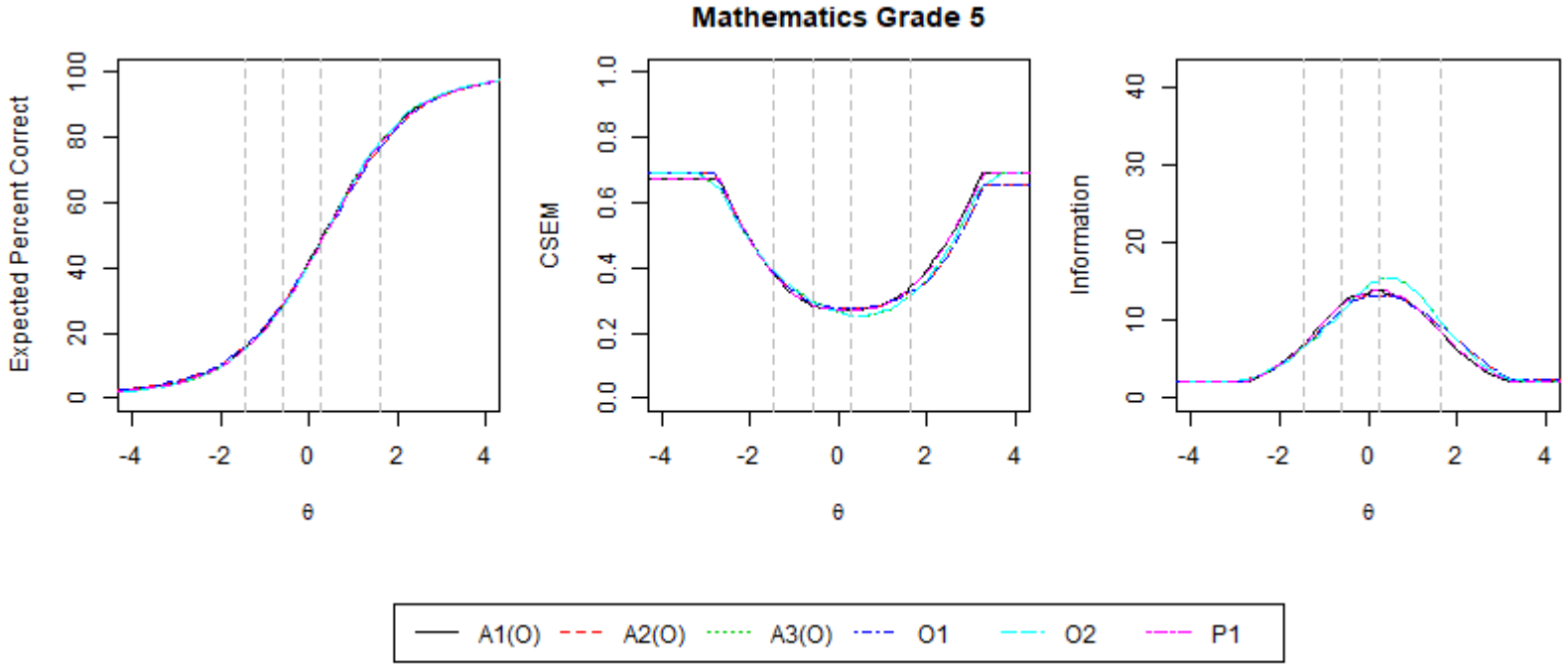

Figure A.12.11 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 5

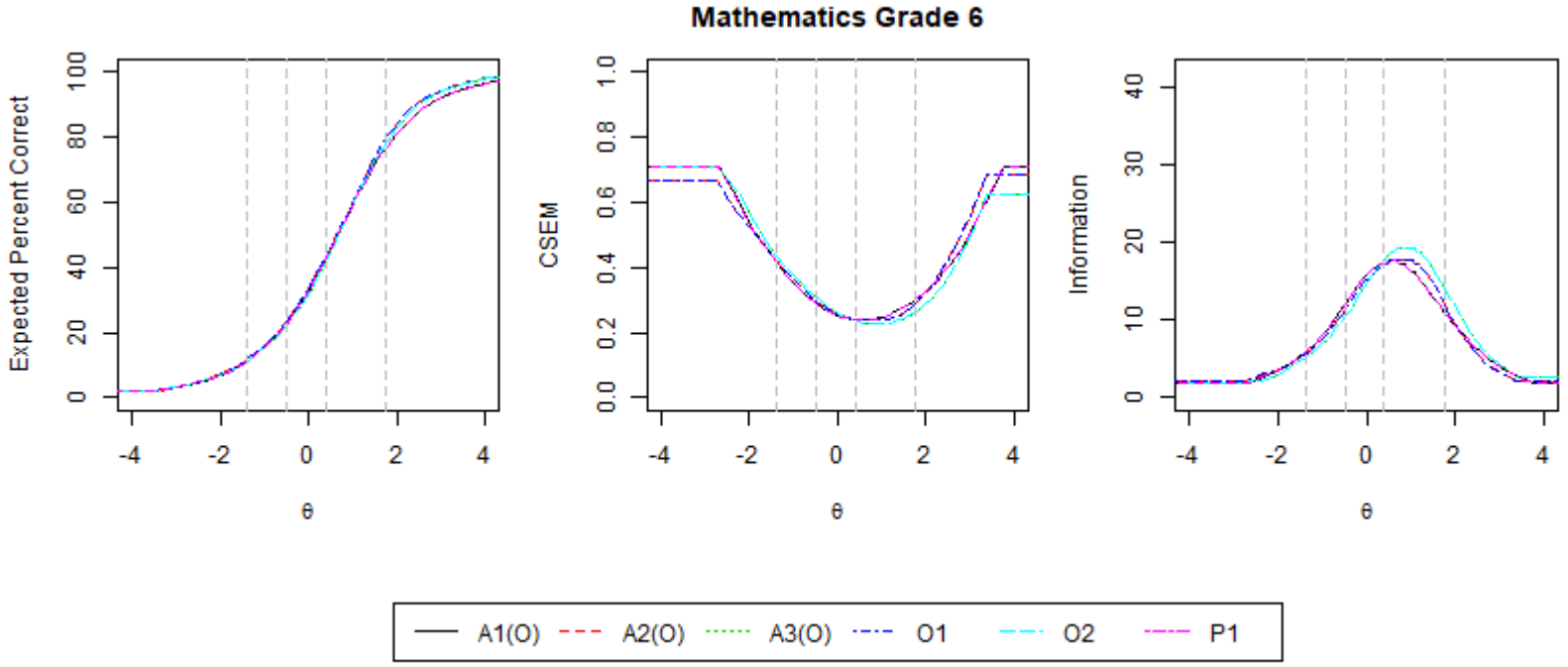

Figure A.12.12 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 6

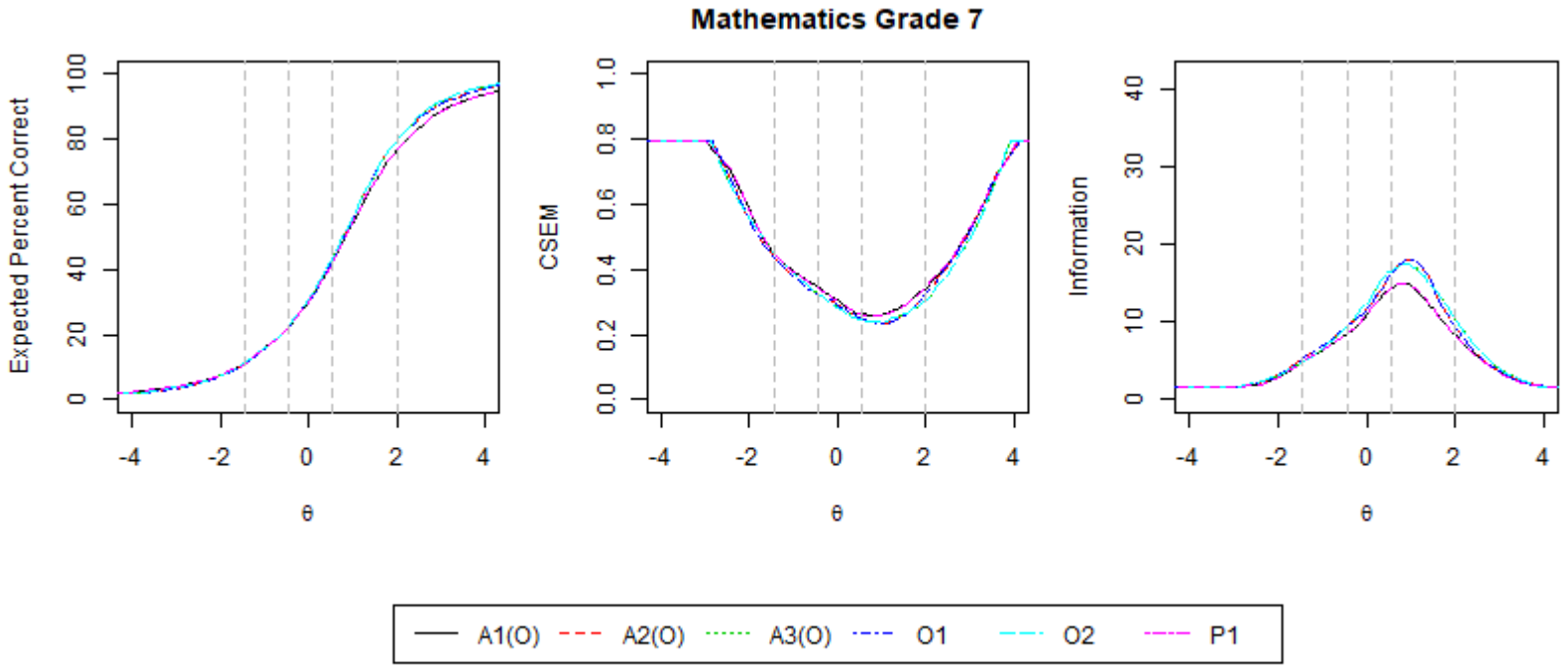

Figure A.12.13 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 7

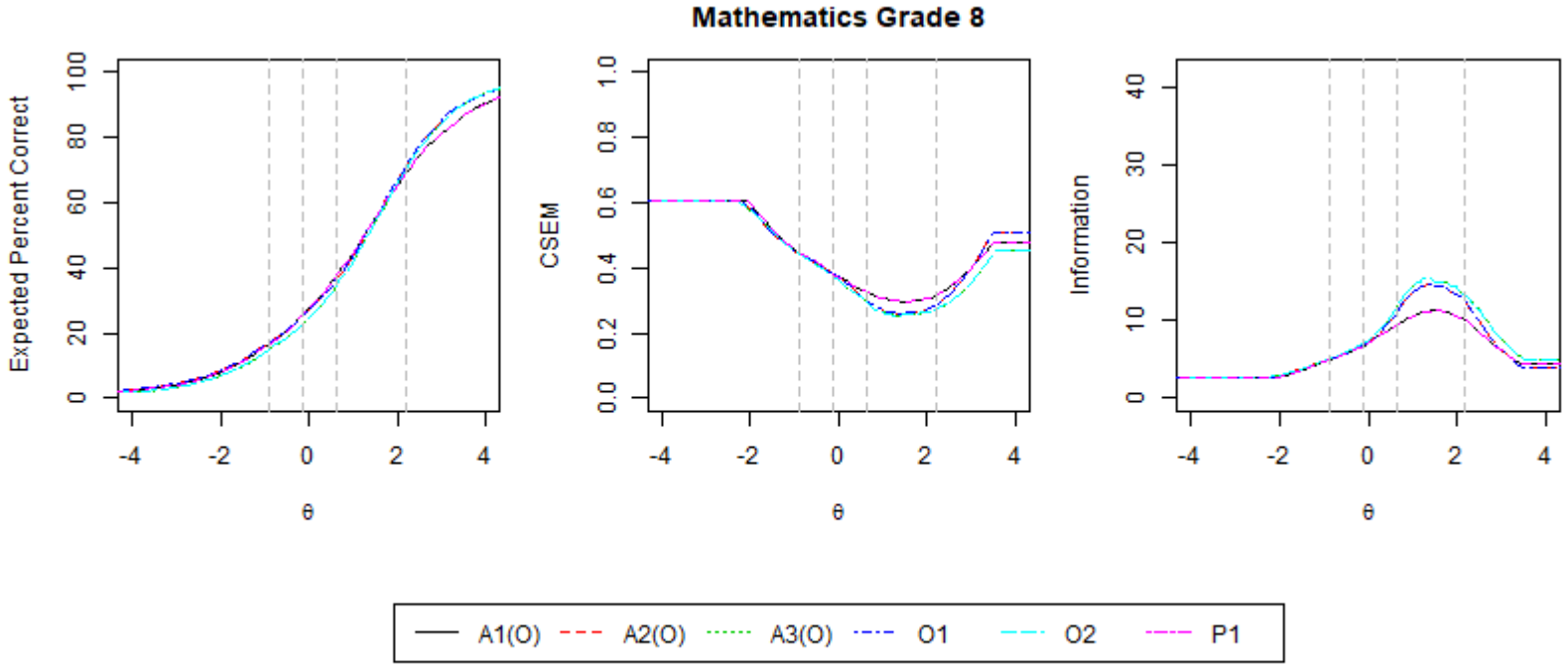

Figure A.12.14 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Mathematics Grade 8

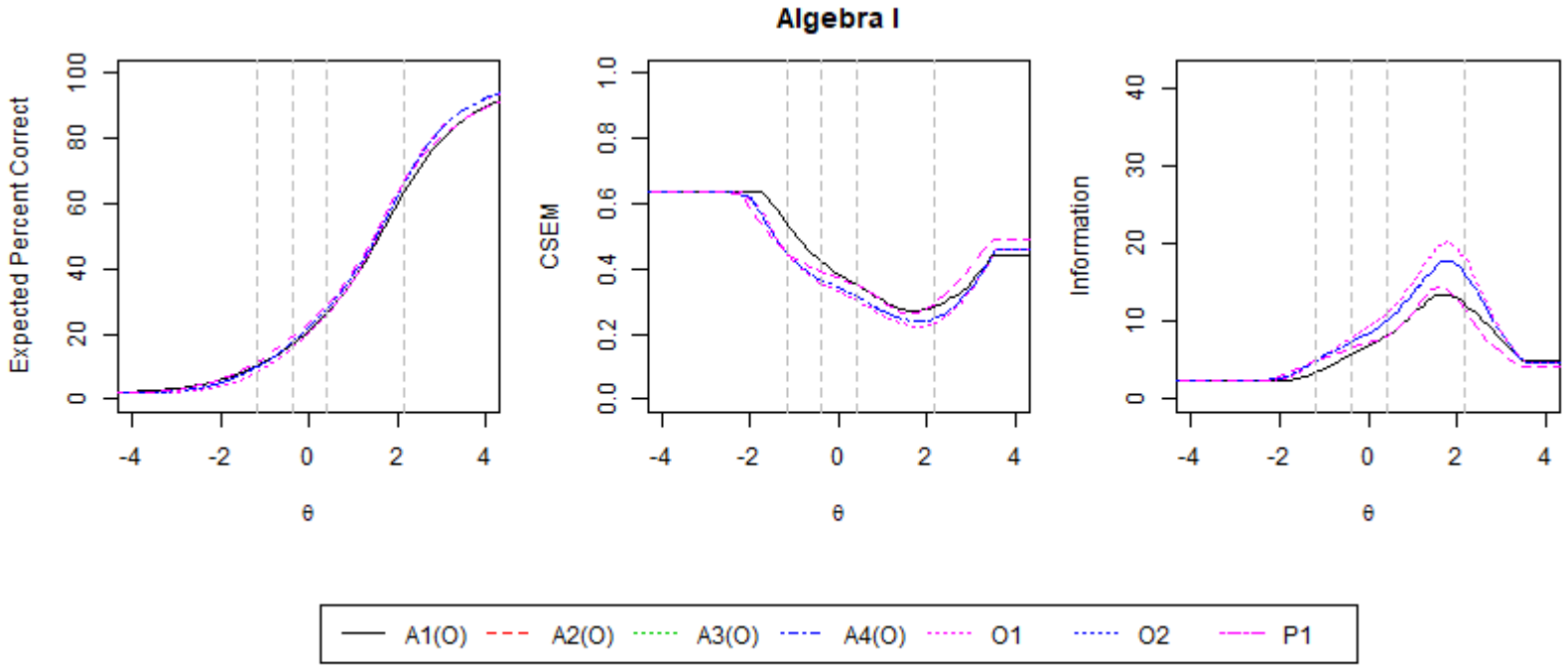

Figure A.12.15 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Algebra I

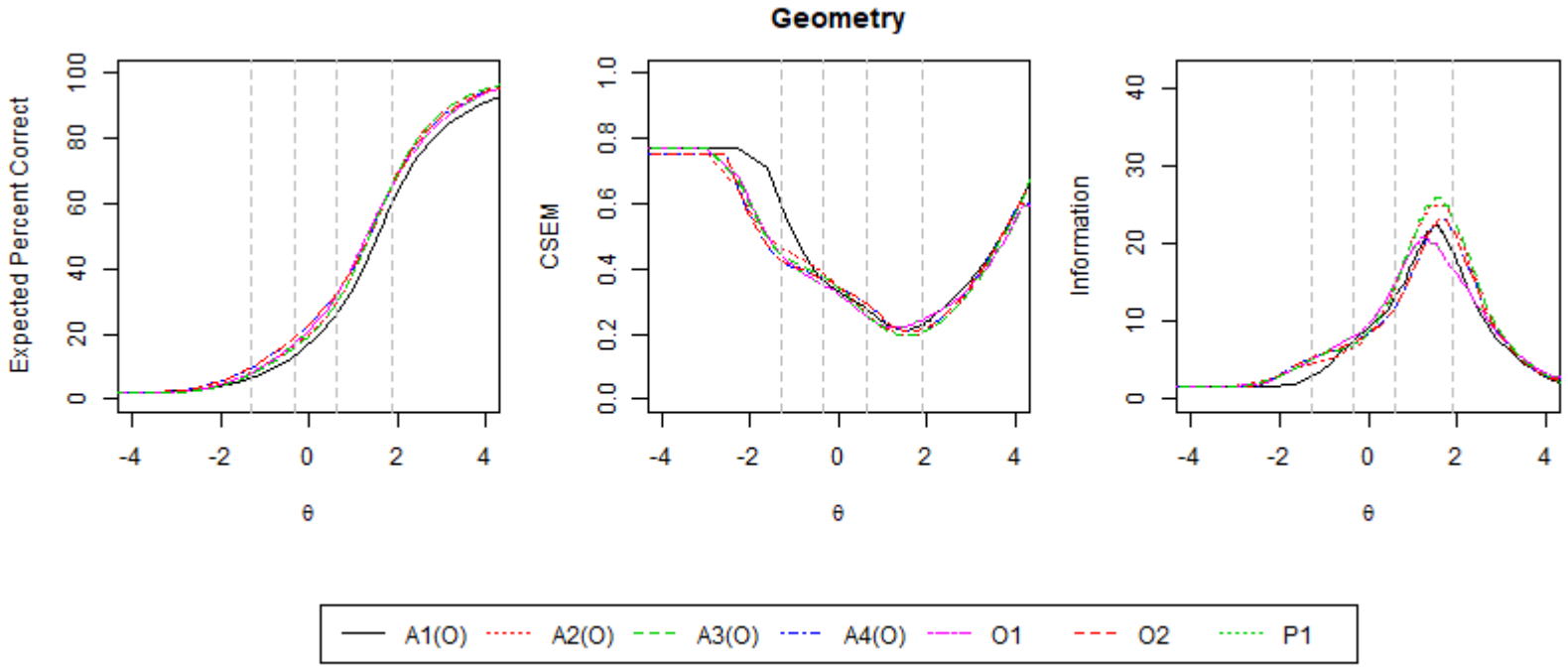

Figure A.12.16 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Geometry

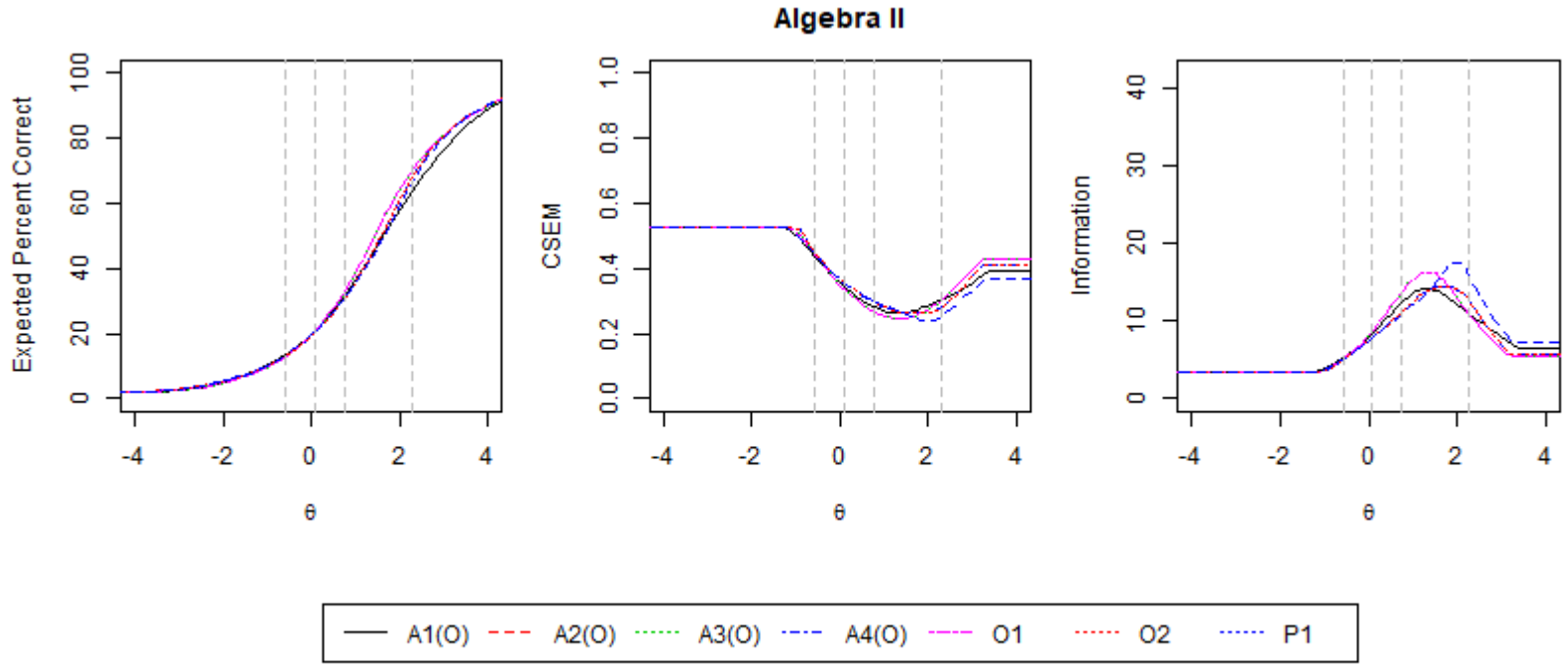

Figure A.12.17 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Algebra II

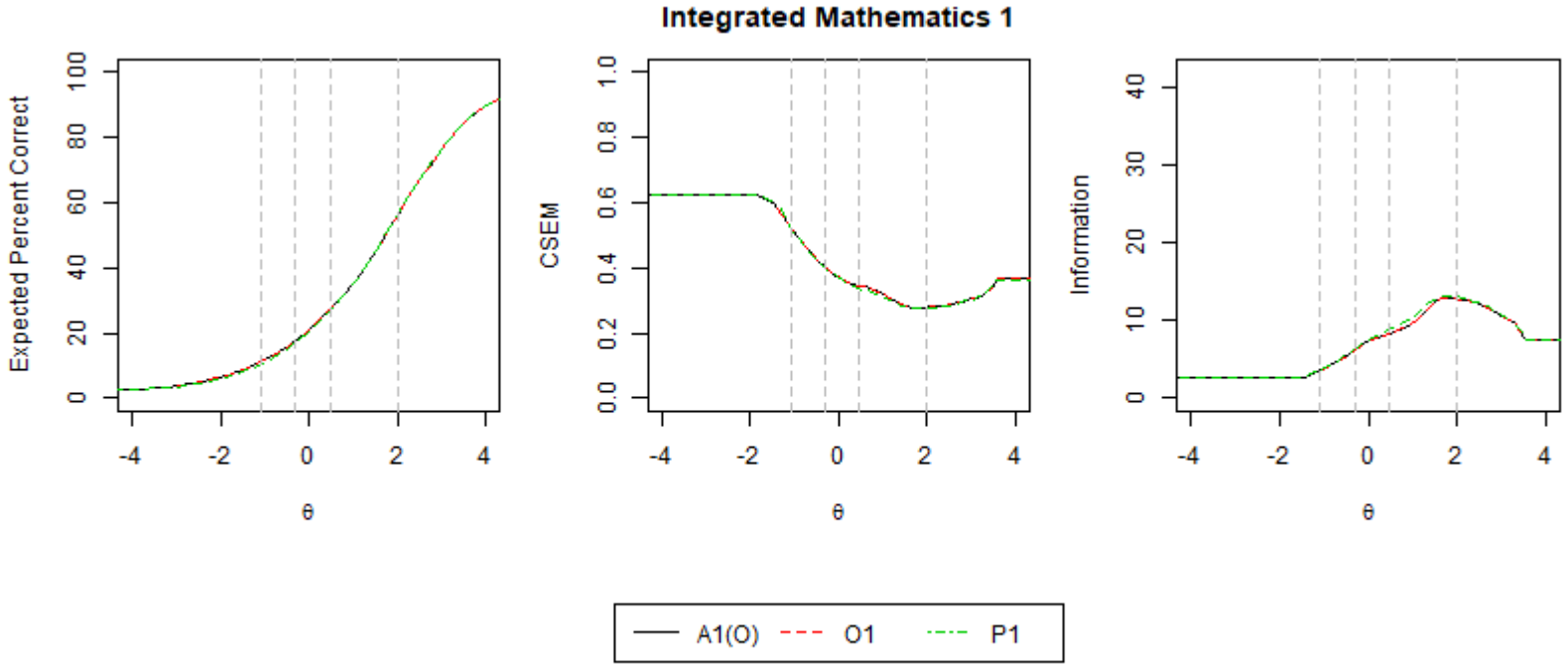

Figure A.12.18 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Integrated Mathematics I

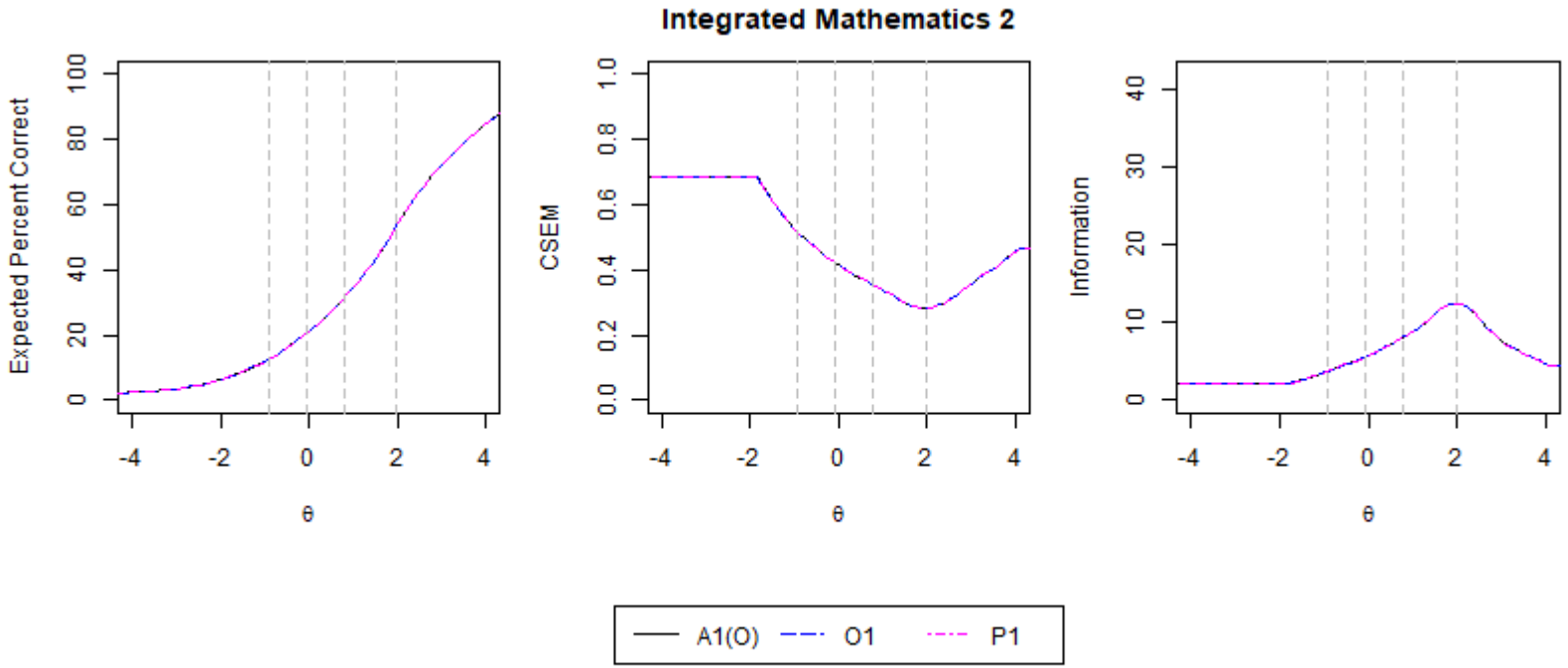

Figure A.12.19 IRT Test Characteristic Curves, Information Curves, and CSEM Curves Integrated Mathematics II

Appendix 12.4: Scale Score Cumulative Frequencies

Table A.12.25 Scale Score Cumulative Frequencies: ELA/L Grade 3

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	6,445	6.65	6,445	6.65
655-659	0	0	6,445	6.65
660-664	3,248	3.35	9,693	10
665-669	1,669	1.72	11,362	11.72
670-674	2,006	2.07	13,368	13.79
675-679	1,620	1.67	14,988	15.46
680-684	3,288	3.39	18,276	18.86
685-689	3,097	3.2	21,373	22.05
690-694	2,870	2.96	24,243	25.01
695-699	2,959	3.05	27,202	28.06
700-704	2,945	3.04	30,147	31.1
705-709	5,561	5.74	35,708	36.84
710-714	2,853	2.94	38,561	39.78
715-719	4,288	4.42	42,849	44.21
720-724	4,392	4.53	47,241	48.74
725-729	4,407	4.55	51,648	53.28
730-734	4,531	4.67	56,179	57.96
735-739	4,594	4.74	60,773	62.7
740-744	4,523	4.67	65,296	67.37
745-749	3,002	3.1	68,298	70.46
750-754	5,890	6.08	74,188	76.54
755-759	2,932	3.02	77,120	79.56
760-764	3,809	3.93	80,929	83.49
765-769	2,473	2.55	83,402	86.05
770-774	4,124	4.25	87,526	90.3
775-779	1,601	1.65	89,127	91.95
780-784	1,232	1.27	90,359	93.22
785-789	1,529	1.58	91,888	94.8
790-794	958	0.99	92,846	95.79
795-799	866	0.89	93,712	96.68
800-804	734	0.76	94,446	97.44
805-809	521	0.54	94,967	97.98
810-814	513	0.53	95,480	98.51
815-819	368	0.38	95,848	98.89
820-824	306	0.32	96,154	99.2
825-829	238	0.25	96,392	99.45
830-834	123	0.13	96,515	99.57
835-839	88	0.09	96,603	99.66
840-844	0	0	96,603	99.66
845-850	325	0.34	96,928	100

Table A.12.26 Scale Score Cumulative Frequencies: ELA/L Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,228	2.25	2,228	2.25
655-659	843	0.85	3,071	3.1
660-664	948	0.96	4,019	4.06
665-669	1,052	1.06	5,071	5.12
670-674	2,370	2.39	7,441	7.52
675-679	2,634	2.66	10,075	10.18
680-684	1,170	1.18	11,245	11.36
685-689	2,766	2.79	14,011	14.15
690-694	4,155	4.2	18,166	18.35
695-699	3,917	3.96	22,083	22.3
700-704	2,790	2.82	24,873	25.12
705-709	5,185	5.24	30,058	30.36
710-714	4,076	4.12	34,134	34.48
715-719	4,030	4.07	38,164	38.55
720-724	5,344	5.4	43,508	43.94
725-729	5,504	5.56	49,012	49.5
730-734	5,307	5.36	54,319	54.86
735-739	5,381	5.44	59,700	60.3
740-744	5,160	5.21	64,860	65.51
745-749	5,025	5.08	69,885	70.59
750-754	4,758	4.81	74,643	75.39
755-759	4,413	4.46	79,056	79.85
760-764	4,026	4.07	83,082	83.92
765-769	3,510	3.55	86,592	87.46
770-774	3,021	3.05	89,613	90.51
775-779	2,415	2.44	92,028	92.95
780-784	1,980	2	94,008	94.95
785-789	1,163	1.17	95,171	96.13
790-794	1,182	1.19	96,353	97.32
795-799	667	0.67	97,020	97.99
800-804	593	0.6	97,613	98.59
805-809	432	0.44	98,045	99.03
810-814	371	0.37	98,416	99.4
815-819	201	0.2	98,617	99.61
820-824	104	0.11	98,721	99.71
825-829	139	0.14	98,860	99.85
830-834	39	0.04	98,899	99.89
835-839	41	0.04	98,940	99.93
840-844	18	0.02	98,958	99.95
845-850	48	0.05	99,006	100

Table A.12.27 Scale Score Cumulative Frequencies: ELA/L Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,320	1.32	1,320	1.32
655-659	0	0	1,320	1.32
660-664	1,410	1.42	2,730	2.74
665-669	4	0	2,734	2.74
670-674	2,085	2.09	4,819	4.84
675-679	2,487	2.5	7,306	7.33
680-684	1,596	1.6	8,902	8.93
685-689	3,005	3.02	11,907	11.95
690-694	3,069	3.08	14,976	15.03
695-699	2,995	3.01	17,971	18.04
700-704	5,968	5.99	23,939	24.03
705-709	2,866	2.88	26,805	26.9
710-714	4,119	4.13	30,924	31.04
715-719	5,495	5.52	36,419	36.55
720-724	5,430	5.45	41,849	42
725-729	5,424	5.44	47,273	47.45
730-734	6,547	6.57	53,820	54.02
735-739	5,271	5.29	59,091	59.31
740-744	4,934	4.95	64,025	64.26
745-749	4,833	4.85	68,858	69.11
750-754	4,707	4.72	73,565	73.84
755-759	5,270	5.29	78,835	79.13
760-764	3,958	3.97	82,793	83.1
765-769	4,311	4.33	87,104	87.43
770-774	2,939	2.95	90,043	90.38
775-779	2,491	2.5	92,534	92.88
780-784	1,485	1.49	94,019	94.37
785-789	1,272	1.28	95,291	95.64
790-794	1,369	1.37	96,660	97.02
795-799	1,010	1.01	97,670	98.03
800-804	580	0.58	98,250	98.61
805-809	454	0.46	98,704	99.07
810-814	312	0.31	99,016	99.38
815-819	243	0.24	99,259	99.63
820-824	151	0.15	99,410	99.78
825-829	72	0.07	99,482	99.85
830-834	67	0.07	99,549	99.92
835-839	37	0.04	99,586	99.95
840-844	16	0.02	99,602	99.97
845-850	30	0.03	99,632	100

Table A.12.28 Scale Score Cumulative Frequencies: ELA/L Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	705	0.72	705	0.72
655-659	466	0.47	1,171	1.19
660-664	1	0	1,172	1.19
665-669	1,329	1.35	2,501	2.54
670-674	2	0	2,503	2.54
675-679	1,836	1.86	4,339	4.4
680-684	2,270	2.3	6,609	6.7
685-689	2,543	2.58	9,152	9.28
690-694	2,547	2.58	11,699	11.87
695-699	2,554	2.59	14,253	14.46
700-704	4,715	4.78	18,968	19.24
705-709	3,459	3.51	22,427	22.75
710-714	3,653	3.71	26,080	26.45
715-719	5,544	5.62	31,624	32.08
720-724	6,000	6.09	37,624	38.16
725-729	4,867	4.94	42,491	43.1
730-734	5,962	6.05	48,453	49.15
735-739	7,178	7.28	55,631	56.43
740-744	4,707	4.77	60,338	61.2
745-749	6,928	7.03	67,266	68.23
750-754	5,434	5.51	72,700	73.74
755-759	5,108	5.18	77,808	78.92
760-764	4,663	4.73	82,471	83.65
765-769	4,125	4.18	86,596	87.83
770-774	3,371	3.42	89,967	91.25
775-779	2,238	2.27	92,205	93.52
780-784	1,794	1.82	93,999	95.34
785-789	1,437	1.46	95,436	96.8
790-794	1,034	1.05	96,470	97.85
795-799	800	0.81	97,270	98.66
800-804	485	0.49	97,755	99.15
805-809	314	0.32	98,069	99.47
810-814	207	0.21	98,276	99.68
815-819	104	0.11	98,380	99.79
820-824	74	0.08	98,454	99.86
825-829	61	0.06	98,515	99.92
830-834	32	0.03	98,547	99.96
835-839	14	0.01	98,561	99.97
840-844	6	0.01	98,567	99.98
845-850	23	0.02	98,590	100

Table A.12.29 Scale Score Cumulative Frequencies: ELA/L Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,818	1.88	1,818	1.88
655-659	1,502	1.55	3,320	3.42
660-664	952	0.98	4,272	4.41
665-669	974	1	5,246	5.41
670-674	2,147	2.21	7,393	7.63
675-679	2,336	2.41	9,729	10.04
680-684	153	0.16	9,882	10.19
685-689	3,235	3.34	13,117	13.53
690-694	2,312	2.38	15,429	15.91
695-699	3,408	3.52	18,837	19.43
700-704	2,331	2.4	21,168	21.83
705-709	4,307	4.44	25,475	26.28
710-714	3,297	3.4	28,772	29.68
715-719	4,478	4.62	33,250	34.3
720-724	4,571	4.71	37,821	39.01
725-729	4,568	4.71	42,389	43.72
730-734	4,376	4.51	46,765	48.24
735-739	5,534	5.71	52,299	53.94
740-744	5,479	5.65	57,778	59.6
745-749	5,465	5.64	63,243	65.23
750-754	5,201	5.36	68,444	70.6
755-759	4,044	4.17	72,488	74.77
760-764	4,528	4.67	77,016	79.44
765-769	4,020	4.15	81,036	83.59
770-774	3,600	3.71	84,636	87.3
775-779	2,489	2.57	87,125	89.87
780-784	2,588	2.67	89,713	92.54
785-789	1,694	1.75	91,407	94.28
790-794	1,364	1.41	92,771	95.69
795-799	889	0.92	93,660	96.61
800-804	938	0.97	94,598	97.57
805-809	743	0.77	95,341	98.34
810-814	449	0.46	95,790	98.8
815-819	226	0.23	96,016	99.04
820-824	355	0.37	96,371	99.4
825-829	153	0.16	96,524	99.56
830-834	109	0.11	96,633	99.67
835-839	45	0.05	96,678	99.72
840-844	93	0.1	96,771	99.82
845-850	179	0.18	96,950	100

Table A.12.30 Scale Score Cumulative Frequencies: ELA/L Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,678	1.75	1,678	1.75
655-659	1,192	1.24	2,870	2.99
660-664	160	0.17	3,030	3.16
665-669	1,488	1.55	4,518	4.7
670-674	1,935	2.02	6,453	6.72
675-679	1,036	1.08	7,489	7.8
680-684	2,042	2.13	9,531	9.93
685-689	1,983	2.07	11,514	11.99
690-694	3,078	3.21	14,592	15.2
695-699	2,897	3.02	17,489	18.21
700-704	2,060	2.15	19,549	20.36
705-709	4,124	4.29	23,673	24.65
710-714	4,185	4.36	27,858	29.01
715-719	3,210	3.34	31,068	32.35
720-724	4,402	4.58	35,470	36.94
725-729	5,544	5.77	41,014	42.71
730-734	5,741	5.98	46,755	48.69
735-739	4,547	4.74	51,302	53.42
740-744	5,433	5.66	56,735	59.08
745-749	4,505	4.69	61,240	63.77
750-754	4,379	4.56	65,619	68.33
755-759	5,329	5.55	70,948	73.88
760-764	3,945	4.11	74,893	77.99
765-769	5,306	5.53	80,199	83.52
770-774	3,119	3.25	83,318	86.76
775-779	2,712	2.82	86,030	89.59
780-784	2,268	2.36	88,298	91.95
785-789	1,849	1.93	90,147	93.88
790-794	1,125	1.17	91,272	95.05
795-799	1,346	1.4	92,618	96.45
800-804	951	0.99	93,569	97.44
805-809	547	0.57	94,116	98.01
810-814	485	0.51	94,601	98.51
815-819	386	0.4	94,987	98.92
820-824	214	0.22	95,201	99.14
825-829	165	0.17	95,366	99.31
830-834	146	0.15	95,512	99.46
835-839	126	0.13	95,638	99.59
840-844	112	0.12	95,750	99.71
845-850	278	0.29	96,028	100

Table A.12.31 Scale Score Cumulative Frequencies: ELA/L Grade 10

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	33	1.19	33	1.19
655-659	18	0.65	51	1.84
660-664	6	0.22	57	2.06
665-669	26	0.94	83	3
670-674	0	0	83	3
675-679	38	1.37	121	4.37
680-684	29	1.05	150	5.42
685-689	30	1.08	180	6.51
690-694	8	0.29	188	6.79
695-699	32	1.16	220	7.95
700-704	72	2.6	292	10.55
705-709	47	1.7	339	12.25
710-714	41	1.48	380	13.73
715-719	64	2.31	444	16.05
720-724	68	2.46	512	18.5
725-729	126	4.55	638	23.06
730-734	76	2.75	714	25.8
735-739	96	3.47	810	29.27
740-744	157	5.67	967	34.95
745-749	161	5.82	1,128	40.77
750-754	105	3.79	1,233	44.56
755-759	151	5.46	1,384	50.02
760-764	139	5.02	1,523	55.04
765-769	153	5.53	1,676	60.57
770-774	152	5.49	1,828	66.06
775-779	152	5.49	1,980	71.56
780-784	130	4.7	2,110	76.26
785-789	74	2.67	2,184	78.93
790-794	103	3.72	2,287	82.65
795-799	94	3.4	2,381	86.05
800-804	89	3.22	2,470	89.27
805-809	43	1.55	2,513	90.82
810-814	70	2.53	2,583	93.35
815-819	35	1.26	2,618	94.62
820-824	40	1.45	2,658	96.06
825-829	20	0.72	2,678	96.78
830-834	8	0.29	2,686	97.07
835-839	29	1.05	2,715	98.12
840-844	10	0.36	2,725	98.48
845-850	42	1.52	2,767	100

Table A.12.32 Scale Score Cumulative Frequencies: ELA/L Grade 11

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	7	1.69	7	1.69
655-659	13	3.15	20	4.84
660-664	1	0.24	21	5.08
665-669	12	2.91	33	7.99
670-674	10	2.42	43	10.41
675-679	15	3.63	58	14.04
680-684	18	4.36	76	18.4
685-689	24	5.81	100	24.21
690-694	32	7.75	132	31.96
695-699	11	2.66	143	34.62
700-704	27	6.54	170	41.16
705-709	28	6.78	198	47.94
710-714	32	7.75	230	55.69
715-719	17	4.12	247	59.81
720-724	25	6.05	272	65.86
725-729	25	6.05	297	71.91
730-734	19	4.6	316	76.51
735-739	15	3.63	331	80.15
740-744	12	2.91	343	83.05
745-749	15	3.63	358	86.68
750-754	11	2.66	369	89.35
755-759	11	2.66	380	92.01
760-764	10	2.42	390	94.43
765-769	4	0.97	394	95.4
770-774	6	1.45	400	96.85
775-779	7	1.69	407	98.55
780-784	1	0.24	408	98.79
785-789	1	0.24	409	99.03
790-794	2	0.48	411	99.52
795-799	0	0	411	99.52
800-804	1	0.24	412	99.76
805-809	0	0	412	99.76
810-814	0	0	412	99.76
815-819	0	0	412	99.76
820-824	0	0	412	99.76
825-829	0	0	412	99.76
830-834	1	0.24	413	100
835-839	0	0	413	100
840-844	0	0	413	100
845-850	0	0	413	100

Table A.12.33 Scale Score Cumulative Frequencies: Mathematics Grade 3

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,794	1.87	1,794	1.87
655-659	71	0.07	1,865	1.94
660-664	1,683	1.75	3,548	3.7
665-669	1,184	1.23	4,732	4.93
670-674	1,381	1.44	6,113	6.37
675-679	3,009	3.13	9,122	9.5
680-684	3,259	3.39	12,381	12.9
685-689	1,484	1.55	13,865	14.44
690-694	3,461	3.6	17,326	18.05
695-699	3,360	3.5	20,686	21.55
700-704	5,176	5.39	25,862	26.94
705-709	4,645	4.84	30,507	31.77
710-714	3,137	3.27	33,644	35.04
715-719	6,104	6.36	39,748	41.4
720-724	2,936	3.06	42,684	44.46
725-729	5,762	6	48,446	50.46
730-734	2,808	2.92	51,254	53.38
735-739	5,207	5.42	56,461	58.81
740-744	3,695	3.85	60,156	62.66
745-749	4,848	5.05	65,004	67.7
750-754	4,524	4.71	69,528	72.42
755-759	4,052	4.22	73,580	76.64
760-764	2,892	3.01	76,472	79.65
765-769	3,464	3.61	79,936	83.26
770-774	3,150	3.28	83,086	86.54
775-779	1,555	1.62	84,641	88.16
780-784	2,831	2.95	87,472	91.11
785-789	2,453	2.55	89,925	93.66
790-794	1,106	1.15	91,031	94.81
795-799	975	1.02	92,006	95.83
800-804	1,208	1.26	93,214	97.09
805-809	707	0.74	93,921	97.82
810-814	591	0.62	94,512	98.44
815-819	484	0.5	94,996	98.94
820-824	223	0.23	95,219	99.18
825-829	143	0.15	95,362	99.32
830-834	179	0.19	95,541	99.51
835-839	111	0.12	95,652	99.63
840-844	110	0.11	95,762	99.74
845-850	249	0.26	96,011	100

Table A.12.34 Scale Score Cumulative Frequencies: Mathematics Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,540	1.58	1,540	1.58
655-659	0	0	1,540	1.58
660-664	1,120	1.15	2,660	2.72
665-669	1,019	1.04	3,679	3.76
670-674	1,674	1.71	5,353	5.48
675-679	3,509	3.59	8,862	9.07
680-684	1,988	2.03	10,850	11.1
685-689	4,267	4.37	15,117	15.47
690-694	4,330	4.43	19,447	19.9
695-699	4,099	4.19	23,546	24.09
700-704	3,985	4.08	27,531	28.17
705-709	3,775	3.86	31,306	32.03
710-714	7,244	7.41	38,550	39.44
715-719	3,638	3.72	42,188	43.16
720-724	6,616	6.77	48,804	49.93
725-729	4,658	4.77	53,462	54.7
730-734	5,672	5.8	59,134	60.5
735-739	5,315	5.44	64,449	65.94
740-744	4,661	4.77	69,110	70.71
745-749	4,297	4.4	73,407	75.1
750-754	3,865	3.95	77,272	79.06
755-759	3,373	3.45	80,645	82.51
760-764	3,896	3.99	84,541	86.5
765-769	2,561	2.62	87,102	89.12
770-774	2,270	2.32	89,372	91.44
775-779	1,932	1.98	91,304	93.42
780-784	1,332	1.36	92,636	94.78
785-789	1,514	1.55	94,150	96.33
790-794	1,029	1.05	95,179	97.38
795-799	795	0.81	95,974	98.19
800-804	452	0.46	96,426	98.66
805-809	368	0.38	96,794	99.03
810-814	323	0.33	97,117	99.36
815-819	264	0.27	97,381	99.63
820-824	0	0	97,381	99.63
825-829	158	0.16	97,539	99.79
830-834	0	0	97,539	99.79
835-839	112	0.11	97,651	99.91
840-844	0	0	97,651	99.91
845-850	89	0.09	97,740	100

Table A.12.35 Scale Score Cumulative Frequencies: Mathematics Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	961	0.98	961	0.98
655-659	0	0	961	0.98
660-664	863	0.88	1,824	1.86
665-669	664	0.68	2,488	2.53
670-674	1,337	1.36	3,825	3.89
675-679	1,375	1.4	5,200	5.29
680-684	3,969	4.04	9,169	9.33
685-689	2,219	2.26	11,388	11.58
690-694	5,043	5.13	16,431	16.71
695-699	5,229	5.32	21,660	22.03
700-704	5,219	5.31	26,879	27.34
705-709	4,994	5.08	31,873	32.42
710-714	6,582	6.7	38,455	39.12
715-719	6,172	6.28	44,627	45.4
720-724	6,754	6.87	51,381	52.27
725-729	4,327	4.4	55,708	56.67
730-734	5,161	5.25	60,869	61.92
735-739	4,543	4.62	65,412	66.54
740-744	4,134	4.21	69,546	70.74
745-749	3,805	3.87	73,351	74.61
750-754	3,548	3.61	76,899	78.22
755-759	3,840	3.91	80,739	82.13
760-764	2,945	3	83,684	85.13
765-769	2,657	2.7	86,341	87.83
770-774	2,375	2.42	88,716	90.24
775-779	2,221	2.26	90,937	92.5
780-784	1,912	1.94	92,849	94.45
785-789	1,690	1.72	94,539	96.17
790-794	665	0.68	95,204	96.84
795-799	931	0.95	96,135	97.79
800-804	500	0.51	96,635	98.3
805-809	450	0.46	97,085	98.76
810-814	388	0.39	97,473	99.15
815-819	290	0.29	97,763	99.45
820-824	125	0.13	97,888	99.57
825-829	94	0.1	97,982	99.67
830-834	0	0	97,982	99.67
835-839	161	0.16	98,143	99.83
840-844	0	0	98,143	99.83
845-850	163	0.17	98,306	100

Table A.12.36 Scale Score Cumulative Frequencies: Mathematics Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	986	1.02	986	1.02
655-659	0	0	986	1.02
660-664	2,303	2.38	3,289	3.39
665-669	0	0	3,289	3.39
670-674	0	0	3,289	3.39
675-679	4,255	4.39	7,544	7.78
680-684	0	0	7,544	7.78
685-689	5,308	5.48	12,852	13.26
690-694	6,079	6.27	18,931	19.53
695-699	2,981	3.08	21,912	22.61
700-704	5,596	5.77	27,508	28.38
705-709	5,244	5.41	32,752	33.79
710-714	4,897	5.05	37,649	38.84
715-719	6,480	6.69	44,129	45.53
720-724	5,753	5.94	49,882	51.47
725-729	5,140	5.3	55,022	56.77
730-734	4,393	4.53	59,415	61.3
735-739	5,236	5.4	64,651	66.7
740-744	6,549	6.76	71,200	73.46
745-749	3,759	3.88	74,959	77.34
750-754	4,086	4.22	79,045	81.55
755-759	3,425	3.53	82,470	85.09
760-764	3,004	3.1	85,474	88.19
765-769	2,985	3.08	88,459	91.27
770-774	2,118	2.19	90,577	93.45
775-779	1,465	1.51	92,042	94.96
780-784	1,562	1.61	93,604	96.57
785-789	1,005	1.04	94,609	97.61
790-794	652	0.67	95,261	98.28
795-799	684	0.71	95,945	98.99
800-804	247	0.25	96,192	99.24
805-809	256	0.26	96,448	99.51
810-814	173	0.18	96,621	99.69
815-819	81	0.08	96,702	99.77
820-824	115	0.12	96,817	99.89
825-829	0	0	96,817	99.89
830-834	64	0.07	96,881	99.96
835-839	0	0	96,881	99.96
840-844	0	0	96,881	99.96
845-850	43	0.04	96,924	100

Table A.12.37 Scale Score Cumulative Frequencies: Mathematics Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	255	0.28	255	0.28
655-659	0	0	255	0.28
660-664	447	0.49	702	0.77
665-669	312	0.34	1,014	1.11
670-674	0	0	1,014	1.11
675-679	1,632	1.79	2,646	2.9
680-684	0	0	2,646	2.9
685-689	2,741	3	5,387	5.9
690-694	1,934	2.12	7,321	8.02
695-699	1,837	2.01	9,158	10.03
700-704	4,504	4.93	13,662	14.96
705-709	4,940	5.41	18,602	20.37
710-714	4,982	5.46	23,584	25.83
715-719	9,404	10.3	32,988	36.13
720-724	4,185	4.58	37,173	40.71
725-729	7,555	8.27	44,728	48.98
730-734	6,438	7.05	51,166	56.03
735-739	5,681	6.22	56,847	62.25
740-744	4,953	5.42	61,800	67.68
745-749	5,136	5.62	66,936	73.3
750-754	5,147	5.64	72,083	78.94
755-759	4,308	4.72	76,391	83.66
760-764	3,526	3.86	79,917	87.52
765-769	2,555	2.8	82,472	90.32
770-774	2,582	2.83	85,054	93.14
775-779	1,463	1.6	86,517	94.75
780-784	1,300	1.42	87,817	96.17
785-789	1,165	1.28	88,982	97.45
790-794	679	0.74	89,661	98.19
795-799	413	0.45	90,074	98.64
800-804	381	0.42	90,455	99.06
805-809	305	0.33	90,760	99.39
810-814	246	0.27	91,006	99.66
815-819	73	0.08	91,079	99.74
820-824	67	0.07	91,146	99.81
825-829	61	0.07	91,207	99.88
830-834	0	0	91,207	99.88
835-839	42	0.05	91,249	99.93
840-844	25	0.03	91,274	99.96
845-850	41	0.04	91,315	100

Table A.12.38 Scale Score Cumulative Frequencies: Mathematics Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,007	2.16	2,007	2.16
655-659	2,614	2.81	4,621	4.97
660-664	46	0.05	4,667	5.02
665-669	4,014	4.32	8,681	9.34
670-674	62	0.07	8,743	9.41
675-679	5,231	5.63	13,974	15.03
680-684	61	0.07	14,035	15.1
685-689	5,890	6.34	19,925	21.44
690-694	6,109	6.57	26,034	28.01
695-699	54	0.06	26,088	28.07
700-704	6,067	6.53	32,155	34.6
705-709	5,643	6.07	37,798	40.67
710-714	5,276	5.68	43,074	46.34
715-719	2,450	2.64	45,524	48.98
720-724	4,629	4.98	50,153	53.96
725-729	4,096	4.41	54,249	58.37
730-734	5,409	5.82	59,658	64.19
735-739	3,086	3.32	62,744	67.51
740-744	2,708	2.91	65,452	70.42
745-749	2,562	2.76	68,014	73.18
750-754	4,289	4.61	72,303	77.79
755-759	2,784	3	75,087	80.79
760-764	2,360	2.54	77,447	83.32
765-769	2,829	3.04	80,276	86.37
770-774	2,356	2.53	82,632	88.9
775-779	2,009	2.16	84,641	91.06
780-784	1,741	1.87	86,382	92.94
785-789	1,127	1.21	87,509	94.15
790-794	1,265	1.36	88,774	95.51
795-799	838	0.9	89,612	96.41
800-804	973	1.05	90,585	97.46
805-809	586	0.63	91,171	98.09
810-814	540	0.58	91,711	98.67
815-819	396	0.43	92,107	99.1
820-824	224	0.24	92,331	99.34
825-829	187	0.2	92,518	99.54
830-834	130	0.14	92,648	99.68
835-839	95	0.1	92,743	99.78
840-844	39	0.04	92,782	99.82
845-850	164	0.18	92,946	100

Table A.12.39 Scale Score Cumulative Frequencies: Algebra I

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	10	0.29	10	0.29
655-659	5	0.15	15	0.44
660-664	0	0	15	0.44
665-669	14	0.41	29	0.85
670-674	22	0.64	51	1.49
675-679	1	0.03	52	1.52
680-684	32	0.93	84	2.45
685-689	43	1.26	127	3.71
690-694	44	1.29	171	4.99
695-699	62	1.81	233	6.8
700-704	130	3.8	363	10.6
705-709	164	4.79	527	15.39
710-714	0	0	527	15.39
715-719	178	5.2	705	20.59
720-724	200	5.84	905	26.43
725-729	190	5.55	1,095	31.98
730-734	176	5.14	1,271	37.12
735-739	184	5.37	1,455	42.49
740-744	424	12.38	1,879	54.88
745-749	166	4.85	2,045	59.73
750-754	173	5.05	2,218	64.78
755-759	262	7.65	2,480	72.43
760-764	109	3.18	2,589	75.61
765-769	220	6.43	2,809	82.04
770-774	154	4.5	2,963	86.54
775-779	117	3.42	3,080	89.95
780-784	84	2.45	3,164	92.41
785-789	93	2.72	3,257	95.12
790-794	49	1.43	3,306	96.55
795-799	62	1.81	3,368	98.36
800-804	18	0.53	3,386	98.89
805-809	17	0.5	3,403	99.39
810-814	7	0.2	3,410	99.59
815-819	5	0.15	3,415	99.74
820-824	1	0.03	3,416	99.77
825-829	6	0.18	3,422	99.94
830-834	2	0.06	3,424	100
835-839	0	0	3,424	100
840-844	0	0	3,424	100
845-850	0	0	3,424	100

Table A.12.40 Scale Score Cumulative Frequencies: Geometry

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	4	0.14	4	0.14
655-659	1	0.03	5	0.17
660-664	0	0	5	0.17
665-669	10	0.34	15	0.51
670-674	0	0	15	0.51
675-679	9	0.31	24	0.82
680-684	17	0.58	41	1.4
685-689	62	2.12	103	3.52
690-694	0	0	103	3.52
695-699	86	2.94	189	6.47
700-704	122	4.18	311	10.64
705-709	139	4.76	450	15.4
710-714	162	5.54	612	20.94
715-719	172	5.89	784	26.83
720-724	163	5.58	947	32.41
725-729	171	5.85	1,118	38.26
730-734	293	10.03	1,411	48.29
735-739	135	4.62	1,546	52.91
740-744	255	8.73	1,801	61.64
745-749	167	5.72	1,968	67.35
750-754	183	6.26	2,151	73.61
755-759	227	7.77	2,378	81.38
760-764	154	5.27	2,532	86.65
765-769	145	4.96	2,677	91.62
770-774	94	3.22	2,771	94.83
775-779	78	2.67	2,849	97.5
780-784	24	0.82	2,873	98.32
785-789	27	0.92	2,900	99.25
790-794	14	0.48	2,914	99.73
795-799	2	0.07	2,916	99.79
800-804	4	0.14	2,920	99.93
805-809	2	0.07	2,922	100
810-814	0	0	2,922	100
815-819	0	0	2,922	100
820-824	0	0	2,922	100
825-829	0	0	2,922	100
830-834	0	0	2,922	100
835-839	0	0	2,922	100
840-844	0	0	2,922	100
845-850	0	0	2,922	100

Table A.12.41 Scale Score Cumulative Frequencies: Algebra II

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	56	2.05	56	2.05
655-659	0	0	56	2.05
660-664	66	2.42	122	4.48
665-669	41	1.5	163	5.98
670-674	0	0	163	5.98
675-679	137	5.03	300	11.01
680-684	0	0	300	11.01
685-689	169	6.2	469	17.2
690-694	0	0	469	17.2
695-699	170	6.24	639	23.44
700-704	114	4.18	753	27.62
705-709	75	2.75	828	30.37
710-714	175	6.42	1,003	36.79
715-719	145	5.32	1,148	42.11
720-724	136	4.99	1,284	47.1
725-729	127	4.66	1,411	51.76
730-734	129	4.73	1,540	56.49
735-739	165	6.05	1,705	62.55
740-744	117	4.29	1,822	66.84
745-749	168	6.16	1,990	73
750-754	92	3.37	2,082	76.38
755-759	93	3.41	2,175	79.79
760-764	115	4.22	2,290	84.01
765-769	106	3.89	2,396	87.89
770-774	79	2.9	2,475	90.79
775-779	49	1.8	2,524	92.59
780-784	47	1.72	2,571	94.31
785-789	45	1.65	2,616	95.96
790-794	41	1.5	2,657	97.47
795-799	20	0.73	2,677	98.2
800-804	12	0.44	2,689	98.64
805-809	7	0.26	2,696	98.9
810-814	10	0.37	2,706	99.27
815-819	6	0.22	2,712	99.49
820-824	2	0.07	2,714	99.56
825-829	1	0.04	2,715	99.6
830-834	2	0.07	2,717	99.67
835-839	2	0.07	2,719	99.74
840-844	2	0.07	2,721	99.82
845-850	5	0.18	2,726	100

Appendix 12.5: Subgroup Scale Score Performance

Table A.12.42 Subgroup Performance for ELA/L Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		96928	724.36	41.06	650	850
Gender	Female	47502	729.22	41.61	650	850
	Male	49342	719.66	39.96	650	850
Ethnicity	American Indian/Alaska Native	349	708.90	42.14	650	829
	Asian	5146	747.36	39.10	650	850
	Black/African American	11883	701.50	36.74	650	850
	Hispanic/Latino	21157	709.13	38.57	650	850
	Native Hawaiian/Pacific Islander	180	740.79	37.88	650	850
	Two or more races	4773	729.37	40.86	650	850
	White	52318	733.43	38.77	650	850
Economic status*	Not economically disadvantaged	50287	737.58	38.44	650	850
	Economically disadvantaged	40130	705.88	36.94	650	850
English learner status	Non English learner	75834	728.03	40.47	650	850
	English learner	15238	701.17	35.39	650	850
Disabilities	Students without disabilities	79922	729.48	39.88	650	850
	Students with disabilities	15824	698.60	37.17	650	850
Reading summative score		96928	41.75	16.81	10	90
Gender	Female	47502	43.32	16.88	10	90
	Male	49342	40.23	16.59	10	90
Ethnicity	American Indian/Alaska Native	349	35.49	16.86	10	90
	Asian	5146	50.87	16.14	10	90
	Black/African American	11883	32.92	15.23	10	90
	Hispanic/Latino	21157	35.65	15.78	10	90
	Native Hawaiian/Pacific Islander	180	47.10	15.09	10	87
	Two or more races	4773	43.71	16.77	10	90
	White	52318	45.32	15.95	10	90
Economic status*	Not economically disadvantaged	50287	47.16	15.87	10	90
	Economically disadvantaged	40130	34.29	15.03	10	90
English learner status	Non English learner	75834	43.33	16.60	10	90
	English learner	15238	32.03	14.13	10	90
Disabilities	Students without disabilities	79922	43.79	16.31	10	90
	Students with disabilities	15824	31.47	15.49	10	90
Writing summative score		96928	25.28	12.53	10	60

Group Type	Group	N	Mean	SD	Min	Max
Gender	Female	47502	26.98	12.59	10	60
	Male	49342	23.63	12.26	10	60
Ethnicity	American Indian/Alaska Native	349	21.45	12.53	10	53
	Asian	5146	31.45	11.80	10	60
	Black/African American	11883	18.93	11.05	10	60
	Hispanic/Latino	21157	21.37	11.81	10	60
	Native Hawaiian/Pacific Islander	180	30.56	12.13	10	60
	Two or more races	4773	26.66	12.50	10	60
	White	52318	27.68	12.20	10	60
Economic status*	Not economically disadvantaged	50287	28.56	12.14	10	60
	Economically disadvantaged	40130	20.52	11.51	10	60
English learner status	Non English learner	75834	26.04	12.51	10	60
	English learner	15238	19.91	11.20	10	60
Disabilities	Students without disabilities	79922	26.59	12.40	10	60
	Students with disabilities	15824	18.67	11.07	10	60

Note. ELA/L = English language arts/literacy, SD = standard deviation. This table is identical to Table 12.5 in Section 12. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.43 Subgroup Performance for ELA/L Scale Scores: Grade 4

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		99006	728.57	35.38	650	850
Gender	Female	48440	733.06	35.61	650	850
	Male	50476	724.25	34.61	650	850
Ethnicity	American Indian/Alaska Native	306	711.17	37.04	650	832
	Asian	5140	747.87	33.86	650	850
	Black/African American	12185	707.60	32.05	650	850
	Hispanic/Latino	21597	715.53	33.06	650	850
	Native Hawaiian/Pacific Islander	178	735.97	35.76	650	839
	Two or more races	4768	732.66	35.53	650	850
	White	53810	736.73	33.11	650	850
Economic status*	Not economically disadvantaged	51618	740.04	32.88	650	850
	Economically disadvantaged	41011	712.39	32.01	650	850
English learner status	Non English learner	78645	731.82	34.79	650	850
	English learner	14571	706.15	29.46	650	825
Disabilities	Students without disabilities	80997	733.59	33.85	650	850
	Students with disabilities	16865	704.53	32.81	650	850
Reading summative score		99006	43.65	14.67	10	90
Gender	Female	48440	44.97	14.58	10	90
	Male	50476	42.38	14.64	10	90
Ethnicity	American Indian/Alaska Native	306	36.69	15.38	10	87
	Asian	5140	51.42	14.08	10	90
	Black/African American	12185	35.52	13.43	10	90
	Hispanic/Latino	21597	38.45	13.71	10	90
	Native Hawaiian/Pacific Islander	178	45.99	15.14	10	90
	Two or more races	4768	45.24	14.72	10	90
	White	53810	46.84	13.86	10	90
Economic status*	Not economically disadvantaged	51618	48.31	13.76	10	90
	Economically disadvantaged	41011	37.13	13.25	10	90
English learner status	Non English learner	78645	45.03	14.47	10	90
	English learner	14571	34.31	11.93	10	81
Disabilities	Students without disabilities	80997	45.67	14.03	10	90
	Students with disabilities	16865	33.94	13.81	10	90
Writing summative score		99006	25.54	11.69	10	60
Gender	Female	48440	27.51	11.57	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	50476	23.64	11.50	10	60
Ethnicity	American Indian/Alaska Native	306	20.82	11.41	10	57
	Asian	5140	31.04	10.76	10	60
	Black/African American	12185	19.11	10.51	10	57
	Hispanic/Latino	21597	21.75	11.17	10	60
	Native Hawaiian/Pacific Islander	178	28.85	11.08	10	51
	Two or more races	4768	26.72	11.76	10	60
	White	53810	28.00	11.17	10	60
Economic status*	Not economically disadvantaged	51618	28.75	11.07	10	60
	Economically disadvantaged	41011	20.92	10.95	10	57
English learner status	Non English learner	78645	26.34	11.60	10	60
	English learner	14571	19.72	10.47	10	54
Disabilities	Students without disabilities	80997	26.95	11.43	10	60
	Students with disabilities	16865	18.80	10.59	10	60

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.44 Subgroup Performance for ELA/L Scale Scores: Grade 5

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		99632	731.20	34.06	650	850
Gender	Female	48401	736.28	34.54	650	850
	Male	51141	726.35	32.87	650	850
Ethnicity	American Indian/Alaska Native	303	710.42	34.67	650	830
	Asian	5145	752.32	33.31	650	850
	Black/African American	12258	711.88	30.58	650	850
	Hispanic/Latino	21887	720.49	31.96	650	850
	Native Hawaiian/Pacific Islander	159	741.55	34.08	662	819
	Two or more races	4570	735.38	34.18	650	846
	White	54294	737.89	32.31	650	850
Economic status*	Not economically disadvantaged	52355	741.83	31.83	650	850
	Economically disadvantaged	41395	715.99	30.78	650	842
English learner status	Non English learner	82839	734.01	33.19	650	850
	English learner	11368	704.20	26.46	650	819
Disabilities	Students without disabilities	81371	736.41	32.25	650	850
	Students with disabilities	17191	706.35	31.45	650	842
Reading summative score		99632	44.50	14.11	10	90
Gender	Female	48401	45.94	14.20	10	90
	Male	51141	43.12	13.88	10	90
Ethnicity	American Indian/Alaska Native	303	36.18	14.37	10	87
	Asian	5145	53.20	14.16	10	90
	Black/African American	12258	37.14	12.91	10	90
	Hispanic/Latino	21887	40.16	13.22	10	90
	Native Hawaiian/Pacific Islander	159	47.23	13.42	17	78
	Two or more races	4570	46.27	14.26	10	90
	White	54294	47.08	13.44	10	90
Economic status*	Not economically disadvantaged	52355	48.84	13.34	10	90
	Economically disadvantaged	41395	38.38	12.74	10	90
English learner status	Non English learner	82839	45.71	13.80	10	90
	English learner	11368	33.30	10.75	10	81
Disabilities	Students without disabilities	81371	46.55	13.37	10	90
	Students with disabilities	17191	34.72	13.44	10	88
Writing summative score		99632	25.61	12.30	10	60
Gender	Female	48401	28.01	12.09	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	51141	23.33	12.06	10	60
Ethnicity	American Indian/Alaska Native	303	19.16	11.76	10	49
	Asian	5145	32.07	10.86	10	60
	Black/African American	12258	19.06	11.08	10	60
	Hispanic/Latino	21887	22.41	11.91	10	60
	Native Hawaiian/Pacific Islander	159	29.86	12.47	10	53
	Two or more races	4570	26.76	12.25	10	60
	White	54294	27.79	11.90	10	60
Economic status*	Not economically disadvantaged	52355	28.83	11.64	10	60
	Economically disadvantaged	41395	20.92	11.61	10	58
English learner status	Non English learner	82839	26.37	12.15	10	60
	English learner	11368	17.92	10.46	10	53
Disabilities	Students without disabilities	81371	27.25	11.97	10	60
	Students with disabilities	17191	17.77	10.75	10	58

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.45 Subgroup Performance for ELA/L Scale Scores: Grade 6

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		98590	733.68	31.38	650	850
Gender	Female	48053	738.33	31.47	650	850
	Male	50445	729.23	30.63	650	850
Ethnicity	American Indian/Alaska Native	339	721.24	31.54	650	850
	Asian	5002	753.80	30.93	650	850
	Black/African American	12128	716.35	28.50	650	825
	Hispanic/Latino	21624	724.19	30.01	650	839
	Native Hawaiian/Pacific Islander	187	742.92	28.68	668	806
	Two or more races	4418	735.96	32.15	650	850
	White	53826	739.69	29.51	650	850
Economic status*	Not economically disadvantaged	52467	743.03	29.21	650	850
	Economically disadvantaged	40404	720.30	29.06	650	850
English learner status	Non English learner	84004	736.26	30.30	650	850
	English learner	9231	704.41	24.24	650	819
Disabilities	Students without disabilities	80872	738.65	29.43	650	850
	Students with disabilities	16660	709.42	29.27	650	839
Reading summative score		98590	45.01	13.14	10	90
Gender	Female	48053	46.39	13.00	10	90
	Male	50445	43.68	13.12	10	90
Ethnicity	American Indian/Alaska Native	339	39.95	12.94	10	90
	Asian	5002	53.00	13.14	10	90
	Black/African American	12128	38.31	12.14	10	90
	Hispanic/Latino	21624	41.17	12.57	10	90
	Native Hawaiian/Pacific Islander	187	48.11	11.87	18	78
	Two or more races	4418	45.97	13.44	10	90
	White	53826	47.37	12.45	10	90
Economic status*	Not economically disadvantaged	52467	48.80	12.37	10	90
	Economically disadvantaged	40404	39.61	12.17	10	90
English learner status	Non English learner	84004	46.10	12.74	10	90
	English learner	9231	32.84	9.90	10	85
Disabilities	Students without disabilities	80872	46.96	12.40	10	90
	Students with disabilities	16660	35.46	12.45	10	84
Writing summative score		98590	25.45	12.49	10	60
Gender	Female	48053	27.83	12.19	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	50445	23.17	12.34	10	60
Ethnicity	American Indian/Alaska Native	339	21.17	12.47	10	60
	Asian	5002	32.62	10.94	10	60
	Black/African American	12128	18.70	11.26	10	52
	Hispanic/Latino	21624	22.28	12.10	10	60
	Native Hawaiian/Pacific Islander	187	29.87	11.34	10	52
	Two or more races	4418	26.19	12.56	10	60
	White	53826	27.65	12.04	10	60
Economic status*	Not economically disadvantaged	52467	28.70	11.83	10	60
	Economically disadvantaged	40404	20.76	11.84	10	60
English learner status	Non English learner	84004	26.23	12.33	10	60
	English learner	9231	16.28	9.86	10	49
Disabilities	Students without disabilities	80872	27.16	12.13	10	60
	Students with disabilities	16660	17.09	10.71	10	60

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.46 Subgroup Performance for ELA/L Scale Scores: Grade 7

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		96950	733.47	37.42	650	850
Gender	Female	47049	739.61	37.32	650	850
	Male	49784	727.63	36.57	650	850
Ethnicity	American Indian/Alaska Native	343	716.10	38.97	650	847
	Asian	4854	757.98	36.32	650	850
	Black/African American	12258	712.96	34.42	650	850
	Hispanic/Latino	20780	723.13	35.23	650	850
	Native Hawaiian/Pacific Islander	149	745.01	35.47	650	828
	Two or more races	4088	736.02	37.22	650	850
	White	53459	740.24	35.65	650	850
Economic status*	Not economically disadvantaged	52152	744.47	34.98	650	850
	Economically disadvantaged	39620	717.78	34.82	650	850
English learner status	Non English learner	83666	736.19	36.41	650	850
	English learner	8355	699.67	29.23	650	809
Disabilities	Students without disabilities	79634	739.56	34.85	650	850
	Students with disabilities	16268	703.55	35.27	650	850
Reading summative score		96950	45.73	15.79	10	90
Gender	Female	47049	47.38	15.61	10	90
	Male	49784	44.15	15.80	10	90
Ethnicity	American Indian/Alaska Native	343	38.80	16.29	10	90
	Asian	4854	55.46	15.64	10	90
	Black/African American	12258	37.68	14.63	10	90
	Hispanic/Latino	20780	41.49	14.91	10	90
	Native Hawaiian/Pacific Islander	149	49.54	15.15	10	90
	Two or more races	4088	46.97	15.71	10	90
	White	53459	48.42	15.14	10	90
Economic status*	Not economically disadvantaged	52152	50.28	14.91	10	90
	Economically disadvantaged	39620	39.26	14.61	10	90
English learner status	Non English learner	83666	46.90	15.39	10	90
	English learner	8355	31.40	11.95	10	90
Disabilities	Students without disabilities	79634	48.17	14.77	10	90
	Students with disabilities	16268	33.73	15.17	10	90
Writing summative score		96950	27.34	12.01	10	60
Gender	Female	47049	29.94	11.57	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	49784	24.87	11.89	10	60
Ethnicity	American Indian/Alaska Native	343	22.13	12.25	10	60
	Asian	4854	34.47	10.22	10	60
	Black/African American	12258	21.17	11.38	10	60
	Hispanic/Latino	20780	24.62	11.62	10	60
	Native Hawaiian/Pacific Islander	149	31.44	11.01	10	51
	Two or more races	4088	27.77	11.99	10	60
	White	53459	29.28	11.56	10	60
Economic status*	Not economically disadvantaged	52152	30.37	11.26	10	60
	Economically disadvantaged	39620	23.00	11.66	10	60
English learner status	Non English learner	83666	28.02	11.81	10	60
	English learner	8355	18.64	10.36	10	53
Disabilities	Students without disabilities	79634	29.11	11.41	10	60
	Students with disabilities	16268	18.63	11.02	10	60

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.47 Subgroup Performance for ELA/L Scale Scores: Grade 8

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		96028	734.91	37.11	650	850
Gender	Female	46187	742.52	36.84	650	850
	Male	49747	727.82	35.94	650	850
Ethnicity	American Indian/Alaska Native	274	718.88	36.83	650	826
	Asian	4478	761.10	36.20	650	850
	Black/African American	12501	715.40	34.10	650	850
	Hispanic/Latino	20816	725.67	35.24	650	850
	Native Hawaiian/Pacific Islander	161	748.71	33.26	659	850
	Two or more races	3942	737.27	37.28	650	850
	White	52837	741.20	35.45	650	850
Economic status*	Not economically disadvantaged	51714	745.01	35.13	650	850
	Economically disadvantaged	39318	720.47	34.86	650	850
English learner status	Non English learner	84167	737.19	36.23	650	850
	English learner	7110	700.41	29.04	650	822
Disabilities	Students without disabilities	78523	741.09	34.57	650	850
	Students with disabilities	16502	705.33	34.51	650	850
Reading summative score		96028	46.24	15.70	10	90
Gender	Female	46187	48.60	15.56	10	90
	Male	49747	44.03	15.50	10	90
Ethnicity	American Indian/Alaska Native	274	39.28	15.66	10	83
	Asian	4478	56.85	15.55	10	90
	Black/African American	12501	38.86	14.71	10	90
	Hispanic/Latino	20816	42.45	15.03	10	90
	Native Hawaiian/Pacific Islander	161	51.28	14.48	10	90
	Two or more races	3942	47.42	15.78	10	90
	White	52837	48.66	15.06	10	90
Economic status*	Not economically disadvantaged	51714	50.38	14.97	10	90
	Economically disadvantaged	39318	40.34	14.73	10	90
English learner status	Non English learner	84167	47.24	15.34	10	90
	English learner	7110	31.47	11.88	10	88
Disabilities	Students without disabilities	78523	48.72	14.71	10	90
	Students with disabilities	16502	34.37	14.82	10	90
Writing summative score		96028	27.36	12.30	10	60
Gender	Female	46187	30.42	11.66	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	49747	24.52	12.20	10	60
Ethnicity	American Indian/Alaska Native	274	23.25	12.36	10	55
	Asian	4478	35.07	10.37	10	60
	Black/African American	12501	20.83	11.64	10	60
	Hispanic/Latino	20816	24.91	11.95	10	60
	Native Hawaiian/Pacific Islander	161	31.96	11.25	10	60
	Two or more races	3942	27.82	12.29	10	60
	White	52837	29.33	11.83	10	60
Economic status*	Not economically disadvantaged	51714	30.29	11.61	10	60
	Economically disadvantaged	39318	23.15	12.02	10	60
English learner status	Non English learner	84167	27.91	12.17	10	60
	English learner	7110	18.65	10.56	10	53
Disabilities	Students without disabilities	78523	29.18	11.73	10	60
	Students with disabilities	16502	18.65	11.17	10	60

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.48 Subgroup Performance for ELA/L Scale Scores: Grade 10

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		2767	757.37	40.11	650	850
Gender	Female	1347	764.34	38.37	650	850
	Male	1376	749.99	40.68	650	850
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	757.54	41.64	650	850
	Black/African American	263	740.13	40.24	650	850
	Hispanic/Latino	605	750.58	39.38	650	850
	Native Hawaiian/Pacific Islander	54	752.93	39.47	650	838
	Two or more races	371	760.51	40.15	650	850
	White	1139	763.81	38.70	650	850
Economic status*	Not economically disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner					
	English learner	143	714.02	38.25	650	795
Disabilities	Students without disabilities	2358	761.89	38.25	650	850
	Students with disabilities	365	726.06	38.58	650	850
Reading summative score		2767	54.81	17.24	10	90
Gender	Female	1347	56.74	16.79	10	90
	Male	1376	52.74	17.51	10	90
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	53.92	17.75	10	90
	Black/African American	263	47.46	16.37	10	90
	Hispanic/Latino	605	52.37	16.89	10	90
	Native Hawaiian/Pacific Islander	54	50.98	15.66	11	90
	Two or more races	371	55.54	17.12	10	90
	White	1139	57.85	16.93	10	90
Economic status*	Not economically disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner					
	English learner	143	36.67	16.20	10	85
Disabilities	Students without disabilities	2358	56.57	16.64	10	90
	Students with disabilities	365	42.73	16.48	10	90
Writing summative score		2767	33.95	11.48	10	60
Gender	Female	1347	36.37	10.40	10	60

Group Type	Group	N	Mean	SD	Min	Max
	Male	1376	31.41	12.01	10	60
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	211	34.68	11.45	10	60
	Black/African American	263	29.80	12.09	10	55
	Hispanic/Latino	605	32.19	11.42	10	60
	Native Hawaiian/Pacific Islander	54	34.28	11.50	10	57
	Two or more races	371	35.11	11.21	10	60
	White	1139	35.15	11.29	10	60
Economic status*	Not economically disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner					
	English learner	143	24.06	11.73	10	42
Disabilities	Students without disabilities	2358	35.18	10.81	10	60
	Students with disabilities	365	25.35	12.23	10	57

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.49 Subgroup Performance for ELA/L Scale Scores: Grade 11

Group Type	Group	N	Mean	SD	Min	Max
Full summative score		413	712.76	31.35	650	833
Gender	Female	211	715.52	32.97	650	803
	Male	193	709.84	29.50	650	833
Ethnicity	American Indian/Alaska Native	61	705.08	26.15	650	760
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	n/a	n/a	n/a	n/a	n/a
	Hispanic/Latino	n/r	n/r	n/r	n/r	n/r
	Native Hawaiian/Pacific Islander	n/a	n/a	n/a	n/a	n/a
	Two or more races	n/r	n/r	n/r	n/r	n/r
	White	n/a	n/a	n/a	n/a	n/a
Economic status*	Not economically disadvantaged	n/a	n/a	n/a	n/a	n/a
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner					
	English learner	29	703.86	37.01	650	803
Disabilities	Students without disabilities	345	716.59	30.87	650	833
	Students with disabilities	68	693.35	26.31	650	792
Reading summative score		413	37.72	13.18	10	77
Gender	Female	211	38.36	13.62	10	77
	Male	193	37.16	12.82	10	77
Ethnicity	American Indian/Alaska Native	61	34.30	11.51	10	60
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	n/a	n/a	n/a	n/a	n/a
	Hispanic/Latino	n/r	n/r	n/r	n/r	n/r
	Native Hawaiian/Pacific Islander	n/a	n/a	n/a	n/a	n/a
	Two or more races	n/r	n/r	n/r	n/r	n/r
	White	n/a	n/a	n/a	n/a	n/a
Economic status*	Not economically disadvantaged	n/a	n/a	n/a	n/a	n/a
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner	n/r	n/r	n/r	n/r	n/r
	English learner	29	32.48	15.30	10	77
Disabilities	Students without disabilities	345	39.24	13.01	10	77
	Students with disabilities	68	30.00	11.31	10	70
Writing summative score		413	18.76	11.53	10	56
Gender	Female	211	20.35	12.08	10	43

Group Type	Group	N	Mean	SD	Min	Max
	Male	193	16.92	10.66	10	56
Ethnicity	American Indian/Alaska Native	61	17.05	10.15	10	36
	Asian	n/r	n/r	n/r	n/r	n/r
	Black/African American	n/a	n/a	n/a	n/a	n/a
	Hispanic/Latino	n/r	n/r	n/r	n/r	n/r
	Native Hawaiian/Pacific Islander	n/a	n/a	n/a	n/a	n/a
	Two or more races	n/r	n/r	n/r	n/r	n/r
	White	n/a	n/a	n/a	n/a	n/a
Economic status*	Not economically disadvantaged	n/a	n/a	n/a	n/a	n/a
	Economically disadvantaged	n/r	n/r	n/r	n/r	n/r
English learner status	Non English learner					
	English learner	29	19.97	11.94	10	43
Disabilities	Students without disabilities	345	19.90	11.83	10	56
	Students with disabilities	68	12.94	7.61	10	42

Note. ELA/L = English language arts/literacy, SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.50 Subgroup Performance for Mathematics Scale Scores: Grade 3

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		96011	730.93	38.66	650	850
Gender	Female	47029	729.42	37.71	650	850
	Male	48909	732.36	39.50	650	850
Ethnicity	American Indian/Alaska Native	339	714.08	38.53	650	830
	Asian	5130	760.21	39.24	650	850
	Black/African American	11609	703.07	32.33	650	850
	Hispanic/Latino	20914	714.39	33.84	650	850
	Native Hawaiian/Pacific Islander	179	738.60	34.18	670	814
	Two or more races	4739	734.44	38.81	650	850
	White	52020	741.11	35.37	650	850
Economic Status*	Not Economically Disadvantaged	49969	745.31	36.08	650	850
	Economically Disadvantaged	39577	711.07	32.85	650	850
English Learner Status	Non English Learner	75112	734.10	38.50	650	850
	English Learner	15095	710.97	32.90	650	850
Disabilities	Students without Disabilities	79199	734.96	37.82	650	850
	Students with Disabilities	15632	710.76	36.67	650	850
Language Form	Spanish	1729	700.02	26.93	650	810

Note. This table is identical to Table 12.7 in Section 12. SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.51 Subgroup Performance for Mathematics Scale Scores: Grade 4

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		97740	726.00	34.73	650	850
Gender	Female	47855	724.80	33.75	650	850
	Male	49816	727.14	35.61	650	850
Ethnicity	American Indian/Alaska Native	310	710.96	35.03	650	818
	Asian	5094	754.31	35.42	650	850
	Black/African American	11813	701.45	28.56	650	836
	Hispanic/Latino	21286	711.49	30.37	650	850
	Native Hawaiian/Pacific Islander	176	733.07	33.30	650	825
	Two or more races	4707	730.37	35.47	650	850
	White	53377	734.56	31.96	650	850
Economic Status*	Not Economically Disadvantaged	51165	738.47	32.57	650	850
	Economically Disadvantaged	40244	708.11	29.25	650	850
English Learner Status	Non English Learner	77643	728.55	34.53	650	850
	English Learner	14357	706.90	28.61	650	849
Disabilities	Students without Disabilities	80032	729.89	34.02	650	850
	Students with Disabilities	16588	707.46	32.24	650	850
Language Form	Spanish	1504	694.98	24.22	650	796

Note. SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.52 Subgroup Performance for Mathematics Scale Scores: Grade 5

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		98306	726.90	33.87	650	850
Gender	Female	47761	726.38	32.45	650	850
	Male	50477	727.38	35.15	650	850
Ethnicity	American Indian/Alaska Native	297	709.60	30.36	650	824
	Asian	5086	755.80	36.96	650	850
	Black/African American	11880	704.22	25.83	650	837
	Hispanic/Latino	21467	714.69	28.93	650	850
	Native Hawaiian/Pacific Islander	157	735.67	34.16	651	819
	Two or more races	4548	730.56	35.20	650	850
	White	53926	734.08	32.25	650	850
Economic Status*	Not Economically Disadvantaged	51937	738.31	33.07	650	850
	Economically Disadvantaged	40571	710.34	27.29	650	850
English Learner Status	Non English Learner	81824	729.00	33.69	650	850
	English Learner	11138	704.61	24.20	650	835
Disabilities	Students without Disabilities	80387	730.75	33.47	650	850
	Students with Disabilities	16887	708.62	29.63	650	850
Language Form	Spanish	1305	700.90	24.19	650	799

Note. SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.53 Subgroup Performance for Mathematics Scale Scores: Grade 6

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		96924	724.92	32.04	650	850
Gender	Female	47233	724.62	31.06	650	850
	Male	49626	725.18	32.94	650	850
Ethnicity	American Indian/Alaska Native	337	711.70	29.75	650	806
	Asian	4940	754.10	34.39	650	850
	Black/African American	11696	703.43	25.78	650	834
	Hispanic/Latino	21166	713.24	27.89	650	834
	Native Hawaiian/Pacific Islander	183	731.60	28.96	650	802
	Two or more races	4347	726.51	33.39	650	850
	White	53282	731.79	29.97	650	850
Economic Status*	Not Economically Disadvantaged	51875	735.74	30.93	650	850
	Economically Disadvantaged	39508	709.68	27.14	650	850
English Learner Status	Non English Learner	82732	727.29	31.63	650	850
	English Learner	9008	698.41	22.94	650	808
Disabilities	Students without Disabilities	79552	729.22	30.92	650	850
	Students with Disabilities	16358	703.99	29.09	650	850
Language Form	Spanish	961	700.27	23.65	650	793

Note. SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.54 Subgroup Performance for Mathematics Scale Scores: Grade 7

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		91315	732.06	28.12	650	850
Gender	Female	44279	731.65	27.30	650	850
	Male	47019	732.46	28.86	650	850
Ethnicity	American Indian/Alaska Native	316	718.41	27.12	650	814
	Asian	4534	759.05	31.10	650	850
	Black/African American	11429	714.08	23.14	650	840
	Hispanic/Latino	19406	723.19	24.64	650	827
	Native Hawaiian/Pacific Islander	87	734.46	30.83	666	811
	Two or more races	3476	731.87	29.43	650	850
	White	51244	737.45	26.43	650	850
Economic Status*	Not Economically Disadvantaged	51505	741.56	27.16	650	850
	Economically Disadvantaged	38688	719.92	24.30	650	850
English Learner Status	Non English Learner	82303	734.36	27.84	650	850
	English Learner	7805	710.00	20.02	650	822
Disabilities	Students without Disabilities	74932	736.11	26.86	650	850
	Students with Disabilities	15447	712.45	25.78	650	850
Language Form	Spanish	367	707.19	19.30	650	774

Note. SD = standard deviation. *Economic status was based on participation in National School Lunch Program, which provides for receipt of free or reduced-price lunch.

Table A.12.55 Subgroup Performance for Mathematics Scale Scores: Grade 8

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		92946	723.42	39.41	650	850
Gender	Female	44707	724.54	38.37	650	850
	Male	48187	722.38	40.32	650	850
Ethnicity	American Indian/Alaska Native	261	707.71	35.02	650	824
	Asian	4325	762.02	43.60	650	850
	Black/African American	11898	698.62	31.44	650	850
	Hispanic/Latino	20119	712.29	34.46	650	850
	Native Hawaiian/Pacific Islander	143	733.70	35.12	650	807
	Two or more races	3703	723.19	39.77	650	850
	White	51577	730.75	37.77	650	850
Economic Status*	Not Economically Disadvantaged	50913	736.41	38.97	650	850
	Economically Disadvantaged	38331	706.66	33.60	650	850
English Learner Status	Non English Learner	82612	726.08	39.42	650	850
	English Learner	6825	693.14	26.93	650	850
Disabilities	Students without Disabilities	75962	728.92	38.18	650	850
	Students with Disabilities	16045	697.40	34.44	650	850
Language Form	Spanish	290	694.47	31.25	650	804

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

Table A.12.56 Subgroup Performance for Mathematics Scale Scores: Algebra I

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		3424	742.39	29.14	650	833
Gender	Female	1626	741.62	28.33	650	829
	Male	1756	743.20	29.90	650	833
Ethnicity	American Indian/Alaska Native	25	716.00	30.81	668	793
	Asian	252	748.23	30.70	650	832
	Black/African American	321	729.98	24.43	668	797
	Hispanic/Latino	679	737.02	27.88	650	833
	Native Hawaiian/Pacific Islander	65	733.88	28.28	659	787
	Two or more races	493	744.62	28.55	674	828
	White	1376	749.01	27.93	650	829
Economic Status*	Not Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
English Learner Status	Non English Learner					
	English Learner	244	726.47	30.09	650	829
Disabilities	Students without Disabilities	2956	745.85	27.56	650	832
	Students with Disabilities	427	718.73	28.97	650	833

Note. This table is identical to Table 12.8 in Section 12. SD = standard deviation. *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 low sample size (n < 100) or missing demographic information.

Table A.12.57 Subgroup Performance for Mathematics Scale Scores: Geometry

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		2922	736.48	24.37	650	806
Gender	Female	1388	735.93	23.61	650	802
	Male	1510	737.01	25.05	667	806
Ethnicity	American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r
	Asian	230	745.56	24.40	680	790
	Black/African American	272	724.43	22.91	650	781
	Hispanic/Latino	611	730.80	23.31	667	806
	Native Hawaiian/Pacific Islander	55	736.24	22.62	695	778
	Two or more races	392	740.76	23.40	667	802
	White	1200	740.71	22.78	650	802
Economic Status*	Not Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
English Learner Status	Non English Learner	n/r	n/r	n/r	n/r	n/r
	English Learner	167	725.43	24.58	675	789
Disabilities	Students without Disabilities	2531	739.01	23.56	650	806
	Students with Disabilities	370	719.00	22.76	650	796

Note. SD = standard deviation. *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 low sample size (n < 100) or missing demographic information.

Table A.12.58 Subgroup Performance for Mathematics Scale Scores: Algebra II

Group Type	Group	N	Mean	SD	Min	Max
Full Summative Score		2726	727.71	35.90	650	850
Gender	Female	1354	725.21	33.40	650	833
	Male	1343	730.13	38.09	650	850
Ethnicity	American Indian/Alaska Native	57	690.49	27.00	650	777
	Asian	225	737.30	36.11	650	844
	Black/African American	227	714.48	30.01	650	793
	Hispanic/Latino	526	720.99	33.43	650	802
	Native Hawaiian/Pacific Islander	45	731.27	32.40	650	782
	Two or more races	343	731.72	35.20	650	850
	White	1058	737.14	34.51	650	850
Economic Status*	Not Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
	Economically Disadvantaged	n/r	n/r	n/r	n/r	n/r
English Learner Status	Non English Learner	n/r	n/r	n/r	n/r	n/r
	English Learner	126	711.29	32.88	650	813
Disabilities	Students without Disabilities	2415	730.49	35.07	650	850
	Students with Disabilities	285	703.54	33.68	650	802

Note. SD = standard deviation. *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 low sample size (n < 100) or missing demographic information.

Appendix 13.1: Reliability 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	3.73	0.86	1739	0.77	44314	0.87
Gender							
Male	54	3.6	0.86	1142	0.76	506	0.87
Female	54	3.85	0.86	595	0.79	21802	0.86
Ethnicity							
White	54	3.9	0.84	918	0.79	141	0.85
Black/African American	54	3.28	0.85	250	0.77	5432	0.86
Asian/Pacific Islander	54	4.25	0.82	2377	0.81	2407	0.82
American Indian/Alaska Native	53	3.31	0.89	144	0.88	152	0.9
Hispanic/Latino	54	3.44	0.86	392	0.73	9686	0.87
Multiple	53	4.03	0.84	1927	0.84	1931	0.84
Special instruction needs							
Economically disadvantaged	54	3.36	0.85	1080	0.74	19208	0.86
Not economically disadvantaged	54	3.76	0.85	594	0.81	24602	0.86
English learner	54	3.33	0.83	353	0.65	6967	0.85
Non-English learner	54	3.65	0.87	1319	0.79	36834	0.87
Students with disabilities	54	3.15	0.86	1684	0.76	6320	0.87
Students without disabilities	53	3.85	0.85	301	0.84	37501	0.85
Students taking accommodated forms							
American Sign Language	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	54	2.66	0.75	1487	0.75	1487	0.75

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

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	4.41	0.86	1901	0.75	44291	0.87
Gender							
Male	70	4.26	0.86	1193	0.75	23125	0.87
Female	70	4.55	0.86	707	0.76	21866	0.87
Ethnicity							
White	70	4.57	0.84	1028	0.77	24671	0.85
Black/African American	70	3.88	0.85	266	0.66	5579	0.85
Asian/Pacific Islander	71	4.89	0.84	2376	0.83	2444	0.84
American Indian/Alaska Native	70	3.89	0.89	126	0.89	117	0.89
Hispanic/Latino	70	4.14	0.85	428	0.7	9880	0.86
Multiple	70	4.86	0.84	1935	0.84	1942	0.84
Special instruction needs							
Economically disadvantaged	70	3.92	0.85	1200	0.7	19675	0.85
Not economically disadvantaged	70	4.33	0.86	630	0.8	24973	0.87
English learner	70	3.94	0.8	352	0.62	6758	0.81
Non-English learner	70	4.22	0.87	1477	0.77	37536	0.88
Students with disabilities	70	3.73	0.85	1842	0.74	6882	0.87
Students without disabilities	70	4.56	0.85	266	0.85	37308	0.85
Students taking accommodated forms							
American Sign Language	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	67	3.27	0.74	1666	0.74	1666	0.74

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

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	70	4.37	0.87	1951	0.72	45896	0.88
Gender							
Male	70	4.2	0.86	1271	0.71	23454	0.88
Female	70	4.53	0.87	679	0.74	22436	0.88
Ethnicity							
White	70	4.51	0.85	1006	0.76	25433	0.87
Black/African American	70	3.86	0.84	299	0.67	5678	0.87
Asian/Pacific Islander	69	4.8	0.85	2446	0.84	2361	0.86
American Indian/Alaska Native	69	3.86	0.88	132	0.85	130	0.91
Hispanic/Latino	70	4.17	0.85	452	0.6	9994	0.87
Multiple	70	4.75	0.85	1836	0.84	1919	0.86
Special instruction needs							
Economically disadvantaged	70	3.94	0.84	1227	0.68	19713	0.87
Not economically disadvantaged	70	4.33	0.86	639	0.75	25729	0.88
English learner	70	3.75	0.76	376	0.58	5093	0.81
Non-English learner	70	4.23	0.87	1493	0.74	40321	0.89
Students with disabilities	70	3.65	0.85	1895	0.71	7069	0.89
Students without disabilities	70	4.53	0.86	288	0.81	38381	0.87
Students taking accommodated forms							
American Sign Language	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	3.3	0.71	1762	0.71	1762	0.71

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

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	4.52	0.88	1823	0.76	238	0.89
Gender							
Male	72	4.33	0.88	1163	0.75	23130	0.89
Female	72	4.71	0.88	660	0.76	104	0.91
Ethnicity							
White	72	4.66	0.87	949	0.78	25173	0.87
Black/African American	72	3.97	0.86	281	0.57	5517	0.87
Asian/Pacific Islander	72	5.11	0.86	2366	0.85	2340	0.87
American Indian/Alaska Native	72	4.18	0.89	147	0.88	152	0.9
Hispanic/Latino	72	4.31	0.87	403	0.65	9942	0.88
Multiple	72	4.87	0.87	1846	0.87	1839	0.88
Special instruction needs							
Economically disadvantaged	72	4.05	0.87	1102	0.67	19300	0.88
Not economically disadvantaged	72	4.5	0.88	638	0.81	25602	0.88
English learner	72	3.65	0.79	327	0.6	4151	0.81
Non-English learner	72	4.38	0.88	1416	0.77	40698	0.89
Students with disabilities	72	3.68	0.87	1782	0.73	6851	0.89
Students without disabilities	72	4.71	0.87	38582	0.86	38064	0.87
Students taking accommodated forms							
American Sign Language	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	70	3.09	0.73	1668	0.73	1668	0.73

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

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	4.71	0.89	1890	0.84	127	0.92
Gender							
Male	72	4.51	0.89	1200	0.82	23120	0.9
Female	72	4.89	0.88	689	0.85	22004	0.9
Ethnicity							
White	72	4.81	0.88	1007	0.83	25364	0.89
Black/African American	72	4.25	0.87	299	0.86	5573	0.88
Asian/Pacific Islander	72	5.21	0.87	2283	0.85	2256	0.89
American Indian/Alaska Native	72	4.41	0.89	129	0.88	156	0.91
Hispanic/Latino	72	4.54	0.87	377	0.81	9617	0.88
Multiple	72	5.13	0.87	1684	0.86	1762	0.88
Special instruction needs							
Economically disadvantaged	72	4.32	0.87	1099	0.8	18917	0.88
Not economically disadvantaged	72	4.65	0.89	694	0.86	25723	0.9
English learner	72	4.02	0.78	296	0.71	3838	0.79
Non-English learner	72	4.57	0.89	1498	0.85	105	0.92
Students with disabilities	72	3.93	0.88	1840	0.8	127	0.92
Students without disabilities	72	4.88	0.88	37970	0.86	37869	0.89
Students taking accommodated forms							
American Sign Language	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-caption	74	4.03	0.92	127	0.92	127	0.92
Screen reader	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-speech	74	3.41	0.79	1738	0.79	1738	0.79

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

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	4.82	0.88	242	0.85	130	0.93
Gender							
Male	72	4.61	0.88	141	0.84	23166	0.89
Female	72	5.04	0.87	101	0.82	21641	0.89
Ethnicity							
White	72	4.92	0.87	24869	0.85	25088	0.89
Black/African American	72	4.36	0.87	5826	0.86	5808	0.88
Asian/Pacific Islander	72	5.43	0.85	2062	0.81	2124	0.88
American Indian/Alaska Native	72	4.52	0.88	125	0.87	118	0.89
Hispanic/Latino	72	4.66	0.87	9788	0.85	9625	0.89
Multiple	72	5.23	0.86	1676	0.84	1650	0.88
Special instruction needs							
Economically disadvantaged	72	4.43	0.88	138	0.8	18882	0.89
Not economically disadvantaged	72	4.7	0.88	25285	0.87	25478	0.89
English learner	72	4.13	0.8	3154	0.79	3217	0.82
Non-English learner	72	4.64	0.89	138	0.86	114	0.93
Students with disabilities	72	4.12	0.88	166	0.84	130	0.93
Students without disabilities	72	4.98	0.86	37441	0.84	37510	0.88
Students taking accommodated forms							
American Sign Language	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Closed-caption	74	4.51	0.93	130	0.93	130	0.93
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

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.7 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	70	5.51	0.82	1888	0.8	868	0.87
Gender							
Male	70	5.43	0.83	939	0.81	430	0.87
Female	70	5.54	0.81	913	0.79	430	0.86
Ethnicity							
White	70	5.49	0.82	790	0.8	344	0.87
Black/African American	68	5.7	0.8	184	0.8	184	0.8
Asian/Pacific Islander	68	5.72	0.82	137	0.82	137	0.82
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	70	5.53	0.81	405	0.8	199	0.83
Multiple	70	5.59	0.81	246	0.78	124	0.88
Special instruction needs							
Economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
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	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students with disabilities	70	4.93	0.81	240	0.78	118	0.88
Students without disabilities	70	5.55	0.81	1612	0.79	742	0.86
Students taking accommodated forms							
American Sign Language	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

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.8 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	4.04	0.79	211	0.78	131	0.8
Gender							
Male	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Female	70	4.5	0.74	108	0.74	108	0.74
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	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Multiple	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Special instruction needs							
Economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
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	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students with disabilities	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students without disabilities	72	4.11	0.78	180	0.77	117	0.79
Students taking accommodated forms							
American Sign Language	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

Note. ELA/L = English language arts/literacy, SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.9 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	3.03	0.92	654	0.9	33023	0.93
Gender							
Male	52	3.02	0.93	401	0.9	16735	0.93
Female	52	3.04	0.92	344	0.89	5097	0.92
Ethnicity							
White	52	3.13	0.91	362	0.9	3724	0.93
Black/African American	52	2.72	0.9	102	0.81	2907	0.91
Asian/Pacific Islander	52	3.17	0.92	1879	0.92	530	0.93
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	52	2.86	0.9	368	0.83	5324	0.92
Multiple	52	3.06	0.93	1865	0.92	345	0.93
Special instruction needs							
Economically disadvantaged	52	2.82	0.9	506	0.83	11048	0.91
Not economically disadvantaged	52	3.16	0.91	285	0.91	4027	0.93
English learner	52	2.81	0.9	261	0.83	3383	0.91
Non-English learner	52	3.07	0.92	481	0.89	6675	0.93
Students with disabilities	52	2.81	0.91	2977	0.88	3805	0.93
Students without disabilities	52	3.07	0.92	161	0.88	7830	0.93
Students taking accommodated forms							
American Sign Language	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	2.88	0.93	9438	0.92	9841	0.93
Students taking translated forms							
Spanish language form	52	2.59	0.86	1505	0.86	1505	0.86

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.10 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	3.07	0.92	572	0.85	34974	0.92
Gender							
Male	52	3.06	0.92	319	0.86	17563	0.93
Female	52	3.07	0.91	273	0.82	4981	0.92
Ethnicity							
White	52	3.19	0.91	385	0.85	4106	0.93
Black/African American	52	2.64	0.88	2338	0.86	3074	0.9
Asian/Pacific Islander	52	3.29	0.92	1781	0.91	518	0.94
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	52	2.82	0.89	274	0.69	5714	0.91
Multiple	52	3.13	0.92	410	0.92	371	0.94
Special instruction needs							
Economically disadvantaged	52	2.77	0.88	347	0.74	11548	0.89
Not economically disadvantaged	52	3.23	0.91	284	0.87	4122	0.93
English learner	52	2.72	0.87	205	0.72	3273	0.9
Non-English learner	52	3.11	0.92	509	0.85	8008	0.93
Students with disabilities	52	2.74	0.88	4103	0.84	3690	0.92
Students without disabilities	52	3.13	0.92	224	0.85	7030	0.93
Students taking accommodated forms							
American Sign Language	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	2.81	0.92	11729	0.91	10890	0.92
Students taking translated forms							
Spanish language form	52	2.37	0.81	1394	0.81	1394	0.81

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.11 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	3.26	0.91	557	0.83	35732	0.91
Gender							
Male	52	3.23	0.91	305	0.8	18140	0.92
Female	52	3.29	0.9	252	0.85	17562	0.9
Ethnicity							
White	52	3.38	0.9	298	0.84	3948	0.92
Black/African American	52	2.79	0.85	104	0.76	3045	0.87
Asian/Pacific Islander	52	3.48	0.92	1846	0.91	448	0.93
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	52	3.06	0.87	280	0.73	5338	0.89
Multiple	52	3.29	0.92	1610	0.92	333	0.93
Special instruction needs							
Economically disadvantaged	52	2.96	0.86	347	0.75	11875	0.87
Not economically disadvantaged	52	3.43	0.91	136	0.85	4168	0.92
English learner	52	2.79	0.81	167	0.7	2333	0.85
Non-English learner	52	3.31	0.91	315	0.82	8308	0.91
Students with disabilities	52	2.86	0.86	3981	0.81	3839	0.91
Students without disabilities	52	3.33	0.91	141	0.81	6271	0.91
Students taking accommodated forms							
American Sign Language	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	2.96	0.91	10269	0.9	11100	0.91
Students taking translated forms							
Spanish language form	52	2.7	0.82	1206	0.82	1206	0.82

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.12 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	2.94	0.92	395	0.83	9475	0.93
Gender							
Male	52	2.94	0.92	240	0.85	5225	0.93
Female	52	2.94	0.92	155	0.78	4249	0.93
Ethnicity							
White	52	3.09	0.91	229	0.83	3620	0.94
Black/African American	52	2.44	0.87	2000	0.86	1859	0.88
Asian/Pacific Islander	52	3.34	0.93	1114	0.93	396	0.95
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	52	2.69	0.89	3936	0.87	3801	0.9
Multiple	52	3.01	0.93	1117	0.92	293	0.94
Special instruction needs							
Economically disadvantaged	52	2.59	0.88	216	0.73	7113	0.89
Not economically disadvantaged	52	3.15	0.92	154	0.86	3627	0.94
English learner	52	2.26	0.8	1832	0.75	1280	0.88
Non-English learner	52	2.99	0.92	328	0.82	7536	0.93
Students with disabilities	52	2.39	0.86	124	0.82	2304	0.92
Students without disabilities	52	3.05	0.92	107	0.83	5598	0.93
Students taking accommodated forms							
American Sign Language	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	2.66	0.92	9795	0.92	9475	0.93
Students taking translated forms							
Spanish language form	52	2.44	0.82	870	0.82	870	0.82

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.13 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	3.23	0.91	274	0.82	9704	0.93
Gender							
Male	52	3.22	0.91	171	0.78	5284	0.93
Female	52	3.23	0.9	103	0.86	4420	0.93
Ethnicity							
White	52	3.38	0.9	135	0.86	3613	0.94
Black/African American	52	2.65	0.87	2676	0.85	2030	0.89
Asian/Pacific Islander	52	3.79	0.91	1660	0.89	448	0.95
American Indian/Alaska Native	52	3	0.89	105	0.89	105	0.89
Hispanic/Latino	52	2.94	0.88	5148	0.87	3545	0.9
Multiple	52	3.23	0.92	1311	0.91	252	0.95
Special instruction needs							
Economically disadvantaged	52	2.85	0.88	137	0.65	5830	0.9
Not economically disadvantaged	52	3.47	0.9	124	0.86	3817	0.94
English learner	52	2.49	0.79	1995	0.76	1698	0.82
Non-English learner	52	3.29	0.91	218	0.83	7947	0.93
Students with disabilities	52	2.61	0.87	196	0.74	3286	0.9
Students without disabilities	52	3.34	0.9	28156	0.9	5979	0.93
Students taking accommodated forms							
American Sign Language	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	2.79	0.93	9585	0.93	9704	0.93
Students taking translated forms							
Spanish language form	52	2.35	0.82	315	0.82	315	0.82

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.14 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	2.74	0.91	333	0.71	8995	0.92
Gender							
Male	52	2.71	0.91	210	0.67	4947	0.92
Female	52	2.78	0.9	123	0.74	4048	0.91
Ethnicity							
White	52	2.85	0.9	157	0.73	3151	0.93
Black/African American	52	2.39	0.84	2149	0.81	1911	0.85
Asian/Pacific Islander	52	3.15	0.93	1231	0.92	348	0.95
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	52	2.56	0.88	4120	0.87	3190	0.88
Multiple	52	2.75	0.91	1054	0.91	229	0.94
Special instruction needs							
Economically disadvantaged	52	2.49	0.86	179	0.52	5659	0.89
Not economically disadvantaged	52	2.92	0.91	133	0.78	3284	0.93
English learner	52	2.24	0.75	1721	0.72	925	0.79
Non-English learner	52	2.78	0.91	265	0.73	7446	0.92
Students with disabilities	52	2.32	0.83	241	0.65	2419	0.9
Students without disabilities	52	2.83	0.91	21087	0.9	5402	0.92
Students taking accommodated forms							
American Sign Language	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	2.57	0.91	9360	0.91	8995	0.92
Students taking translated forms							
Spanish language form	52	2.37	0.85	254	0.85	254	0.85

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.15 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	2.79	0.87	1932	0.86	1179	0.87
Gender							
Male	55	2.79	0.88	998	0.87	603	0.89
Female	55	2.79	0.85	907	0.85	565	0.86
Ethnicity							
White	55	2.86	0.86	831	0.85	462	0.88
Black/African American	55	2.5	0.8	109	0.8	182	0.8
Asian/Pacific Islander	55	2.94	0.88	100	0.88	134	0.88
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	55	2.66	0.86	359	0.86	254	0.86
Multiple	55	2.87	0.87	284	0.86	167	0.88
Special instruction needs							
Economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Not economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
English learner	55	2.69	0.87	139	0.87	139	0.87
Non-English learner	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students with disabilities	55	2.43	0.85	144	0.83	216	0.87
Students without disabilities	55	2.82	0.86	1689	0.86	1025	0.87
Students taking accommodated forms							
American Sign Language	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

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.16 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	2.96	0.87	1626	0.87	883	0.87
Gender							
Male	55	2.97	0.88	826	0.88	450	0.88
Female	55	2.94	0.86	785	0.85	428	0.86
Ethnicity							
White	55	3.01	0.86	374	0.85	683	0.86
Black/African American	55	2.63	0.83	138	0.83	138	0.83
Asian/Pacific Islander	55	3.22	0.89	115	0.89	115	0.89
American Indian/Alaska Native	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Hispanic/Latino	55	2.83	0.86	341	0.86	184	0.87
Multiple	55	2.98	0.87	229	0.87	113	0.87
Special instruction needs							
Economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
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	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students with disabilities	55	2.49	0.84	186	0.82	114	0.88
Students without disabilities	55	3.01	0.86	1426	0.86	766	0.87
Students taking accommodated forms							
American Sign Language	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

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Table A.13.17 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	3.05	0.86	1601	0.86	974	0.87
Gender							
Male	55	3.11	0.88	768	0.87	490	0.89
Female	55	2.98	0.83	810	0.83	480	0.83
Ethnicity							
White	55	3.22	0.86	368	0.85	644	0.86
Black/African American	55	2.67	0.76	130	0.76	130	0.76
Asian/Pacific Islander	55	3.19	0.87	134	0.87	134	0.87
American Indian/Alaska Native							
Hispanic/Latino	55	2.89	0.83	294	0.82	206	0.84
Multiple	55	3.07	0.86	217	0.85	107	0.9
Special instruction needs	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Economically disadvantaged	n/r	n/r	n/r	n/r	n/r	n/r	n/r
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	n/r	n/r	n/r	n/r	n/r	n/r	n/r
Students with disabilities	55	2.53	0.85	149	0.83	105	0.87
Students without disabilities	55	3.09	0.86	1430	0.85	867	0.86
Students taking accommodated forms							
American Sign Language	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

Note. SEM = standard error of measurement, n/r = not reported because the form type was not administered, a low sample size (n<100), or missing demographic information.

Appendix 13.2: Reliability of Classification by Content and Grade/Subject

Table A.13.18 Reliability of Classification: Grade 3 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.23**	0.04	0.00	0.00	0.00	**0.23**
	700-724	0.04	**0.11**	0.05	0.00	0.00	0.04
	725-749	0.00	0.05	**0.11**	0.05	0.00	0.00
	750-809	0.00	0.00	0.05	**0.22**	0.02	0.00
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision consistency	650-699	**0.22**	0.06	0.01	0.00	0.00	**0.22**
	700-724	0.05	**0.08**	0.05	0.01	0.00	0.05
	725-749	0.01	0.05	**0.08**	0.06	0.00	0.01
	750-809	0.00	0.02	0.07	**0.20**	0.02	0.00
	810-850	0.00	0.00	0.00	0.01	**0.00**	0.00

Note. ELA/L = English language arts/literacy.

Table A.13.19 Reliability of Classification: Grade 4 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.18**	0.03	0.00	0.00	0.00	0.21
	700-724	0.04	**0.13**	0.05	0.00	0.00	0.23
	725-749	0.00	0.05	**0.16**	0.05	0.00	0.27
	750-809	0.00	0.00	0.05	**0.20**	0.03	0.29
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01
Decision consistency	650-699	**0.17**	0.05	0.01	0.00	0.00	0.23
	700-724	0.05	**0.10**	0.07	0.01	0.00	0.22
	725-749	0.01	0.06	**0.12**	0.06	0.00	0.25
	750-809	0.00	0.01	0.07	**0.16**	0.03	0.27
	810-850	0.00	0.00	0.00	0.02	**0.01**	0.03

Note. ELA/L = English language arts/literacy.

Table A.13.20 Reliability of Classification: Grade 5 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.14**	0.03	0.00	0.00	0.00	0.17
	700-724	0.04	**0.15**	0.05	0.00	0.00	0.24
	725-749	0.00	0.06	**0.17**	0.05	0.00	0.28
	750-809	0.00	0.00	0.05	**0.23**	0.02	0.30
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision consistency	650-699	**0.13**	0.05	0.01	0.00	0.00	0.19
	700-724	0.04	**0.12**	0.07	0.01	0.00	0.24
	725-749	0.00	0.06	**0.13**	0.06	0.00	0.26
	750-809	0.00	0.01	0.07	**0.21**	0.02	0.30
	810-850	0.00	0.00	0.00	0.01	**0.01**	0.02

Note. ELA/L = English language arts/literacy.

Table A.13.21 Reliability of Classification: Grade 6 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.11**	0.02	0.00	0.00	0.00	0.13
	700-724	0.03	**0.16**	0.05	0.00	0.00	0.24
	725-749	0.00	0.05	**0.20**	0.05	0.00	0.31
	750-809	0.00	0.00	0.05	**0.23**	0.02	0.31
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01
Decision consistency	650-699	**0.11**	0.04	0.00	0.00	0.00	0.15
	700-724	0.04	**0.13**	0.07	0.00	0.00	0.24
	725-749	0.00	0.06	**0.16**	0.06	0.00	0.29
	750-809	0.00	0.01	0.07	**0.20**	0.02	0.30
	810-850	0.00	0.00	0.00	0.02	**0.01**	0.03

Note. ELA/L = English language arts/literacy.

Table A.13.22 Reliability of Classification: Grade 7 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.16**	0.03	0.00	0.00	0.00	0.19
	700-724	0.03	**0.12**	0.05	0.00	0.00	0.21
	725-749	0.00	0.05	**0.16**	0.05	0.00	0.25
	750-809	0.00	0.00	0.05	**0.20**	0.03	0.29
	810-850	0.00	0.00	0.00	0.02	**0.04**	0.06
Decision consistency	650-699	**0.15**	0.04	0.01	0.00	0.00	0.20
	700-724	0.04	**0.10**	0.06	0.01	0.00	0.20
	725-749	0.00	0.05	**0.12**	0.06	0.00	0.24
	750-809	0.00	0.01	0.07	**0.17**	0.04	0.28
	810-850	0.00	0.00	0.00	0.04	**0.04**	0.08

Note. ELA/L = English language arts/literacy.

Table A.13.23 Reliability of Classification: Grade 8 ELA/L

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.14**	0.02	0.00	0.00	0.00	0.17
	700-724	0.04	**0.12**	0.05	0.00	0.00	0.20
	725-749	0.00	0.05	**0.16**	0.05	0.00	0.26
	750-809	0.00	0.00	0.06	**0.25**	0.03	0.33
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision consistency	650-699	**0.14**	0.04	0.01	0.00	0.00	0.18
	700-724	0.04	**0.09**	0.06	0.01	0.00	0.20
	725-749	0.01	0.05	**0.12**	0.06	0.00	0.24
	750-809	0.00	0.01	0.07	**0.21**	0.03	0.32
	810-850	0.00	0.00	0.00	0.03	**0.02**	0.05

Note. ELA/L = English language arts/literacy.

Table A.13.24 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	**0.05**	0.01	0.00	0.00	0.00	0.06
	700-724	0.03	**0.05**	0.04	0.00	0.00	0.12
	725-749	0.00	0.04	**0.11**	0.06	0.00	0.21
	750-809	0.00	0.01	0.07	**0.31**	0.06	0.45
	810-850	0.00	0.00	0.00	0.05	**0.12**	0.16
Decision consistency	650-699	**0.05**	0.02	0.01	0.00	0.00	0.08
	700-724	0.02	**0.04**	0.05	0.02	0.00	0.13
	725-749	0.01	0.03	**0.08**	0.07	0.00	0.20
	750-809	0.00	0.01	0.08	**0.25**	0.06	0.40
	810-850	0.00	0.00	0.00	0.08	**0.11**	0.19

Note. ELA/L = English language arts/literacy.

Table A.13.25 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	**0.27**	0.06	0.00	0.00	0.00	0.34
	700-724	0.07	**0.19**	0.07	0.00	0.00	0.33
	725-749	0.00	0.06	**0.12**	0.05	0.00	0.23
	750-809	0.00	0.00	0.02	**0.07**	0.01	0.11
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision consistency	650-699	**0.26**	0.09	0.01	0.00	0.00	0.36
	700-724	0.08	**0.14**	0.06	0.01	0.00	0.29
	725-749	0.01	0.07	**0.09**	0.04	0.00	0.21
	750-809	0.00	0.01	0.04	**0.07**	0.00	0.13
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.01

Note. ELA/L = English language arts/literacy.

Table A.13.26 Reliability of Classification: Grade 3 Mathematics

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.18**	0.03	0.00	0.00	0.00	0.21
	700-724	0.03	**0.15**	0.04	0.00	0.00	0.23
	725-749	0.00	0.05	**0.15**	0.04	0.00	0.24
	750-809	0.00	0.00	0.04	**0.20**	0.02	0.26
	810-850	0.00	0.00	0.00	0.01	**0.04**	0.06
Decision consistency	650-699	**0.17**	0.05	0.00	0.00	0.00	0.22
	700-724	0.04	**0.12**	0.06	0.00	0.00	0.22
	725-749	0.00	0.05	**0.12**	0.05	0.00	0.23
	750-809	0.00	0.00	0.05	**0.18**	0.02	0.26
	810-850	0.00	0.00	0.00	0.03	**0.04**	0.07

Table A.13.27 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	**0.20**	0.03	0.00	0.00	0.00	0.23
	700-724	0.04	**0.18**	0.04	0.00	0.00	0.27
	725-749	0.00	0.04	**0.17**	0.04	0.00	0.25
	750-809	0.00	0.00	0.03	**0.18**	0.01	0.23
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.02
Decision consistency	650-699	**0.19**	0.05	0.00	0.00	0.00	0.25
	700-724	0.04	**0.15**	0.06	0.00	0.00	0.26
	725-749	0.00	0.06	**0.14**	0.05	0.00	0.24
	750-809	0.00	0.00	0.05	**0.16**	0.01	0.23
	810-850	0.00	0.00	0.00	0.01	**0.01**	0.02

Table A.13.28 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	**0.18**	0.04	0.00	0.00	0.00	0.22
	700-724	0.04	**0.21**	0.04	0.00	0.00	0.30
	725-749	0.00	0.05	**0.15**	0.04	0.00	0.24
	750-809	0.00	0.00	0.03	**0.17**	0.01	0.21
	810-850	0.00	0.00	0.00	0.01	**0.03**	0.03
Decision consistency	650-699	**0.17**	0.06	0.00	0.00	0.00	0.23
	700-724	0.05	**0.17**	0.06	0.00	0.00	0.28
	725-749	0.00	0.07	**0.12**	0.05	0.00	0.24
	750-809	0.00	0.00	0.04	**0.15**	0.01	0.21
	810-850	0.00	0.00	0.00	0.02	**0.03**	0.04

Table A.13.29 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	**0.19**	0.03	0.00	0.00	0.00	0.22
	700-724	0.04	**0.21**	0.04	0.00	0.00	0.29
	725-749	0.00	0.05	**0.18**	0.04	0.00	0.27
	750-809	0.00	0.00	0.03	**0.16**	0.01	0.20
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.02
Decision consistency	650-699	**0.18**	0.05	0.00	0.00	0.00	0.23
	700-724	0.04	**0.18**	0.06	0.00	0.00	0.28
	725-749	0.00	0.06	**0.15**	0.05	0.00	0.26
	750-809	0.00	0.00	0.05	**0.14**	0.01	0.20
	810-850	0.00	0.00	0.00	0.01	**0.01**	0.03

Table A.13.30 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	**0.07**	0.02	0.00	0.00	0.00	0.10
	700-724	0.03	**0.23**	0.05	0.00	0.00	0.31
	725-749	0.00	0.05	**0.24**	0.04	0.00	0.34
	750-809	0.00	0.00	0.03	**0.19**	0.01	0.23
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision consistency	650-699	**0.07**	0.04	0.00	0.00	0.00	0.11
	700-724	0.03	**0.20**	0.07	0.00	0.00	0.30
	725-749	0.00	0.07	**0.20**	0.05	0.00	0.32
	750-809	0.00	0.00	0.05	**0.17**	0.01	0.24
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03

Table A.13.31 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	**0.24**	0.04	0.00	0.00	0.00	0.28
	700-724	0.04	**0.17**	0.04	0.00	0.00	0.25
	725-749	0.00	0.05	**0.12**	0.04	0.00	0.21
	750-809	0.00	0.00	0.03	**0.19**	0.01	0.23
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03
Decision consistency	650-699	**0.23**	0.06	0.01	0.00	0.00	0.30
	700-724	0.05	**0.13**	0.05	0.01	0.00	0.23
	725-749	0.01	0.06	**0.09**	0.05	0.00	0.20
	750-809	0.00	0.01	0.05	**0.17**	0.01	0.23
	810-850	0.00	0.00	0.00	0.01	**0.02**	0.03

.

Table A.13.32 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	**0.04**	0.01	0.00	0.00	0.00	0.06
	700-724	0.02	**0.13**	0.04	0.00	0.00	0.20
	725-749	0.00	0.05	**0.23**	0.06	0.00	0.34
	750-809	0.00	0.00	0.06	**0.33**	0.01	0.40
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision consistency	650-699	**0.04**	0.03	0.00	0.00	0.00	0.07
	700-724	0.02	**0.11**	0.07	0.00	0.00	0.20
	725-749	0.00	0.06	**0.18**	0.07	0.00	0.31
	750-809	0.00	0.00	0.08	**0.31**	0.01	0.40
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.01

Table A.13.33 Reliability of Classification: Geometry

	Full Summative Scale Score	Level 1	Level 2	Level 3	Level 4	Level 5	Category Total
Decision accuracy	650-699	**0.04**	0.01	0.00	0.00	0.00	0.06
	700-724	0.02	**0.19**	0.04	0.00	0.00	0.25
	725-749	0.00	0.06	**0.26**	0.06	0.00	0.38
	750-809	0.00	0.00	0.04	**0.24**	0.01	0.30
	810-850	0.00	0.00	0.00	0.00	**0.01**	0.01
Decision consistency	650-699	**0.04**	0.03	0.00	0.00	0.00	0.07
	700-724	0.02	**0.16**	0.06	0.00	0.00	0.25
	725-749	0.00	0.07	**0.22**	0.07	0.00	0.36
	750-809	0.00	0.00	0.07	**0.22**	0.01	0.30
	810-850	0.00	0.00	0.00	0.01	**0.01**	0.02

Table A.13.34 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	**0.19**	0.03	0.00	0.00	0.00	0.22
	700-724	0.05	**0.14**	0.06	0.00	0.00	0.25
	725-749	0.00	0.06	**0.15**	0.05	0.00	0.26
	750-809	0.00	0.00	0.05	**0.21**	0.01	0.27
	810-850	0.00	0.00	0.00	0.00	**0.00**	0.00
Decision consistency	650-699	**0.18**	0.05	0.01	0.00	0.00	0.24
	700-724	0.05	**0.11**	0.07	0.01	0.00	0.24
	725-749	0.01	0.06	**0.12**	0.06	0.00	0.24
	750-809	0.00	0.01	0.07	**0.19**	0.01	0.28
	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			Matched		
	Current Form 1	Original Form 1	DIFF*	Current Form 1	Original Form 1	DIFF*
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

Note. ELA/L = English language arts/literacy, *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

Note. *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

Note. ELA/L = English language arts/literacy, *DIFF = current percent – original percent.

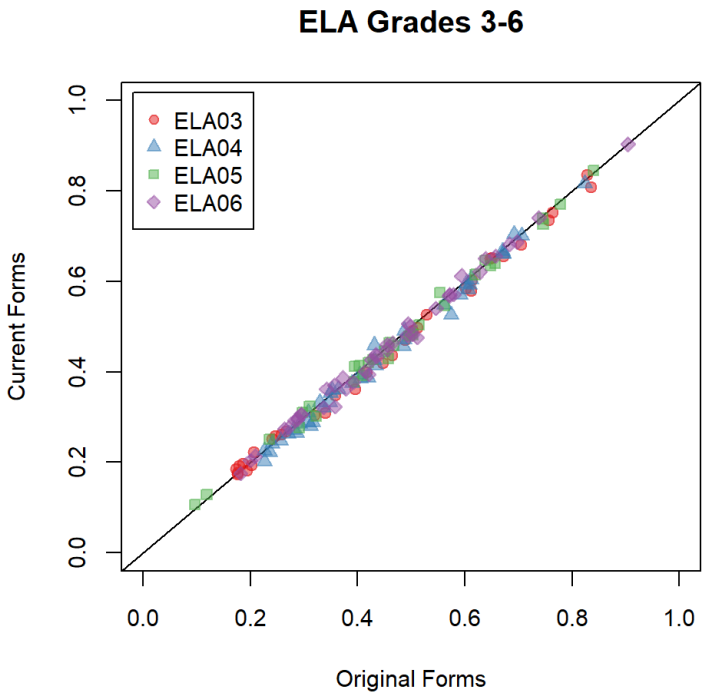

Figure A.14.1 ELA/L Grades 3-6 P-Values

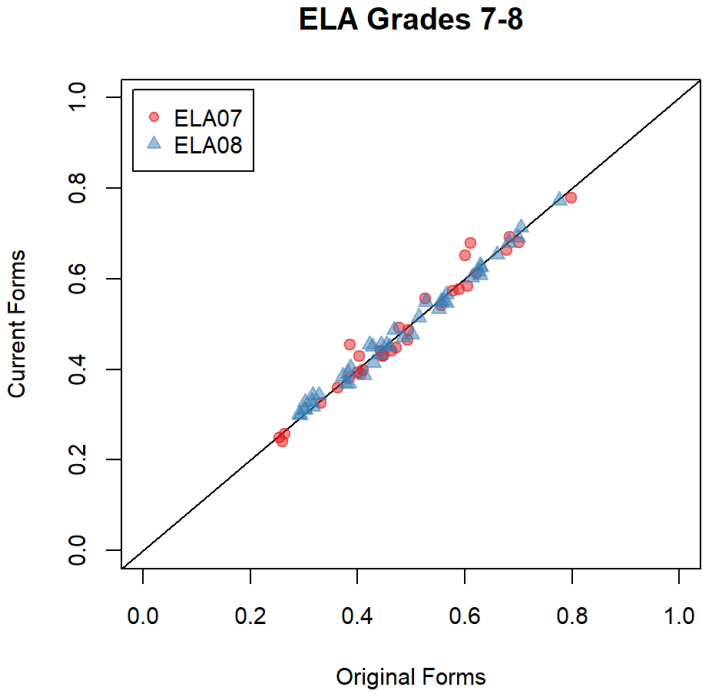

Figure A.14.2 ELA/L Grades 7-8 P-Values

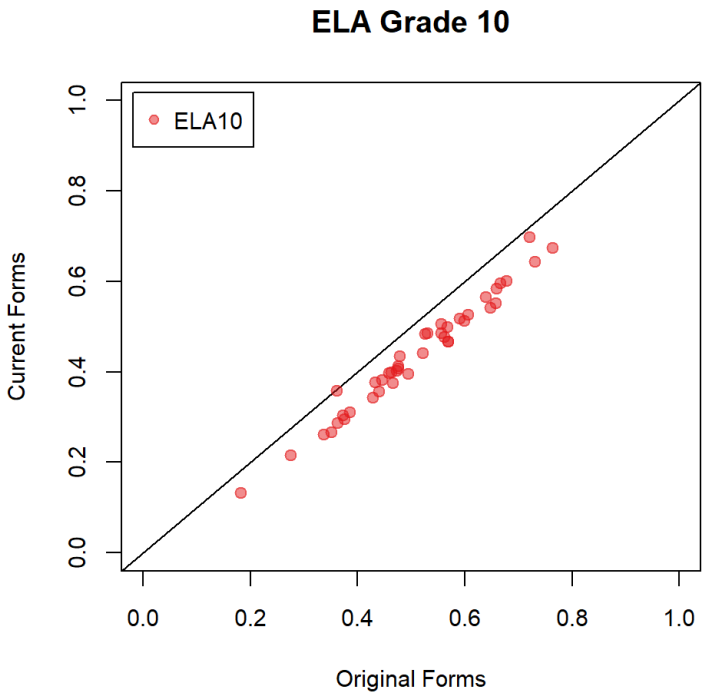

Figure A.14.3 ELA/L Grade 10 P-Values

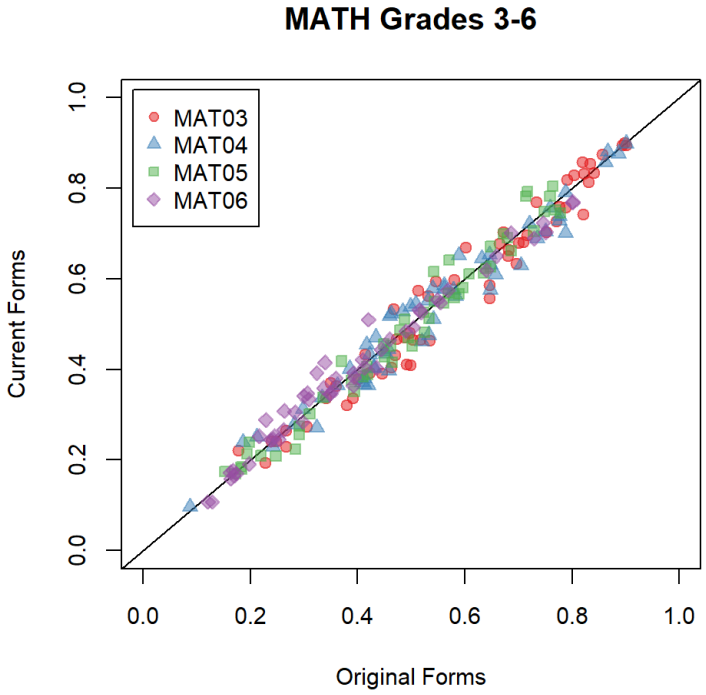

Figure A.14.4 Mathematics Grades 3-6 P-Values

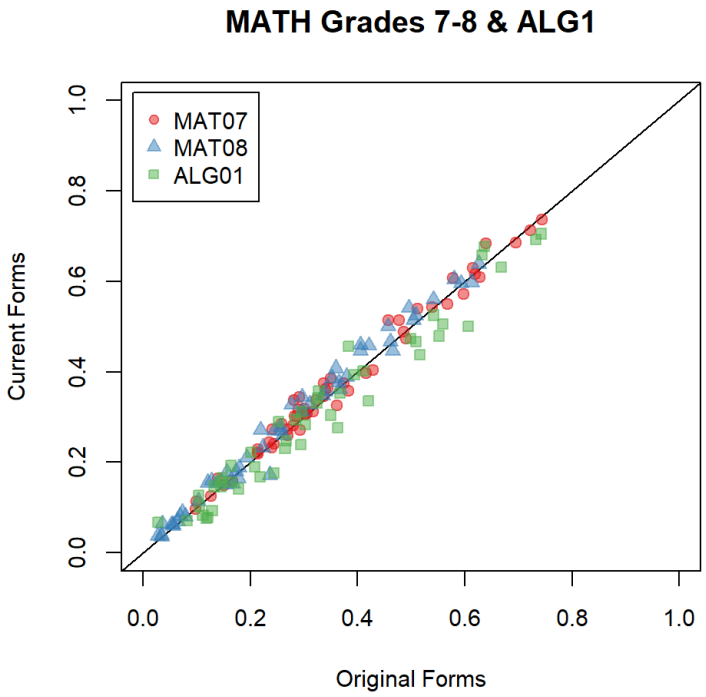

Figure A.14.5 Mathematics Grade 7-8 and Algebra I P-Values

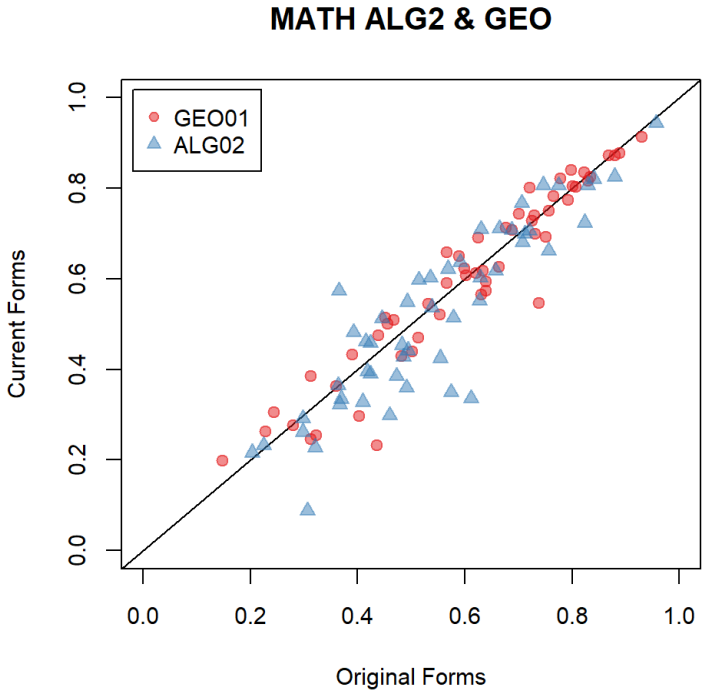

Figure A.14.6 Algebra II and Geometry P-Values

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

Note. ELA/L = English language arts/literacy,
*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

Note. *Difference = current p-value – original p-value

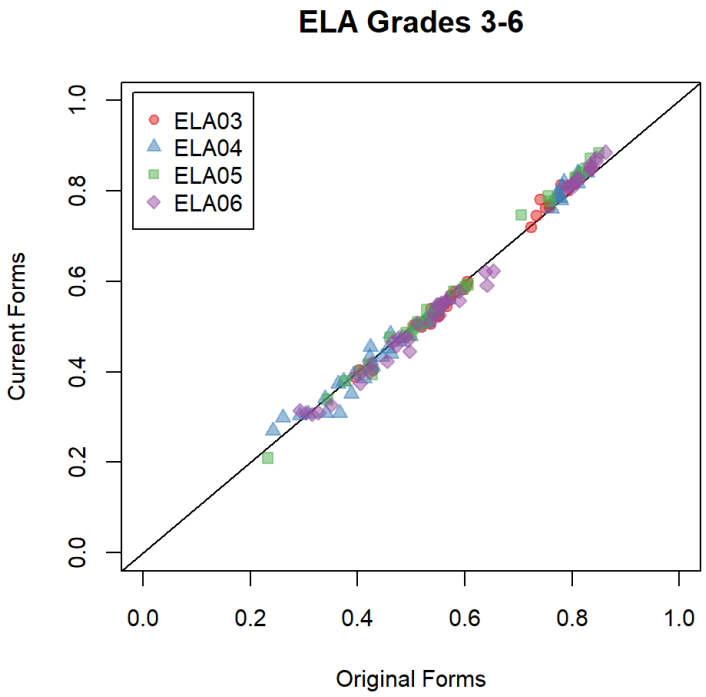

Figure A.14.7 Polyserial Correlations ELA/L Grades 3-6

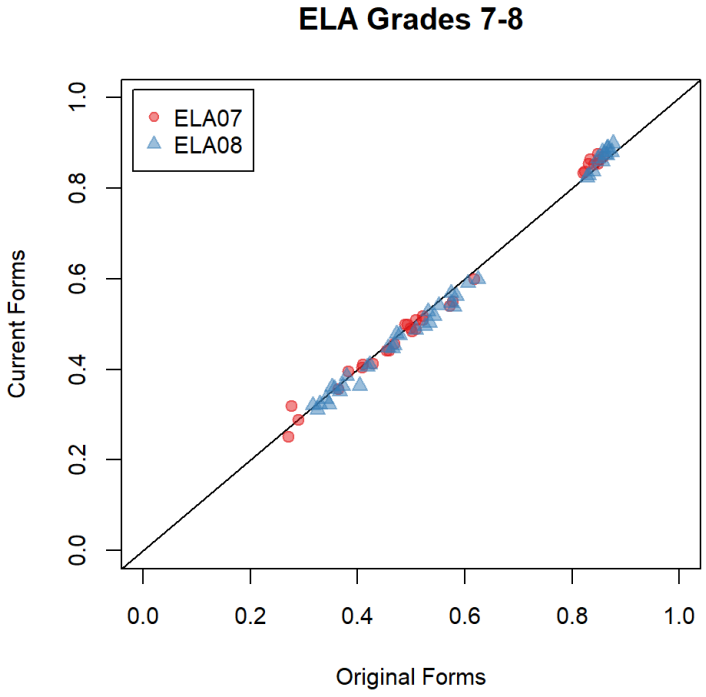

Figure A.14.8 Polyserial Correlations ELA/L Grades 7-8

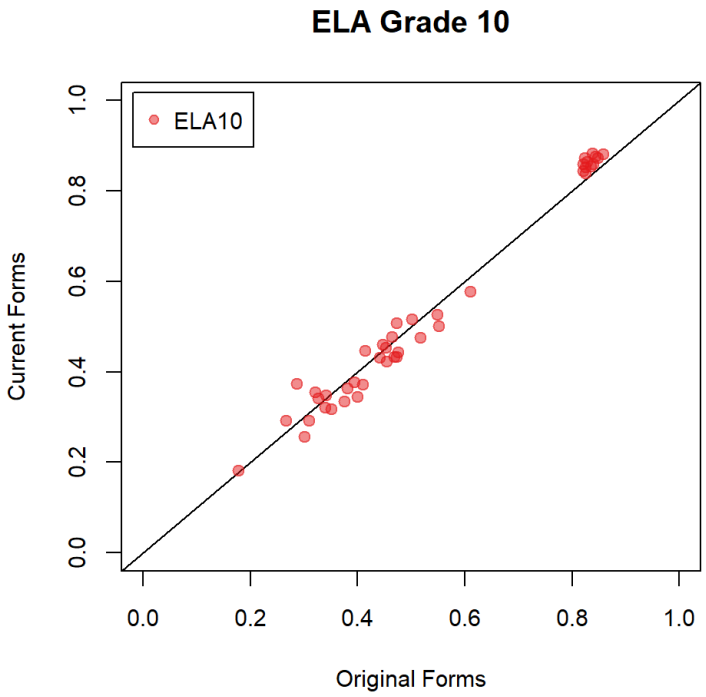

Figure A.14.9 Polyserial Correlations ELA/L Grade 10

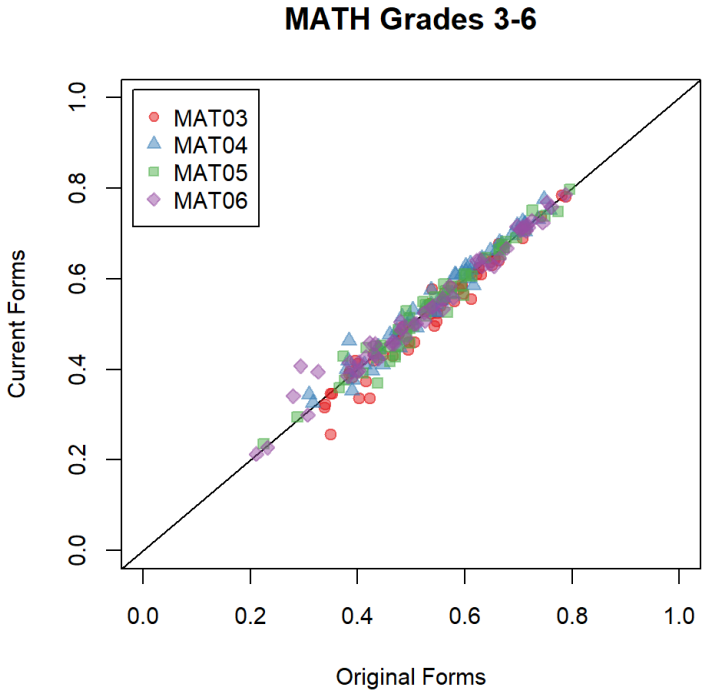

Figure A.14.10 Polyserial Correlations Mathematics Grades 3-6

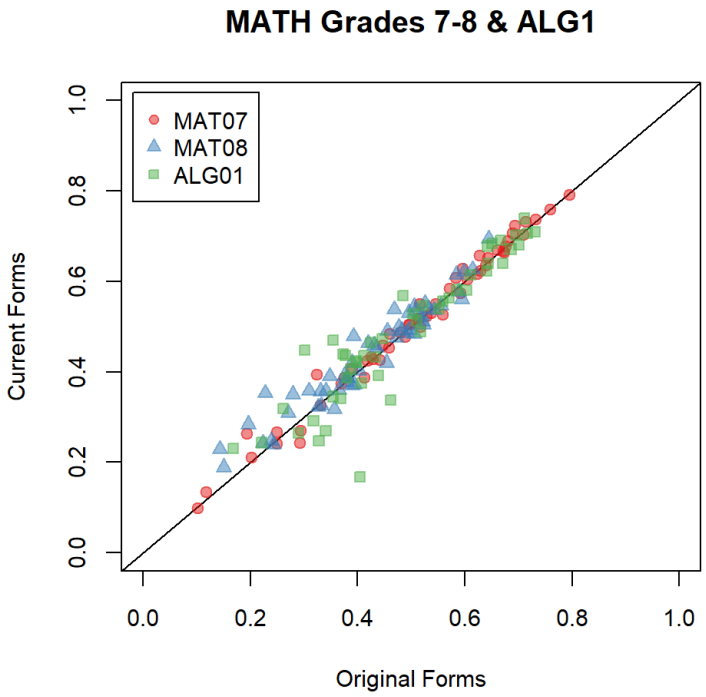

Figure A.14.11 Polyserial Correlations Mathematics Grades 7-8 and Algebra I

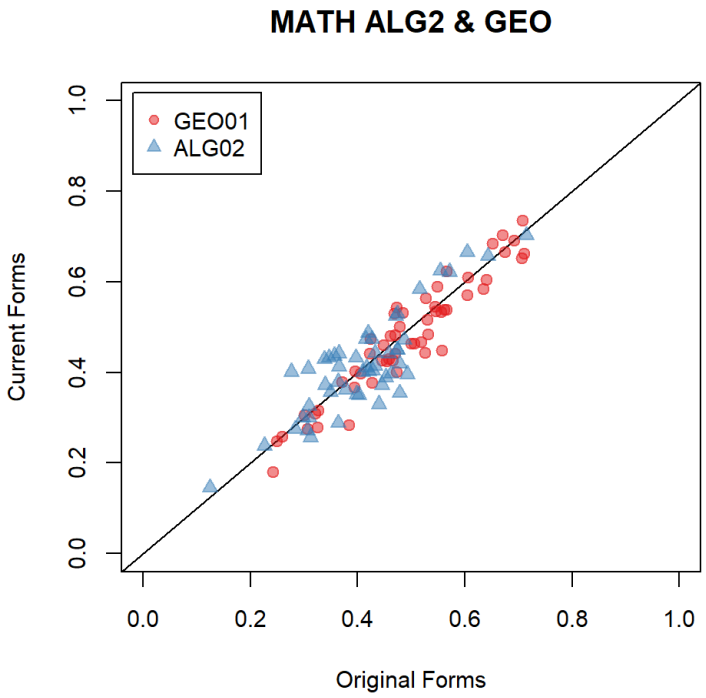

Figure A.14.12 Polyserial Correlations Algebra II and Geometry

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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, *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%

Note. ELA/L = English language arts/literacy.

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%

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, RS = raw score, SEM = standard error of measurement.

Table A.14.13 ELA/L Scale Score Standard Error of Measurement

Grade	Original Form 1		Original Form 2		Current Form 1		Current Form 2	
	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

Note. ELA/L = English language arts/literacy, SS = scale score, SEM = standard error of measurement.

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

Note. **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

Note. RS = raw score, SEM = standard error of measurement.

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

Note. SS = scale score, SEM = standard error of measurement.

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

Note. ELA/L = English language arts/literacy, SD = standard deviation, *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

Note. SD = standard deviation, *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

Note. ELA/L = English language arts/literacy, SD = standard deviation, *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

Note. ELA/L = English language arts/literacy, SD = standard deviation, *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

Note. ELA/L = English language arts/literacy, ES = effect size.

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

Note. ES = effect size.

Table A.14.23 ELA/L Subclaim Distribution Comparison: Effect Size

Grade	Subclaim Distribution Effect Size				
	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

Note. ELA/L = English language arts/literacy.

Table A.14.24 Mathematics Subclaim Distribution Comparison: Effect Size

Grade/ Course	Subclaim Distribution Effect Size			
	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

Grade	2018 Original SS			2019 Current SS			2019-2018		
	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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, SD = standard deviation, *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

Note. *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

Note. *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			R^2		
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

Note. ELA/L = English language arts/literacy.

Table A.14.30 Mathematics Longitudinal Regression

	Sample Size			R^2		
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

Note. ***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					

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy, **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

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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		

Note. ELA/L = English language arts/literacy.

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

Note. ELA/L = English language arts/literacy.

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

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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

Note. ELA/L = English language arts/literacy. *Diff: Current Alpha – Spearman Brown (SB) Prophecy.

Table A.14.44 ELA/L Reading Information (RI) Subclaim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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

Note. ELA/L = English language arts/literacy, *Diff: Current Alpha – Spearman Brown (SB) Prophecy.

Table A.14.45 ELA/L Reading Literature (RL) Subclaim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, *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

Note. ELA/L = English language arts/literacy, *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

Appendix 15: Growth

Appendix 15 provides the summary growth results for subgroups for grade 4 – 11 ELA/L and mathematics 4 – 8 and high school. Grade 9 ELA, Algebra II, Integrated mathematics I and II do not have sufficient sample sizes for subgroup summary analysis.

Table A.15.1 Summary of SGP Estimates for Subgroups: Grade 5 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,491	47.37	13.53	46
Female	43,832	52.67	12.99	54
Ethnicity				
White	50,369	53.25	13.04	54
African American	11,404	38.24	13.70	34
Asian/Pacific Islander	4,571	60.74	12.42	65
American Indian/Alaska Native	151	45.05	13.72	44
Hispanic	20,082	45.85	13.78	44
Multiple	3,683	50.50	13.28	51
Special instruction needs				
Economically Disadvantaged	40,139	43.63	13.76	41
Not-economically disadvantaged	50,184	54.99	12.87	57
English learner	10,139	43.31	14.67	40
Non-English learner	80,184	50.78	13.09	51
Students with disabilities	15,804	43.17	14.64	40
Students without disabilities	74,519	51.38	12.98	52

Note. ELA/L = English language arts/literacy, SGP = student growth percentile.

Table A.15.2 Summary of SGP Estimates for Subgroups: Grade 6 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,973	48.01	14.10	47
Female	43,915	52.14	13.63	53
Ethnicity				
White	50,113	51.70	13.63	52
African American	11,350	42.89	14.49	40
Asian/Pacific Islander	4,523	58.32	13.42	62
American Indian/Alaska Native	193	52.45	13.82	52
Hispanic	20,047	48.25	14.21	47
Multiple	3,605	48.81	13.90	49
Special instruction needs				
Economically Disadvantaged	39,509	46.65	14.21	45
Not-economically disadvantaged	50,379	52.68	13.61	54
English learner	8,218	45.31	15.35	43.5
Non-English learner	81,670	50.50	13.72	51
Students with disabilities	15,339	45.18	15.17	43
Students without disabilities	74,549	51.03	13.60	51

Note. ELA/L = English language arts/literacy, SGP = student growth percentile.

Table A.15.3 Summary of SGP Estimates for Subgroups: Grade 7 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,660	48.26	14.05	48
Female	43,046	51.87	13.70	53
Ethnicity				
White	49,916	51.22	13.71	52
African American	11,397	44.71	14.21	43
Asian/Pacific Islander	4,379	57.34	13.78	60
American Indian/Alaska Native	202	46.28	13.74	45.5
Hispanic	19,399	48.68	14.12	48
Multiple	3,368	48.45	13.99	47
Special instruction needs				
Economically Disadvantaged	38,584	46.84	14.00	45
Not-economically disadvantaged	50,122	52.45	13.79	54
English learner	7,434	48.58	14.42	48
Non-English learner	81,272	50.14	13.83	50
Students with disabilities	15,157	45.62	14.30	44
Students without disabilities	73,549	50.91	13.79	51

Note. ELA/L = English language arts/literacy, SGP = student growth percentile.

Table A.15.4 Summary of SGP Estimates for Subgroups: Grade 8 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	47,254	48.43	14.01	48
Female	43,817	52.35	13.80	53
Ethnicity				
White	50,745	50.14	13.79	50
African American	11,992	47.41	14.38	47
Asian/Pacific Islander	4,309	58.80	13.58	63
American Indian/Alaska Native	185	45.58	13.66	43
Hispanic	20,169	50.67	14.04	51
Multiple	3,606	50.48	13.77	51
Special instruction needs				
Economically Disadvantaged	38,569	48.52	14.06	48
Not-economically disadvantaged	52,568	51.65	13.80	52
English learner	6,477	49.13	14.55	49
Non-English learner	84,660	50.42	13.86	51
Students with disabilities	15,877	45.75	14.61	44
Students without disabilities	75,260	51.29	13.76	52

Note. ELA/L = English language arts/literacy, SGP = student growth percentile.

Table A.15.5 Summary of SGP Estimates for Subgroups: Grade 10 ELA/L

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	794	50.85	14.72	50.5
Female	767	48.21	14.57	48
Ethnicity				
White	616	52.09	14.54	53
African American	150	42.05	14.10	39
Asian/Pacific Islander	187	53.69	15.16	56
American Indian/Alaska Native	--	--	--	--
Hispanic	350	46.58	14.93	43
Multiple	233	49.82	14.40	49
Special instruction needs				
Economically Disadvantaged	--	--	--	--
Not-economically disadvantaged	1,597	49.76	14.64	50
English learner	75	46.79	13.33	43
Non-English learner	1,522	49.90	14.70	50
Students with disabilities	237	43.63	14.60	42
Students without disabilities	1,360	50.83	14.64	51

Note. ELA/L = English language arts/literacy, SGP = student growth percentile. "--" = insufficient sample.

Table A.15.6 Summary of SGP Estimates for Subgroups: Grade 5 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,555	50.23	12.95	50
Female	43,910	49.95	13.04	50
Ethnicity				
White	51,038	53.86	12.30	55
African American	11,317	37.32	15.09	33
Asian/Pacific Islander	4,764	61.15	11.12	66
American Indian/Alaska Native	137	49.40	13.71	46
Hispanic	19,147	44.45	14.07	42
Multiple	3,990	52.18	12.82	53
Special instruction needs				
Economically Disadvantaged	38,499	42.39	14.49	39
Not-economically disadvantaged	52,008	55.81	11.88	58
English learner	9,259	44.62	15.61	42
Non-English learner	81,248	50.72	12.69	51
Students with disabilities	15,811	47.95	15.12	47
Students without disabilities	74,696	50.55	12.54	51
Spanish language form	1,206	37.51	15.48	32

Note. SGP = student growth percentile.

Table A.15.7 Summary of SGP Estimates for Subgroups: Grade 6 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,168	50.01	14.91	50
Female	44,009	50.40	14.99	51
Ethnicity				
White	50,631	53.02	14.11	54
African American	11,272	40.81	17.33	37
Asian/Pacific Islander	4,654	61.22	12.60	66
American Indian/Alaska Native	184	46.76	15.67	43.5
Hispanic	19,460	45.87	16.30	44
Multiple	3,909	49.41	15.00	50
Special instruction needs				
Economically Disadvantaged	38,146	45.13	16.55	43
Not-economically disadvantaged	52,072	53.91	13.77	56
English learner	7,738	44.40	18.47	42
Non-English learner	82,480	50.74	14.62	51
Students with disabilities	15,420	47.47	17.36	46
Students without disabilities	74,798	50.76	14.45	51
Spanish language form	857	37.27	16.63	33

Note. SGP = student growth percentile.

Table A.15.8 Summary of SGP Estimates for Subgroups: Grade 7 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	44,872	50.87	15.26	51
Female	42,293	49.13	15.36	49
Ethnicity				
White	49,439	51.81	14.71	53
African American	11,104	42.79	17.20	40
Asian/Pacific Islander	4,310	59.16	13.69	63
American Indian/Alaska Native	193	46.35	15.51	44
Hispanic	18,766	47.83	16.08	47
Multiple	3,312	48.60	15.46	48
Special instruction needs				
Economically Disadvantaged	37,632	46.38	16.43	45
Not-economically disadvantaged	49,533	52.80	14.45	54
English learner	7,023	46.96	17.80	46
Non-English learner	80,142	50.30	15.09	50
Students with disabilities	14,841	45.61	17.18	43
Students without disabilities	72,324	50.93	14.92	51
Spanish language form	252	44.27	18.08	41.5

Note. SGP = student growth percentile.

Table A.15.9 Summary of SGP Estimates for Subgroups: Grade 8 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,915	49.59	15.87	49
Female	42,654	50.45	15.89	51
Ethnicity				
White	49,776	51.36	15.30	52
African American	11,554	43.76	17.59	41
Asian/Pacific Islander	4,139	58.46	13.92	62
American Indian/Alaska Native	178	42.42	16.66	38
Hispanic	19,413	48.67	16.72	48
Multiple	3,458	49.26	15.97	49
Special instruction needs				
Economically Disadvantaged	37,544	46.71	17.05	45
Not-economically disadvantaged	51,051	52.43	15.01	54
English learner	6,102	47.70	18.67	47
Non-English learner	82,493	50.18	15.67	50
Students with disabilities	15,453	45.89	17.78	44
Students without disabilities	73,142	50.88	15.47	51
Spanish language form	175	48.73	17.48	45

Note. SGP = student growth percentile.

Table A.15.10 Summary of SGP Estimates for Subgroups: Algebra II

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	659	50.34	15.90	51
Female	661	48.67	16.31	48
Ethnicity				
White	539	52.60	15.59	55
African American	107	43.69	15.86	41
Asian/Pacific Islander	155	52.33	16.24	51
Hispanic	29	33.17	19.42	31
Multiple	269	44.91	16.50	45
Special Instruction Needs	204	50.11	16.62	49.5
Economically disadvantaged	--	--	--	--
Not-economically disadvantaged	1,337	49.56	16.10	50
English learner	67	47.10	16.98	43
Non-English learner	1,270	49.69	16.05	50
Students with disabilities	166	38.14	16.65	34.5
Students without disabilities	1,171	51.18	16.02	52

Note. SGP = student growth percentile. "--" = insufficient sample.