



New Meridian

Technical Report New Meridian Summative Assessments Spring 2023: New Jersey



Table of Contents

Table of Contents.....	2
List of Tables	7
List of Figures	9
Chapter 1. Introduction.....	10
1.1. Background.....	10
1.2. Purpose of the Summative Tests.....	11
1.3. Composition of Operational Tests	11
1.4. Intended Population.....	12
1.5. Overview of the Technical Report.....	12
Chapter 2. Test Design and Development.....	15
2.1. Overview of the Test	15
2.2. ELA Claims and Subclaims.....	15
2.3. Mathematics Claims and Subclaims.....	17
2.4. Test Development Activities	18
2.4.1. Item Development Process	19
2.4.1.1. Bank Analysis and Item Development Plan.....	19
2.4.1.2. Text Selection for ELA.....	19
2.4.1.3. Item Development	19
2.4.2. Item and Text Review Committees	20
2.4.2.1. Text Review	20
2.4.2.2. Content Item Review	20
2.4.2.3. Bias and Sensitivity Review.....	20
2.4.2.4. Editorial Review.....	20
2.4.2.5. Data Review.....	20
2.4.3. Operational Test Construction.....	21
2.4.3.1. Test Construction Activities.....	21
2.4.3.2. Test Form Verification Meeting to Review Test Construction Inputs	22
2.4.3.3. Accommodated Form Review Process	22
2.4.3.4. Spanish-Language Assessments for Mathematics.....	23
2.4.4. Linking Design of the Operational Test	23

2.4.5. Field Test Data Collection Overview	23
Chapter 3. Assessment Administration	24
3.1 Test Security and Administration Policies	24
3.1.1. Secure vs. Nonsecure Materials.....	24
3.1.2. Scorable vs. Nonscorable Materials	24
3.2. Accessibility Features and Accommodations.....	25
3.2.1. Participation Guidelines for Assessments.....	25
3.2.2. Accessibility System.....	26
3.2.3. What are Accessibility Features?.....	26
3.2.4. Accommodations for Students with Disabilities and English Learners.....	26
3.2.5. Unique Accommodations.....	28
3.2.6. Emergency Accommodations	28
3.2.7. Student Refusal Form	28
3.3. Testing Irregularities and Security Breaches	28
3.4. Data Forensics Analyses	30
3.4.1. Response Change Analysis	30
3.4.2. Aberrant Response Analysis	31
3.4.3. Plagiarism Analysis	31
3.4.4. Longitudinal Performance Monitoring.....	31
3.4.5. Internet and Social Media Monitoring.....	31
3.4.6. Off-Hours Testing Monitoring.....	32
3.5. Quality Control of Test Administration	32
Chapter 4. Item Scoring.....	34
4.1. Machine-scored Items	34
4.1.1. Key-based Items	34
4.1.2. Rule-based Items	34
4.2. Human or Hand-scored Items	35
4.2.1. Scorer Training	36
4.2.2. Scorer Qualification	39
4.2.3. Managing Scoring.....	40
4.2.4. Monitoring Scoring.....	40

4.2.4.1. Second Scoring.....	40
4.2.4.2. Backreading	41
4.2.4.3. Validity.....	41
4.2.4.4. Calibration Sets.....	42
4.2.4.5. Inter-rater Agreement.....	42
4.3. Automated Scoring for PCRs.....	43
4.3.1. Concepts Related to Automated Scoring.....	43
4.3.1.1. Continuous Flow.....	43
4.3.1.2. Training of IEA Using Operational Data.....	43
4.3.1.3. Smart Routing	43
4.3.1.4. Quality Criteria for Evaluating Automated Scoring.....	44
4.3.1.5. Hierarchy of Assigned Scores for Reporting.....	44
4.3.2. Sampling Responses Used for Training IEA.....	45
4.3.3. Primary Criteria for Evaluating IEA Performance.....	45
4.3.4. Contingent Primary Criteria for Evaluating IEA Performance.....	45
4.3.5. Applying Smart Routing	46
4.3.6. Evaluation of Secondary Criteria for Evaluating IEA Performance	47
4.3.7. Inter-rater Agreement for Prose Constructed-Response	49
4.4. Quality Control of Scoring	50
Chapter 5. Performance Standards and Standards Validation	52
5.1. Performance Standards	52
5.2. Performance Levels and Definitions.....	52
5.3. Performance Level Setting (PLS) Process	53
5.3.1. Research Studies	53
5.3.2. Pre-Policy Meeting.....	54
5.3.3. Performance Level Setting Meetings.....	54
5.3.4. Post-Policy Reasonableness Review	55
Chapter 6. Item Analysis.....	56
6.1. Overview.....	56
6.2. Data Screening Criteria.....	56
6.3. Classical Item Analysis	57
6.3.1. Item Difficulty	57

6.3.2. Response Option or Score Point Proportions	57
6.3.3. Item-Total Correlations	58
6.3.4. Results of Classical Item Analysis	59
Chapter 7. Item Response Theory Analysis, Calibration and Scaling	62
7.1. Overview	62
7.2. IRT Models	62
7.3. Summary Statistics and Distributions from IRT Analyses	62
7.4. Scale Scores	65
7.4.1. Summative Score Scale and Performance Levels	65
7.4.2. ELA Reading and Writing Claim Scales	67
7.4.3. Subclaims Scales	69
7.4.4. Creating Conversion Tables	69
7.5. Scale Score Distributions	72
7.5.1. ELA Score Distributions	72
7.5.2. Mathematics Score Distributions	75
7.5.3. ELA Major Claims Score Distributions	77
7.5.4. Scale Score Distributions for Student Demographic Groups of Interest	78
7.6. Interpreting Claim Scores and Subclaim Scores	81
7.6.1. Interpreting Claim Scores	81
7.6.2. Interpreting Subclaim Scores	81
Chapter 8. Student Demographics and Differential Item Functioning (DIF)	83
8.1. Overview of Test-Taking Population	83
8.2. Rules for Inclusion of Students in Analyses	83
8.3. Demographics	83
8.4. Differential Item Functioning	86
8.4.1. Dichotomous Items: Mantel-Haenszel	86
8.4.2. Polytomous Items: Standardized Mean Difference	87
8.4.3. DIF Classification	87
8.4.4. Differential Item Functioning Results	88
Chapter 9. Reliability	91
9.1. Overview	91

9.2. Reliability and SEM Estimation	92
9.3 Scale Score Reliability Estimation	92
9.3. Reliability Results	93
9.3.1. Raw Score Reliability Results	93
9.3.2. Scale Score Reliability Results.....	95
9.4. Reliability Results for Demographic Groups of Interest	95
9.5. Reliability Estimates of Subclaim Scores.....	98
9.6. Reliability of Classification.....	101
Chapter 10. Validity.....	105
10.1. Overview	105
10.2. Evidence Based on Test Content.....	105
10.3. Evidence Based on Internal Structure	107
10.3.1. Intercorrelations	108
10.3.2. Reliability	110
10.3.3. Local Item Dependence	110
10.4. Evidence from Special Studies.....	115
10.4.1. Content Alignment Studies.....	115
10.4.2. Benchmarking Study.....	118
10.4.3. Longitudinal Study of External Validity of Performance Levels (Phase 1).....	118
10.4.4. Mode and Device Comparability Studies.....	119
10.4.5. Quality Testing Standards	121
10.4.5.1. Scale Score Comparability: Item-Level Analysis	125
10.4.5.2. Scale Score Comparability: Longitudinal Analysis.....	127
10.4.5.3. Performance Level Comparability: Test-Level Analyses	129
10.4.5.4. Performance Level Comparability: Classification Analyses	129
10.4.5.5. Performance Level Comparability: Longitudinal Analyses.....	129
10.4.5.6. Quality Testing Standards Summary.....	130
10.5. Evidence Based on Response Processes	130
10.6. Interpretations of Test Scores.....	131
10.7. Evidence Based on the Consequences to Testing.....	132
10.8. Summary	133

Chapter 11. Student Growth Measures	135
Chapter 12. References.....	143
Appendix 2	146
Appendix 6	153
A.6.1. Classical Item Analysis Statistics	154
Appendix 7	190
A.7.2. IRT Test Characteristic, Information and Standard Error of Measurement Curves.....	193
A.7.3. ELA Score Distributions.....	198
A.8.2. Post-Administration Differential Item Functioning Results	274
Appendix 9	288
Appendix 10.....	303
Appendix 11.....	334

List of Tables

Table 2.2.1 Form Composition for ELA Grade 3	16
Table 2.2.2 Contribution of Prose Constructed-Response Items to ELA	17
Table 2.3.1 Mathematics Form Composition for Grade 3.....	18
Table 4.2.1.1 Training Materials Used During Scoring.....	37
Table 4.2.2 Mathematics Qualification Requirements	40
Table 4.2.3 Scoring Hierarchy Rules.....	41
Table 4.2.4 Scoring Validity Agreement Requirements	42
Table 4.2.5 Inter-Rater Agreement Expectations and Results	42
Table 4.3.1 Comparison Groups.....	47
Table 4.3.2 Prose Constructed-Response Average Agreement Indices by Test.....	49
Table 6.3.1 Summary of Post-Administration P-values for ELA Operational Items by Grade.....	59
Table 6.3.2 Summary of Post-Administration P-values for Mathematics Operational Items by Grade	60
Table 6.3.3 Summary of Post-Administration ITC for ELA Operational Items by Grade.....	60
Table 6.3.4 Summary of Post-Administration ITC for Mathematics Operational Items by Grade.....	61
Table 7.3.1 IRT Parameter Estimates Summary for All Items for ELA by Grade.....	63
Table 7.3.2 IRT Parameter Estimates Summary for All Items for Mathematics by Grade.....	64

Table 7.3.3 IRT Parameter Distribution by Year for All Items for ELA Assessments.....	64
Table 7.3.4 IRT Parameter Estimates Summary for All Items for Mathematics Assessments	65
Table 7.4.1 Calculating Scaling Constants for Reading and Writing Claims	69
Table 7.5.1 Grade 3 ELA Scale Score Cumulative Frequencies	73
Table 7.5.2 Grade 3 Mathematics Scale Score Cumulative Frequencies	75
Table 7.5.3 Grade 3 ELA Subgroup Performance for Scale Scores	78
Table 7.5.4 Grade 3 Mathematics Subgroup Performance for Scale Scores	80
Table 8.3.1 ELA Test Takers	84
Table 8.3.2 Mathematics Test Takers	84
Table 8.3.3 Grade 3 ELA Test Taker Demographic Information	85
Table 8.3.4 Grade 3 Mathematics Test Taker Demographic Information	85
Table 8.4.1 DIF Comparison Groups.....	86
Table 8.4.2 Mantel-Haenszel Contingency Table.....	87
Table 8.4.3 DIF Classifications	88
Table 8.4.4 Grade 3 ELA Post-Administration Differential Item Functioning.....	88
Table 8.4.5 Grade 3 Mathematics Post-Administration Differential Item Functioning.....	89
Table 9.3.1 Summary of ELA Test Reliability Estimates for Total Group	94
Table 9.3.2 Summary of Mathematics Test Reliability Estimates for Total Group	94
Table 9.3.3 ELA Scale Score Reliability	95
Table 9.3.4 Mathematics Scale Score Reliability Results.....	95
Table 9.4.1 Grade 3 ELA Summary of Test Reliability Estimates for Subgroups.....	96
Table 9.4.2 Grade 3 Mathematics Summary of Test Reliability Estimates for Subgroups.....	97
Table 9.5.1 Average ELA Reliability Estimates for Subscores.....	99
Table 9.5.2 Average Math Reliability Estimates for Subscores.....	100
Table 9.6.1 Classification Accuracy Indices at Cut Score Level for ELA.....	102
Table 9.6.2 Classification Consistency Indices at Cut Score Level for ELA.....	102
Table 9.6.3 Classification Accuracy Indices at Cut Score Level for Mathematics.....	103
Table 9.6.4 Classification Consistency Indices at Cut Score Level for Mathematics.....	103
Table 10.3.1 Grade 3 ELA Average Intercorrelations between Subclaims	109
Table 10.3.2 Grade 3 Mathematics Average Intercorrelations between Subclaims	109
Table 10.3.3 Conditions Used in LID Investigation and Results	113

Table 10.3.4 Summary of Q3 Values for ELA Grade 4 and Integrated Mathematics II (Spring 2015)....	114
Table 10.4.1 Prior Grades Used in ELA Matching	122
Table 10.4.2 Prior Grades/Courses Used in Mathematics Matching.....	122
Table 10.4.3 ELA Matching Sample Size Results	123
Table 10.4.4 Mathematics Matching Sample Size Results	124

List of Figures

Figure 7.4.1 ELA Grade 3 Test Characteristic Curves, CSEM Curves, and Information Curves	72
Figure 7.4.2 Mathematics Grade 3 Test Characteristic Curves, CSEM Curves, and Information Curves...	72
Figure 7.5.1 Grade 3 ELA Score Distribution	73
Figure 7.5.2 Grade 3 Mathematics Score Distribution	75
Figure 7.5.3 Grade 3 ELA Reading Score Distribution.....	77
Figure 7.5.4 Grade 3 ELA Writing Score Distribution	78
Figure 10.3.1 Comparison of Internal Consistency by Item and Cluster (Testlet).....	113
Figure 10.3.2 Distribution of Q3 Values for Grade 4 ELA (Spring 2015).....	114
Figure 10.3.3 Distribution of Q3 Values for Integrated Mathematics II (Spring 2015).....	115
Figure 10.4.1 ELA Grades 3-6 P-Values.....	126
Figure 10.4.2 Mathematics Grades 3-6 P-Values.....	126

Chapter 1. Introduction

The purpose of this technical report is to describe the operational administration of the New Meridian Affiliate program's summative assessments in the 2022–2023 academic year, including test form construction, test administration, item scoring, student characteristics, classical item analysis results, reliability results, evidence of validity, item response theory (IRT) calibrations and scaling, performance level-setting procedure, growth measures, and quality control procedures. Throughout this Technical Report, only New Jersey student data is included in all analyses, descriptions, and data summaries.

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 tied into more rigorous academic expectations and help prepare students for success in college and the workforce, as well as 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 was 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 developed and administered next-generation assessments that, compared to traditional K–12 assessments, more accurately measured student progress toward college and career readiness. The PARCC-developed assessments were aligned to the Common Core State Standards (CCSS) and included both English language arts (ELA) assessments (grades 3 through 11) and mathematics assessments (grades 3 through 8 and high school). Compared to traditional standardized tests, these assessments were intended to measure more complex skills like critical thinking, persuasive writing, and problem-solving.

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

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

Department of Defense Education Activity, and the District of Columbia participated in the fourth administration in school year 2017–2018.

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

New Meridian, in coordination with multiple states and vendors, developed an alternate form of the summative assessment to meet the needs for shorter testing times desired by several states. Through extensive research and guidance from the Technical Advisory Committee, the alternate blueprint was available in spring 2019, in addition to the original blueprint. New Meridian’s state-centric solution to educational assessment provided states the flexibility of selecting the assessment solution that best fits their specific needs. For the academic year 2018–2019, participating states and agencies included the Bureau of Indian Education, the District of Columbia, Illinois, New Jersey, and New Mexico.

Most testing in spring 2020 was cancelled due to the COVID-19 global pandemic with the exception of a small number of students in Illinois who completed testing prior to the closure of schools. Some states further cancelled administration in spring 2021.

For the academic year 2022–2023, participating states and agencies included the Department of Defense Education Activity, Illinois, and New Jersey. All participating states and agencies administered assessments in spring 2023. This document reports the summative assessment results from New Jersey only.

1.2. Purpose of the Summative 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 the New Jersey Student Learning Standards (NJSLS) 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, which are included to provide data to support horizontal linking across test forms within a grade and content area, are proportionally representative of the operational test blueprint. The summative assessment is a mixed-format test: the current summative assessments are administered in either computer-based (CBT) or paper-based (PBT) format.

The ELA assessments focus on students’ ability to independently read and comprehend a range of sufficiently complex texts and to write effectively when analyzing text. These assessments contain literary and informational texts, and each passage set has four to eight brief comprehension and vocabulary

questions. ELA 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 New Jersey students taking ELA in grades 3 through 9 and/or mathematics in grades 3 through 8, as well as students taking high school mathematics (i.e., Algebra I, Geometry, Algebra II). 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. Overview of the Technical Report

This Technical Report presents the results of analyses for the operational administration of the spring 2023 forms. In this document, the term “operational items” refers to the scorable items that contribute to students' raw scores and scale scores. The term “field-test items” refers to the nonscorable items that were newly developed to collect statistics for future test administrations in New Jersey, calibrated by using the field-test spring 2023 administration data. The report begins by providing explanations of the test form construction process, test administration and scoring of the test items. Subsequent chapters of the report present descriptions of student characteristics and results of statistical analyses, including classical test theory statistics, item response theory (IRT) scaling and parameters, and differential item functioning (DIF) analyses.

The Technical Report contains the following chapters:

Chapter 2 – Test Development

This chapter describes the test design and procedures followed during the development of operational test forms, as well as characteristics for the 2023 forms.

Chapter 3 – Test Administration

This chapter presents the operational administration schedule, information regarding test security and confidentiality, accessibility features and accommodations, testing irregularities and security breaches, and administration quality control procedures.

Chapter 4 – Item Scoring

This chapter explains the key-based and rule-based processes for machine-scored items, as well as the training and monitoring processes for human-scored items.

Chapter 5 – Standard Setting

This chapter describes the performance levels and the processes previously followed to establish the performance thresholds for each subject area.

Chapter 6 – Item Analysis

This chapter describes the classical item-level statistics calculated for the operational test data and the flagging criteria used to identify items that performed differently than expected. Results of these analyses are presented in this section.

Chapter 7 – IRT Analysis, Calibration and Scoring

This chapter presents the information related to the item response theory (IRT) models and the descriptive statistics of the item parameters. The development of the reporting scales, conversion tables, and scale score distributions are also presented. Note that all tests delivered in 2023 employed a pre-equated model, in which previously estimated item parameters are used to generate scoring tables.

Chapter 8 – Student Demographics and Differential Item Functioning (DIF)

This chapter describes the rules for inclusion of students in analyses, distributions of students by grade, mode and gender, and distributions of demographic variables of interest. Also, methods for conducting DIF 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.

Chapter 9 – Reliability

This chapter presents the results of scale score reliability, internal consistency reliability analyses, and corresponding standard errors of measurement for content area and Mode (CBT or PBT) for all students, and subgroups of interest. This is followed by reliability of classification (i.e., decision accuracy and decision consistency).

Chapter 10 – Validity

This chapter presents the validity evidence in support of test score claims and based on analyses of the internal structure of the tests. Correlations between subscores are reported by content area and mode (CBT or PBT) for all students.

Chapter 11 – Student Growth Measures

This chapter provides details on student growth percentiles (SGP). Information about the model, model fit, and SGP averages at the overall level for all students, and for subgroups of interest, are provided.

Appendices

To facilitate utility, tables in the appendices are numbered sequentially according to the chapter represented by the tables. For example, the first appendix table for Chapter 6 is numbered A.6.1, the second appendix table for Chapter 6 is numbered A.6.2, and so on.

Chapter 2. Test Design and Development

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 to be college- and career-ready. The summative assessments are designed to examine 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. Throughout this ongoing process, test developers create 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.

2.1. Overview of the Test

The summative assessments include both ELA and mathematics assessments in grades 3 through 8 and high school. Assessments contain selected-response, short and extended constructed-response, technology-enabled and technology-enhanced items (TEIs), 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; the act of performing the task is the way in which data is collected. Students may be asked, among other interactions, to categorize information, organize or classify data, order a series of events, plot data, generate equations, or highlight text. 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, and American Sign Language [ASL] video versions of the test, as well as response accommodations that allow students to respond to test items using different formats). English learners who are native Spanish speakers may take the mathematics assessments in Spanish; and both LP and text-to-speech versions of the test are available in Spanish. (Please refer to the Accessibility Features and Accommodations Manual for in-depth information.)

2.2. ELA Claims and Subclaims

The ELA summative assessment contains literary and informational reading passages with comprehension and vocabulary questions. The ELA 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 either read or view a video version of 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 claim structure, grounded in the NJSLS, undergirds the design and development of the ELA summative assessments. These claims are referred to as the Master Claim, Major Claims, and Subclaims.

Master Claim. The Master Claim is the overall performance goal for the ELA 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. Major Claims describe the evidence needed to sufficiently yield scale scores for making longitudinal comparisons. These Major Claims consist of 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 the standards and evidence statements outlined in the evidence tables for reading and writing (refer to the test specifications documents). The claims and evidence statements are grouped into the following categories:

- Reading: Vocabulary
- Reading: Literary Text
- Reading: Informational Text
- Writing: Written Expression
- Writing: Knowledge of Language and Conventions

The structure of the grade 3 ELA exam is outlined in Table 2.2.1. Corresponding information is provided in Appendix 2.1 for all grades.

Table 2.2.1 Form Composition for ELA Grade 3

Major Claims	Subclaims	Number of Items	Number of Points
Reading	Literary Text	4–6	11–12
	Informational Text	4	11
	Vocabulary	4	8
	Claim Total	12–14	30–31
Writing	Written Expression	2	18
	Knowledge of Language and Conventions	2	6
	Claim Total	4	24
Summative Total		14–16*	54–55

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Each ELA 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 assessments include two PCR tasks. Table 2.2.2 details the number of possible points and the associated subclaims for the three PCR tasks:

- Literary Analysis

- Research Simulation
- Narrative Writing

All ELA assessments include the Research Simulation task and either the Literary Analysis or the Narrative Writing tasks. The Literary Analysis task and the Research Simulation task are scored for two traits: Reading Comprehension and Written Expression, and Knowledge of Language and Conventions. The Narrative Writing task is scored for two traits: Written Expression and Knowledge of Language and 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 PCR items are provided in Table 2.2.

Table 2.2.2 Contribution of Prose Constructed-Response Items to ELA

Grade	Score	Possible Points		
		Literary Analysis Task*	Research Simulation Task*	Narrative Writing Task*
3	Reading Comprehension	3	3	0
	Written Expression	9	9	9
	Knowledge of Language and Conventions	3	3	3
	Total	15	15	12
4–5	Reading Comprehension	4	4	0
	Written Expression	12	12	9
	Knowledge of Language and Conventions	3	3	3
	Total	19	19	12
6–11	Reading Comprehension	4	4	0
	Written Expression	12	12	12
	Knowledge of Language and Conventions	3	3	3
	Total	19	19	15

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

2.3. Mathematics 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 NJSLs with a focus on modeling and reasoning with precision.

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

Master Claim: The master claim indicates the degree to which a student is on track to being college- or career-ready. The student solves grade-level/course-level problems aligned to the New Jersey Student Learning Standards, which include 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 the information outlined in the evidence statement tables for mathematics (refer to the test specifications documents). The subclaims 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 Practices with Connections to Content: modeling/application by solving real-world problems by applying knowledge and skills articulated in the standards.

Table 2.3.1 includes the numbers of items and points associated with subclaim scores for mathematics grade 3, as an example of the composition of the mathematics tests. Because there is substantial variation in the composition of the tests corresponding information is provided in the tables in Appendix 2.1 for all mathematics grades/courses.

Table 2.3.1 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

2.4. Test Development Activities

PARCC test development activities began with the standards and model content frameworks. From these, more than 2,000 educators, researchers, and psychometricians created 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, and the Text

Selection Guidelines. Refer to the New Meridian Resource [website](#) for further information about these documents.

2.4.1. Item Development Process

Test and item development activities were conducted under the guidance and oversight of the Technical Advisory Committee, the New Jersey state content leads, the Content and Bias/Sensitivity Text and Item Review Committees, and staff members from New Meridian.

Developing high-quality assessment content with authentic stimuli that measures rigorous learning standards is a complex process involving the services of many experts. New Meridian accomplishes this by overseeing robust teams of assessment designers, psychometricians, managers, trainers, content providers, content experts, editors, artists, programmers, technicians, human scorers, advisors and members of the state content leads and Accessibility, Accommodations, and Fairness (AAF) leads or experts.

2.4.1.1. 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 the state content leads for ELA and mathematics, evaluated the strengths of the bank and considered the needs for future tests to establish an item development plan.

2.4.1.2. Text Selection for ELA

Using the Passage Selection Guidelines, ELA 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 passage types to allow a range of standards/evidences that meet the assessment claims. ELA tests are based on authentic texts, including multi-media stimuli. Authentic texts are previously published grade-appropriate texts that reflect the original language of the authors; authentic texts are not developed for assessment purposes or to achieve a particular readability metric. 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. ELA item development was not conducted until after texts were approved by the Text Review Committee.

2.4.1.3. Item Development

Guided by foundational documents described above, New Meridian managed the item writing process to develop the number of items specified in the annual asset development plan. Prior to further committee reviews, the assessment content teams at New Meridian 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), and bias and sensitivity to enable accurate measurement of the standards.

2.4.2. Item and Text Review Committees

State content leads for ELA and mathematics, as well as state-level experts and local educators, conducted rigorous reviews of every item and passage being developed for the summative assessment system. These reviews ensured all test items are of the highest quality, aligned to the standards, and fair for all student populations. All reviewers were nominated by their state education agency. The purpose of the educator reviews was to provide feedback to New Meridian 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.

2.4.2.1. Text Review

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

2.4.2.2. Content Item Review

During Content Item Review meetings, committees reviewed and edited test items for adherence to the foundational documents, basic universal design principles, accessibility guidelines, and associated item metadata. Committees accessed the item content within the Pearson Assessment Banking for Building and Interoperability (ABBI) system that previews how the passages and items will be displayed in an operational online environment. Committees also verified that the appropriate scoring rule had been applied to each item. The Content Item Review Committees were made up of state content leads and state approved subject matter experts (SMEs).

2.4.2.3. 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.

2.4.2.4. Editorial Review

NMC and Pearson editorial staff reviewed all items and tasks. The committee reviewed the items for grammar, punctuation, clarity and adherence to the Style Guide.

2.4.2.5. Data Review

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

2.4.3. Operational Test Construction

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

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

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

Multiple core forms were constructed for a given assessment to enhance test security and to support an item release plan. Core forms were the operational test forms consisting of only those items that counted toward a student's total score. Core forms were constructed of a set of items unique to each core and a subset of items shared between cores used for linking. 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 including genre, topics, word count and text complexity.

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

2.4.3.1. Test Construction Activities

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

1. Online forms were constructed to match the blueprint and test construction specifications.
2. Paper forms were constructed to match the blueprint and test construction specifications.
3. Accommodated and accessibility forms were constructed 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 New Meridian 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. Linking across administrations for operational forms was accomplished by including prior operational items on the current operational test forms.

NMC assessment specialists identified forms for each grade/subject suitable for use as the accommodated forms. The content of these forms was also reviewed by accessibility specialists allowing for content changes prior to the Test Construction Committee meetings.

These test construction activities provided the meeting materials necessary to conduct test form verification meetings. These meeting materials included:

- 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.

2.4.3.2. Test Form Verification Meeting to Review Test Construction Inputs

Members of the Content Item Review Committees and the AAF experts participated in the building of operational core forms that met the summative assessment requirements. During this process, members met in a virtual meeting to review and make recommendations for changes so that test forms conformed to both the content and psychometric requirements of the assessment.

2.4.3.3. 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 experts reviewed the proposed accommodated forms at the Test Form Verification meeting for accessibility to make sure that the content can be accommodated for students with disabilities and English learners without changing the underlying measured construct.

Forms were identified to support the following accommodations:

Accommodated Base 1

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

Accommodated Base 2

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

Accommodated Base 3 (mathematics only)

- Text-to-speech second form

Spanish is used for mathematics only. Closed captioning is used for ELA only.

At the conclusion of the meetings, all test forms were constructed to meet test blueprints and requirements, and if necessary, reflect the operational linking design. Each test form reflected the test blueprint in terms of content, item types, and test length, as well as expected difficulty and performance

along the ability continuum. Linking sets were proportionally representative of the operational test blueprint. The operational core forms, linking item set, and field-test item sets were reviewed during the test form verification meeting and approved prior to the test administration.

2.4.3.4. Spanish-Language Assessments for Mathematics

For English learners, the mathematics assessments are offered in Spanish, as well as in Spanish-language large print and text-to-speech versions. Once the operational form was approved, the items were transadapted. 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 in grade-level and culturally appropriate transadaptation. For the Spanish language text-to-speech 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 text-to-speech form. Phonetic mark-up, which guides how the text-to-speech 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 experts with assistance from content experts at the test construction meetings, the transadapted forms underwent additional quality checks: a Pearson Spanish copy edit services review and approval, and an AAF experts review and approval.

2.4.4. Linking Design of the Operational Test

To support the goal of score comparability within and across administrations and years, a hybrid linking approach was implemented that incorporated the strengths of common-item linking and randomly equivalent groups. The use of items shared across 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 (Kolen & Brennan, 2004).

Across-administration linking, or year-to-year linking, consisted of common items included in two different administrations. This approach 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.

2.4.5. Field Test Data Collection Overview

Field test items were embedded in the spring operational mathematics forms. Field test items for ELA operational forms were administered to the students from a stratified random sample of New Jersey schools. As described in Chapter 2.4.2.5, field tested items undergo a thorough data review process. Field tested items are not used in determining student performance on the assessment, but instead can be used in future operational administrations if approved by the Data Review Committee.

Chapter 3. Assessment Administration

3.1 Test Security and Administration Policies

The administration of the summative assessments is a secure testing event. Maintaining the security of test materials before, during, and after the test administration is crucial to obtaining valid and reliable results. School Test Coordinators are responsible for ensuring that all personnel with authorized access to secure materials are trained in and subsequently act in accordance with all security requirements.

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

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

3.1.1. Secure vs. Nonsecure Materials

Participating states and agencies define secure materials as those that must be closely monitored and tracked to prevent unauthorized access to or prohibited use or distribution of secure content such as test items, reading passages, student work, and so on. For paper-based tests (PBTs), secure materials include both used and unused test booklets and used scratch paper, while for computer-based tests (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 on, and so on.

3.1.2. Scorable vs. Nonscorable Materials

Paper-based assessments have both scorable and nonscorable materials, while computer-based assessments have only nonscorable materials. Scorable materials for paper-based assessments consist of used test booklets (for grade 3, and which may include student work) and answer documents (grades 4 and above). 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, including printed student testing tickets, prior to testing. 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 paper-based testing students include large-print test booklets, braille test booklets, scratch paper (paper used by students to take notes and work through items), and printed mathematics reference sheets (grades 5 through 8 and high school).

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

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

All test security and administration policies are found in the *Test Coordinator Manual* and the *Test Administrator Manual*.

3.2. Accessibility Features and Accommodations

3.2.1. Participation Guidelines for Assessments

All students, including students with disabilities (SWDs) and English learners (ELs), are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for ELs in their first year in a U.S. school, and certain SWDs 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 Individuals with Disabilities Education Act of 2004 (IDEA), Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008), and the Elementary and Secondary Education Act (ESEA) of 1965, as amended by the Every Student Succeeds Act (ESSA). All students can receive accessibility features on the summative assessments.

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

1. SWDs 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 ELs.
4. Students who are ELs with disabilities and have an IEP or 504 plan.

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

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 and through item development, field testing, and implementation of the assessments for all students, including SWDs, ELs, and ELs with disabilities. Accommodations may still be needed for some SWDs and ELs to assist in demonstrating what they know and can do. 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 computer-based assessments, accessibility features are tools or preferences that are either built into the assessment system or provided externally by Test Administrators and may be used by any student taking the summative assessments (i.e., students with and without disabilities, gifted students, ELs, and ELs 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 the accessibility features 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 NJGPA, accommodations are considered adjustments to the testing conditions, test format, or test administration which provide equitable access during assessments for SWDs and students who are ELs. In general, the administration of the assessment should not be the first occasion on which an accommodation is introduced to the student. To the extent possible, accommodations should do the following:

- Provide equitable access during instruction and assessments.
- Mitigate the effects of a student's disability.
- Not reduce learning or performance expectations.
- Not change the construct being assessed.

- Not compromise the integrity or validity of the assessment.

Accommodations are intended to reduce and/or eliminate the effects of a student's disability and/or English language proficiency level; however, accommodations should never reduce learning expectations by reducing the scope, complexity or rigor of an assessment. Moreover, accommodations provided to a student on the summative assessments must be generally consistent with those provided for classroom instruction and classroom assessments. There are some accommodations that may be used for instruction and for formative assessments that are not allowed for a 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 NJDOE 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 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 SWDs 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 ELs should be described and documented.
- Students who are ELs with disabilities are eligible to receive accommodations for both SWDs and ELs.
- Accommodations should become part of the student's program of daily instruction as soon as possible after completion and approval of the appropriate plan.
- Accommodations should not be introduced for the first time during the testing of a student.
- Accommodations should be monitored for effectiveness.
- Accommodations used for instruction should also be used, if allowable, on local district assessments and state assessments.
- In the following scenarios, the school must follow each state's policies and procedures for notifying the state assessment office:
 - A student was provided a test accommodation that was not listed in his or her IEP/504 plan/documentation for an English learner, or
 - A student was not provided a test accommodation that was listed in his or her IEP/504 plan/documentation for an English learner.

3.2.5. Unique Accommodations

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

3.2.6. Emergency Accommodations

Emergency accommodation may be appropriate for a student who incurs a temporary disabling condition that interferes with test performance shortly before or during the assessment window. A student, whether or not they already have an IEP or 504 plan, may require accommodation as a result of a recently occurring accident or illness. Cases include a student who has a recently fractured limb (e.g., arm, wrist, or shoulder); a student whose only pair of eyeglasses has broken; or a student returning to school after a serious or prolonged illness or injury. Emergency accommodation should be given only if the accommodation will result in a valid score for the student (i.e., does not change the construct being measured by the test[s]). If the principal (or designee) determines that a student requires an emergency accommodation on the summative assessment, an Emergency Accommodation Form must be completed and maintained in the student's assessment file. The parent must be notified that 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 EL 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, or (if required by the member state) EL 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 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 distributed or while students are testing.
- Deviating from testing time procedures.
- Allowing cheating of any kind.
- Providing unauthorized persons with access to secure materials.
- Unlocking a test in PearsonAccess^{next} during non-testing times.
- Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore not appropriate.
- Allowing students to test before or after the state's test administration window.

Test Materials:

- Losing a student test booklet or answer document.
- Losing a student testing ticket.
- Leaving test materials unattended or failing to keep test materials secure at all times.
- Reading or viewing the passages or test items before, during, or after testing (*Exception:* Administration of a human reader/signer accessibility feature for mathematics or accommodation for English language arts, 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 are 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 integral components 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.” 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 content area. This analysis was conducted for prose constructed-response tasks administered online using some of the same artificial intelligence techniques that are applied in automated essay scoring. Specifically, this method was based on latent semantic analysis (LSA) technology to detect possible plagiarism. Using LSA, the content of each constructed response was compared against the content of every other constructed response, and a measure that indicated the degree 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 is a process that checks for suspicious activity occurring outside computer-based testing windows. Participating states and agencies have set start and end times for administering computer-based assessments. Authorized users in the state role were allowed to override these start and end times. The off-hours testing monitoring tracked such overrides and logged them in an operational report. States could use this report to follow up with the organizations.

3.5. Quality Control of Test Administration

Pearson provided high-quality materials in a timely and efficient manner to meet the needs of the test administration. 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. Chapter 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 prepress and

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

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

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

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

Chapter 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 New Meridian content staff 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 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 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 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 scored item or task 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.

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. These processes are the same for both paper and online modes of testing.

4.2. Human or Hand-scored Items

Constructed-response items were scored by human scorers in a process referred to as hand scoring. Online training units were used to train all scorers. The online training units included prompts (items), passages, rubrics, training sets and qualification sets. Scorers who successfully completed the training and demonstrated 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 (ELA) 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 range-finding and rangefinder review meetings where educators and administrators from states met to interpret the scoring rubrics and determine consensus scores for student responses. Range-finding meetings were held prior to scoring field-test items, and range-finding review meetings were held prior to scoring operational items.

At range-finding 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 range-finding 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, there was one prototype training set per task type (i.e., Research Simulation, Literary Analysis, Narrative Writing). 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 task type or mathematics item grouping was based on the prototype. Once a prototype was chosen, full training materials were developed for that item. Then, scorers at each grade level 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 4.2.1 details the composition of the anchor sets, practice sets, and qualification sets.

Table 4.2.1.1 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 comprises one to three annotated responses per score point.
	The anchor sets for ELA 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 for mathematics prototype and abbreviated items include two to three sets of ten annotated responses.
The practice sets provide guidance and practice for trainees in defining the line between score categories, as well as applying the scoring criteria to a wider range of types of responses.	ELA practice sets for prototype items include two sets of five annotated responses and two sets of 10 annotated responses. The subsequent ELA practice sets for abbreviated items include two sets of ten annotated responses.

Training Set Development

Description	Specification
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 predetermined 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 10 responses each (not annotated). The subsequent mathematics abbreviated items for mathematics do not include qualification sets.
	The qualification sets for ELA prototype items include three sets of 10 responses each (not annotated). The subsequent ELA abbreviated items do not include qualification sets.

4.2.2. Scorer Qualification

To score items, scorers were required to show that they were able to apply scoring methodology accurately through a qualification process. Scorers were asked to apply scores to three qualification sets consisting of 10 responses each. ELA 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 Knowledge of Language and Conventions trait. The Narrative Writing Task had two traits: Written Expression and Knowledge of Language 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 to qualify.

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

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

For mathematics qualification, the requirements were based on the item type and score point range. Because mathematics items can have one or more scoring traits, a scorer needed to achieve the requirements as set forth in Table 4.2.2 separately for each scoring trait (when applicable to the

item).

Table 4.2.2 Mathematics Qualification Requirements

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

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

4.2.3. Managing Scoring

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

4.2.4. Monitoring Scoring

4.2.4.1. Second Scoring

During scoring, Pearson’s ePEN2 scoring system automatically and randomly distributed a minimum of 10 percent of student responses for second scoring; scorers had no indication whether a response had been scored previously. Humans applied the second score for all mathematics items. The second scoring for ELA was performed either by human scorers or by Pearson’s Intelligent Essay Assessor. If the first and second scores applied were nonadjacent, a third and occasionally a fourth score were assigned to resolve scorer disagreements. When a resolution score (i.e., third score) was nonadjacent to one or both of the first and second scores, the content specialist or scoring director would apply an adjudication score (fourth score). If a response was scored more than once, the rules in Table 4.2.3 were applied to determine the final score.

Table 4.2.3 Scoring Hierarchy Rules

Score Type	Rank	Final Score Calculation
Adjudication	1	If an adjudication score is assigned, this is the final score.
Resolution	2	If no adjudication score is assigned, this is the final score.
Backread	3	If no adjudication or resolution score is assigned, the latest backreading score is the final score.
Human First Score	4	If no adjudication, resolution, or backreading score is assigned, this is the final score.
Human Second Score	5	If no adjudication, resolution, backreading, or human first score is assigned, this is the final score.
Intelligent Essay Assessor Score	6	If no human score is assigned, this is the final score.

4.2.4.2. Backreading

Backreading was one of the major responsibilities of Pearson Scoring Supervisors and a primary tool for proactively guarding against scorer drift, where scorers score responses in comparison to one another instead of in comparison to the training responses. Scoring supervisory staff used the ePEN2 backreading tool to review scores assigned to individual student responses by any given scorer in order to confirm that the scores were correctly assigned and to give feedback and remediation to individual scorers. Pearson backread approximately 5 percent of the hand-scored responses. Backreading scores did not override the original score but were used to monitor scorer performance.

4.2.4.3. Validity

Validity responses are prescored responses strategically interspersed in the pool of live responses. These responses were not distinguishable from any other responses so that scorers were not aware they were scoring validity responses rather than live responses. The use of validity responses provided an objective measure that helped ensure that scorers were applying the same standards throughout the project. In addition, validity was at times shared with scorers in a process known as “validity as review.” Validity as review provided scorers automated, immediate feedback: a chance to review responses they mis-scored, with reference to the correct score and a brief explanation of that score. One validity response was sent to scorers for every 25 “live” responses scored.

Validity agreement requirements for scorers are listed in Table 4.2.4. 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 4.2.4 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	Multi-trait	65%	96%

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

4.2.4.4. 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 range-finding standards, introduce scoring decisions, or address scoring issues and trends. Calibration was used either to correct a scoring issue or trend or continue scorer training by introducing a scoring decision. Calibration was administered regularly throughout scoring.

4.2.4.5. 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 4.2.5.

Table 4.2.5 Inter-Rater Agreement Expectations and Results

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

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

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

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

The results for the ELA PCR are provided in Chapter 4.3.7, “Inter-rater Agreement for Prose Constructed-Response.”

4.3. Automated Scoring for PCRs

Automated scoring performed by Pearson’s Intelligent Essay Assessor (IEA) was the default option for scoring the summative assessment’s online prose constructed-response (PCR) tasks. Under the default option, it was assumed that operational scores for approximately 90 percent of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. Human scoring was applied to responses that were scored while IEA was being trained as well as to additional responses routed to human scoring when there was uncertainty about the automated scores.

For 10 percent of responses, a second “reliability” score was assigned. The purpose of the reliability score was to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. When IEA provided the first score of record, the second reliability score was a human score.

4.3.1. Concepts Related to Automated Scoring

The sections below describe concepts related to automated scoring.

4.3.1.1. 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.

4.3.1.2. Training of IEA Using Operational Data

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

4.3.1.3. 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.

4.3.1.4. 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-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- Quadratic-weighted kappa between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA-human should be less than 0.15.

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

- Primary Criteria — Based on responses to validity papers: With smart routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.
- Contingent Primary Criteria — Based on the training responses if validity responses are not available: With smart routing applied as needed, IEA-human exact agreement is within 5.25 percent of human-human exact agreement for each trait score.
- Secondary Criteria — Based on the training responses: With smart routing applied as needed, IEA-human differences on statistical measures for each trait score are within the Williamson et al. (2012) tolerances for subgroups with at least 50 responses.

4.3.1.5. 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 2023 operational data, the early performance of human scoring was closely monitored to verify that an appropriate set of data would be available for training IEA. In particular, several characteristics of the human scoring data were monitored, including:

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

Although the desired characteristics of the training data were easily achieved for some prompts, they were more challenging to achieve for others. For some prompts, a subset of scores were reset and clarifying directions were provided to scorers to improve human-human agreement. For other prompts, special sampling approaches were used to increase the numbers of responses that received top scores. In addition, a healthy percentage of responses were backread during the sampling period and these scores as well as double human scores were all part of the data used to train IEA.

4.3.3. Primary Criteria for Evaluating IEA Performance

The primary criteria for evaluating IEA performance are based on evaluating validity papers and are 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 the overall validity agreement, conditional agreement was also examined. In general, it was desirable for IEA to exceed 65 percent agreement at every score point as well as be close to or exceed the human validity agreement at each score point.

4.3.4. Contingent Primary Criteria for Evaluating IEA Performance

For many of the prompts trained in 2023, it was not possible to utilize human-scored validity

responses in evaluating IEA performance. In these cases, IEA was evaluated based on IEA-human exact agreement for each trait score and compared to agreement based on responses that were double-scored by humans. A portion of the data was held out for evaluating IEA-human exact agreement according to the following steps:

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

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

4.3.5. Applying Smart Routing

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

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

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

1. The continuous IEA score for each of the two trait scores was rounded to the nearest score interval of 0.2, starting from zero. For example, IEA scores between 0 and 0.1 were rounded to an interval score of 0, scores between 0.1 and 0.3 were rounded to an interval score of 0.2, scores between 0.3 and 0.5 were rounded to an interval score of 0.4, and so on.
2. Within each of these intervals, the percentage of exact agreement between IEA integer

scores and the human scores was calculated for each trait.

3. For each prompt, agreement rates were evaluated by rounding interval. Those intervals for which the agreement rates were below a designated threshold for either trait were identified.
4. Once IEA scoring was implemented, responses within intervals for which IEA-human agreement was below the designated threshold were routed for additional human scoring.

In training IEA, the scoring models without smart routing were evaluated first by applying either the primary validity criteria or the contingent criteria as described in Chapter 4.3.4. For those prompts that did not meet these criteria, increasing smart routing thresholds were applied in an iterative fashion to filter scores and evaluate the remaining scores against the criteria. That is, in any one iteration a particular smart routing threshold was applied such that only scores falling in intervals for which exact agreement exceeded the threshold were included in evaluating the criteria. If the primary or contingent criteria were not met with this level of smart routing, an increased smart routing threshold was applied iteratively until the primary or contingent criteria were met, or the maximum threshold reached. If the criteria were still not met after a maximum threshold was applied, different models were investigated and/or additional human scoring data utilized until an IEA scoring model was found that met the criteria.

4.3.6. Evaluation of Secondary Criteria for Evaluating IEA Performance

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

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

Once IEA was trained and deployed, scoring sets used in training were matched to demographic information so that agreement between IEA and human scorers could be evaluated across subgroups. The analysis was conducted for the eleven comparison groups outlined in Table 4.3.1.

Table 4.3.1 Comparison Groups

Group Type	Comparison Groups
Sex	Female
	Male

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

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

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

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

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

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

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

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

4.3.7. Inter-rater Agreement for Prose Constructed-Response

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

For 10 percent of responses, a second “reliability” score was assigned. The purpose of the reliability score is to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. Inter-rater agreement is the agreement between the first and second scores assigned to student responses and is the measure of how often scorers agree with each other. Pearson scoring staff used inter-rater agreement indices as one factor in determining the needs for continuing training and intervention on both individual and group levels. Inter-rater agreement expectations are provided in Table 4.5 in Chapter 4.2.4. For ELA PCR traits, the expectation for agreement is an inter-rater agreement of 65 percent or higher between two scorers. When IEA provided the first score of record, the second reliability score was a human score. For a subset of responses, the first and second score were both human scores.

Table 4.3.2 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 4.3.2 presents the average count and the average for the agreement indices.

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

Table 4.3.2 Prose Constructed-Response Average Agreement Indices by Test

			Written Expression				Conventions			
Grade	N-PCRs	Count	Exact	Kappa	QW Kappa	r	Exact	Kappa	QW Kappa	r
3	4	26,875	74.43	0.57	0.74	0.75	75.1	0.6	0.78	0.78
4	4	26,078	73.63	0.61	0.81	0.81	73.08	0.6	0.8	0.8
5	4	28,513	73.45	0.6	0.81	0.81	73.55	0.6	0.81	0.82
6	3	51,516	75.93	0.65	0.86	0.86	76.5	0.66	0.85	0.85
7	3	74,197	74.17	0.64	0.87	0.87	73.83	0.64	0.86	0.86
8	4	29,737	75.3	0.66	0.9	0.9	75.03	0.66	0.88	0.88
9	4	4,954	81.95	0.75	0.91	0.91	80.95	0.74	0.9	0.9

4.4. Quality Control of Scoring

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

For the 2015 operational administration, Pearson's validation team prepared test plans used throughout the scoring process. Test plan preparation was organized around detailed specifications that continued to the 2023 operational administration.

Based on lessons learned from previous administrations, the following quality steps were implemented:

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

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

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

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

- Pearson's psychometrics team conducted empirical analyses based on preliminary data files and flagged items based on statistical criteria.
- Pearson's content team reviewed the flagged items and provided feedback on the accuracy of content, answer keys, and scoring.
- Items potentially requiring changes were added to the product validation log for further investigation by other Pearson teams.
- Staff was notified of items for which keys or scoring changes were recommended.

- Participating states and agencies approved or rejected scoring changes.
- All approved scoring changes were implemented and validated prior to the generation of the data files used for psychometric processing.

Chapter 5. Performance Standards and Standards Validation

5.1. 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.

5.2. Performance Levels and Definitions

For the NJSLA 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, may be found below.

Level 5: Exceeded Expectations:

Students performing at this level exceed academic expectations for the knowledge, skills, and practices contained in the standards for ELA or mathematics assessed at their grade level. They are academically well prepared to engage successfully in further studies in this content area.

Level 4: Met Expectations:

Students performing at this level met academic expectations for the knowledge, skills, and practices contained in the standards for ELA or mathematics assessed at their grade level. They are academically prepared to engage successfully in further studies in this content area.

Level 3: Approached Expectations:

Students performing at this level approach academic expectations for the knowledge, skills, and practices contained in the standards for ELA or mathematics assessed at their grade level. They are likely prepared to engage successfully in further studies in this content area.

Level 2: Partially Met Expectations:

Students performing at this level partially meet academic expectations for the knowledge, skills, and practices contained in the standards for ELA or mathematics assessed at their grade level. They will likely need academic support to engage successfully in further studies in this content area.

Level 1: Did Not Yet Meet Expectations:

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

5.3. Performance Level Setting (PLS) Process

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 to establish performance standards for the summative assessments are:

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

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

5.3.1. 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 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. Additional details about the benchmarking study can be found in the *Performance Level Setting Technical Report* as well as in the *PARCC Benchmarking Study Report*. Additional details 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*.

5.3.2. Pre-Policy Meeting

Prior to the PLS meetings, a pre-policy meeting was convened to determine reasonable ranges that would be shown to panelists during the high school PLS meetings. Pre-policy meeting participants included representatives from both K–12 and higher education who served in roles such as commissioner/superintendent, deputy/assistant commissioner, state board member, director of assessment, director of academic affairs, senior policy associate, and so on. The reasonable ranges recommended by the pre-policy meeting defined the minimum and maximum percentage of students that would be expected to be classified as college- and career-ready. The pre-policy meeting participants reviewed the test purpose, how the performance standards will be used, and the results of the research studies to provide the recommendations for the reasonable ranges without viewing any student performance data.

5.3.3. 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 Affiliate 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 Angoff standard setting method (Plake et al., 2005) allowed for incorporation of the multipoint items by asking educators to evaluate (Yes or No) whether a borderline student would earn the maximum number of points on an item, a lesser number of points on an item, or no points on the item. In the case of a single point or multiple-choice item, this task simplifies to the standard Yes/No method.

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

A dry run of the PLS meeting process was held for grade 11 ELA and Algebra II in order to evaluate the implementation of the PLS method with the innovative characteristics of the summative assessments. 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*.

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

5.3.4. 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 grade 11 assessments) the Advisory Committee on College Readiness as part of a post-policy reasonableness review. This group reviewed both the median threshold score recommendations from each committee and the variability in the threshold scores as represented by the standard error of judgment (SEJ) of the committee. Adjustments to the median threshold scores that were within 2 SEJ were considered to be consistent with the PLS panels' recommendation.

In addition to voting to adopt the performance standards based on the committee's recommendations, this group also voted to conduct a shift in the performance levels to better meet the intended inferences about student performance. Holding the college- and career-ready (or on-track) expectations (i.e., the current level 4) constant, performance levels above this expectation were combined and performance levels below this expectation were expanded to create the final system of performance levels with three below and two above the college- and career-ready (or on-track) expectation. The shift in performance levels was accomplished using a scale anchoring process that involved two primary steps. In the first step, the top two performance levels, above college- and career-ready (or on track), were combined into a single performance level and an additional performance level below college- and career-ready (or on track) was created by empirically determining the midpoint between the existing two levels. In the second step, the performance level descriptors (PLDs) were updated using items that discriminated student performance well at this level to create a PLD aligned with the new empirically determined performance level. At this same time, PLDs for all performance levels were reviewed for consistency and continuity. Members of the original PLS committees were recruited to participate in this process. Additional information about this process can be found in the *Performance Level Setting Technical Report*.

Chapter 6. Item Analysis

6.1. Overview

This chapter describes the results of the classical item analyses conducted for the operational items on the spring 2023 operational NJSLA forms. Item analysis serves two purposes: to inform item exclusion decisions for IRT analysis and to provide item statistics for the item bank.

An operational item may appear on multiple core forms. The tables below list unique item counts for the assessments. Some items may have been used on more than one core form; in this case, item statistics are based on all student responses. PCR items in the NJSLA ELA are scored on two traits: (1) Written Expression and (2) Knowledge of Language and Conventions. NJSLA ELA item counts include both scored PCR traits.

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. Spoiled or “do not score” items, if any, are excluded from the total test score in item analysis. These items are removed from scoring because of item performance, technical scoring issues, content concerns, or multiple/no correct answers.

Item-level analyses were conducted for each form. These analyses included difficulty (p-value and pseudo p-value) and discrimination (item-total correlation).

6.2. Data Screening Criteria

Analyses were performed on an Incomplete Data Matrix (IDM) that was generated from the scored student results files. Item analyses were conducted by form. Student records were removed prior to running the analyses if the records met any of the criteria indicated in the test calibration specifications.

Within the IDM for a given test form, data can be missing due to students omitting responses to one or more items or students not reaching one or more items. Items are counted as “omitted” (i.e., skipped) when a student did not provide a response when items coming before and after have student responses. Omitted responses to items in students’ response strings are re-coded to a value of “0” and are treated as being incorrect during classical test theory (CTT) analyses, as well as during IRT item calibration and student scoring. Missing data from not reaching items occurs when there is a contiguous series of items with missing responses, which continues to the last item administered to a given student. Item response data in earlier portions of student response strings exists prior to the first missing value from not reaching a given item. Missing data due to not reaching items is not re-coded in any way, however. Missing data in student response strings due to not reaching one or more items, or not being administered one or more items, is excluded from classical test theory (CTT) analysis and IRT item calibration and student scoring, and therefore do not contribute to the item statistics or scores.

6.3. Classical Item Analysis

6.3.1. Item Difficulty

The p -value for dichotomous items—that is, one-point items scored as either correct or incorrect—is the mean item score, which is calculated as the proportion of examinees who answer the item correctly among the total number of examinees having scored data for the item. The formula to calculate the p -value for dichotomous items is:

Equation 6.1

$$p\text{-value} = \bar{x}_i = \frac{1}{n} \cdot \sum_1^n x_i$$

where x_i are the individual student item scores on item i and n is the total number of students having scored data for the item.

For polytomous items, the pseudo p -value is calculated by dividing the average score on the item by the maximum obtainable points possible for the item

Equation 6.2

$$\text{pseudo } p\text{-value} = \frac{\bar{x}}{T}$$

where $\bar{x}$ is the mean item score and T is the maximum obtainable points possible for the item.

P -value and pseudo p -value item statistics are bounded between 0 and 1. For these statistics, higher values indicate easier items while lower values indicate more difficult items. Frequently, the p -value and pseudo p -value are reported as percentages that are calculated by multiplying the proportions by 100. For instance, a p -value of 0.67 means that 67 percent of the students answered the dichotomous item correctly. On the other hand, a pseudo p -value of 0.67 means the average score obtained for the item among all students with scored data is 67 percent of the total maximum obtainable points possible for the polytomous item.

6.3.2. Response Option or Score Point Proportions

A dichotomous item's alternate response options (i.e., commonly known as being distractors) are plausible but incorrect options that are included to test common misconceptions or miscalculations. Ideally, all response options should garner a proportion of student responses. For a given response option, the proportion is calculated by the simple formula

Equation 6.3

$$\text{proportion} = \frac{N_o}{N_T}$$

where N_o is the number of students selecting the specific option and N_T is the total number of students having scored data for the item.

In the case of polytomous items, the proportion is calculated for each score point as the number of students obtaining the specific score point (N_{sp}) divided by the total number of students having scored data for the item (N_T)

Equation 6.4

$$proportion = \frac{N_{sp}}{N_T}.$$

6.3.3. Item-Total Correlations

The item-total correlation is the relationship between students' performances on a specific item and students' performances on the assessment overall. Possible values for the item-total correlation are bounded between -1 and +1. The correlation will be positive when the mean total test score of the students answering the item correctly is greater than the mean total test score of the students having an incorrect or omitted response for the item. A negative item-total correlation indicates that students with lower mean total test scores were more likely to answer the question correctly than students with higher total test scores. A negative item-total correlation may indicate that an item has multiple correct answers or an incorrect answer key.

The point-biserial correlation (Crocker & Algina, 1986) is one possible item-total correlation for dichotomously scored items. However, the correlation will be spuriously high because the item of interest is also included in the calculation of the total test score (i.e., correlating with itself; Henrysson, 1963). Therefore, a correction is made by calculating the means after excluding the item from the calculation of the total test score (i.e., the total operational test score not including the item of interest for the calculation)

Equation 6.5

$$r_{pbis} = \frac{(\bar{M}'_+ - \bar{M}')}{S'} \sqrt{p/(1-p)}$$

where $\bar{M}'_+$ is the mean score with the item excluded for students who answered the item correctly, $\bar{M}'$ is the mean score with the item excluded for all students who either answered incorrectly or have an omitted response, S' is the standard deviation of the distribution of students' scores (with the item excluded for all students), and p is the item p -value (difficulty).

The Pearson correlation (polyserial), calculated after excluding the item of interest, is typically computed for polytomous items by this equation:

Equation 6.6

$$r = \frac{\sum(x_i - \bar{x})(y'_i - \bar{y}')}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y'_i - \bar{y}')^2}};$$

where x_i is the student score point on the item, $\bar{x}$ is the mean score for the item, y'_i is the total score with the item excluded for the student, and $\bar{y}'$ is the mean total score with the item excluded for all students (Lemke & Wiersma, 1976).

6.3.4. Results of Classical Item Analysis

Item analysis results are reported in the below are computed as the weighted average across test forms. Each form is weighted according to its contribution to the total population. Therefore, operational core forms, which are administered to larger numbers of students, contribute more to the computations than accommodated forms, which are administered to smaller numbers of students.

Table 6.3.1 Summary of Post-Administration P-values for ELA Operational Items by Grade

Grade	N Items	Mean P-Value	SD P-Value	Min. P-Value	Max. P-Value	Median P-Value
3	23	0.50	0.15	0.24	0.71	0.49
4	31	0.52	0.17	0.26	0.81	0.48
5	31	0.52	0.16	0.26	0.92	0.49
6	31	0.54	0.16	0.29	0.90	0.57
7	31	0.54	0.13	0.31	0.83	0.52
8	31	0.53	0.14	0.26	0.80	0.53
9	20	0.49	0.14	0.23	0.79	0.51

Note: SD = standard deviation

Table 6.3.1 presents post-administration summary statistics for item p-values for the operational items for ELA. The weighted mean p-values range from 0.49 and 0.53, and the standard deviations range from 0.14 and 0.17, indicating that most items were of moderate difficulty. Difficulty tended to be relatively consistent across grades.

Table 6.3.2 Summary of Post-Administration P-values for Mathematics Operational Items by Grade

Grade	N Items	Mean P-Value	SD P-Value	Min. P-Value	Max. P-Value	Median P-Value
3	57	0.60	0.21	0.11	0.95	0.63
4	54	0.55	0.20	0.17	0.94	0.57
5	56	0.51	0.24	0.12	0.92	0.52
6	51	0.45	0.23	0.08	0.94	0.43
7	55	0.41	0.22	0.08	0.87	0.37
8	53	0.30	0.18	0.04	0.71	0.26
A1	49	0.35	0.16	0.07	0.73	0.34
GO	53	0.43	0.20	0.12	0.81	0.45
A2	51	0.45	0.19	0.13	0.92	0.42

Note. A1 = Algebra I, GO = Geometry, A2 = Algebra II; SD = standard deviation.

Table 6.3.2 presents post-administration summary statistics for item p-values for the operational items for mathematics. The weighted mean p-values range from 0.30 and 0.60, and the standard deviations range from 0.16 and 0.24. P-values tended to be relatively consistent across grades, except for grade 8 and Algebra I, which tended to be more difficult on average.

Table 6.3.3 Summary of Post-Administration ITC for ELA Operational Items by Grade

Grade	N Items	Mean ITC	SD ITC	Min. ITC	Max. ITC	Median ITC
3	23	0.65	0.15	0.29	0.92	0.60
4	31	0.59	0.15	0.32	0.89	0.54
5	31	0.57	0.14	0.30	0.90	0.55
6	31	0.60	0.14	0.37	0.92	0.56
7	31	0.60	0.14	0.18	0.93	0.58
8	31	0.57	0.15	0.33	0.93	0.54
9	20	0.58	0.17	0.28	0.92	0.54

Note. SD = standard deviation.

Table 6.3.3 presents post-administration summary statistics for item-total correlations for operational items for ELA. The weighted mean correlations range from 0.57 to 0.65 and standard deviations range from 0.14 to 0.17. Correlations tended to be robust and relatively consistent across grades, indicating the items discriminated well between students that performed better overall than students that performed worse overall.

Table 6.3.4 Summary of Post-Administration ITC for Mathematics Operational Items by Grade

Grade	N Items	Mean ITC	SD ITC	Min. ITC	Max. ITC	Median ITC
3	57	0.71	0.10	0.41	0.85	0.74
4	54	0.70	0.10	0.44	0.84	0.72
5	56	0.69	0.13	0.30	0.85	0.73
6	51	0.67	0.14	0.31	0.85	0.71
7	55	0.67	0.15	0.30	0.85	0.74
8	53	0.55	0.16	0.21	0.77	0.57
A1	49	0.60	0.18	0.24	0.85	0.65
GO	53	0.62	0.14	0.36	0.84	0.66
A2	51	0.67	0.12	0.35	0.83	0.70

Note. SD = standard deviation. Note: A1=Algebra I, A2=Algebra II, GO=Geometry.

Table 6.3.4 presents post-administration summary statistics for item-total correlations for operational items for mathematics. The weighted mean correlations range from 0.55 to 0.71 and standard deviations range from 0.10 to 0.18. Correlations tended to be robust, indicating the items discriminated well between students that performed better overall than students that performed worse overall. Mean item-total correlations for grade 8 and Algebra I, where mean p-values tended to be lower compared to other grades, demonstrate that while items tended to be harder on average, the forms still demonstrated good discrimination between students.

Chapter 7. Item Response Theory Analysis, Calibration and Scaling

7.1. Overview

The items on ELA and mathematics core forms are linked to their respective base reporting scales via pre-equating. This linking or equating allows test forms within and across years to be directly comparable. This section of the technical report describes the item response theory (IRT) models used for pre-equating, item calibration and calculation and reporting of New Jersey students' scale scores. Descriptive statistics of the distributions of item parameter estimates for each assessment component are also included in this chapter.

7.2. IRT Models

The items on mathematics operational core forms were calibrated using the IRT two-parameter logistic (2PL) model (Birnbbaum, 1968) for dichotomously scored items and the Generalized Partial Credit Model (GPCM) (Muraki, 1997) for polytomously scored items. ELA items were calibrated using the GPCM. To be concise, the two models are expressed using a single formula since the 2PL model can be considered algebraically nested within the GPCM, which is denoted as:

Equation 7.1

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

where $a_i(\theta_j - b_i + d_{ik}) \equiv 0$; $P_{im}(\theta_j)$ is the probability of the j^{th} 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 parameter 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 zero to $M_i - 1$. For items with just two response categories, with one category being scored as a correct response and the other category scored as an incorrect response, the d_{ik} parameter becomes zero since no step parameters are needed, making the 2PL model a special case of the GPCM.

7.3. Summary Statistics and Distributions from IRT Analyses

Table 7.3.1 and Table 7.3.2 present summary statistics for the pre-equated IRT (a- and b-) parameter estimates for the New Jersey Student Learning Assessment (NJSLA) ELA and mathematics assessments. The summary statistics for IRT parameter estimates include the operational items

administered in the spring 2023 administration. Pre-equated parameters are presented since these were parameters used for scoring students in spring 2023.

Table 7.3.1 shows the *a*- and *b*-parameter estimates for the ELA assessments. In IRT, the *a*-parameter refers to the item’s ability to discriminate the performance of test takers. The *b*-parameter represents the item’s level of difficulty, with larger, positive values reflecting a harder item. The items summarized in Table 7.1 include all unique items on online general forms, paper forms and online accommodated forms. For PCR items, the item traits are included in the item count since IRT parameters are calculated for each trait separately. Thus, each PCR item contributes twice to the number of items: once for each trait. Score points for PCR items are based on the weighted scores for each trait.

Table 7.3.1 IRT Parameter Estimates Summary for All Items for ELA by Grade

Grade	No. of Score Points	No. of Items	b Estimates Summary				a Estimates Summary			
			Mean	SD	Min	Max	Mean	SD	Min	Max
3	94	42	0.44	0.99	-0.85	3.48	0.58	0.25	0.15	1.06
4	125	56	0.48	0.97	-1.42	2.66	0.49	0.25	0.17	1.06
5	126	56	0.48	1.03	-1.92	3.59	0.47	0.24	0.13	0.99
6	111	50	0.34	0.87	-1.93	1.93	0.51	0.23	0.21	1.16
7	125	56	0.28	0.73	-1.52	1.60	0.51	0.27	0.13	1.20
8	125	56	0.27	0.81	-1.42	2.38	0.48	0.25	0.08	1.10
9	98	44	0.74	0.94	-0.91	2.87	0.51	0.29	0.15	1.22

Note. SD = standard deviation.

Table 7.3.2 shows the a- and b-parameter estimates for the mathematics assessments.

Table 7.3.2 IRT Parameter Estimates Summary for All Items for Mathematics by Grade

Grade	No. of Score Points	No. of Items	b Estimates Summary				a Estimates Summary			
			Mean	SD	Min	Max	Mean	SD	Min	Max
3	139	88	-0.37	1.14	-3.22	2.05	0.80	0.25	0.28	1.31
4	134	80	-0.03	1.05	-2.65	2.35	0.72	0.21	0.32	1.35
5	147	85	0.05	1.28	-2.57	4.53	0.68	0.28	0.17	1.50
6	140	80	0.36	1.15	-4.46	2.24	0.69	0.27	0.25	1.35
7	140	84	0.60	1.13	-2.23	3.42	0.73	0.34	0.19	1.72
8	134	79	0.99	1.16	-2.04	3.54	0.61	0.30	0.10	1.44
A1	138	75	1.30	1.28	-0.77	7.71	0.58	0.31	0.12	1.62
GO	143	81	1.00	1.05	-1.30	3.83	0.73	0.37	0.19	1.68
A2	145	77	1.24	1.06	-1.84	3.80	0.70	0.34	0.19	1.56

Note: A1=Algebra I, GO=Geometry, A2=Algebra II, SD = standard deviation.

Items on spring 2023 forms used pre-equated IRT parameters that were calculated during item calibration in previous years. Table 7.3.3 shows the source years for the IRT item parameters for the 2023 ELA assessments.

Table 7.3.3 IRT Parameter Distribution by Year for All Items for ELA Assessments

Grade	No. of Items	2014	2015	2016	2017	2018	2019	2022
3	42	0	0	0	1	1	19	21
4	56	0	0	0	8	3	22	23
5	56	0	0	0	3	11	15	27
6	50	0	4	0	1	12	17	16
7	56	0	5	10	4	15	14	8
8	56	0	0	1	5	6	36	8
9	44	0	0	6	10	16	12	0

Table 7.3.4 shows the source years for the IRT item parameters for the 2023 mathematics assessments.

Table 7.3.4 IRT Parameter Estimates Summary for All Items for Mathematics Assessments

Grade	No. of Items	2014	2015	2016	2017	2018	2019	2022
3	88	0	16	12	8	9	27	16
4	80	0	17	18	10	11	22	2
5	85	0	24	10	15	5	24	7
6	80	0	15	12	12	11	22	8
7	84	0	18	13	11	12	20	10
8	79	0	16	13	12	7	19	12
A1	75	0	10	5	12	9	16	23
GO	81	0	10	15	13	6	20	17
A2	77	1	9	18	7	10	20	12

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

7.4. Scale Scores

Reporting scales designate student performance into one of five performance levels with Level 1 indicating the lowest level of performance and Level 5 indicating the highest level of performance. Threshold or cut scores associated with performance levels were initially expressed as raw scores on the 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.

7.4.1. Summative Score Scale and Performance Levels

There are 201 defined summative scale score points for both ELA and mathematics, ranging from 650 to 850. The lowest obtainable scale score is 650 and the highest obtainable scale score is 850. The thresholds for summative performance levels on the scale score metric recommended by the scale score task force are Level 2 (*Partially Met Expectations*) and Level 4 (*Met Expectations*). 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 on a scoring table.

For spring 2015, scale scores were defined for each test as a linear transformation of the theta (θ_{2015}) scale. The theta values associated with the Level 2 and Level 4 performance levels were identified using the test characteristic curve associated with the performance level setting form. With Levels 2 and 4 scale

scores fixed at 700 and 750, respectively, the relationship between theta (θ_{2015}) and scale scores ($ScaleScore_{2015}$) was established as:

Equation 7.2

$$ScaleScore_{2015} = A_{2015} \times \theta_{2015} + B_{2015}$$

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

Equation 7.3

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

and

Equation 7.4

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

As indicated by these formulas, the slope and intercept for the summative scale scores were based on the theta scale, and by default the IRT parameter scale, established in 2015. Because 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 ($ScaleScore_{2015}$)

Equation 7.5

$$ScaleScore_{2015} = SA_{2016} \times \theta_{2016} + SB_{2016}$$

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:

Equation 7.6

$$SA_{2016} = \frac{A_{2015}}{slope_{2015\_to\_2016}}$$

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:

Equation 7.7

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

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, respectively. This is described in more detail in section 7.4.2, below.

A and B values resulting from these calculations as well as the theta values associated with the threshold performance levels are included in Appendix 7.1. Also, the 2015–2016 technical report includes raw to scale score conversion tables for the performance level setting forms.

7.4.2. ELA Reading and Writing Claim Scales

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 on 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 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 7.4.1.

Table 7.4.1 Calculating Scaling Constants for Reading and Writing Claims

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

A and B values resulting from these calculations are included in Appendix 7.1.

7.4.3. Subclaims Scales

The Level 4 cut is defined as *Meets or Exceeds Expectations*, students meeting these standards are prepared to engage successfully in further studies in the assessed content area. 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 Level 3 and Level 4 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*. Students whose raw subclaim score fell between the Level 3 and 4 thresholds were designated as *Nearly Meets Expectations*.

7.4.4. Creating Conversion Tables

A conversion table relates the number of points earned by a student for the ELA 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 IRT ability parameters or theta scores. In carrying out the calculations, estimates of item parameters and thetas are substituted for parameters in the formulas in each of the following steps:

Step 1: The expected item score (i.e., estimated item true score) is calculated for every theta in the selected range (between -15 and +15, in 0.0001 increments) based on the GPCM for both dichotomous and polytomous items

Equation 7.8

$$s_i(\theta_j) = \sum_{m=0}^{M_i-1} m P_{im}(\theta_j)$$

Equation 7.9

$$P_{im}(\theta_j) = \frac{\exp[\sum_{k=0}^m D a_i(\theta_j - b_i + d_{ik})]}{\sum_{v=0}^{M_i-1} \exp[\sum_{k=0}^v D a_i(\theta_j - b_i + d_{iv})]}$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $s_i(\theta_j)$ is the expected item score for item i on theta, θ_j ; $P_{im}(\theta_j)$ is the probability of a student, j , with θ_j getting score m on item i ; m_i is the number of score categories of item i ; with possible item scores as consecutive integers from 0 to $M_i - 1$; D is the IRT scale constant (1.7); a_i is the item slope parameter; b_i is the item 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. Since the 2PL model can be considered a special case of the GPCM, the latter can be used to calibrate dichotomously scored items, resulting in 2PL model item parameters.

Step 2: The expected (weighted) test score for every theta in the selected range is calculated as

Equation 7.10

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

Where T_j is the expected (weighted) test score on theta, θ_j ; w_i is the item weight for item i I is the total number of items in the test form.

Step 3: The estimated conditional standard error of measurement (CSEM) is calculated for each theta in the selected range as

Equation 7.11

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

Equation 7.12

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

Equation 7.13

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

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

Step 4: Every raw score is matched with a theta, where θ_j is the theta for a raw score r_h , if $T_j - r_h$ is minimum across all T_j .

Step 5: The reported scale score is calculated. Using the A and B scaling constants in Appendix 7.1, each theta value is converted to a scale score and each theta CSEM to a scale score CSEM:

Equation 7.14

$$ScaleScore = A \times \theta + B$$

Equation 7.15

$$CSEM = CSEM_{\theta} \times A$$

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 7.4.4.1 contains the TCC, estimated CSEM and estimated test information function (TIF) curves for all operational forms of the Grade 3 ELA examinations. Within each figure, the curve is reported on the theta scale. Figure 7.4.4.2 contains TCC, CSEM, and TIF curves for all operational Grade 3 mathematics forms. Similar curves for the other grades and mathematics courses may be found in Appendix 7.2.

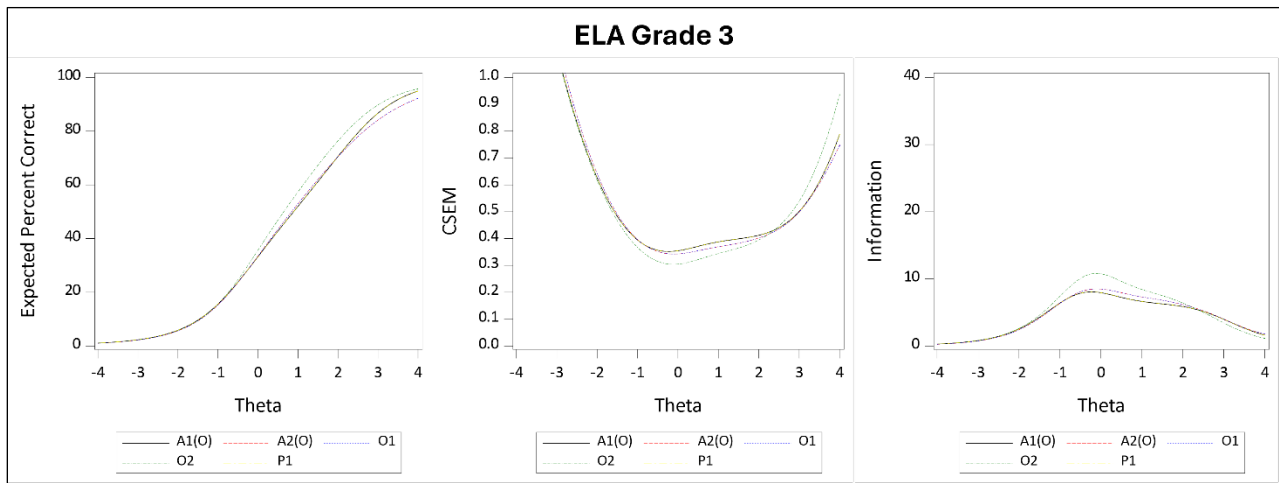


Figure 7.4.1 ELA Grade 3 Test Characteristic Curves, CSEM Curves, and Information Curves

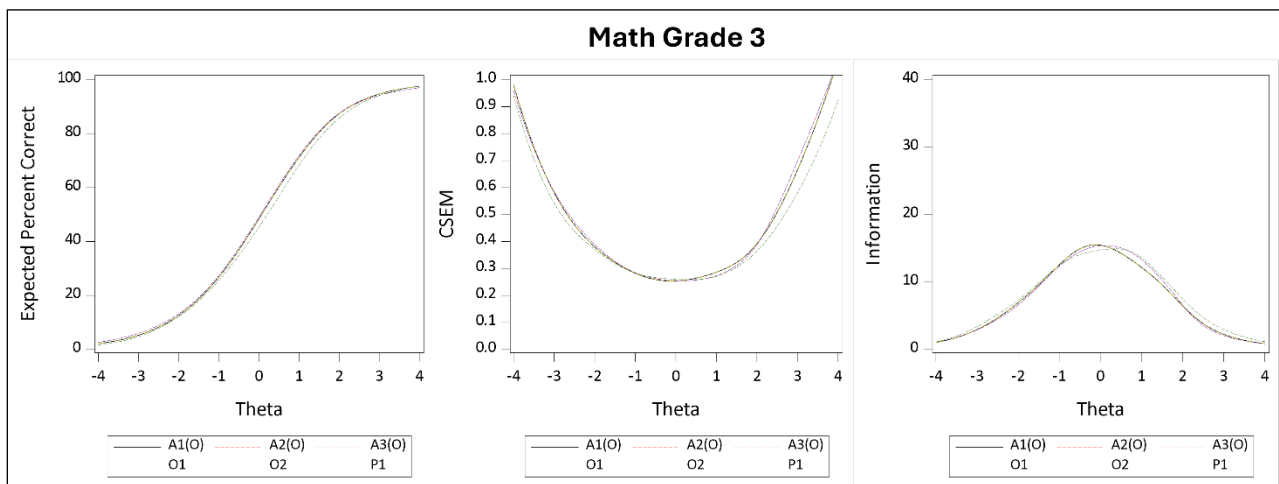


Figure 7.4.2 Mathematics Grade 3 Test Characteristic Curves, CSEM Curves, and Information Curves

7.5. Scale Score Distributions

7.5.1. ELA Score Distributions

Score distribution information for Grade 3 ELA forms is illustrated in Figure 7.5.1. The vertical axis of the graph represents the proportion of students earning the scale score point indicated along the horizontal axis. For the summative score distribution, the y-axis ranges from 0 to 0.1 and the x-axis from 650 to 850. The score distributions for all grades are provided in Appendix 7.3.

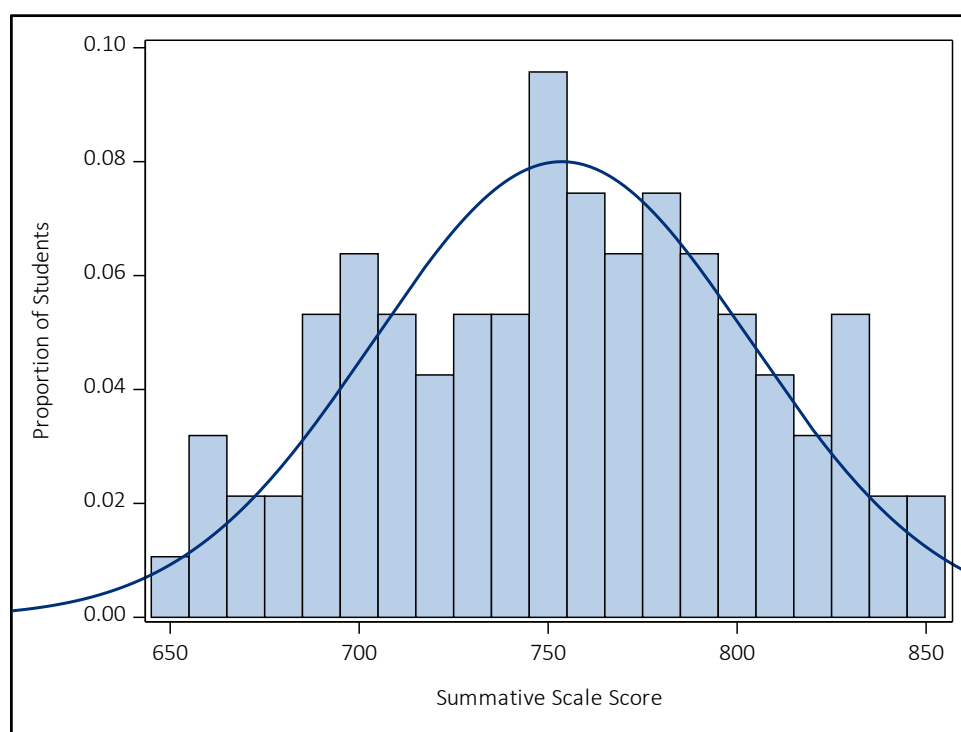


Figure 7.5.1 Grade 3 ELA Score Distribution

The performance of students taking 2023 Grade 3 ELA is reported in the cumulative score distributions in Table 7.5.1. Note that since this table reports actual student performance, not all scale score bands may be realized.

Table 7.5.1 Grade 3 ELA Scale Score Cumulative Frequencies

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	3,185	3.45	3,185	3.45
655-659	862	0.94	4,047	4.39
660-664	1,747	1.89	5,794	6.28
665-669	—	—	5,794	6.28
670-674	2,860	3.10	8,654	9.38
675-679	1,242	1.35	9,896	10.73
680-684	1,712	1.86	11,608	12.59
685-689	2,691	2.92	14,299	15.51
690-694	2,441	2.65	16,740	18.16
695-699	2,231	2.42	18,971	20.58
700-704	3,122	3.39	22,093	23.97

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
705-709	2,090	2.27	24,183	26.24
710-714	3,300	3.58	27,483	29.82
715-719	3,048	3.31	30,531	33.13
720-724	2,199	2.39	32,730	35.52
725-729	4,559	4.95	37,289	40.47
730-734	3,467	3.76	40,756	44.23
735-739	5,005	5.43	45,761	49.66
740-744	2,569	2.79	48,330	52.45
745-749	5,151	5.59	53,481	58.04
750-754	3,913	4.24	57,394	62.28
755-759	4,205	4.56	61,599	66.84
760-764	3,749	4.07	65,348	70.91
765-769	2,635	2.86	67,983	73.77
770-774	4,778	5.18	72,761	78.95
775-779	2,272	2.46	75,033	81.41
780-784	2,896	3.14	77,929	84.55
785-789	2,639	2.86	80,568	87.41
790-794	2,167	2.35	82,735	89.76
795-799	1,300	1.41	84,035	91.17
800-804	1,778	1.93	85,813	93.10
805-809	1,504	1.63	87,317	94.73
810-814	859	0.93	88,176	95.66
815-819	735	0.80	88,911	96.46
820-824	586	0.64	89,497	97.10
825-829	579	0.63	90,076	97.73
830-834	522	0.57	90,598	98.30
835-839	400	0.43	90,998	98.73
840-844	292	0.32	91,290	99.05
845-849	108	0.12	91,398	99.17
850	794	0.86	92,192	100.00

7.5.2. Mathematics Score Distributions

Score distribution information for Grade 3 mathematics forms is illustrated in Figure 7.5.2. The vertical axis of the graph represents the proportion of students earning the scale score point indicated along the horizontal axis. For the summative score distribution, the y-axis ranges from 0 to 0.08 and the x-axis from 650 to 850. The scores distributions for all grades are provided in Appendix 7.4.

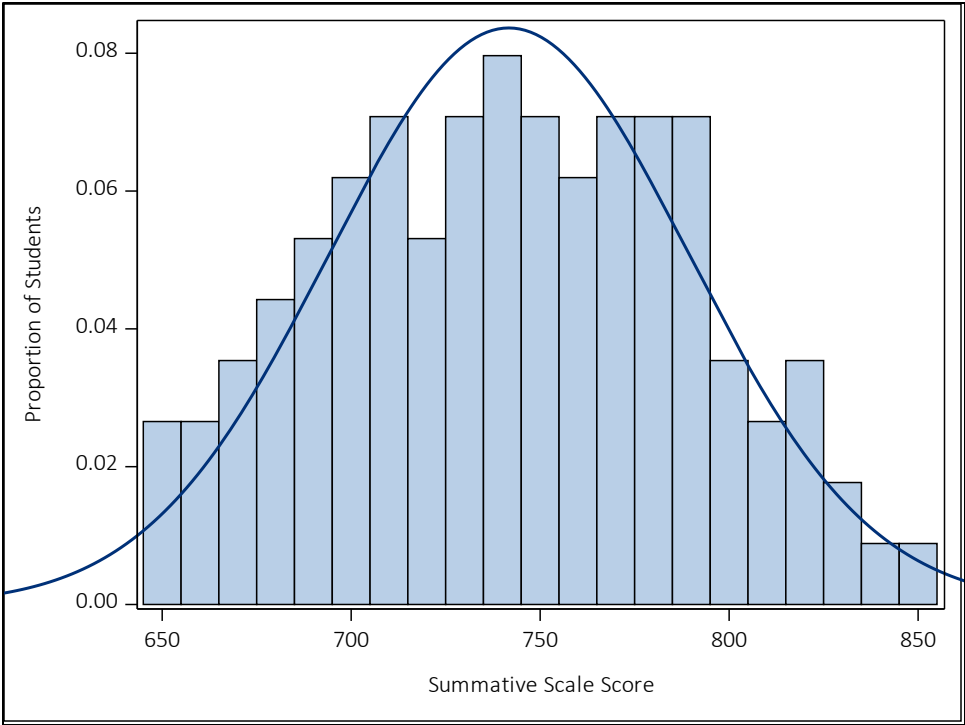


Figure 7.5.2 Grade 3 Mathematics Score Distribution

The performance of students taking 2023 Grade 3 mathematics is reported by form in the cumulative score distributions given in Table 7.5.2. Note that since this table reports actual student performance, not all scale score bands may be realized.

Table 7.5.2 Grade 3 Mathematics Scale Score Cumulative Frequencies

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	698	0.74	698	0.74
655-659	824	0.88	1,522	1.62
660-664	12	0.01	1,534	1.63
665-669	1,180	1.26	2,714	2.89
670-674	766	0.82	3,480	3.71

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
675-679	1,541	1.64	5,021	5.35
680-684	900	0.96	5,921	6.31
685-689	1,851	1.97	7,772	8.28
690-694	1,933	2.06	9,705	10.34
695-699	2,000	2.13	11,705	12.47
700-704	2,139	2.28	13,844	14.75
705-709	4,388	4.67	18,232	19.42
710-714	2,359	2.51	20,591	21.93
715-719	4,518	4.81	25,109	26.74
720-724	2,300	2.45	27,409	29.19
725-729	4,738	5.04	32,147	34.23
730-734	3,616	3.85	35,763	38.08
735-739	4,873	5.19	40,636	43.27
740-744	5,096	5.42	45,732	48.69
745-749	5,024	5.35	50,756	54.04
750-754	3,809	4.05	54,565	58.09
755-759	3,814	4.06	58,379	62.15
760-764	4,828	5.14	63,207	67.29
765-769	4,790	5.10	67,997	72.39
770-774	4,503	4.79	72,500	77.18
775-779	4,386	4.67	76,886	81.85
780-784	2,107	2.24	78,993	84.09
785-789	3,753	3.99	82,746	88.08
790-794	1,723	1.83	84,469	89.91
795-799	2,356	2.51	86,825	92.42
800-804	1,492	1.59	88,317	94.01
805-809	1,343	1.43	89,660	95.44
810-814	1,218	1.30	90,878	96.74
815-819	457	0.49	91,335	97.23
820-824	950	1.01	92,285	98.24
825-829	427	0.45	92,712	98.69
830-834	309	0.33	93,021	99.02
835-839	—	—	93,021	99.02

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
840-844	496	0.53	93,517	99.55
845-849	—	—	93,517	99.55
850	451	0.48	93,968	100.00

7.5.3. ELA Major Claims Score Distributions

Score distributions are also presented for the Reading and Writing subscores. In Figure 7.5.3 and Figure 7.5.4, the respective Major Claim score distributions for Grade 3 ELA Reading and Writing are presented. For the Reading distribution, the y-axis ranges from 0 to 0.1 and the x-axis from 10 to 90. For the Writing distribution, the y-axis ranges from 0 to 0.15 and the x-axis from 10 to 60. ELA Major Claim score distributions for all grades are presented in Appendix 7.5.

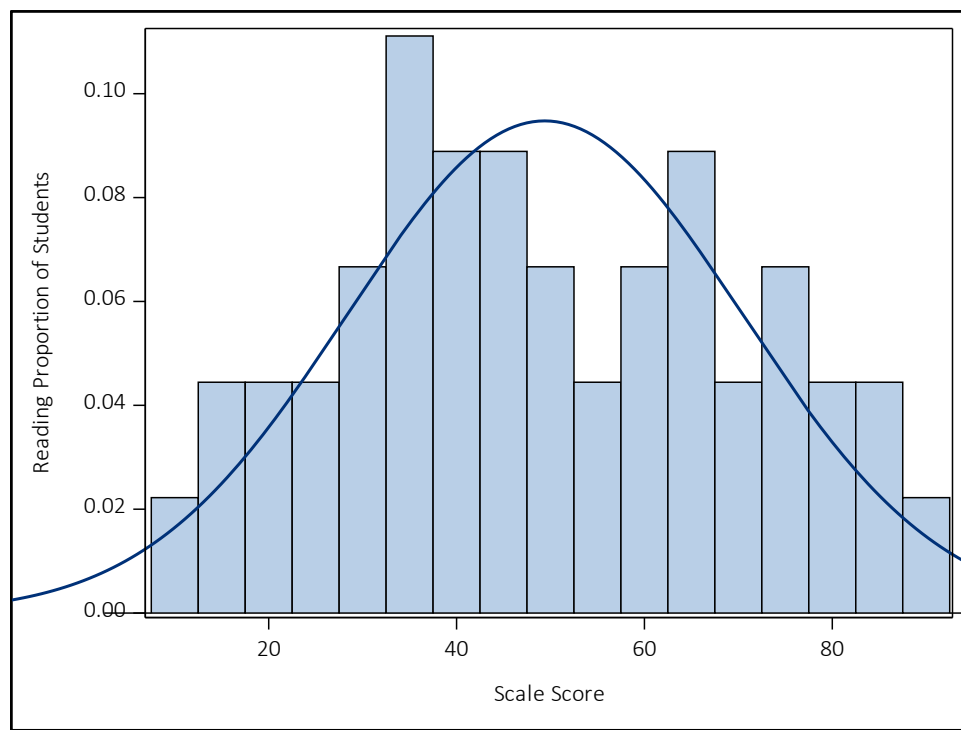


Figure 7.5.3 Grade 3 ELA Reading Score Distribution

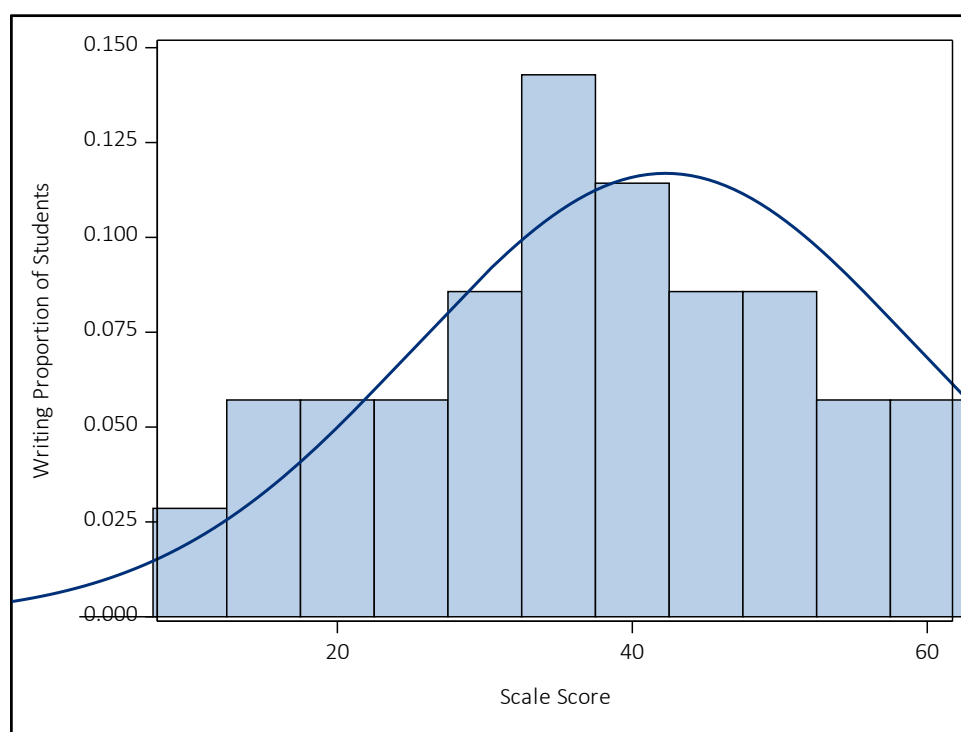


Figure 7.5.4 Grade 3 ELA Writing Score Distribution

7.5.4. Scale Score Distributions for Student Demographic Groups of Interest

The performance of demographic groups of students is summarized in Table 7.5.3 and Table 7.5.4 for Grade 3 ELA and mathematics, respectively. Each table reports the number of students in ethnicity groups, and other demographic variables including gender, economic status, English Learner status, and students with disabilities. For each demographic grouping the mean scale score and standard deviation is presented along with the minimum and maximum scale scores for reference. Table 7.5.4.1 also summarizes group performance on the Reading and Writing summative scores. Subgroup performance for all grades presented in Appendix 7.6.

Table 7.5.3 Grade 3 ELA Subgroup Performance for Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		85,471	742.54	43.65	650	850
Gender	Female	43,085	746.85	44.05	650	850
	Male	42,382	738.16	42.80	650	850
Ethnicity	American Indian/Alaska Native	186	749.32	42.26	650	850

Group Type	Subgroup	N	Mean	SD	Min.	Max.
	Asian	9,764	769.95	39.16	650	850
	Black or African American	12,039	725.47	41.99	650	850
	Hispanic/Latino	27,034	726.58	42.24	650	850
	Native Hawaiian or Pacific Islander	143	745.87	42.30	650	843
	Two or more races	3,027	752.29	42.29	650	850
	White	33,258	752.72	39.13	650	850
Economic Status*	Not Economically Disadvantaged	54,601	753.81	41.19	650	850
	Economically Disadvantaged	30,870	722.63	40.67	650	850
English Learner Status	Non-English Learner	75,734	747.33	42.11	650	850
	English Learner	9,737	705.33	37.06	650	850
Disabilities	Students without Disabilities	73,870	745.41	43.07	650	850
	Student with Disability (SWD)	11,601	724.28	42.92	650	850
Reading Summative Score		85,471	46.66	17.51	10	90
Gender	Female	43,085	47.87	17.58	10	90
	Male	42,382	45.43	17.35	10	90
Ethnicity	American Indian/Alaska Native	186	49.17	16.66	10	90
	Asian	9,764	57.30	16.07	10	90
	Black or African American	12,039	40.11	16.53	10	90
	Hispanic/Latino	27,034	39.99	16.45	10	90
	Native Hawaiian or Pacific Islander	143	47.48	16.73	10	90
	Two or more races	3,027	51.00	17.08	10	90
	White	33,258	50.92	16.10	10	90
Economic Status*	Not Economically Disadvantaged	54,601	51.24	16.78	10	90
	Economically Disadvantaged	30,870	38.56	15.77	10	90
English Learner Status	Non-English Learner	75,734	48.61	16.98	10	90
	English Learner	9,737	31.48	13.75	10	90
Disabilities	Students without Disabilities	73,870	47.64	17.32	10	90
	Student with Disability (SWD)	11,601	40.40	17.43	10	90
Writing Summative Score		85,471	31.69	13.06	10	60

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Gender	Female	43,085	33.24	12.82	10	60
	Male	42,382	30.11	13.11	10	60
Ethnicity	American Indian/Alaska Native	186	33.58	12.57	10	60
	Asian	9,764	38.77	10.73	10	60
	Black or African American	12,039	27.00	13.28	10	60
	Hispanic/Latino	27,034	27.94	13.27	10	60
	Native Hawaiian or Pacific Islander	143	32.85	12.69	10	60
	Two or more races	3,027	33.71	12.65	10	60
	White	33,258	34.16	11.87	10	60
Economic Status	Not Economically Disadvantaged	54,601	34.49	12.16	10	60
	Economically Disadvantaged	30,870	26.75	13.13	10	60
English Learner Status	Non-English Learner	75,734	32.86	12.64	10	60
	English Learner	9,737	22.62	12.73	10	60
Disabilities	Students without Disabilities	73,870	32.60	12.77	10	60
	Student with Disability (SWD)	11,601	25.90	13.39	10	60

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

Table 7.5.4 Grade 3 Mathematics Subgroup Performance for Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		69,710	750.75	37.48	650	850
Gender	Female	35,180	748.09	36.29	650	850
	Male	34,527	753.46	38.47	650	850
Ethnicity	American Indian/Alaska Native	147	755.44	34.90	668	850
	Asian	8,181	778.78	33.40	650	850
	Black or African American	9,326	729.83	35.37	650	850
	Hispanic/Latino	21,417	735.78	34.79	650	850
	Native Hawaiian or Pacific Islander	113	749.88	39.15	650	850
	Two or more races	2,522	757.97	36.29	650	850
	White	27,983	760.32	32.47	650	850

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Economic Status*	Not Economically Disadvantaged	46,095	760.49	35.24	650	850
	Economically Disadvantaged	23,615	731.73	34.28	650	850
English Learner Status	Non-English Learner	62,531	753.86	36.65	650	850
	English Learner	7,179	723.61	33.43	650	850
Disabilities	Students without Disabilities	61,461	752.00	37.08	650	850
	Student with Disability (SWD)	8,249	741.38	39.08	650	850
Language Form	Spanish	1,645	711.74	29.55	650	820

7.6. Interpreting Claim Scores and Subclaim Scores

7.6.1. Interpreting Claim Scores

ELA assessments provide separate claim scale scores for both Reading and Writing. The claim scale scores and summative scale score are on different scales; therefore, the sum of the scale scores for each claim will not equal the summative scale score. Reading scale scores range from 10 to 90 and Writing scale scores range from 10 to 60.

The claim scores can be interpreted by comparing a student's claim scale score to the average performance for the school, district, and state. The Individual Student Report (ISR) provides the student scale score results and the average scale score results for the school, district, and state.

7.6.2. 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 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.

Chapter 8. Student Demographics and Differential Item Functioning (DIF)

8.1. Overview of Test-Taking Population

More than 700,000 New Jersey students attempted New Meridian summative assessment forms in the spring of 2023. Chapter 8 reports important demographic descriptions of the testing population in each grade and subject as well as results of Differential Item Functioning (DIF) analysis. DIF analyses are utilized to detect systematic differences in performance on the test items between demographic groups.

8.2. Rules for Inclusion of Students in Analyses

Criteria for inclusion of students were implemented prior to all operational analyses. These rules were established by NMC psychometrics in consultation with NJDOE 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 when:

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

Additionally, in cases where students had more than one valid record, the record with the higher raw score was chosen. Records for students with administration issues or anomalies were excluded from analyses.

8.3. Demographics

Table 8.3.1 presents the number and percentage of students who took ELA test forms by mode, either computer-based test (CBT) or paper-based test (PBT). Table 8.3.2 presents breakdowns of test-taking populations for mathematics forms by grade and mode, including those students who took Spanish language mathematics forms.

Markedly more students tested online than on paper across all grades for both content areas. For ELA, the percentage of online students by grade level was greater than 99 percent. For all mathematics students, the percentage of students testing online was greater than 99 percent. The percentage of students taking Spanish-language mathematics paper forms tended to be greater than the percentage of students taking paper forms in ELA or English mathematics forms, but the percentage of students taking online Spanish-language mathematics forms was still greater than 97 percent.

Table 8.3.1 ELA Test Takers

Grade	% of All Students	N Students	N CBT	% CBT	N PBT	% PBT
3	100.0	92,192	92,138	99.9	54	0.1
4	100.0	93,412	93,354	99.9	58	0.1
5	100.0	94,841	94,757	99.9	84	0.1
6	100.0	96,143	96,097	100.0	46	0.0
7	100.0	97,017	96,962	99.9	55	0.1
8	100.0	100,115	100,065	100.0	50	0.0
9	100.0	97,174	97,036	99.9	138	0.1

Note. n/r = not reported due to n<20. CBT=Computer-based test. PBT=Paper-based test.

Table 8.3.2 Mathematics Test Takers

Grade	Language	% of All Students	N Students	N CBT	% CBT	N PBT	% PBT
3	English	100.0	93,968	93,848	99.9	120	0.1
	Spanish	100.0	2,859	2,786	97.4	73	2.6
4	English	100.0	95,000	94,863	99.9	137	0.1
	Spanish	100.0	2,861	2,785	97.3	76	2.7
5	English	100.0	96,372	96,247	99.9	125	0.1
	Spanish	100.0	2,683	2,614	97.4	69	2.6
6	English	100.0	97,653	97,565	99.9	88	0.1
	Spanish	100.0	2,699	2,673	99.0	26	1.0
7	English	100.0	93,463	93,369	99.9	94	0.1
	Spanish	100.0	2,700	2,680	99.3	20	0.7
8	English	100.0	66,693	66,604	99.9	89	0.1
	Spanish	100.0	2,332	2,311	99.1	21	0.9
A1	English	100.0	105,535	105,336	99.8	199	0.2
	Spanish	100.0	3,716	3,700	99.6	n/r	n/r
A2	English	100.0	10,089	10,087	100.0	n/r	n/r
	Spanish	100.0	302	302	100.0	n/r	n/r
GO	English	100.0	30,240	30,213	99.9	27	0.1
	Spanish	100.0	370	370	100.0	n/r	n/r

Note. n/r = not reported due to n<20. CBT=Computer-based test. PBT=Paper-based test. Note: A1=Algebra I, A2=Algebra II, GO=Geometry.

Table 8.3.3 summarizes demographic information for students with valid ELA scores, and Table 8.3.4 presents demographics for students with valid mathematics scores. Percentages are not reported in instances where fewer than 20 students were tested.

Table 8.3.3 Grade 3 ELA Test Taker Demographic Information

Demographic	N	Percent
Economically Disadvantaged	34,169	37.1
Student with Disability (SWD)	18,311	19.9
English Learner	10,626	11.5
Male	46,592	50.5
Female	45,595	49.5
American Indian/ Alaska Native	196	0.2
Asian	10,017	10.9
Black or African American	13,190	14.3
Hispanic/Latino	29,658	32.2
White/Caucasian	35,732	38.8
Native Hawaiian or Pacific Islander	149	0.2
Two or more races	3,228	3.5
Unknown Ethnicity	22	0.0

Table 8.3.4 Grade 3 Mathematics Test Taker Demographic Information

Demographic	N	Percent
Economically Disadvantaged	35,118	37.4
Student with Disability (SWD)	18,330	19.5
English Learner	12,444	13.2
Male	47,505	50.6
Female	46,457	49.4
American Indian/ Alaska Native	198	0.2
Asian	10,226	10.9
Black or African American	13,231	14.1
Hispanic/Latino	30,958	32.9
White/Caucasian	35,948	38.3
Native Hawaiian or Pacific Islander	153	0.2
Two or more races	3,230	3.4
Unknown Ethnicity	24	0.0

8.4. Differential Item Functioning

Differential item functioning (DIF) is a procedure that matches students based on total test scores to compare the performance of similarly able students across subgroups. The procedure identifies two contrasting groups called focal and reference for which differences in item performances are computed. Table 8.4.1 indicates the focal and comparison groups used in DIF comparisons. At least 100 students in each group (focal and reference) and 300 total students across the two groups are required for DIF procedures to be conducted. For the procedures described next, positive DIF values indicate that, for students of similar ability, the focal group has a higher mean item score than the reference group. Negative DIF values indicate that, for students of similar ability, the focal group has a lower mean item score than the reference group.

Table 8.4.1 DIF Comparison Groups

Comparison Type	Focal Group (N≥100)	Reference Group (N≥100)
Gender	Female	Male
Ethnicity	African American	White
	Asian	White
	American Indian/Alaska Native	White
	Hispanic	White
	Pacific Islander	White
	Multiple	White
Economic Status	Economically Disadvantaged	Not Economically Disadvantaged
English Learners	English Learner	English Proficient (including former English learners)
Students with an IEP	IEP	No IEP

8.4.1. Dichotomous Items: Mantel-Haenszel

The Mantel-Haenszel (MH) chi-square approach (Mantel & Haenszel, 1959) is used to detect DIF in dichotomously scored, one-point items. The range of total scores is divided into 10 stratifications (*S*) based on raw score performance, and those strata are used to match samples from each group. Contingency tables (such as in Table 8.4.2) for each stratum are constructed for the responses to the item in which *S* represents the strata, *W_{rs}* and *W_{fs}* represent the number of students (in the reference and focal groups, respectively) who answer the item incorrectly, *R_{rs}* and *R_{fs}* represent the number of students (in the reference and focal groups, respectively) who answer the item correctly, and *N_{ts}* represents the total number of students (*W_{rs}* + *R_{rs}* + *W_{fs}* + *R_{fs}*).

Table 8.4.2 Mantel-Haenszel Contingency Table

Score Stratum (S)	Incorrect/Wrong (O)	Correct/Right (1)	Total
Reference	Wrs	Rrs	$Wrs + Rrs$
Focal	Wfs	Rfs	$Wfs + Rfs$
Total	$Wrs + Wfs$	$Rrs + Rfs$	Nts

A common odds ratio is computed across all intervals of matched groups using the following formula (Dorans & Holland, 1993):

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^S R_{rs}W_{fs} / N_{ts}}{\sum_{s=1}^S R_{fs}W_{rs} / N_{ts}}.$$

Furthermore, the Mantel-Haenszel delta statistic (MHD-DIF) (Holland & Thayer, 1988) is computed to measure the degree and magnitude of DIF using the formula

$$MH_{D-DIF} = -2.35 \ln(\hat{\alpha}_{MH}).$$

8.4.2. Polytomous Items: Standardized Mean Difference

For polytomous items, the MHD-DIF is not calculated. Rather, a standardized mean difference (SMD) is calculated using a contingency table that extends the possible items scores beyond 1 point using this formula,

$$SMD = \sum_s w_{Fs} m_{Fs} - \sum_s w_{Rs} m_{Rs}$$

where $w_{Fs} = n_{F+s}/n_{F++}$ is the focal group proportion at the s th stratification variable; $m_{Fs} = (1/n_{F+s})F_s$ is the focal group's mean item score in the s th stratum; and $m_{Rs} = (1/n_{R+s})R_s$ is the reference group's mean item score in the s th stratum. Because the focal group proportion is used in both terms of the equation, the reference group's item mean is weighted, whereas the focal group's item mean is unweighted.

The effect size (ES) is then computed by dividing by the total group standard deviation (SD) using this equation:

$$ES = \frac{SMD}{SD}.$$

By using Mantel's chi-square statistic (1963), the magnitude of the ES is interpreted using Golia's (2012) rules.

8.4.3. DIF Classification

Based on the DIF statistics and significance tests, items are classified into three categories: A, B, or C (as in Table 8.4.3). Category A items contain negligible difference in performance and Category B items exhibit slight to moderate difference in performance, while Category C items exhibit moderate to large difference in performance. Items flagged with C-DIF during the preliminary analyses were provided to both the ELA and mathematics test development managers and the accessibility, accommodations, and fairness (AAF) specialist as part of the preliminary analysis communication plan.

Table 8.4.3 DIF Classifications

Analysis		Criteria
Differential Item Functioning (DIF)	+ Favors the focal group – Favors the reference group	
Mantel-Haenszel		A. Negligible – MH is not significantly different from 0 OR (MH is significantly different from 0 AND has a delta absolute value < 1). B. Slight to Moderate – MH is significantly different from 0 AND the absolute value of delta is < 1.5 AND has a delta absolute value greater than or equal to 1. C. Moderate to Large – MH is significantly different from 1 AND delta has an absolute value greater than or equal to 1.5.
Standardized Mean Difference		A. Negligible – is not significantly different from 0 OR has an absolute value ≤ 0.17 . B. Slight to Moderate – is significantly different from 0 AND $0.17 < ES \leq 0.25$. C. Moderate to Large – is significantly different from 0 AND has an absolute value > 0.25 .

8.4.4. Differential Item Functioning Results

The tables below present DIF results for grade 3 ELA and mathematics tests, offering insights into potential disparities in item performance across different demographic groups. Similar tables for all grades are presented in Appendix 8.2. DIF analysis helps identify whether certain test items exhibit differential difficulty or functioning for distinct subgroups.

Table 8.4.4 Grade 3 ELA Post-Administration Differential Item Functioning

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	23					23	100				
White vs. Black/ African American	23			1	4	22	96				
White vs. Hispanic/ Latino	23			2	9	21	91				
White vs. Asian	23					23	100				
White vs. American Indian/ Alaska Native	23	1	4			22	96				
White vs. Native Hawaiian or Pacific Islander	23					23	100				
White vs. Two or more races	23					23	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	23					23	100				
Non-English Learner vs. English Learner	23	2	9			21	91				

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Student without Disability vs. Student with Disability	23					23	100				

Table 8.4.4 presents DIF results for the grade 3 ELA assessment. For the majority of DIF comparisons, all 23 items exhibit A level DIF, indicating no significant difference in performance on any items between these groups.

For the White vs Black/African American comparison, 1 item exhibited B- (B minus) level DIF, while the remaining 22 items exhibited A level DIF. Similarly, in the White vs Hispanic/Latino comparison, two items display B-level DIF, with the rest showing A level DIF. The White vs American Indian/Alaska Native comparison exhibited one item at C- level DIF, and Non-English Learner vs English Learner had 2 items at C- level DIF. Notably, the C- level DIF items in both comparisons are negative, suggesting that American Indian/Alaska Native and Non-English Learner groups perform less favorably on these specific items.

Table 8.4.5 Grade 3 Mathematics Post-Administration Differential Item Functioning

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	57			3	5	54	95				
White vs. Black/ African American	57			1	2	56	98				
White vs. Hispanic/ Latino	57					57	100				
White vs. Asian	57					55	96	2	4		
White vs. American Indian/ Alaska Native	57					57	100				
White vs. Native Hawaiian or Pacific Islander	57					57	100				
White vs. Two or more races	57					57	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	57					57	100				
Non-English Learner vs. English Learner	57			1	2	56	98				
Student without Disability vs. Student with Disability	57			1	2	56	98				

Table 8.4.5 presents a comprehensive view of Post-Administration DIF results for the grade 3 mathematics assessment. For the Female vs Male comparison, 3 items show B- level DIF, while the remaining 54 items exhibit A level DIF. For the White vs Black/African American comparison, 1 item shows B- level DIF, while the remaining 56 items exhibit A level DIF. For the White vs. Asian comparison 2 items show B+ level DIF, meaning that the Asian student group performed more favorably. For the Non-English Learner vs English Learner and Student without Disability vs. Student with Disability comparisons, 1 item shows B- level DIF, and 57 items show A level DIF.

Chapter 9. Reliability

9.1. Overview

Reliability focuses on the extent to which differences in scores reflect true differences in the level of knowledge, skills, and abilities being assessed rather than chance fluctuations. Thus, reliability measures the level of consistency of the scores that would result if the assessment were to be repeatedly administered under the same conditions. Any degree of inconsistency is assumed due to random fluctuations that occur during administration. The sources of random fluctuations can be internal or external for the students, internal for the assessment, and/or from other phenomena that randomly occur during administration. For example, random fluctuation can be due to the use of multiple forms of the assessment administered to the students, or the assignment of raters assigned to score students' responses to constructed-response item prompts. In statistical terms, the variance in the distribution of scores, essentially the observed differences among students, is partly due to real differences in the levels of knowledge, skills, and abilities being assessed (true variance) and partly due to differences caused by random errors that customarily occurs in the measurement process (error variance). Reliability is the proportion of the total variance that is true variance. Psychometricians use statistical formulas to estimate the level of reliability of students' scores.

There are several different ways to estimate 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 performances of students among items within an assessment. An internal consistency reliability estimate is used because it serves as a good estimate of reliability when using multiple items within a test form, but it does not consider form-to-form variation due to lack of test form parallelism, nor can it provide information regarding score reliability across repeated administrations due to day-to-day variations, for example, day-to-day changes of students' states of health or of the administration environment.

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 administrations if the students do not change in their level of the knowledge or skills measured by the assessment. Acceptable ranges of reliability tend to exceed 0.6, with values over 0.8 considered good to excellent (Cortina, 1993; Schmitt, 1996). Estimates lower than 0.5 may indicate a lack of internal consistency.

In classical test theory, standard errors of measurement (SEM) quantify the amount of error in the 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 amount of measurement error increases, and the

variability of students' observed scores is likely to increase across hypothetical repeated administrations. Observed scores with large SEMs pose a challenge to the valid interpretation of a single score. Classical test theory reliability and SEM estimates were calculated for each NJSLA test form and the weighted average score is reported here.

9.2. Reliability and SEM Estimation

Coefficient alpha (Cronbach, 1951) is a reliability measure for use when there are dichotomously or polytomously scored items (Brennan, 2001). The coefficient is calculated by using both the items' variances and observed variance of the total raw scores in the following formula:

Equation 9.1

$$\alpha_x = \frac{n}{n-1} \left(1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right)$$

in which n is the number of items, σ_i^2 is the variance of scores on each item, and σ_x^2 is the variance of the total raw scores. Coefficient alpha is a lower bound estimate of the reliability of the distribution of total raw scores. For example, if coefficient alpha has a value of .90 for the estimated level of reliability, the level of reliability (as a theoretical quantity) would be even higher in value. When other administration conditions and contexts are held constant, the more items the test form includes, the greater the value of coefficient alpha, and in turn, the greater the reliability of scores. Additionally, when sample sizes become smaller and more homogeneous, lower reliability estimates are obtained.

The formula for calculating the classical test theory SEM is

Equation 9.2

$$SEM = \sigma_x \sqrt{1 - \alpha_x}$$

where σ_x is the standard deviation of the total raw scores and α_x is the value of coefficient alpha as computed above.

9.3 Scale Score Reliability Estimation

Like the level of classical test theory reliability, the level of scale score reliability can range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain similar scores upon repeated testing occasions. Because the scale scores are computed via a procedure that differs from the calculation of the total raw scores, coefficient alpha cannot be computed for scale scores. Instead, Kolen et al.'s (1996) method for scale score reliability is calculated. For details of scale score calculation, please review Chapter 7.4.

The general formula for the reliability coefficient,

Equation 9.3

$$\rho = 1 - \frac{\sigma^2(E)}{\sigma^2(X)},$$

involves the error variance $\sigma^2(E)$ and the total observed 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. Denoting X as the raw total score ranging from 0 to X , and s as a resulting scale score after transformation, the conditional distribution of scale scores is written as $P(X = x|\theta)$. The mean and variance $\sigma^2[s(X)]$ of this distribution can be computed using these scores and their associated probabilities.

The average error variance of the scale scores is computed as

Equation 9.4

$$\sigma^2(Error_{scale}) = \int_{\theta} \sigma^2(s(X)|\theta) g(\theta) d\theta$$

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.

Equation 9.5

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

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

9.3. Reliability Results

9.3.1. Raw Score Reliability Results

Table 9.1 and Table 9.2 summarize test reliability estimates for the total testing group for ELA and mathematics, respectively. Please note that during operational test form construction, described in section 2.2.3 of this report, multiple parallel operational forms of the accommodated core form may be constructed to facilitate delivery and scoring. The tables below report the weighted average statistics for all operational forms created for the ELA and mathematics assessments. Estimates were calculated for the total group, and only for subgroups of 100 or more students, who were administered a specific test form.

Table 9.3.1 provides a summary of test reliability estimates for the ELA test forms based on data from the total group. Since reliabilities were averaged across test forms, the minimum and maximum reliabilities are also provided. Minimum reliability reports the reliability for the test form with the lowest value and maximum reliability reports the reliability for the test form with the largest value. Typically, test forms administered to larger numbers of students (operational online forms) tend to have higher reliability compared to test forms administered to smaller numbers of students (accommodated forms).

Mean reliabilities tended to be robust for each grade level, with the mean reliabilities ranging from 0.86 to 0.88. The lowest minimum reliability for a test form was 0.81 in grade 9 ELA.

Table 9.3.1 Summary of ELA 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	3.80	0.86	6,375	0.83	42,655	0.88
4	4	70	4.53	0.87	368	0.85	42,848	0.89
5	4	72	4.64	0.87	7,116	0.84	43,955	0.89
6	4	72	4.90	0.87	7,161	0.85	44,571	0.88
7	4	72	5.30	0.88	250	0.83	44,673	0.90
8	4	72	5.35	0.87	185	0.82	46,549	0.89
9	4	70	5.16	0.86	4,911	0.81	46,751	0.87

Table 9.3.2 provides a summary of test reliability estimates for the mathematics test forms based on data from the total group. Mean reliabilities tended to be robust for each grade level, with the mean reliabilities ranging from 0.84 to 0.91. The lowest minimum reliability for a test form was 0.78 in Algebra I.

Table 9.3.2 Summary of Mathematics Test Reliability Estimates for Total Group

Grade Level	Number of Forms	Avg. Max Possible Score	Avg. Raw Score SEM	Average Reliability	Minimum Reliability		Maximum Reliability	
					N	Alpha	N	Alpha
3	5	52	3.39	0.91	487	0.89	11,306	0.92
4	5	52	3.59	0.90	470	0.86	13,822	0.91
5	5	52	3.58	0.89	430	0.85	13,344	0.90
6	5	52	3.49	0.89	257	0.84	37,449	0.90
7	5	52	3.29	0.89	264	0.88	9,093	0.90
8	5	52	2.98	0.84	7,381	0.82	9,686	0.84
A1	5	56	3.32	0.87	259	0.78	44,010	0.88

Grade Level	Number of Forms	Avg. Max Possible Score	Avg. Raw Score SEM	Average Reliability	Minimum Reliability		Maximum Reliability	
					N	Alpha	N	Alpha
A2	5	56	3.76	0.88	43,948	0.87	541	0.92
GO	5	56	3.41	0.89	43,948	0.87	1,545	0.90

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

9.3.2. Scale Score Reliability Results

Table 9.3.3 and Table 9.3.4 report the weighted average scale score reliability and SEM estimates across test forms for ELA and mathematics, respectively.

Table 9.3.3 ELA Scale Score Reliability

Grade Level	Number of Forms	Avg. Scale Score SEM	Avg. Scale Score Reliability	Min. Scale Score Reliability	Max. Scale Score Reliability
3	4	13.88	0.92	0.87	0.94
4	4	13.22	0.91	0.83	0.94
5	4	13.19	0.90	0.83	0.92
6	4	12.25	0.90	0.83	0.92
7	4	12.79	0.93	0.87	0.95
8	4	13.24	0.93	0.89	0.95
9	3	13.74	0.91	0.84	0.94

Table 9.3.4 Mathematics Scale Score Reliability Results

Grade Level	Number of Forms	Avg. Scale Score SEM	Avg. Scale Score Reliability	Min. Scale Score Reliability	Max. Scale Score Reliability
3	5	10.38	0.95	0.93	0.96
4	5	10.33	0.94	0.89	0.94
5	5	11.30	0.92	0.86	0.93
6	5	11.71	0.90	0.81	0.92
7	5	11.83	0.86	0.80	0.89
8	5	13.31	0.87	0.85	0.89
A1	5	15.64	0.88	0.73	0.90
A2	4	12.47	0.95	0.93	0.95
GO	4	11.86	0.86	0.84	0.87

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

9.4. Reliability Results for Demographic Groups of Interest

Raw score reliability statistics were also calculated for student demographic groups. The results for these gender, ethnic and student needs groups, along with similar statistics calculated for the students taking various accommodated forms, are reported in Table 9.4.1 for grade 3 ELA and Table 9.4.2 for grade 3 mathematics. For some accommodation categories, the tables note "n/r" (not reported) for SEM and Alpha

reliability coefficient, indicating that the reliability estimates are not available for those specific accommodation types due to insufficient sample sizes. The reliability statistics for the demographic groups of interest for all grades and courses are available in Appendix 9.3.

Table 9.4.1 Grade 3 ELA Summary of Test Reliability Estimates for Subgroups

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	54	3.80	0.87	346	0.83	42,655	0.88
Gender							
Male	54	3.74	0.86	3,992	0.82	21,168	0.88
Female	54	3.85	0.86	128	0.84	21,485	0.88
Ethnicity							
Black/African American	54	3.58	0.85	1,101	0.78	6,026	0.88
Asian/Pacific Islander	54	4.04	0.84	248	0.84	4,966	0.85
Hispanic/Latino	54	3.63	0.85	2,456	0.80	13,402	0.88
American Indian/Alaska Native	54	3.82	0.88	105	0.88	105	0.88
Two or more races	54	3.96	0.87	193	0.86	1,466	0.87
White	54	3.90	0.84	108	0.80	16,679	0.86
Special Instruction Needs							
Economically Disadvantaged	54	3.56	0.85	3,102	0.77	15,254	0.87
Not Economically Disadvantaged	54	3.92	0.86	149	0.83	27,401	0.87
English Learner	54	3.32	0.82	806	0.75	4,811	0.85
Non-English Learner	54	3.84	0.86	263	0.83	37,844	0.87
Students with Disabilities (SWD)	54	3.58	0.87	335	0.83	5,746	0.89
Students without Disabilities	54	4.14	0.87	36,961	0.87	36,909	0.88
Students Taking Accommodated Forms							
ASL	54	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	54	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	54	3.35	0.84	203	0.84	203	0.84
Non-Screen Reader	54	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	54	2.90	0.82	6,350	0.82	6,350	0.82

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha

Note: n/r = not reported.

Table 9.4.2 Grade 3 Mathematics Summary of Test Reliability Estimates for Subgroups

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	3.39	0.91	487	0.89	14,110	0.92
Gender							
Male	52	3.40	0.91	284	0.89	5,997	0.93
Female	52	3.38	0.91	203	0.89	5,308	0.92
Ethnicity							
Black/African American	52	3.12	0.91	1,888	0.91	1,888	0.91
Asian/Pacific Islander	52	3.51	0.89	4,221	0.87	1,045	0.92
Hispanic/Latino	52	3.21	0.90	277	0.88	4,023	0.91
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.47	0.91	1,294	0.90	331	0.93
White	52	3.47	0.90	13,934	0.88	3,919	0.92
Special Instruction Needs							
Economically Disadvantaged	52	3.15	0.90	270	0.87	5,222	0.91
Not Economically Disadvantaged	52	3.48	0.90	22,864	0.89	6,979	0.92
English Learner	52	3.08	0.90	206	0.87	1,932	0.91
Non-English Learner	52	3.44	0.91	281	0.90	9,338	0.92
Students with Disabilities (SWD)	52	3.28	0.91	335	0.89	4,182	0.92
Students without Disabilities	52	3.41	0.91	152	0.87	9,108	0.92
Students Taking Accommodated Forms							
American Sign Language	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	3.29	0.89	294	0.89	294	0.89
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.37	0.92	22,690	0.92	22,690	0.92

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Students Taking Translated Forms							
Spanish Language	52	2.96	0.88	1,645	0.88	1,081	0.89

Note: n/r = not reported.

9.5. Reliability Estimates of Subclaim Scores

Table 9.5.1 presents the average reliability estimates for the various subclaim scores across different forms in ELA testing, providing insights into the consistency and dependability of these scores. The table includes claim and subclaim scores for Reading (RD), Reading: Literature (RL), Reading: Information (RI), Reading: Vocabulary (RV), Writing (WR), Writing: Written Expression (WE), and Writing: Knowledge of Language and Conventions (WKL). The table includes score points, representing the number of points for each subscore, and alpha, indicating the reliability coefficient for each subscore. The reliability estimates provide insights into the internal consistency and dependability of each subscore within the ELA tests across different forms.

Table 9.5.1 Average ELA Reliability Estimates for Subscores

	Reading: Total		Reading: Literature		Reading: Information		Reading: Vocabulary		Writing: Total		Writing Expression		Writing: Knowledge Language and Conventions	
Grade Level	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability
3	30-31	0.82	11-12	0.64	11-11	0.67	08-08	0.50	24-24	0.78	18-18	0.72	06-06	0.81
4	40-44	0.84	16-18	0.70	12-16	0.63	08-12	0.55	27-30	0.79	21-24	0.78	06-06	0.82
5	40-44	0.83	16-18	0.65	14-14	0.64	10-12	0.59	27-30	0.82	21-24	0.81	06-06	0.84
6	40-44	0.84	16-18	0.74	14-16	0.56	08-14	0.61	30-30	0.82	24-24	0.81	06-06	0.82
7	40-44	0.85	16-16	0.70	14-16	0.69	08-12	0.59	30-30	0.83	24-24	0.84	06-06	0.86
8	40-44	0.84	16-18	0.67	14-16	0.68	08-10	0.53	30-30	0.83	24-24	0.84	06-06	0.85
9	40-40	0.82	12-16	0.65	14-16	0.70	08-14	0.47	30-30	0.81	24-24	0.80	06-06	0.81

Table 9.5.2 presents reliability estimates for the various subscores across mathematics forms. The subscores include Major Content (MC), Additional & Supporting Content (ASC), Expressing Mathematical Reasoning (MR), and Modeling & Applications (M&A). Each form is associated with corresponding score points and alpha reliability coefficients for the mentioned subscores.

Table 9.5.2 Average Math Reliability Estimates for Subscores

	Major Content		Additional & Supporting Content		Mathematics Reasoning		Modeling Practice	
Grade Level	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability
3	20-20	0.85	10-10	0.75	10-10	0.64	12-12	0.65
4	21-21	0.83	09-09	0.67	10-10	0.69	12-12	0.66
5	20-20	0.79	10-10	0.66	10-10	0.62	12-12	0.66
6	20-20	0.78	10-10	0.60	10-10	0.69	12-12	0.65
7	20-20	0.81	10-10	0.63	10-10	0.70	12-12	0.66
8	20-20	0.70	10-10	0.45	10-10	0.57	12-12	0.53
A1	17-17	0.72	09-11	0.56	10-10	0.52	15-15	0.64
A2	16-18	0.80	12-12	0.69	10-10	0.50	15-15	0.67
GO	18-18	0.73	12-12	0.64	10-10	0.60	15-15	0.60

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

9.6. Reliability of Classification

Classification accuracy is defined as the extent to which the actual classifications of test takers (on the basis of their test scores) agree with those that would be made on the basis of their true scores—if their true scores could somehow be known. The term consistency refers to the agreement between classifications based on two nonoverlapping, equally difficult forms of the test (parallel forms) (Livingston & Lewis, 1995).

We used Livingston and Lewis's (1995) approach, which is intended to handle situations where items are not equally weighted and/or some or all the items are polytomously scored. This method is formulated as:

Equation 9.6

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

where X_{min} is the lowest score for X , X_{max} is the highest score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability. This method models the distribution of the true scores and of scores on a parallel form by using a four-parameter beta distribution.

As seen in the above formula, classification accuracy and consistency indices rely on the interaction between several different factors related to test design and standard-setting decisions. These factors include the number of score cuts, test reliability, measurement accuracy at the cut score, distance between adjacent cuts, location of the cut scores on the ability scale, and percentage of students around a cut score (Ercikan & Julian, 2002; Lee et al., 2002). Because these statistics are influenced by the interplay between a variety of factors, only a very limited number of studies to date have investigated the ideal or expected levels of decision consistency and accuracy needed for educational assessments.

Classification accuracy indices quantify the percentage of students who are accurately placed below and/or above a given cut score. For example, a classification accuracy index of 0.88 for cut score L1/L2 means that were students to be classified twice, once according to their observed score and once according to their true score, 88% of those students would be classified in the same category both times.

Similarly, classification consistency indices give the percentage of students classified consistently below and/or above a given cut score. For example, a classification index of 0.84 for cut score L1/L2 means that were two parallel forms to be administered to students, 84% of those students would be classified in the same way for both forms.

Table 9.6.1 through Table 9.6.4 display the weighted mean classification and accuracy indices at scores for ELA and mathematics.

Table 9.6.1 Classification Accuracy Indices at Cut Score Level for ELA

Grade	L1/L2	L2/L3	L3/L4	L4/L5
3	0.94	0.92	0.92	0.96
4	0.95	0.92	0.91	0.94
5	0.95	0.92	0.90	0.95
6	0.95	0.92	0.90	0.94
7	0.96	0.94	0.92	0.93
8	0.96	0.94	0.92	0.94
9	0.95	0.92	0.91	0.93

Table 9.6.1 shows the classification accuracy indices at cut score level for ELA, with columns indicating different cut score levels. For example, L1/L2 indicates classification accuracy at the cut score distinguishing achievement level 1 and achievement level 2. Classification accuracy was above 90% for all cuts. These results indicate that if students were classified twice as described above, they would be classified in the same achievement level over 90% of the time.

Table 9.6.2 Classification Consistency Indices at Cut Score Level for ELA

Form	L1/L2	L2/L3	L3/L4	L4/L5
3	0.91	0.89	0.88	0.95
4	0.93	0.89	0.87	0.91
5	0.93	0.89	0.86	0.93
6	0.93	0.89	0.86	0.92
7	0.94	0.91	0.89	0.90
8	0.94	0.91	0.89	0.91
9	0.93	0.89	0.87	0.91

Table 9.6.2 presents classification consistency indices at different cut score levels for ELA, with columns indicating different cut score levels. Classification consistency indices were above 86% for all cuts. These results indicate that if students were administered two parallel forms, they would be classified the same way for both forms over 85% of the time.

Table 9.6.3 Classification Accuracy Indices at Cut Score Level for Mathematics

Form	L1/L2	L2/L3	L3/L4	L4/L5
3	0.96	0.94	0.93	0.96
4	0.96	0.94	0.93	0.97
5	0.95	0.92	0.92	0.96
6	0.94	0.91	0.91	0.96
7	0.94	0.89	0.89	0.96
8	0.90	0.89	0.92	0.99
A1	0.91	0.89	0.91	0.97
A2	0.96	0.95	0.94	0.96
GO	0.97	0.93	0.88	0.93

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

Table 9.6.3 shows the classification accuracy indices at cut score level for mathematics, with columns indicating different cut score levels. For example, L1/L2 indicates classification accuracy at the cut score distinguishing achievement level 1 and achievement level 2. Classification accuracy was above 88% for all cuts, indicating if students were classified twice as described above, they would be classified in the same achievement level over 80% of the time in almost all instances.

Table 9.6.4 Classification Consistency Indices at Cut Score Level for Mathematics

Form	L1/L2	L2/L3	L3/L4	L4/L5
3	0.95	0.92	0.91	0.94
4	0.94	0.91	0.90	0.95
5	0.93	0.89	0.88	0.95
6	0.92	0.87	0.87	0.94
7	0.92	0.85	0.85	0.95
8	0.86	0.85	0.89	0.99
A1	0.88	0.85	0.87	0.96
A2	0.94	0.92	0.91	0.94
GO	0.96	0.90	0.84	0.90

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

Table 9.6.4 presents classification consistency indices at different cut score levels for mathematics, with columns indicating different cut score levels. Classification consistency indices were above 85% for all

cuts. These results indicate that if students were administered two parallel forms, they would be classified the same way for both forms over 85% of the time in almost all instances.

Chapter 10. Validity

10.1. 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), states the following:

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 spring 2022–2023 NJGPA administration provided an opportunity to gather evidence of validity based on both test content and on the internal structure of the tests.

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

10.2. 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 NJSLS) 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 jointly built spreadsheets at the evidence statement level that incorporated the probability statements from the test blueprints and attrition rates from committee review and data review. The basis of our entire item development is driven by the use of these item development target spreadsheets. Before beginning item development, Pearson uses these target spreadsheets to develop an internal item development plan to correlate with the expectations of the test design. These are reviewed

and approved by state or agency leads and New Meridian. All parties acknowledge that each assessment has multiple parts, and each part specifies the types of tasks and standards eligible for assessment.

In addition to the evidence statements, content is aligned through the articulation of performance in the performance level descriptors. At the policy level, the performance level descriptors include policy claims about the educational achievement of students who attain a particular performance level, as well as a broad description of the grade-level knowledge, skills, and practices that students performing at a particular achievement level are able to demonstrate. Such policy-level descriptors are the foundation for the subject- and grade-specific performance level descriptors, which, along with the evidence frameworks, guide the development of the items and tasks.

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

Gathering construct validity evidence for the assessments is embedded in the process by which the assessment content is developed and validated. At each step in the assessment development process, participating states or agencies involved hundreds of educators, assessment experts and bias and sensitivity experts in review of text, items and tasks for accuracy, appropriateness and freedom from bias. See Chapter 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. 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 in 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 with an electronic version of the accessibility guidelines and the linguistic complexity rubric. Items and passages are reviewed internally by accessibility and fairness experts trained in the principles of universal design and who become well-versed in the accessibility guidelines. Items received internal review for alignment to evidence tables, task generation model, item selection guidelines and accessibility and fairness reviews.

An important consideration when constructing test forms is recognition of items that may introduce construct-irrelevant variance. Such items should not be included on test forms to help ensure fairness to all subgroups of students. New Meridian convened bias and sensitivity committees to review all items. Additionally, content experts facilitated reviews of all items. All reviewers were trained using the bias and sensitivity guidelines, and the guidelines were used to review items and ELA passages. Accommodations were made available based on individual need documented in the student's approved IEP, 504 Plan, or if required by the participating state or agency, an English Learner (EL) Plan. An accessibility specialist worked in consultation with the accessibility specialist to review forms and determine which forms should be used for students with accommodations.

The ELA and mathematics operational test forms, as described in Chapter 2, were carefully constructed to align with the test blueprints and specifications that are based on the Common Core State Standards (CCSS). During the fall of 2016, content experts representing various participating states and agencies, along with other content experts, held a series of meetings to review the operational forms for ELA and mathematics. These meetings provided an 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. Subsequent test forms, including for use in 2023 follow the test forms specified in these meetings.

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

10.3. 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). 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 and for mathematics.

The summative assessments provide a full summative test score, a Reading claim score and a Writing claim score, as well as ELA 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.

10.3.1. Intercorrelations

The ELA 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 of 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 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 summative tests have four subclaim scores—Major Content (MC), Additional and Supporting Content (ASC), Expressing Mathematical Reasoning (EMR), and Modeling and Applications (M&A).

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 Chapter 9 for reliability estimates for the overall population, subgroups of interest, as well as for claims and subclaims for ELA 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 Reading and Writing claim scores and the ELA 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.

Table 10.3.1 presents the weighted average intercorrelation results for grade 3 ELA, and Table 10.3.2 presents the same information for grade 3 mathematics. Intercorrelations for all grades are provided in Appendix 10.1.

Table 10.3.1 Grade 3 ELA Average Intercorrelations between Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	92,192	1						
RL	92,192	0.88	1					
RI	92,192	0.89	0.66	1				
RV	92,192	0.79	0.53	0.60	1			
WR	92,192	0.72	0.60	0.73	0.51	1		
WE	92,192	0.71	0.59	0.71	0.49	0.99	1	
WKL	92,192	0.70	0.57	0.69	0.51	0.92	0.85	1

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

ELA intercorrelations range from moderate to high. The intercorrelations of reading subclaim scores range from 0.51 to 0.89 and likely indicate measurement of a single reading construct, with Reading Vocabulary a possibly related construct. The intercorrelations of writing subscores range from 0.85 to 0.99, indicating there is likely only one writing dimension being measured. The intercorrelations between reading and writing subscores, with the exception of Reading Vocabulary, likely indicate a unidimensional measurement factor. 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 intercorrelations amongst the subclaims are sufficiently high to provide evidence that the ELA tests are unidimensional. The moderate intercorrelations among the subclaims and claims suggest the claims may be sufficient for individual student reporting.

Table 10.3.2 Grade 3 Mathematics Average Intercorrelations between Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	93,968	1			
ASC	93,968	0.72	1		
MR	93,968	0.71	0.67	1	
MP	93,968	0.81	0.68	0.68	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice.

The grade 3 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, EMR, and M&A subclaims; the intercorrelations amongst the ASC, EMR, and M&A subclaims are usually slightly lower. The mathematics intercorrelations are sufficiently high to suggest that mathematics tests are likely to be unidimensional with some minor secondary dimensions.

All observed mathematics intercorrelations between subclaims are moderate. The main observable pattern in the mathematics intercorrelations is that the Major Content subclaim generally has slightly

higher correlations with the Additional and Supporting Content, Mathematical Reasoning, and Modeling Practice subclaims; the intercorrelations among the Additional and Supporting Content, Mathematical Reasoning, and Modeling Practice 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.

10.3.2. Reliability

The reliability analyses presented in Chapter 9 of this technical report provide information about the internal consistency of the summative assessments. Internal consistency is typically measured via correlations amongst the items on an assessment and provides an indication of how much the items measure the same general construct. The reliability estimates, computed using coefficient alpha (Cronbach, 1951) for all grades and courses, are presented in Appendix 9.1. As do the subclaim intercorrelations, the reliability estimates indicate that the items within each assessment are measuring the same construct and provide further evidence of unidimensionality.

10.3.3. Local Item Dependence

In addition to the intercorrelations for ELA and mathematics, local item independence was evaluated. Local independence, one of the primary assumptions of item response theory (IRT), states that 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. When present, local item dependence (LID) essentially overstates the amount of information predicted by the IRT model. LID can exert other undesirable psychometric effects and represents a threat to validity since other factors besides the construct of interest are present. Classical statistics are also affected when LID is present since estimates of test reliability like IRT information can be inflated (Zenisky et al., 2003).

The LID issue affects the choice of item scoring in IRT calibrations. Specifically, if evidence suggests these items indeed have local dependence, then it might be preferable to sum the item scores into clusters or testlets as a method of minimizing LID. However, if these items do not appear to have strong local item dependence, then retaining the scores as individual item scores in an IRT calibration is preferred since more information concerning item properties is retained. During the initial operational administration of the summative assessments in spring 2015, a study that included two methods of investigating the presence of LID was conducted. A description of the methods and study findings are summarized below.

First, analyses of the internal consistency in items and testlets were conducted under classical test theory (Wainer & Thissen, 2001) 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

Equation 10.1

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

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:

Equation 10.2

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

Equation 10.3

$$Q_3 = r(d_i, d_{i'})$$

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 general rule, 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 test levels, administered in spring 2015, were selected:

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

One spring 2015 CBT form for each of the tests was selected that was roughly at the median in terms of test difficulty. For ELA, 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 10.3.4.

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

The results for the internal consistency analysis are shown in Figure 10.3.1. In every instance, the item-level Cronbach's alpha is higher than in the testlet configuration. The greatest difference was for Algebra II, which showed a difference of 0.07. Although this was not unexpected, the magnitude of the differences in the respective alpha coefficients in general do not suggest a concerning level of LID. Table 10.4 shows the summary for the $Q3$ values. Figure 10.3.2 and Figure 10.3.3 show graphs of the distribution of $Q3$ values. Most of the $Q3$ values were small and negative, again suggesting that LID is not at a level of concern. For these two test levels, the difference in the alpha coefficients was .03 and was consistent with the low values of $Q3$.

In summary, this investigation did not find evidence for the existence of pervasive LID. The results of both the internal consistency analyses and $Q3$ methods support a claim of minimal LID. For a multiple-choice-only test containing four reading passages with 5 to 12 items associated with a reading passage, Sireci et al. (1991) reported that testlet alpha was approximately 10 percent lower than the item-level coefficient. In comparison, the tests have complex test structures and exhibited smaller differences in alpha coefficients. In addition, the median $Q3$ values presented in Table 10.3.5 centered around the expectation of $-1/n-1$.

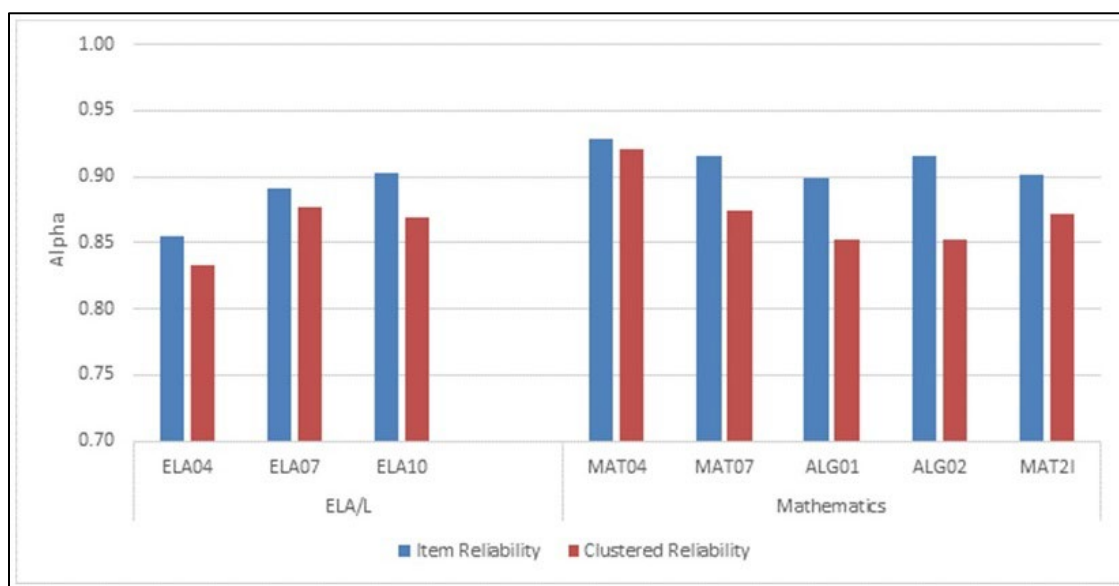


Figure 10.3.1 Comparison of Internal Consistency by Item and Cluster (Testlet)

Table 10.3.3 Conditions Used in LID Investigation and Results

Content	Grade	N Valid	N Complete	Percent Incomplete	No. Items	No. Tasks	Item Rel.	Task Rel.
ELA	4	13,660	13,518	1.04	31	5	0.86	0.83
ELA	7	12,757	12,685	0.56	41	7	0.89	0.88
ELA	10	3,097	3,033	2.07	41	7	0.90	0.87
Math	4	10,332	10,255	0.75	53	4	0.93	0.92
Math	7	10,295	10,188	1.04	50	6	0.92	0.87
Math	A1	5,072	4,885	3.69	52	6	0.90	0.85
Math	A2	4,982	4,769	4.28	54	6	0.92	0.85
Math	M2	2,708	2,645	2.33	51	6	0.90	0.87

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

Table 10.3.4 Summary of Q3 Values for ELA Grade 4 and Integrated Mathematics II (Spring 2015)

Course	Min.	Q1	Median	Mean	Q3	Max.	SD
ELA 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

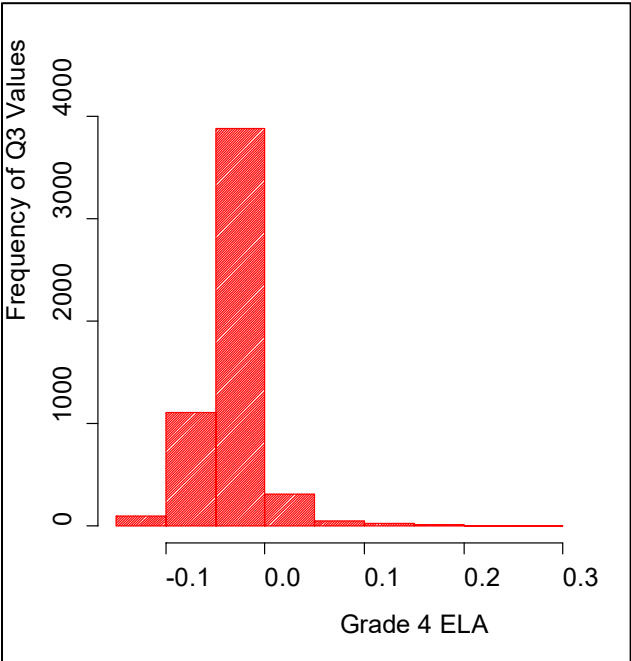


Figure 10.3.2 Distribution of Q3 Values for Grade 4 ELA (Spring 2015)

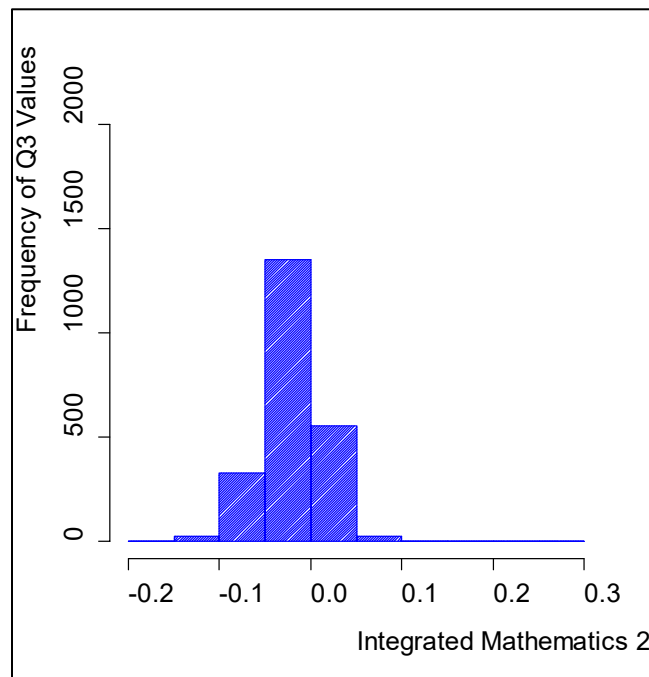


Figure 10.3.3 Distribution of Q3 Values for Integrated Mathematics II (Spring 2015)

10.4. Evidence from Special Studies

Several research studies were conducted to provide additional validity evidence for the participating state and agencies' goals of assessing more rigorous academic expectations, helping to prepare students for college and careers, and providing information back to teachers and parents about their students' progress toward college and career readiness. Some of the special studies conducted include

- 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.

10.4.1. Content Alignment Studies

In 2016, content of the ELA assessments at grades 5, 8, and 11 and the Algebra II and Integrated Mathematics II assessments were evaluated to determine how well the assessments were aligned to the Common Core State Standards (CCSS; Doorey, & Polikoff, 2016; Schultz et al., 2016). These content alignment studies were conducted by the Fordham Institute for grades 5 and 8 and by Human Resources Research Organization (HumRRO) for the high school assessments. Both of these studies used the same

methodology by having content experts review the assessment items and answers (for the constructed-response items the rubrics were reviewed). The content experts then judged how well the items aligned to the CCSS, the depth of knowledge of the items, and the accessibility of the items to all students, including English learners and students with disabilities. The authors of both studies noted that the content experts reviewing the assessments were required to be familiar with the CCSS but could not be employed by participating organizations or be the writers of the CCSS. Therefore, an effort was made to eliminate any potential conflicts of interest.

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

The assessment program was rated as Excellent Match for ELA content and depth and Good Match for mathematics content and depth for grades 5 and 8. However, for grade 11 ELA 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, it was noted that the assessments include complex texts, a range of cognitive demands, and contain a variety of item types. Furthermore, the ELA “assessments require close reading, assess writing to sources, research, and inquiry, and emphasize vocabulary and language skills” (Doorey & Polikoff, 2016). The grade 11 ELA assessment had a smaller range of depth and included items assessing the higher-demand cognitive level. A weakness of the ELA assessments is the lack of a listening and speaking component. It was also suggested that the ELA assessments could be enhanced by the inclusion of a research task that requires the use of two or more sources of information.

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

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

In 2017, HumRRO conducted a study to evaluate the quality and alignment of ELA 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, 99.6 percent of grade 3 and 4 items, 95.5 percent of grade 6 items, and 94.6 percent of grade 7 items were fully aligned. For mathematics, 92.0 percent of grade 3, 91.1 percent of grade 4 items, 83.1 percent of grade 6 items, and 94.0 percent of grade 7 items were fully aligned. The majority of the items that did not fall into fully aligned were considered partially aligned to the standards. CCSS are designed to be measured by multiple items, so items that aligned to multiple CCSS received a partially aligned rating. The overall item-to-CCSS alignment was captured by a holistic alignment rating that indicated if an item captured the identified standards as a set. Holistic ratings (either yes or no) were found by averaging review ratings across clusters for items that included more than one standard. For ELA, for all four grades, at least 93 percent of items had a holistic alignment rating of yes to indicate that the identified standards captured the skills or knowledge required. For mathematics, grade 6 had the lowest percentage for the holistic alignment rating of yes (84.8 percent), and grade 7 had the highest (96.3 percent). Overall, the alignment study suggests that the identified CCSS capture the knowledge and skills required in the items.

In addition to the alignment study, HumRRO also evaluated the CCSSO criteria for content and depth for ELA 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 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 assessments received a composite rating of Excellent Match for assessing the content needed for college and career readiness. There are four criteria for ELA 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 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 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.

10.4.2. Benchmarking Study

The purpose of the benchmarking study (McClarty et al., 2015) was to provide information that would inform the performance level setting (PLS) process. An evidence-based standard setting approach (EBSS; McClarty et al., 2013) was used to establish the performance levels for its assessments. In EBSS, the threshold scores for performance levels are set based on a combination of empirical research evidence and expert judgment. This benchmarking study provided one source of empirical evidence to inform the college- and career-readiness performance level (i.e., Level 4). The study findings were provided to a pre-policy standard-setting committee. The charge of this committee was to suggest a reasonable range for the percentage of students meeting or exceeding the Level 4 threshold score and therefore considered college- and career-ready. Chapter 2.4 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, Algebra II, and Integrated Mathematics III assessments, the grade 8 ELA and mathematics assessments, and the grade 4 ELA and mathematics assessments. The assessments and Level 4 expectations were compared with comparable assessments and expectations for the Programme of International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), National Assessment of Educational Progress (NAEP), ACT, SAT, the Michigan Merit Exam, and the Virginia End-of-Course exams. For each external assessment, the best-matched performance level was determined and the percentage of students reaching that level across the nation and in the participating states and agencies was determined. Across all grades and subjects, the data indicated approximately 25 to 50 percent of students were college- and career-ready or on track to readiness based on the Level 4 expectations.

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

10.4.3. Longitudinal Study of External Validity of Performance Levels (Phase 1)

In 2016–2017, the first phase of a two-part external validity study of claims about the alignment of Level 4 to college readiness was completed (Steedle et al., 2017) using the summative assessment scores from the 2014–2015 and 2015–2016 academic years. Associations between the performance levels and college-readiness benchmarks established by the College Board and ACT were used to study the claim that students who achieve Level 4 have a .75 probability of attaining at least a C in entry-level, credit-

bearing, postsecondary coursework. Regression estimates measured the relationship between the summative assessment scores and external test scores. The Level 4 benchmark was used to estimate the expected score on an external test, and vice versa. Assessment scores were dichotomized for additional analyses. Cross-tabulation tables provided classification agreement among tests. Logistic regression modeled the relationship between students' summative scores and their probabilities of meeting the external assessment benchmark, and vice versa.

These methods were used to make the following comparisons in mathematics: Algebra I and PSAT10 Math; Geometry and PSAT10 Math; Algebra II and PSAT10 Math; Algebra II and PSAT/NMSQT Math; Algebra II and SAT Math; and Algebra II and ACT Math. The classification agreement (meeting the benchmark on both tests, or not meeting the benchmark on both tests), ranged from 62.5 percent to 86.5 percent. The overall trend indicated that students who met the benchmark on a mathematics assessment were likely to meet or exceed the benchmark on an external test (probabilities ranged from .509 to .886). However, students who met the benchmark on the external test had relatively low probabilities of meeting the mathematics benchmark (.097 to .310).

The following comparisons were made in ELA: 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 results was similar to mathematics. The classification agreements ranged from 67.3 percent to 79.7 percent. Students meeting the ELA 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 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 was 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.

10.4.4. Mode and Device Comparability Studies

The summative assessments have been operational since the 2014–2015 school year. In addition to the traditional paper format, the assessments were available for online administration via a variety of

electronic devices, including desktop computers, laptop computers, and tablets. The research agenda includes several studies evaluating the interchangeability of scale scores across modes and devices.

This report describes a two-pronged study consisting of a mode comparability analysis and a device comparability analysis. In the mode comparability analysis, scores arising from the paper administration were compared to those arising from any type of online administration. In the device comparability analysis, online scores arising from tests administered using a tablet are compared with online scores arising from any other type of electronic administration where a tablet was not present (i.e., laptops, desktops, Chromebooks).

The goal of this study was threefold: 1) to investigate whether assessment items were of similar difficulty across the levels of conditions for each analysis (i.e., paper and online for the mode comparability analysis and tablet and non-tablet for the device comparability analysis); 2) to determine whether the psychometric properties of test scores were similar across the levels of conditions for each analysis; and 3) to determine whether overall test performance was similar across the levels of conditions for each analysis.

This study examined performance on 12 assessments, split evenly between mathematics and ELA. 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 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, which was the same as the minimum standard error of theta. Therefore, the adjustments are within measurement precision for each assessment.

The results of the device comparability study revealed consistent evidence supporting the comparability between the tablet condition (TC) and the non-tablet condition (NTC). Specifically, the item means suggested that items were similarly difficult for the TC and NTC, and none of the items were flagged for device effects. The DIF analysis revealed that none of the items had C-level DIF. Consistent with the findings at the item level, an examination of test reliability indicated that the TC and NTC test forms were similarly reliable and that none of the test forms were flagged for device effects. Furthermore, the test-

level adjustment analysis as well as the change of the students' performance levels after the adjustment constants were applied did not indicate strong evidence of device effects.

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

10.4.5. Quality Testing Standards

New Meridian, in coordination with multiple states and vendors, developed an alternate form of its summative assessment to meet the needs for shorter testing times desired by several states. Researchers used 2017 (Boyd et al., 2018) and 2018 (Minchen et al., 2018) student data to evaluate 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 (QTS; 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 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), 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 Table 10.4.1 and Table 10.4.2. 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 grade 10, the prior assessment was ELA grade 8 for the matching process.

Table 10.4.1 Prior Grades Used in ELA Matching

2019 Grade	Prior Grade	Prior Test Year
3	N/A	N/A
4	3	2018
5	4	2018
6	5	2018
7	6	2018
8	7	2018
10	8	2017

Table 10.4.2 Prior Grades/Courses Used in Mathematics Matching

2019 Grade	Prior Grade	Prior Test Year
3	N/A	N/A
4	3	2018
5	4	2018
6	5	2018
7	6	2018
8	7	2018
AI	7 (44%), 8 (56%)	2018
GO	AI	2018
A2	AI (10%), GO (90%)	2018

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

Sample sizes before and after the matching process are listed in Table 10.4.3 for ELA and Table 10.4.4 for mathematics. For the ELA grade 9, Geometry, and Algebra II assessments, matched samples were fairly small, ranging from 75 to 1,540. Due to the small sample for ELA 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 10.4.3 ELA Matching Sample Size Results

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

Table 10.4.4 Mathematics Matching Sample Size Results

	Form	Unmatched		Matched	
		Current Forms N	Original Forms N	Current Forms N	Original Forms N
3	1	88,858	26,531	25,970	25,970
	2	88,919	26,595	25,987	25,987
4	1	87,291	25,941	25,070	25,070
	2	87,488	26,192	25,207	25,207
5	1	91,136	27,333	26,377	26,377
	2	91,739	27,611	26,754	26,754
6	1	95,174	28,514	27,677	27,677
	2	94,800	28,342	27,665	27,665
7	1	93,777	24,547	23,855	23,855
	2	93,265	24,141	23,485	23,485
8	1	83,289	15,293	14,962	14,962
	2	76,135	13,973	13,695	13,695
A1	1	43,232	21,530	16,926	16,926
	2	46,482	23,036	18,157	18,157
GO	1	40,673	3,252	1,540	1,540
	2	40,918	3,360	1,514	1,514
A2	1	27,568	1,037	823	823
	2	27,527	1,066	753	753

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

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:

Equation 10.4

$$D = \frac{\bar{x}_1 - \bar{x}_2}{S_p}$$

where $\bar{x}_1$ and $\bar{x}_2$ 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:

Equation 10.5

$$h = 2(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_2})$$

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:

Equation 10.6

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

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

10.4.5.1. 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. First, p -values were compared. Scatterplots for the current form p -values and original form p -values for ELA grades 3 to 6 and mathematics grades 3 to 6 are presented in Figure 10.4.1 and Figure 10.4.2, respectively.

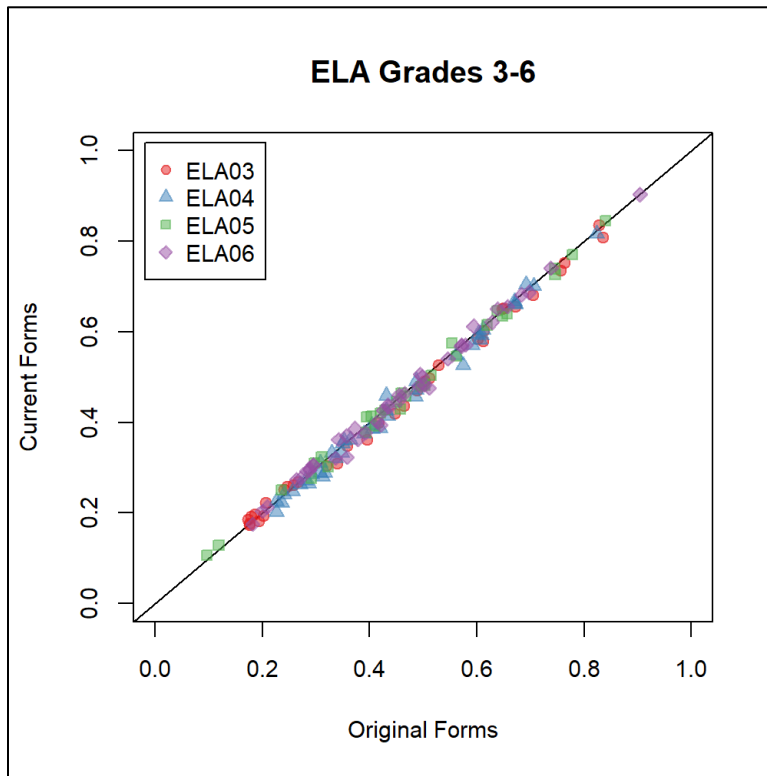


Figure 10.4.1 ELA Grades 3-6 P-Values

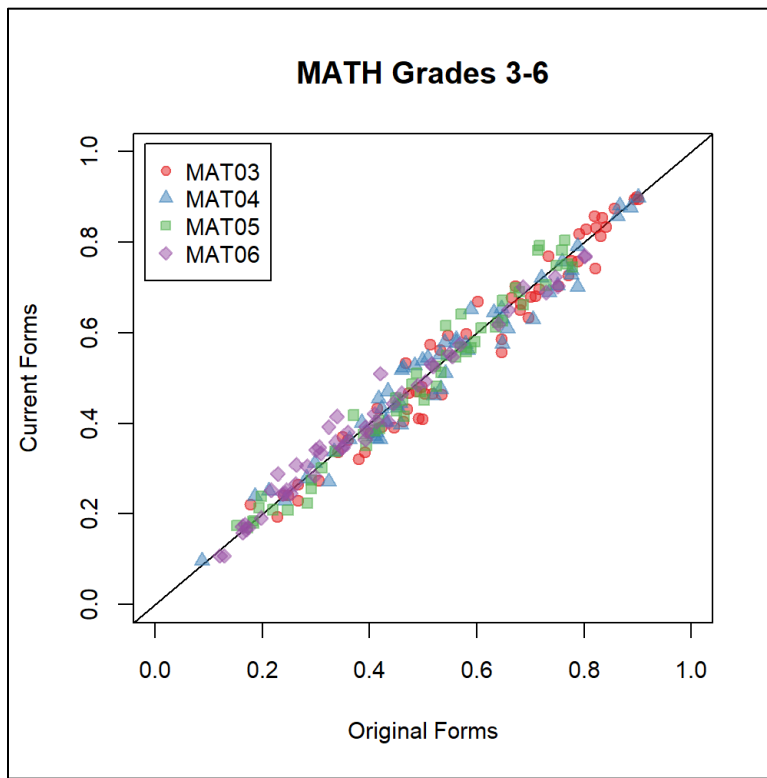


Figure 10.4.2 Mathematics Grades 3-6 P-Values

The scatterplots for all grades and courses are presented in Appendix 10.2., specifically in Figure A.10.2.1-Figure A.10.2.6. 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 grade 10, Algebra II, and Geometry.

The distributions of p -value differences for all grades are presented in Table A.10.2.4-Table A.10.2.5. Differences tend to be small and center around zero, except for ELA grade 10, Algebra II, and Geometry. For ELA grades 3 through 8, differences in item difficulties range from -0.049 to 0.070 . For mathematics grades 3 through 8 and Algebra I, differences in item difficulties range from -0.105 to 0.090 . The high school assessments show larger differences. P -values for ELA 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 Figure A.10.2.7-Figure A.10.2.12, 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 Table A.10.2.6-Table A.10.2.7, tend to be small and center around 0, except for ELA grade 10, Algebra II, and Geometry. For ELA grades 3 through 8, differences in polyserial values range from -0.058 to 0.043 . For Mathematics grades 3 through 8, differences in polyserial values range from -0.090 to 0.125 . The high school assessments show larger differences.

Common items were checked for differential item functioning (DIF) on several categories separately for the current and original forms, using the matched samples. The resulting cross tabulation 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 the current and original forms. DIF results summaries can be found in Table A.10.2.8-Table A.10.2.10.

10.4.5.2. 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 Table A.10.2.25-Table A.10.2.28. 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

Equation 10.7

$$SS_{2019} = \beta_0 + \beta_1 \times SS_{2018} + \beta_2 \times C + \beta_3 \times SS_{2018} \times C$$

Reduced Model

Equation 10.8

$$SS_{2019} = \beta_0 + \beta_1 \times SS_{2018}$$

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 Table A.10.2.29-Table A.10.2.30.

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. Chapter 11 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 and mathematics grades 4–8, differences between the mean SGPs were generally less than 5 percentile points in magnitude. At the overall level, mean differences (measured in percentile points and computed as the current form mean SGP minus original form mean SGP) ranged from –3.0 to 1.3 for ELA 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 grade 4, and mathematics grades 4, 6, and 8 were –5.3, –8.4, –9.1, and –6.5, respectively. Of note is that each of these exceptions occurs when the

sample size is relatively small. For mathematics grade 8, there were only 730 Asian students administered each type of form; all Native American grades contained less than 200 students for each type of form. SGP mean differences for all students and for each of the subgroups for Algebra I tended to be slightly higher than 5 in absolute value but always less than 10. Results for Geometry and Algebra 2 are not included due to small sample sizes.

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

10.4.5.3. 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 grade 10. Detailed results for ELA and mathematics grade 3 can be found in Table A.10.2.31-Table A.10.2.32, respectively. A summary of the effect sizes for all assessments can be found in Table A.10.2.33. 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 0.10 in magnitude, except for ELA grade 10. These results can be found in Table A.10.2.34.

10.4.5.4. 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 -0.04 to -0.23 for performance level classification accuracy (Table A.10.2.35 [ELA] and Table A.10.2.37 [mathematics]), from -0.05 to -0.25 for performance level classification consistency (Table A.10.2.36 [ELA] and Table A.10.2.38 [mathematics]), from -0.02 to -0.10 for CCR classification accuracy (Table A.10.2.35 and Table A.10.2.37), and from -0.02 to -0.12 for CCR classification consistency Tables (Table A.10.2.36 and Table A.10.2.38).

10.4.5.5. 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 and mathematics grade 6 can be found

in Table A.10.2.39 and Table A.10.2.40, while a summary of results across all assessments can be found in Table A.10.2.41.

10.4.5.6. 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 grade 10, the prior assessment used was ELA grade 8 from 2017. A test and results that are two years removed may be less than ideal. Second, high school tests tended to have smaller samples and were obtained from fewer states. Third, high school curriculum and course progressions may vary from state to state.

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

10.5. Evidence Based on Response Processes

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

New Meridian has undertaken research to investigate the quality of the items, tasks, and stimuli, focusing on whether students interact with items/tasks as intended, whether they were given enough time to complete the assessments, and the degree to which scoring rubrics allow accurate and reliable scoring. In addition, the accessibility of the test for students with disabilities and English learners has been examined. This research has included examining students' understanding of the format of the assessments and the use of technology.

One such study involved a series of four component studies that were conducted to evaluate the usability and effect of a drawing tool for online mathematics items. The purpose of these studies was to determine if results could support the use of the drawing tool, which is a way to expand students' ability to demonstrate their understanding and reasoning, thereby enhancing accessibility and construct validity of

the assessment. This goal is in keeping with guidance from the Common Core State Standards (CCSS) and the National Council of Teachers of Mathematics (NCTM) that students should have multiple paths and tools available to express their responses. Additionally, the drawing tool was intended to boost comparability across modes.

The first two studies (Brandt, Bercovitz, McNally, & Zimmerman, 2015; Brandt, Bercovitz, & Zimmerman, 2015) focused on evaluating the usability of the tool itself both in the general population and among students with low-vision and fine motor impairment disabilities. During these studies, detailed information regarding the functionality of the tool was collected and it was determined that the items should be tested operationally.

The third and fourth studies (Steedle & LaSalle, 2016; Minchen et al., 2018) involved evaluating the effect of the tool in the context of the operational assessments. The third study was conducted in grade 3 and the fourth study was conducted in grades 4 and 5. To evaluate the drawing tool in context, a set of items were studied by field testing them with and without the drawing tool. The drawing tool version of each item was randomly assigned to students so that comparisons could be made. The goal was to explore the impact of the drawing tool on item performance. In general, the results showed that the drawing tool usually did not have a significant impact on performance or item statistics. Items with access to the drawing tool, however, did show longer response times for grades 4 and 5, prompting a limitation to be placed on the number of drawing tool items in each unit.

Several other research efforts have investigated questions relevant to response processes evidence. Descriptions of the research conducted can be found online.¹

10.6. Interpretations of Test Scores

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

Students classified as either Level 4 or Level 5 are meeting or exceeding the grade level expectations. Performance level descriptors (PLDs) assist with the understanding and interpretations of the ELA scores (<https://resources.newmeridiancorp.org/ela-test-design/>) and mathematics scores (<https://resources.newmeridiancorp.org/math-test-design/>). Additionally, resource information is available

¹ Various research is described at: <http://resources.newmeridiancorp.org/>.

online to educators, parents, and students (<http://resources.newmeridiancorp.org/>). Chapter 7 of this technical report provides more information on the scale scores and the subclaim scores.

10.7. Evidence Based on the Consequences to Testing

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

The 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.

10.8. Summary

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

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

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

Several studies were conducted as part of the assessment program. They included 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 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.

The mode comparability study indicated that the comparability across modes was inconsistent across content domains and grade levels. The results of the mode comparability analysis were mixed and found to be consistent with prior research. The results of the device comparability study revealed consistent evidence supporting the comparability between the tablet condition (TC) and the non-tablet condition (NTC). In both the mode and device comparability studies, there were little to no items flagged for mode or device effects, the psychometric properties of test scores were similar across assessment conditions, and any adjustments to student performance for the paper or tablet condition were within measurement precision.

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

Chapter 6 provides information concerning the test characteristics based on classical test theory.

Chapter 7 provides detailed information concerning the scores that were reported and the cut scores for ELA and mathematics.

Chapter 8 provides information regarding student characteristics for the spring administration of the ELA and mathematics administration as well as information regarding the differential item functioning (DIF) analyses.

Chapter 9 provides information on the test reliability (total test score and for subclaims) and includes information on the interrater reliability/agreement.

Chapter 11. 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. The following sections describe the norm groups, the estimation procedure, and the results for SGPs based on New Jersey student results.

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 values represent higher growth, and lower values represent lower growth. For example, an SGP of 60 on grade 7 ELA means that the student scored better than 60 percent of the students in the state who took grade 7 ELA during the just-completed school year *and* who had achieved a similar score as this student on the grade 6 ELA assessment during the previous school year and the grade 5 ELA assessment two school years previously². An SGP of 50 represents median student growth for the state. 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 2022–2023 academic year is the eighth year of test administration. Students who do not have a previous test score, which include any new students and all grade 3 students, do not receive an SGP.

11.1 Norm Groups

The norm groups consisted of students with the same prior scores based on grade or content area progressions (termed academic peers). SGPs were based on up to two years of prior test scores from the two previous school years. Tables 11.1–11.5 list the grade or content area progressions required for SGPs based on one prior or two prior test scores for ELA grades 4 through 9, mathematics grades 4 through 8, Algebra I, Geometry and Algebra II, respectively. In general, the progressions of grade levels and content areas are consecutive. The 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

² 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 discussing it directly.

at least 1,000 students. Some progressions did not meet the minimum sample size for SGP calculations. SGPs are not calculated for grade 3 ELA or mathematics as there are no previous year scores.

Table 11.1 ELA 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	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

Table 11.2 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	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

Table 11.3 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

Table 11.4 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

Table 11.5 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

11.2 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 τ th quantile function for Y_t conditional on prior scores ($Y_{t-1}, \dots, Y_1$) is

$$Q_{Y_t}(\tau|Y_{t-1}, \dots, Y_1) = \sum_{u=1}^{t-1} \sum_{j=1}^n \phi_{ju}(Y_u) \beta_{ju}(\tau) \quad (11-1)$$

where ϕ_{ju} ($j=1,2,\dots, n$ students; $u=1, \dots, t-1$ 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 (CSEMs) were incorporated for calculation of SGP standard errors of measurement (SEMs). Goodness of fit results were checked (i.e., uniform distribution of SGPs by prior achievement) for indications of ceiling/floor effects for each SGP norm-group analysis.

11.3 Student Growth Percentile Results/Model Fit for Total Group

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

Tables 15.6 and 15.7 summarize SGP estimates for the total testing group for ELA and mathematics, respectively. Median SGPs were all 50. If the model is a perfect fit, the median is expected to be 50 with norm-referenced data. The minimum SGP is 1 and the maximum SGP is 99. The average standard error for the SGPs is within expectations for these models.

In general, SGPs can be divided into three categories: below 30 indicating that a student is not meeting a year's worth of growth, a SGP of 30–70 indicating that a student did achieve a year's worth of growth, and a SGP over 70 indicating that the student surpassed a year's worth of growth. It is important to note that definitions such as these are not inherent to the SGP method, but rather require expert judgment (Betebenner, 2009). The observed standard errors, ranging from 13.1–16.1, support these interpretations (Betebenner et al., 2016).

Table 11.6 Summary of ELA SGP Estimates for Total Group

Grade	Sample Size	Average SGP	Average Standard Error	Median SGP
4	88,032	49.9	13.5	50
5	89,634	50.0	13.9	50
6	90,846	50.0	13.2	50
7	91,934	50.0	13.1	50
8	95,033	50.0	13.6	50
9	88,782	49.9	13.4	50

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

Table 11.7 Summary of Mathematics SGP Estimates for Total Group

Grade	Sample Size	Average SGP	Average Standard Error	Median SGP
4	89,156	49.9	13.7	50
5	90,714	49.9	13.9	50
6	91,826	49.9	14.6	50
7	87,900	49.9	14.4	50
8	62,056	49.9	16.1	50
A1	87,426	49.8	16.1	50
GO	26,226	49.9	13.7	50
A2	8,522	49.8	14.4	50

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

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

11.4 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 15.8 summarizes SGPs for groups of interest for ELA grade 4. The ELA tables for grades 5–9 are provided in the Appendix (Tables A.11.1–A.11.6). Table 11.9 summarizes SGPs for groups of interest for mathematics grade 4; the other mathematics subgroup results are provided in the Appendix (Tables A.11.7–A.11.14). Median SGPs for subgroups of interest fell within the band of 30–70, which is considered to be adequate growth.

11.4.1 SGP Results for Gender

English Language Arts

The median SGPs for females tend to be higher than the median SGPs for males. The median SGP for females ranges from 49 to 54, whereas the median SGP for males ranges from 46 to 51. 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 47 to 52, and the median SGP for males ranges from 48 to 53. The standard errors for both are similar to the total group.

11.4.2 SGP Results for Ethnicity

English Language Arts

The African American group median SGP ranges from 43 to 48, with students in higher grades at the higher range. Asian American students tend to have the highest median SGPs, ranging from 59 to 63. American Indian/Alaska Native students had median SGPs ranging from 49 to 55.5. The median SGP for Hispanic students ranges from 45 to 49. For all ethnicity groups, standard errors are similar to that of the total group.

Mathematics

The median SGP for African American students ranges from 39 to 46, with the highest growth in mathematics grade 7. Asian American students tend to have the highest SGPs across all tests, with a minimum of 58 and a maximum of 62. American Indian/Alaska Native students had median SGPs ranging from 43 to 54. The median SGP for Hispanic students ranges from 43 to 48. For all ethnicities, the standard errors for all groups are under 20 points.

11.4.3 SGP Results for Special Instructional Needs

English Language Arts

Economically disadvantaged and English language learner students tend to have moderate median SGPs. The median SGP ranges from 44 to 48 for economically disadvantaged students and from 39 to 45 for English language learners. Students with disabilities observed median SGP of 37 to 42. 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 42 to 47 for economically disadvantaged students and from 39 to 47 for English language learners. For students with disabilities, the median SGP ranges from 37.5 to 46, 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 11.8 Summary of SGP Estimates for Subgroups: Grade 4 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	44,622	50.4	13.5	51
Female	43,404	49.5	13.4	49
Ethnicity				
White	34,374	52.0	13.4	53
African American	12,521	45.4	13.7	44
Asian	9,461	58.2	13.3	61
Pacific Islander	149	50.0	13.5	50
American Indian/Alaska Native	177	52.9	13.6	55
Hispanic	28,301	46.4	13.6	45
Multiple	3,038	52.3	13.5	54
Special Instruction Needs				
Economically Disadvantaged	32,481	45.5	13.6	44
Not-economically Disadvantaged	55,549	52.5	13.4	54
English Learner (EL)	7,995	44.2	14.0	42
Non-English Learner	80,036	50.5	13.4	51
Students with Disabilities (SWD)	15,203	44.0	14.0	42
Students without Disabilities	69,183	51.3	13.4	52

Table 11.9 Summary of Student Growth Percentile Estimates for Subgroups: Grade 4 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,198	50.2	13.7	50
Female	43,951	49.7	13.7	49
Ethnicity				
White	34,485	51.4	13.7	52
African American	12,534	45.5	13.7	44
Asian	9,525	57.5	14.2	61
Pacific Islander	152	54.9	14.3	52.5
American Indian/Alaska Native	177	51.0	13.4	48
Hispanic	29,231	47.5	13.6	46
Multiple	3,041	51.6	13.9	52
Special Instruction Needs				
Economically Disadvantaged	33,167	46.7	13.7	45
Not-economically Disadvantaged	55,987	51.9	13.8	53
English Learner (EL)	9,161	46.5	13.7	45
Non-English Learner	79,994	50.3	13.7	50
Students with Disabilities (SWD)	15,209	44.4	13.7	42
Students without Disabilities	70,310	51.1	13.8	52

Chapter 12. References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baldwin, P., Margolis, M.J., Clauser, B.E., Mee, J., Winward, M. (2019). The choice of response probability in bookmark standard setting: An experimental study. *Educational Measurement: Issues and Practice*, 39(1), 37–44.
- Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012). Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments. *Bulletin*, Issue 21. Pearson Education, Inc.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical Theories of Mental Test Scores* (pp 397–472). Reading, MA: Addison Wesley Publishing.
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38(4), 295–317.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). American Council on Education and Praeger.
- Cizek, G. & Bunch, M. (2007). *Standard setting: A guide to establishing performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Dorans, N. J. (2013). *ETS contributions to the quantitative assessment of item, test and score fairness (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04)*. Princeton, NJ: Educational Testing Service.
- Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach. (ETS Research Report No. 91-47)*. Princeton, NJ: Educational Testing Service

- Ercikan, K. & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education*, 15(3), 269–294.
- Henrysson, S. (1963). Correction of Item-Total Correlations in item analysis. *Psychometrika*, 28(2), 211–218.
- Holland, P. W. & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
- Kolen, M. J. (2004). POLYCSEM windows console version [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- 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.
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26, 412–432.
- Lemke, E., & Wiersma, W. (1976). *Principles of psychological measurement*. Chicago, Ill: McNally.
- Lewis, D., Mitzel, H., & Green, D. (1996, June). Standard setting: A bookmark approach. In D. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- Livingston, S. A., (2018). *Test reliability—basic concepts* (Research Memorandum No. RM-18-01). Princeton, NJ: 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, 179–197.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 453–461.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(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, 719–748.
- McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson Educational Measurement, Iowa City, IA: Pearson.

- Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* 16(2), 159–176.
- Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). Setting multiple performance standards using the Yes/No method: An alternative item mapping method. *Meeting of the National Council on Measurement in Education*. Montreal, Canada.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. Synthesis Report.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S.C. (2003). *Effects of local dependence on the validity of IRT item test, and ability statistics*. (Technical Report). American College Admissions Test.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Erlbaum
- Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items* (ETS Research Report RR-97-05). Princeton, NJ: Educational Testing Service.

(Appendices are numbered in reference to the chapter in the body of the text for which they expand or amplify the information reported.)

Appendix 2

A.2.1. Form Composition

Table A.2.1.1 Form Composition for ELA Grade 3

Major Claims	Subclaims	Number of Items	Number of Points
Reading	Literary Text	4–6	11–12
	Informational Text	4	11
	Vocabulary	4	8
	Claim Total	12–14	30–31
Writing	Written Expression	2	18
	Knowledge of Language and Conventions	2	6
	Claim Total	4	24
Summative Total		14–16*	54–55

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.2 Form Composition for ELA Grade 4

Major Claims	Subclaims	Number of Items	Number of Points
Reading	Literary Text	7–8	16–18
	Informational Text	6	16
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing	Written Expression	2	21–24
	Knowledge of Language and Conventions	2	6
	Claim Total	4	27–30
Summative Total		20*	67–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.3 Form Composition for ELA Grade 5

Major Claims	Subclaims	Number of Items	Number of Points
Reading			
	Literary Text	7–8	16–18
	Informational Text	6	16
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing			
	Written Expression	2	21–24
	Knowledge of Language and Conventions	2	6
	Claim Total	4	27–30
Summative Total		20*	67–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.4 Form Composition for ELA Grade 6

Major Claims	Subclaims	Number of Items	Number of Points
Reading			
	Literary Text	7–8	16–18
	Informational Text	6	16
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing			
	Written Expression	2	24
	Knowledge of Language and Conventions	2	6
	Claim Total	4	30
Summative Total		20*	70–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.5 Form Composition for ELA Grade 7

Major Claims	Subclaims	Number of Items	Number of Points
Reading			
	Literary Text	7–8	16–18
	Informational Text	6	16
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing			
	Written Expression	2	24

	Knowledge of Language and Conventions	2	6
	Claim Total	4	30
Summative Total		20*	70–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.6 Form Composition for ELA Grade 8

Major Claims	Subclaims	Number of Items	Number of Points
Reading			
	Literary Text	– 7–8	– 16–18
	Informational Text	6	16
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing			
	Written Expression	2	24
	Knowledge of Language and Conventions	2	6
	Claim Total	4	30
Summative Total		20*	70–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.7 Form Composition for ELA Grade 9

Major Claims	Subclaims	Number of Items	Number of Points
Reading			
	Literary Text	– 4–8	– 12–16
	Informational Text	– 6–9	– 16–22
	Vocabulary	4–5	8–10
	Claim Total	18	40–44
Writing			
	Written Expression	2	24
	Knowledge of Language and Conventions	2	6
	Claim Total	4	30
Summative Total		20*	70–74

*Items serve multiple reporting purposes. ELA PCR items report out to Written Expression, Knowledge of Language and Conventions, and Reading Literature or Informational Text.

Table A.2.1.8 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
Summative Total		33	52

Note. This table is identical to Table 2.3 in Chapter 2.

Table A.2.1.9 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
Summative Total		31	52

Table A.2.1.10 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
Summative Total		31	52

Table A.2.1.11 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
Summative Total		29	52

Table A.2.1.12 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
Summative Total		31	52

Table A.2.1.13 Form Composition for Mathematics Grade 8

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	18-19	20-22
	Additional & Supporting Content	5-6	8-10
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	12
Summative Total		30	52

Table A.2.1.14 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
Summative Total		28	55

Table A.2.1.15 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
Summative Total		30	55

Table A.2.1.16 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^*)	0–1	0–2
Summative Total		29	55

Appendix 6

A.6.1. Classical Item Analysis Statistics

Table A.6.1.1 ELA Grade 3 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.45	0.56
Item 2	0.71	0.75
Item 3	0.52	0.50
Item 4	0.49	0.52
Item 5	0.67	0.64
Item 6	0.51	0.60
Item 7	0.24	0.92
Item 8	0.66	0.73
Item 9	0.42	0.59
Item 10	0.70	0.78
Item 11	0.31	0.88
Item 12	0.27	0.29
Item 13	0.49	0.67
Item 14	0.34	0.60
Item 15	0.52	0.49
Item 16	0.63	0.72
Item 17	0.42	0.60
Item 18	0.28	0.84
Item 19	0.69	0.74
Item 20	0.64	0.55
Item 21	0.44	0.55
Item 22	0.62	0.81
Item 23	0.49	0.57

Table A.6.1.2 ELA Grade 4 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.76	0.63
Item 2	0.43	0.53
Item 3	0.66	0.54
Item 4	0.45	0.66
Item 5	0.50	0.51
Item 6	0.69	0.48
Item 7	0.81	0.76
Item 8	0.61	0.68
Item 9	0.62	0.65
Item 10	0.76	0.44
Item 11	0.26	0.89
Item 12	0.76	0.77
Item 13	0.65	0.69
Item 14	0.36	0.86
Item 15	0.76	0.79
Item 16	0.62	0.60
Item 17	0.34	0.46
Item 18	0.60	0.58
Item 19	0.81	0.64
Item 20	0.41	0.51
Item 21	0.47	0.36
Item 22	0.41	0.50
Item 23	0.52	0.62
Item 24	0.34	0.89

Item	P-value	Polyserial
Item 25	0.48	0.49
Item 26	0.26	0.32
Item 27	0.33	0.43
Item 28	0.48	0.47
Item 29	0.40	0.48
Item 30	0.41	0.43
Item 31	0.29	0.53

Table A.6.1.3 ELA Grade 5 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.69	0.64
Item 2	0.46	0.53
Item 3	0.30	0.41
Item 4	0.78	0.70
Item 5	0.48	0.58
Item 6	0.64	0.51
Item 7	0.55	0.55
Item 8	0.26	0.90
Item 9	0.54	0.55
Item 10	0.56	0.50
Item 11	0.57	0.51
Item 12	0.75	0.65
Item 13	0.92	0.69
Item 14	0.48	0.44
Item 15	0.49	0.54
Item 16	0.38	0.60
Item 17	0.29	0.39
Item 18	0.38	0.61
Item 19	0.69	0.60
Item 20	0.52	0.61
Item 21	0.56	0.54
Item 22	0.65	0.59
Item 23	0.43	0.50
Item 24	0.56	0.45
Item 25	0.71	0.71

Item	P-value	Polyserial
Item 26	0.31	0.89
Item 27	0.43	0.57
Item 28	0.44	0.30
Item 29	0.44	0.86
Item 30	0.45	0.46
Item 31	0.39	0.42

Table A.6.1.4 ELA Grade 6 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.63	0.63
Item 2	0.47	0.51
Item 3	0.41	0.50
Item 4	0.41	0.52
Item 5	0.69	0.66
Item 6	0.90	0.73
Item 7	0.59	0.48
Item 8	0.44	0.54
Item 9	0.76	0.68
Item 10	0.42	0.50
Item 11	0.37	0.92
Item 12	0.58	0.50
Item 13	0.29	0.49
Item 14	0.53	0.45
Item 15	0.34	0.49
Item 16	0.38	0.47
Item 17	0.63	0.48
Item 18	0.37	0.49
Item 19	0.35	0.37
Item 20	0.34	0.88
Item 21	0.66	0.67
Item 22	0.58	0.56
Item 23	0.44	0.52
Item 24	0.79	0.78
Item 25	0.61	0.64

Item	P-value	Polyserial
Item 26	0.58	0.70
Item 27	0.36	0.88
Item 28	0.76	0.63
Item 29	0.67	0.67
Item 30	0.57	0.60
Item 31	0.78	0.60

Table A.6.1.5 ELA Grade 7 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.46	0.57
Item 2	0.60	0.49
Item 3	0.52	0.45
Item 4	0.74	0.72
Item 5	0.44	0.86
Item 6	0.48	0.51
Item 7	0.56	0.61
Item 8	0.44	0.54
Item 9	0.44	0.60
Item 10	0.57	0.67
Item 11	0.69	0.51
Item 12	0.74	0.55
Item 13	0.39	0.46
Item 14	0.59	0.61
Item 15	0.50	0.51
Item 16	0.38	0.55
Item 17	0.46	0.60
Item 18	0.46	0.91
Item 19	0.38	0.18
Item 20	0.31	0.70
Item 21	0.62	0.67
Item 22	0.42	0.61
Item 23	0.63	0.60
Item 24	0.66	0.51
Item 25	0.65	0.51

Item	P-value	Polyserial
Item 26	0.49	0.58
Item 27	0.83	0.74
Item 28	0.62	0.62
Item 29	0.54	0.50
Item 30	0.75	0.54
Item 31	0.34	0.93

Table A.6.1.6 ELA Grade 8 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.42	0.46
Item 2	0.48	0.49
Item 3	0.67	0.71
Item 4	0.26	0.33
Item 5	0.65	0.49
Item 6	0.41	0.42
Item 7	0.45	0.91
Item 8	0.60	0.42
Item 9	0.56	0.54
Item 10	0.80	0.72
Item 11	0.42	0.61
Item 12	0.58	0.68
Item 13	0.39	0.42
Item 14	0.54	0.52
Item 15	0.68	0.54
Item 16	0.64	0.63
Item 17	0.43	0.63
Item 18	0.72	0.54
Item 19	0.39	0.34
Item 20	0.56	0.47
Item 21	0.70	0.59
Item 22	0.56	0.41
Item 23	0.36	0.50
Item 24	0.32	0.56
Item 25	0.52	0.54

Item	P-value	Polyserial
Item 26	0.47	0.93
Item 27	0.38	0.91
Item 28	0.40	0.55
Item 29	0.53	0.42
Item 30	0.69	0.68
Item 31	0.77	0.70

Table A.6.1.7 ELA Grade 9 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.04	0.27
Item 2	0.29	0.37
Item 3	0.01	0.47
Item 4	0.19	0.41
Item 5	0.38	0.44
Item 6	0.28	0.41
Item 7	0.31	0.24
Item 8	0.31	0.50
Item 9	0.12	0.25
Item 10	0.24	0.26
Item 11	0.40	0.40
Item 12	0.48	0.36
Item 13	0.02	0.41
Item 14	0.04	0.51
Item 15	0.14	0.29
Item 16	0.01	0.19
Item 17	0.02	0.55
Item 18	0.01	0.32
Item 19	0.19	0.48
Item 20	0.01	0.34

Table A.6.1.8 Math Grade 3 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.55	0.69
Item 2	0.30	0.64
Item 3	0.40	0.72
Item 4	0.71	0.80
Item 5	0.72	0.80
Item 6	0.62	0.51
Item 7	0.40	0.81
Item 8	0.83	0.48
Item 9	0.64	0.80
Item 10	0.56	0.72
Item 11	0.87	0.71
Item 12	0.68	0.59
Item 13	0.84	0.75
Item 14	0.63	0.61
Item 15	0.64	0.76
Item 16	0.61	0.74
Item 17	0.52	0.74
Item 18	0.67	0.75
Item 19	0.70	0.80
Item 20	0.25	0.68
Item 21	0.49	0.80
Item 22	0.82	0.73
Item 23	0.76	0.59
Item 24	0.26	0.73
Item 25	0.39	0.81

Item	P-value	Polyserial
Item 26	0.57	0.64
Item 27	0.56	0.71
Item 28	0.77	0.71
Item 29	0.83	0.44
Item 30	0.69	0.74
Item 31	0.89	0.69
Item 32	0.49	0.73
Item 33	0.31	0.81
Item 34	0.63	0.80
Item 35	0.77	0.63
Item 36	0.11	0.77
Item 37	0.80	0.74
Item 38	0.46	0.76
Item 39	0.54	0.61
Item 40	0.70	0.77
Item 41	0.59	0.74
Item 42	0.29	0.85
Item 43	0.55	0.67
Item 44	0.95	0.64
Item 45	0.47	0.41
Item 46	0.94	0.66
Item 47	0.70	0.81
Item 48	0.20	0.76
Item 49	0.77	0.75
Item 50	0.25	0.82
Item 51	0.46	0.74

Item	P-value	Polyserial
Item 52	0.33	0.80
Item 53	0.78	0.76
Item 54	0.92	0.66
Item 55	0.54	0.58
Item 56	0.67	0.79
Item 57	0.85	0.75

Table A.6.1.9 Math Grade 4 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.28	0.77
Item 2	0.61	0.61
Item 3	0.70	0.72
Item 4	0.20	0.57
Item 5	0.87	0.66
Item 6	0.68	0.73
Item 7	0.39	0.82
Item 8	0.66	0.75
Item 9	0.54	0.68
Item 10	0.82	0.51
Item 11	0.57	0.62
Item 12	0.65	0.72
Item 13	0.57	0.48
Item 14	0.27	0.77
Item 15	0.44	0.74
Item 16	0.17	0.52
Item 17	0.60	0.77
Item 18	0.30	0.69
Item 19	0.26	0.73
Item 20	0.68	0.79
Item 21	0.69	0.75
Item 22	0.45	0.72
Item 23	0.47	0.79
Item 24	0.32	0.68
Item 25	0.59	0.52

Item	P-value	Polyserial
Item 26	0.48	0.80
Item 27	0.40	0.81
Item 28	0.61	0.78
Item 29	0.79	0.70
Item 30	0.88	0.70
Item 31	0.86	0.68
Item 32	0.46	0.66
Item 33	0.63	0.53
Item 34	0.44	0.44
Item 35	0.40	0.70
Item 36	0.51	0.79
Item 37	0.73	0.66
Item 38	0.48	0.75
Item 39	0.52	0.81
Item 40	0.67	0.69
Item 41	0.41	0.65
Item 42	0.26	0.84
Item 43	0.83	0.71
Item 44	0.37	0.48
Item 45	0.61	0.63
Item 46	0.59	0.79
Item 47	0.39	0.78
Item 48	0.71	0.79
Item 49	0.20	0.82
Item 50	0.94	0.64
Item 51	0.90	0.67

Item	P-value	Polyserial
Item 52	0.80	0.71
Item 53	0.25	0.74
Item 54	0.75	0.74

Table A.6.1.10 Math Grade 5 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.53	0.71
Item 2	0.56	0.47
Item 3	0.40	0.80
Item 4	0.77	0.77
Item 5	0.25	0.82
Item 6	0.86	0.70
Item 7	0.13	0.74
Item 8	0.82	0.73
Item 9	0.12	0.80
Item 10	0.16	0.82
Item 11	0.14	0.81
Item 12	0.50	0.46
Item 13	0.30	0.71
Item 14	0.65	0.77
Item 15	0.31	0.73
Item 16	0.17	0.81
Item 17	0.22	0.68
Item 18	0.80	0.67
Item 19	0.87	0.57
Item 20	0.38	0.82
Item 21	0.37	0.81
Item 22	0.74	0.42
Item 23	0.40	0.78
Item 24	0.50	0.56
Item 25	0.62	0.85

Item	P-value	Polyserial
Item 26	0.76	0.37
Item 27	0.65	0.73
Item 28	0.23	0.68
Item 29	0.26	0.81
Item 30	0.25	0.71
Item 31	0.57	0.80
Item 32	0.75	0.72
Item 33	0.13	0.74
Item 34	0.21	0.68
Item 35	0.72	0.61
Item 36	0.75	0.74
Item 37	0.84	0.73
Item 38	0.56	0.82
Item 39	0.59	0.33
Item 40	0.39	0.77
Item 41	0.24	0.64
Item 42	0.21	0.30
Item 43	0.43	0.82
Item 44	0.77	0.55
Item 45	0.38	0.77
Item 46	0.79	0.71
Item 47	0.73	0.73
Item 48	0.75	0.75
Item 49	0.51	0.58
Item 50	0.87	0.68
Item 51	0.52	0.61

Item	P-value	Polyserial
Item 52	0.68	0.75
Item 53	0.64	0.79
Item 54	0.55	0.50
Item 55	0.36	0.63
Item 56	0.92	0.65

Table A.6.1.11 Math Grade 6 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.10	0.82
Item 2	0.24	0.81
Item 3	0.78	0.35
Item 4	0.37	0.79
Item 5	0.63	0.84
Item 6	0.65	0.66
Item 7	0.61	0.54
Item 8	0.94	0.37
Item 9	0.37	0.60
Item 10	0.24	0.42
Item 11	0.13	0.75
Item 12	0.45	0.63
Item 13	0.79	0.64
Item 14	0.87	0.69
Item 15	0.78	0.75
Item 16	0.79	0.65
Item 17	0.26	0.78
Item 18	0.71	0.64
Item 19	0.74	0.71
Item 20	0.42	0.71
Item 21	0.13	0.76
Item 22	0.48	0.52
Item 23	0.35	0.77
Item 24	0.32	0.80
Item 25	0.66	0.80

Item	P-value	Polyserial
Item 26	0.57	0.55
Item 27	0.34	0.45
Item 28	0.66	0.68
Item 29	0.38	0.80
Item 30	0.34	0.51
Item 31	0.15	0.64
Item 32	0.11	0.80
Item 33	0.32	0.81
Item 34	0.51	0.39
Item 35	0.58	0.77
Item 36	0.51	0.65
Item 37	0.49	0.55
Item 38	0.60	0.57
Item 39	0.08	0.71
Item 40	0.23	0.72
Item 41	0.20	0.85
Item 42	0.15	0.81
Item 43	0.69	0.77
Item 44	0.58	0.57
Item 45	0.61	0.77
Item 46	0.46	0.68
Item 47	0.43	0.56
Item 48	0.16	0.71
Item 49	0.40	0.81
Item 50	0.31	0.76
Item 51	0.32	0.31

Table A.6.1.12 Math Grade 7 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.43	0.76
Item 2	0.15	0.81
Item 3	0.28	0.83
Item 4	0.12	0.84
Item 5	0.45	0.50
Item 6	0.12	0.58
Item 7	0.75	0.49
Item 8	0.22	0.80
Item 9	0.60	0.80
Item 10	0.26	0.63
Item 11	0.35	0.31
Item 12	0.20	0.78
Item 13	0.67	0.76
Item 14	0.72	0.75
Item 15	0.15	0.77
Item 16	0.20	0.73
Item 17	0.30	0.43
Item 18	0.51	0.38
Item 19	0.69	0.61
Item 20	0.29	0.76
Item 21	0.24	0.74
Item 22	0.19	0.77
Item 23	0.73	0.75
Item 24	0.74	0.75
Item 25	0.80	0.68

Item	P-value	Polyserial
Item 26	0.87	0.69
Item 27	0.48	0.73
Item 28	0.37	0.45
Item 29	0.59	0.40
Item 30	0.33	0.81
Item 31	0.65	0.74
Item 32	0.11	0.73
Item 33	0.66	0.76
Item 34	0.85	0.58
Item 35	0.24	0.62
Item 36	0.61	0.47
Item 37	0.22	0.80
Item 38	0.50	0.47
Item 39	0.54	0.78
Item 40	0.17	0.85
Item 41	0.16	0.79
Item 42	0.15	0.75
Item 43	0.63	0.62
Item 44	0.21	0.66
Item 45	0.62	0.37
Item 46	0.45	0.83
Item 47	0.48	0.30
Item 48	0.21	0.59
Item 49	0.24	0.83
Item 50	0.08	0.76
Item 51	0.42	0.65

Item	P-value	Polyserial
Item 52	0.31	0.49
Item 53	0.36	0.82
Item 54	0.59	0.76
Item 55	0.51	0.80

Table A.6.1.13 Math Grade 8 Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.18	0.63
Item 2	0.47	0.58
Item 3	0.51	0.38
Item 4	0.11	0.68
Item 5	0.12	0.74
Item 6	0.23	0.66
Item 7	0.15	0.44
Item 8	0.24	0.51
Item 9	0.68	0.71
Item 10	0.44	0.36
Item 11	0.70	0.63
Item 12	0.29	0.66
Item 13	0.26	0.22
Item 14	0.09	0.53
Item 15	0.20	0.66
Item 16	0.05	0.60
Item 17	0.05	0.33
Item 18	0.19	0.59
Item 19	0.60	0.75
Item 20	0.08	0.69
Item 21	0.35	0.37
Item 22	0.34	0.30
Item 23	0.39	0.59
Item 24	0.11	0.70
Item 25	0.27	0.36

Item	P-value	Polyserial
Item 26	0.12	0.76
Item 27	0.58	0.73
Item 28	0.30	0.55
Item 29	0.20	0.50
Item 30	0.39	0.43
Item 31	0.71	0.57
Item 32	0.16	0.36
Item 33	0.48	0.54
Item 34	0.18	0.27
Item 35	0.18	0.55
Item 36	0.21	0.75
Item 37	0.40	0.41
Item 38	0.36	0.77
Item 39	0.39	0.37
Item 40	0.63	0.77
Item 41	0.07	0.52
Item 42	0.04	0.55
Item 43	0.57	0.61
Item 44	0.49	0.76
Item 45	0.23	0.30
Item 46	0.35	0.21
Item 47	0.17	0.59
Item 48	0.15	0.75
Item 49	0.26	0.64
Item 50	0.13	0.77
Item 51	0.38	0.29

Item	P-value	Polyserial
Item 52	0.50	0.47
Item 53	0.35	0.46

Table A.6.1.14 Algebra I Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.42	0.57
Item 2	0.62	0.71
Item 3	0.08	0.61
Item 4	0.41	0.80
Item 5	0.35	0.77
Item 6	0.33	0.79
Item 7	0.30	0.37
Item 8	0.19	0.75
Item 9	0.45	0.40
Item 10	0.29	0.72
Item 11	0.17	0.85
Item 12	0.26	0.52
Item 13	0.18	0.83
Item 14	0.40	0.82
Item 15	0.26	0.72
Item 16	0.34	0.32
Item 17	0.32	0.24
Item 18	0.58	0.56
Item 19	0.29	0.29
Item 20	0.49	0.68
Item 21	0.21	0.73
Item 22	0.65	0.71
Item 23	0.31	0.36
Item 24	0.56	0.56
Item 25	0.56	0.65

Item	P-value	Polyserial
Item 26	0.39	0.44
Item 27	0.11	0.65
Item 28	0.41	0.36
Item 29	0.33	0.74
Item 30	0.15	0.70
Item 31	0.35	0.79
Item 32	0.60	0.72
Item 33	0.43	0.38
Item 34	0.73	0.70
Item 35	0.07	0.77
Item 36	0.27	0.77
Item 37	0.09	0.69
Item 38	0.44	0.62
Item 39	0.14	0.47
Item 40	0.38	0.43
Item 41	0.52	0.52
Item 42	0.27	0.34
Item 43	0.47	0.40
Item 44	0.34	0.79
Item 45	0.34	0.37
Item 46	0.30	0.36
Item 47	0.37	0.81
Item 48	0.11	0.37
Item 49	0.39	0.74

Table A.6.1.15 Algebra II Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.52	0.46
Item 2	0.50	0.63
Item 3	0.37	0.49
Item 4	0.21	0.57
Item 5	0.35	0.79
Item 6	0.92	0.47
Item 7	0.78	0.79
Item 8	0.76	0.53
Item 9	0.59	0.59
Item 10	0.66	0.80
Item 11	0.64	0.66
Item 12	0.46	0.73
Item 13	0.31	0.82
Item 14	0.53	0.78
Item 15	0.41	0.35
Item 16	0.43	0.83
Item 17	0.42	0.83
Item 18	0.42	0.52
Item 19	0.49	0.52
Item 20	0.34	0.55
Item 21	0.27	0.77
Item 22	0.37	0.55
Item 23	0.43	0.81
Item 24	0.68	0.51
Item 25	0.32	0.73

Item	P-value	Polyserial
Item 26	0.32	0.79
Item 27	0.25	0.71
Item 28	0.53	0.63
Item 29	0.26	0.72
Item 30	0.23	0.67
Item 31	0.20	0.77
Item 32	0.28	0.78
Item 33	0.13	0.75
Item 34	0.59	0.64
Item 35	0.25	0.78
Item 36	0.20	0.69
Item 37	0.32	0.70
Item 38	0.56	0.70
Item 39	0.37	0.59
Item 40	0.77	0.62
Item 41	0.66	0.56
Item 42	0.21	0.70
Item 43	0.71	0.53
Item 44	0.83	0.56
Item 45	0.18	0.75
Item 46	0.48	0.80
Item 47	0.64	0.80
Item 48	0.55	0.76
Item 49	0.34	0.78
Item 50	0.40	0.66
Item 51	0.71	0.72

Table A.6.1.16 Geometry Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.30	0.79
Item 2	0.49	0.46
Item 3	0.69	0.60
Item 4	0.58	0.49
Item 5	0.38	0.70
Item 6	0.52	0.48
Item 7	0.74	0.46
Item 8	0.20	0.59
Item 9	0.33	0.84
Item 10	0.52	0.51
Item 11	0.61	0.68
Item 12	0.66	0.73
Item 13	0.15	0.43
Item 14	0.37	0.72
Item 15	0.79	0.75
Item 16	0.30	0.79
Item 17	0.53	0.57
Item 18	0.53	0.51
Item 19	0.47	0.81
Item 20	0.28	0.67
Item 21	0.74	0.43
Item 22	0.56	0.43
Item 23	0.35	0.69
Item 24	0.45	0.53
Item 25	0.13	0.78

Item	P-value	Polyserial
Item 26	0.51	0.66
Item 27	0.45	0.72
Item 28	0.33	0.69
Item 29	0.21	0.78
Item 30	0.25	0.66
Item 31	0.66	0.74
Item 32	0.25	0.60
Item 33	0.57	0.38
Item 34	0.78	0.46
Item 35	0.21	0.72
Item 36	0.57	0.62
Item 37	0.13	0.84
Item 38	0.21	0.61
Item 39	0.28	0.38
Item 40	0.65	0.39
Item 41	0.38	0.50
Item 42	0.45	0.71
Item 43	0.37	0.66
Item 44	0.14	0.79
Item 45	0.50	0.79
Item 46	0.24	0.73
Item 47	0.74	0.36
Item 48	0.81	0.53
Item 49	0.58	0.60
Item 50	0.55	0.77
Item 51	0.12	0.60

Item	P-value	Polyserial
Item 52	0.25	0.37
Item 53	0.13	0.75

Appendix 7

A.7.1. IRT Threshold Scores and Scaling Constants

Table A.7.1.1 Threshold Scores and Scaling Constants for ELA

Grade	L2		L3		L4		L5		A	B
	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score		
3	-0.9648	700	-0.284	725	0.3968	750	2.036	810	36.722	735.429
4	-1.3004	700	-0.5079	725	0.2846	750	1.5578	790	31.546	741.021
5	-1.3411	700	-0.4924	725	0.3563	750	2.0224	799	29.458	739.505
6	-1.3656	700	-0.4827	725	0.4002	750	1.8133	790	28.316	738.667
7	-1.2488	700	-0.5117	725	0.2254	750	1.2614	785	33.916	742.354
8	-1.273	700	-0.5402	725	0.1925	750	1.4696	794	34.118	743.433
9	-1.1635	700	-0.4329	725	0.2977	750	1.5065	791	34.217	739.812

Table A.7.1.2 Threshold Scores and Scaling Constants for Mathematics

Grade	L2		L3		L4		L5		A	B
	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score		
3	-1.4141	700	-0.6356	725	0.1429	750	1.3931	790	32.113	745.411
4	-1.384	700	-0.5484	725	0.2873	750	1.8323	796	29.916	741.404
5	-1.4571	700	-0.5959	725	0.2653	750	1.6262	790	29.030	742.299
6	-1.3829	700	-0.4948	725	0.3935	750	1.7567	788	28.146	738.925
7	-1.4464	700	-0.4505	725	0.5453	750	1.9919	786	25.103	736.310
8	-0.8851	700	-0.1264	725	0.6323	750	2.1896	801	32.950	729.164
A1	-1.1781	700	-0.3853	725	0.4075	750	2.1651	805	31.532	737.149
A2	-0.5759	700	0.086	725	0.748	750	2.2728	808	37.767	721.750
GO	-1.3013	700	-0.3389	725	0.6235	750	1.894	783	25.977	733.803

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

Table A.7.1.3 Scaling Constants for Reading and Writing of ELA

Grade	Reading		Writing	
	AR	BR	AW	BW
3	14.6891	44.1719	7.3445	32.0859
4	12.6184	46.4086	6.3093	33.2043
5	11.7832	45.8019	5.8916	32.901
6	11.3264	45.4669	5.6632	32.7335
7	13.5664	46.9416	6.7832	33.4708
8	13.6472	47.3732	6.8237	33.6866
9	13.687	45.925	6.8435	32.9625

A.7.2. IRT Test Characteristic, Information and Standard Error of Measurement Curves

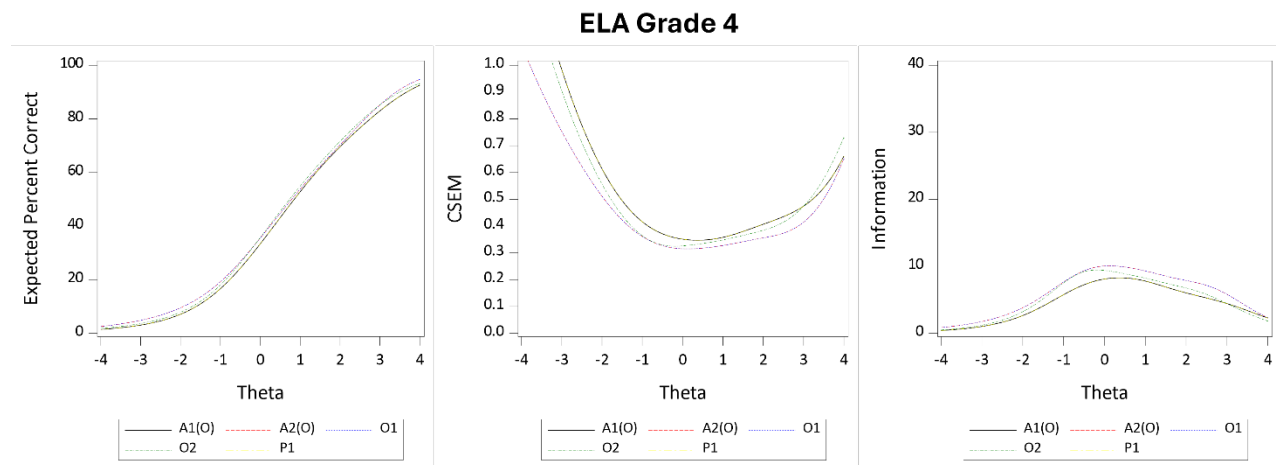


Figure A.7.2.1 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 4

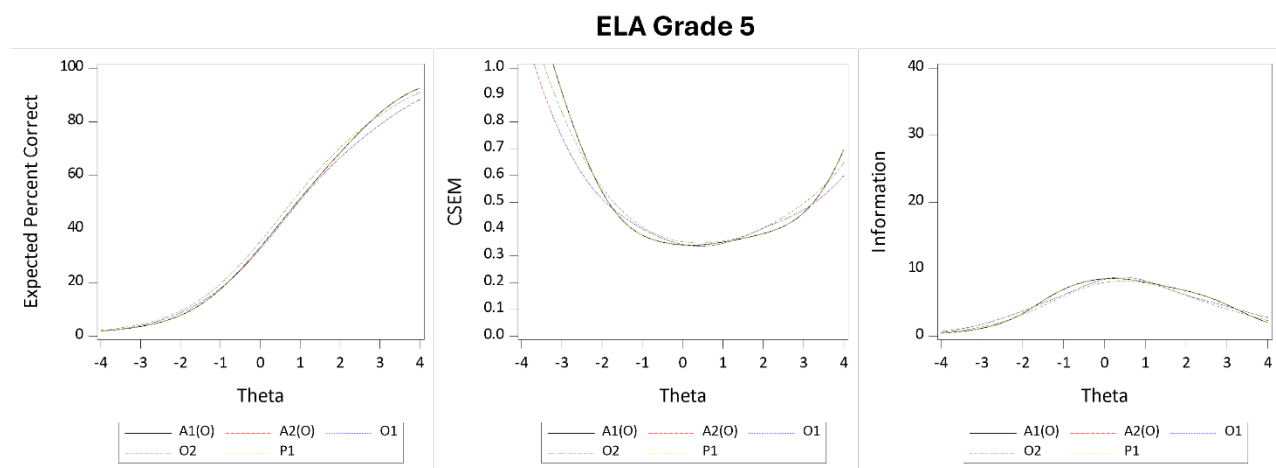


Figure A.7.2.2 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 5

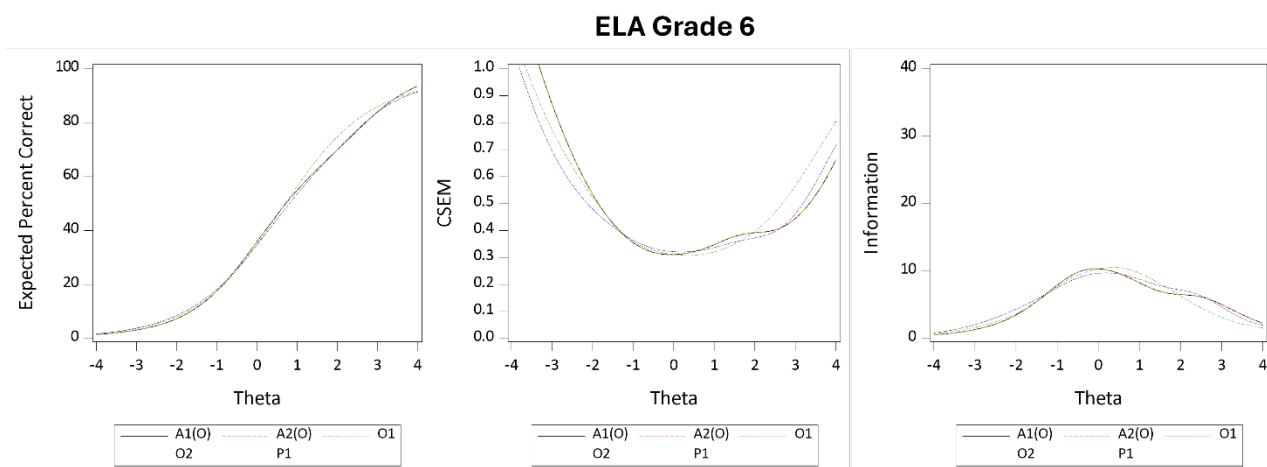


Figure A.7.2.3 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 6

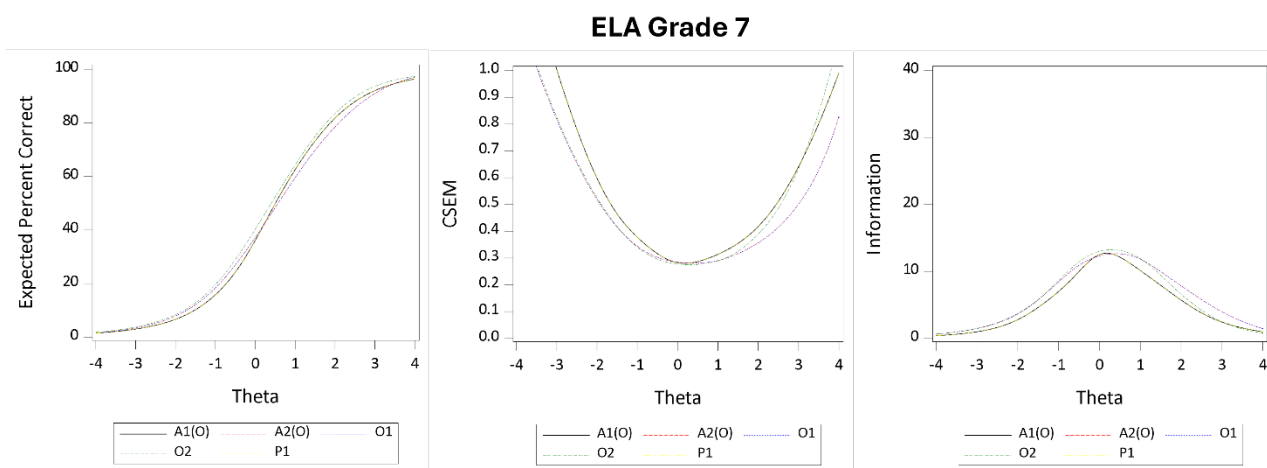


Figure A.7.2.4 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 7

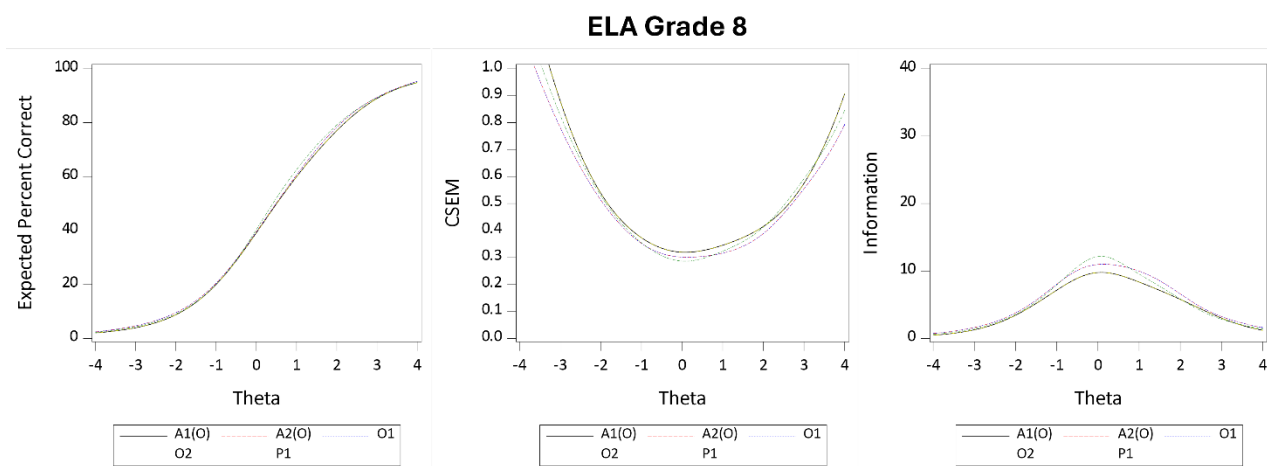


Figure A.7.2.5 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 8

ELA Grade 9

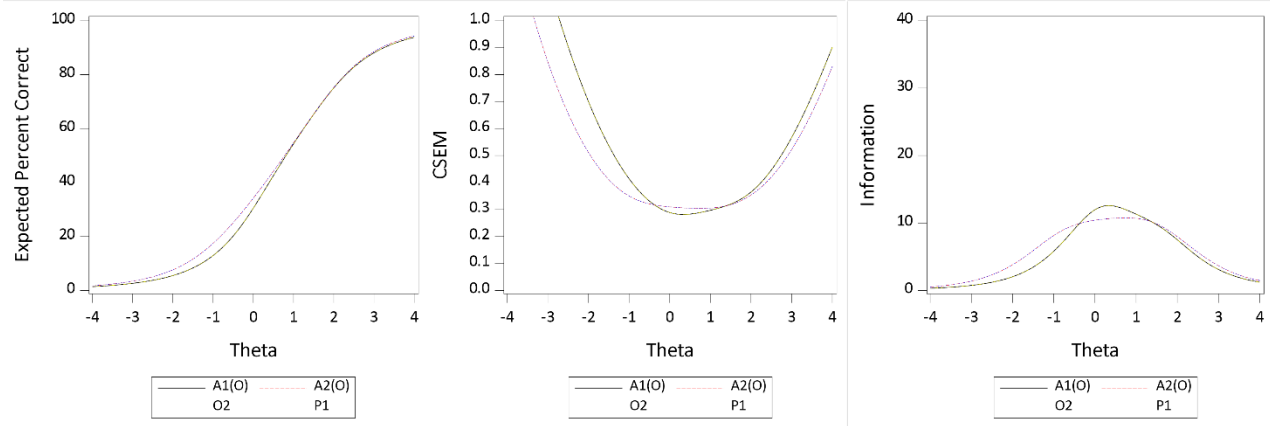


Figure A.7.2.6 Test Characteristic Curves, CSEM Curves, and Information Curves for ELA Grade 9

Math Grade 4

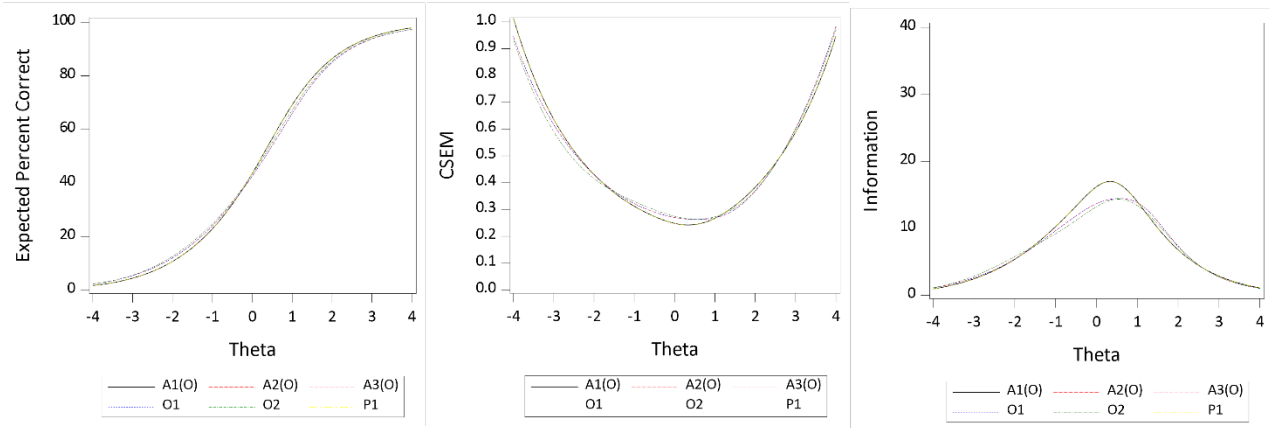


Figure A.7.2.7 Test Characteristic Curves, CSEM Curves, and Information Curves for Math Grade 4

Math Grade 5

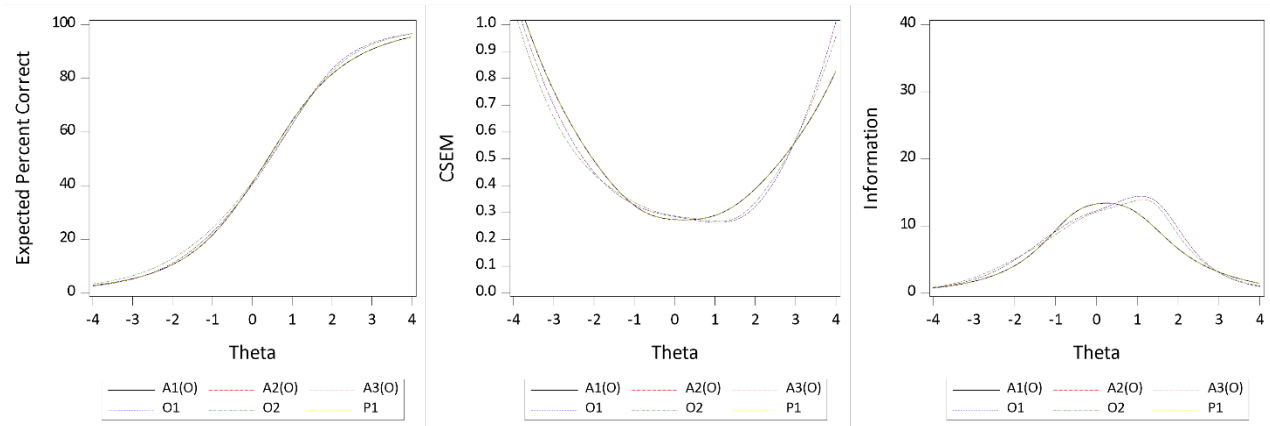


Figure A.7.2.8 Test Characteristic Curves, CSEM Curves, and Information Curves for Math Grade 5

Math Grade 6

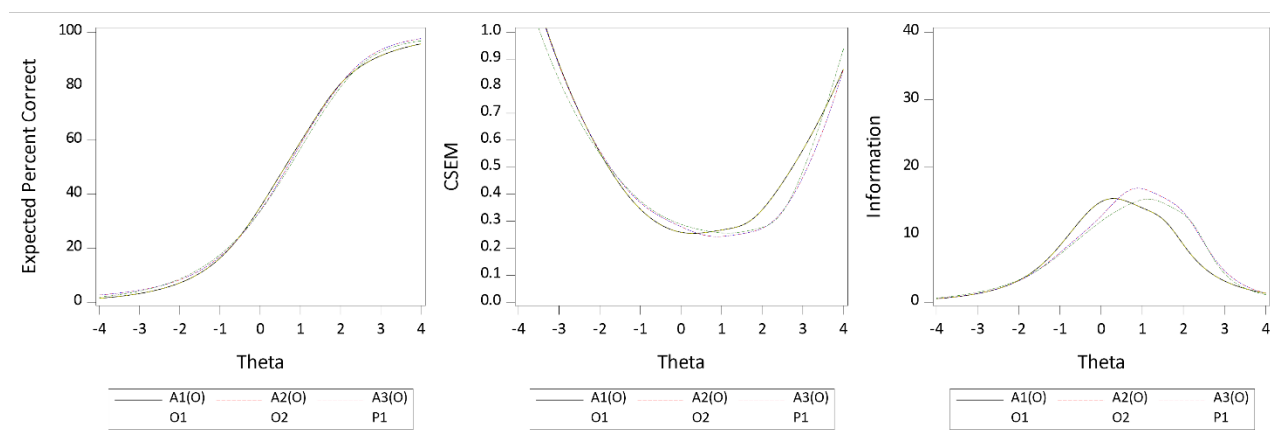


Figure A.7.2.9 Test Characteristic Curves, CSEM Curves, and Information Curves for Math Grade 6

Math Grade 7

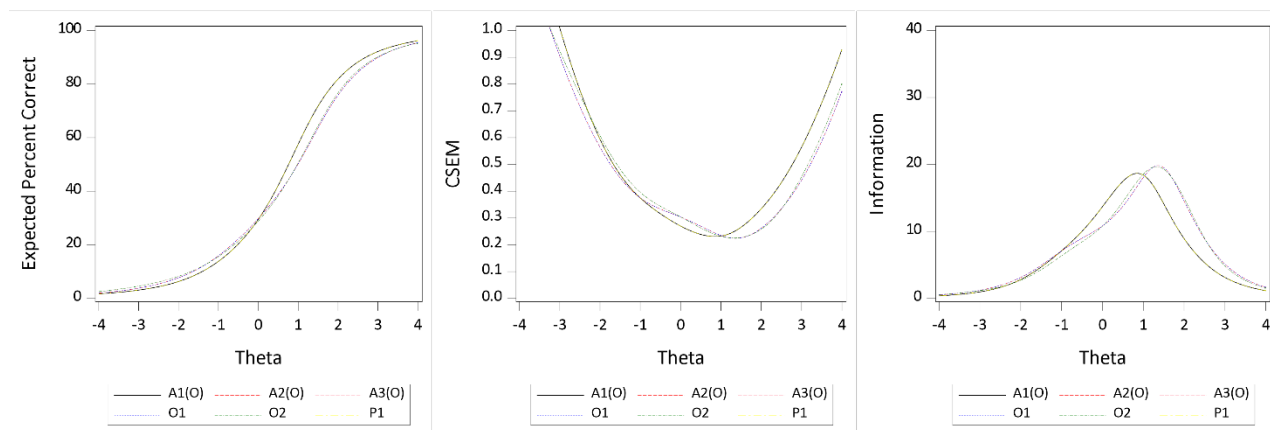


Figure A.7.2.10 Test Characteristic Curves, CSEM Curves, and Information Curves for Math Grade 7

Math Grade 8

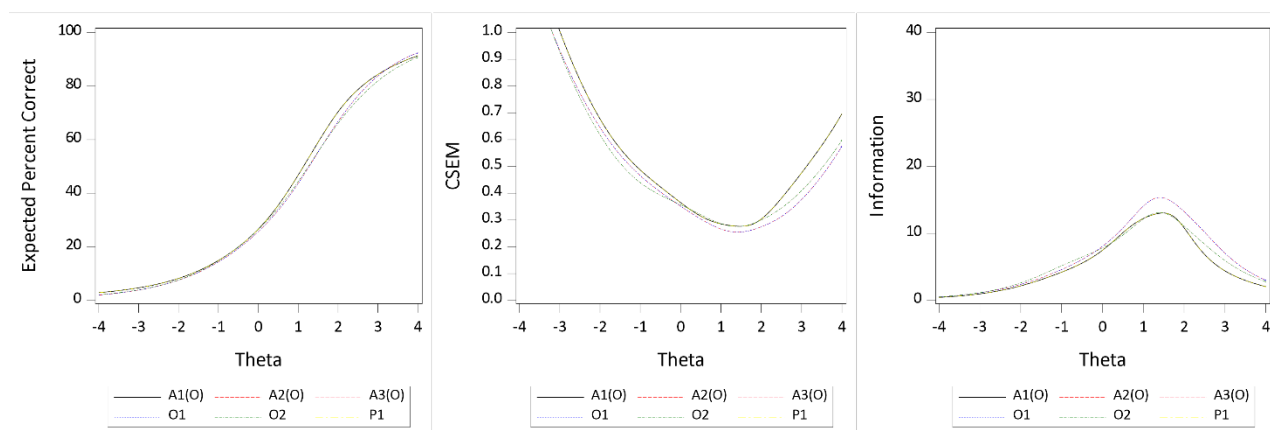


Figure A.7.2.11 Test Characteristic Curves, CSEM Curves, and Information Curves for Math Grade 8

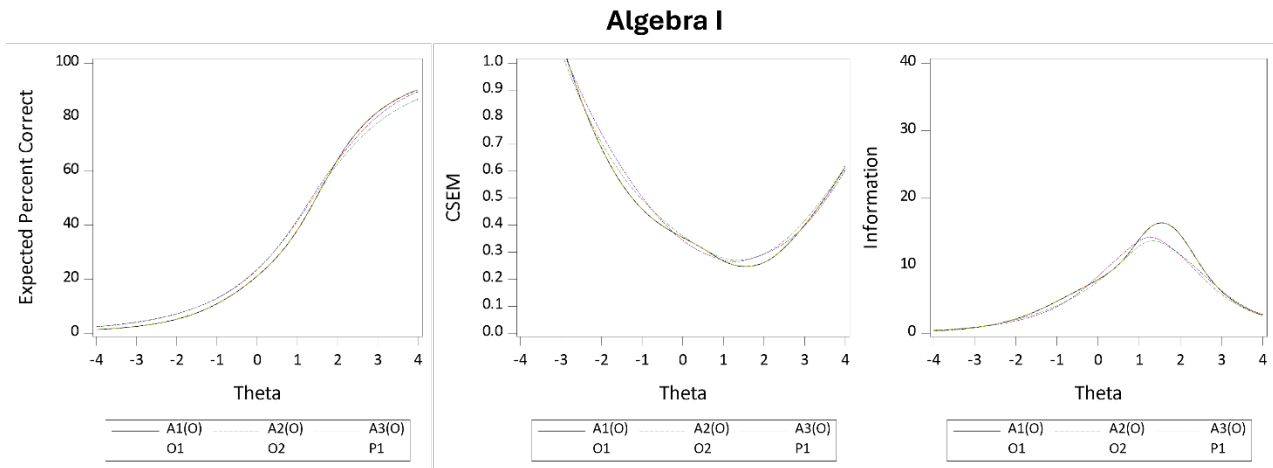


Figure A.7.2.12 Test Characteristic Curves, CSEM Curves, and Information Curves for Algebra I

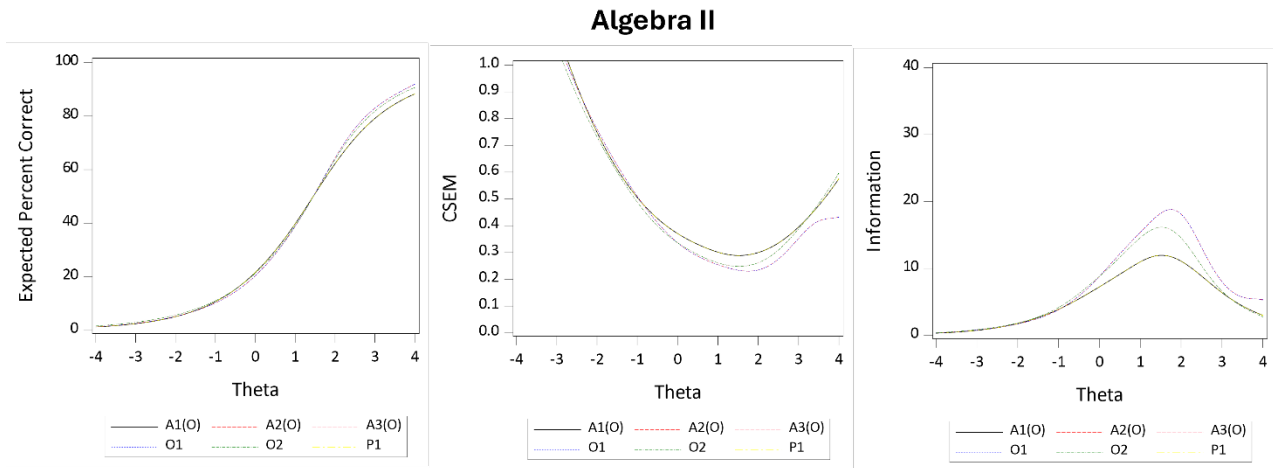


Figure A.7.2.13 Test Characteristic Curves, CSEM Curves, and Information Curves for Algebra II

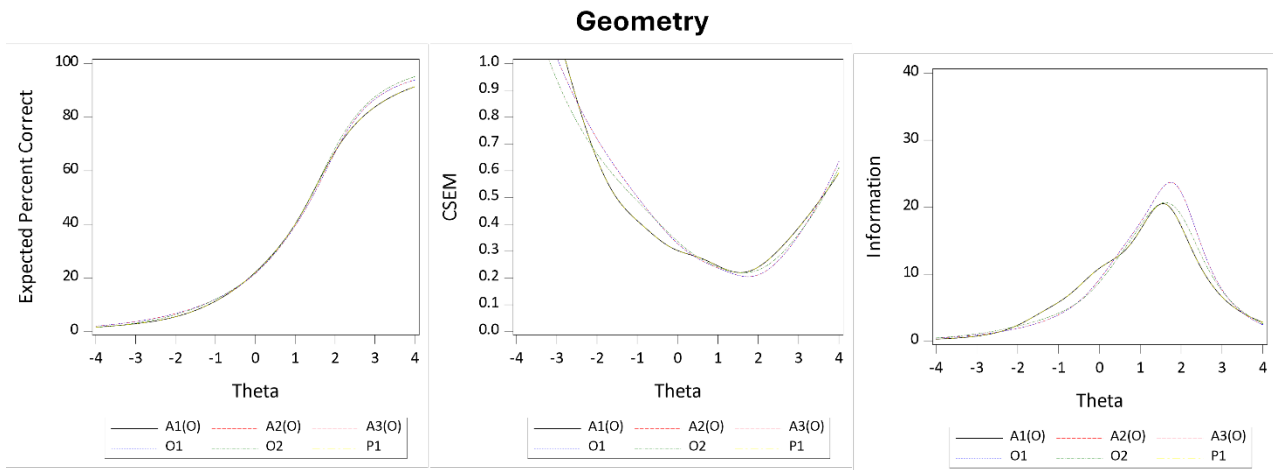


Figure A.7.2.14 Test Characteristic Curves, CSEM Curves, and Information Curves for Geometry

A.7.3. ELA Score Distributions

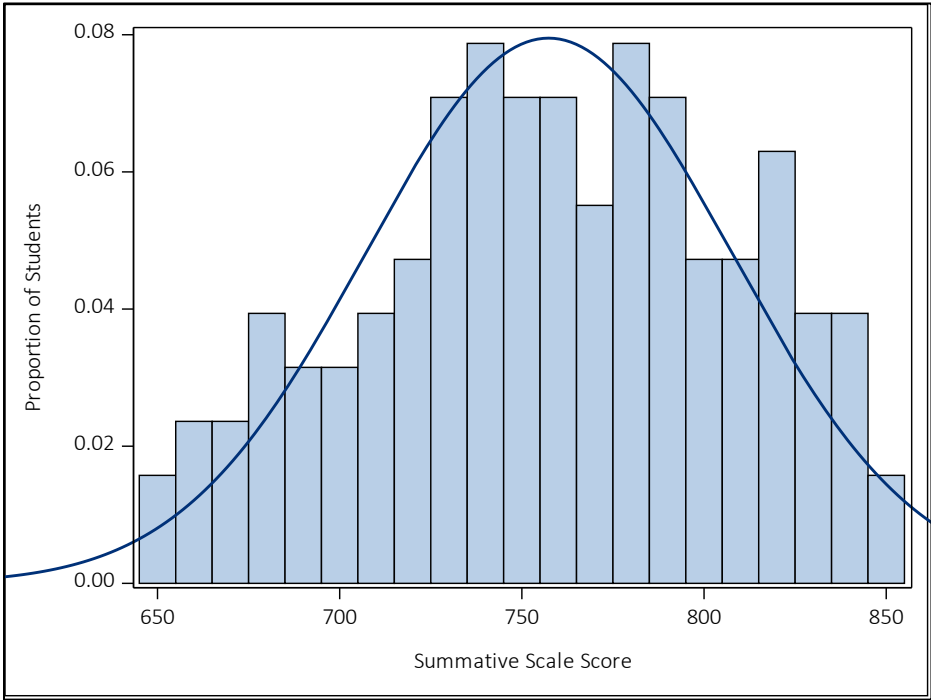


Figure A.7.3.1 Summative Scale Score Distribution for ELA Grade 4

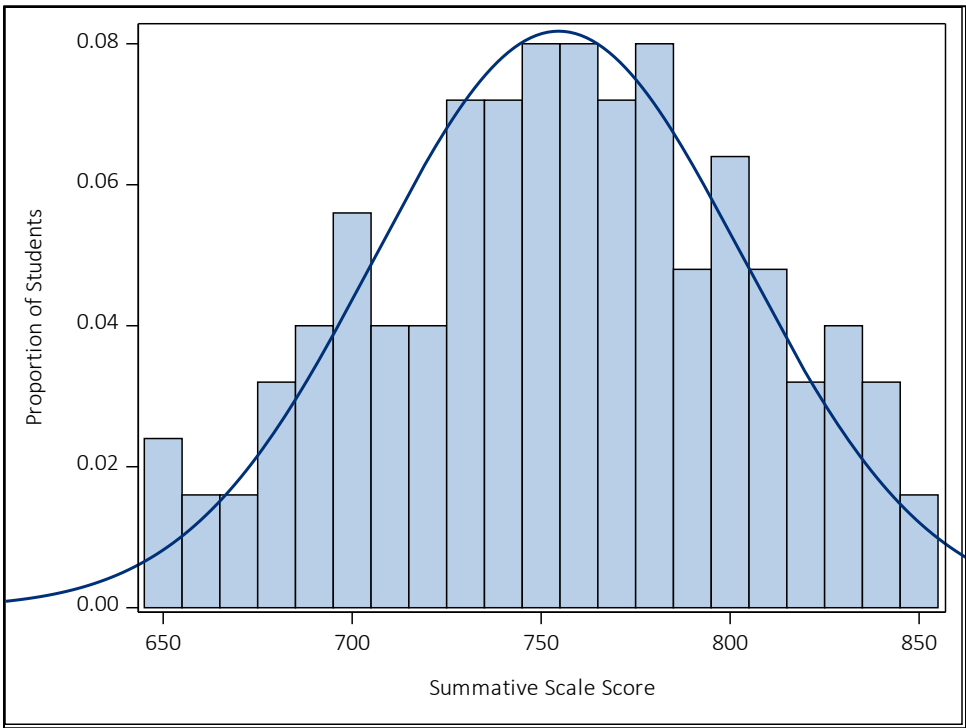


Figure A.7.3.2 Summative Scale Score Distribution for ELA Grade 5

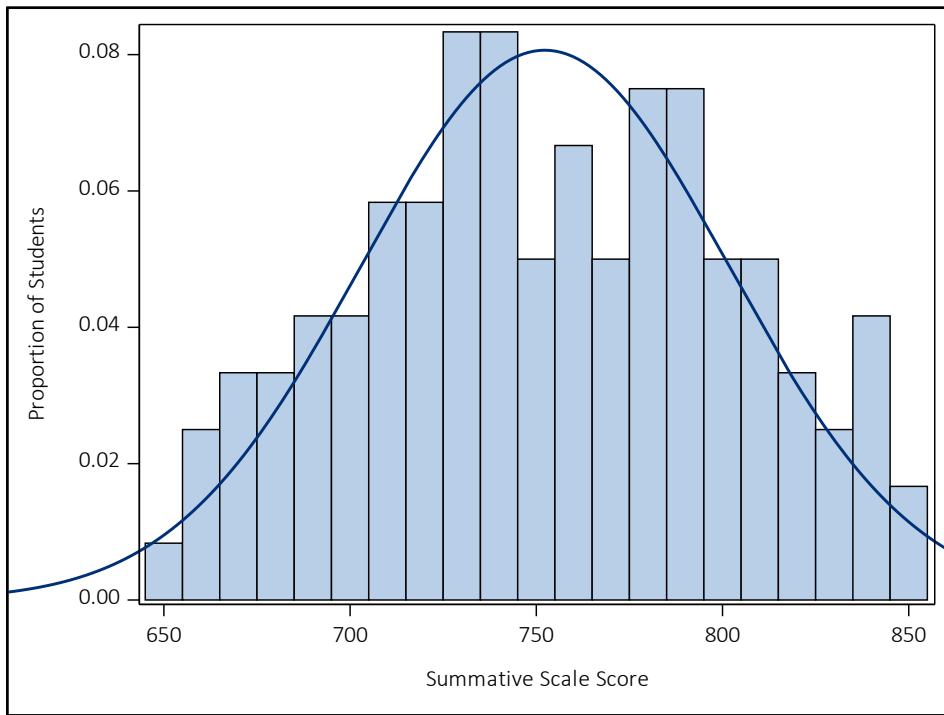


Figure A.7.3.3 Summative Scale Score Distribution for ELA Grade 6

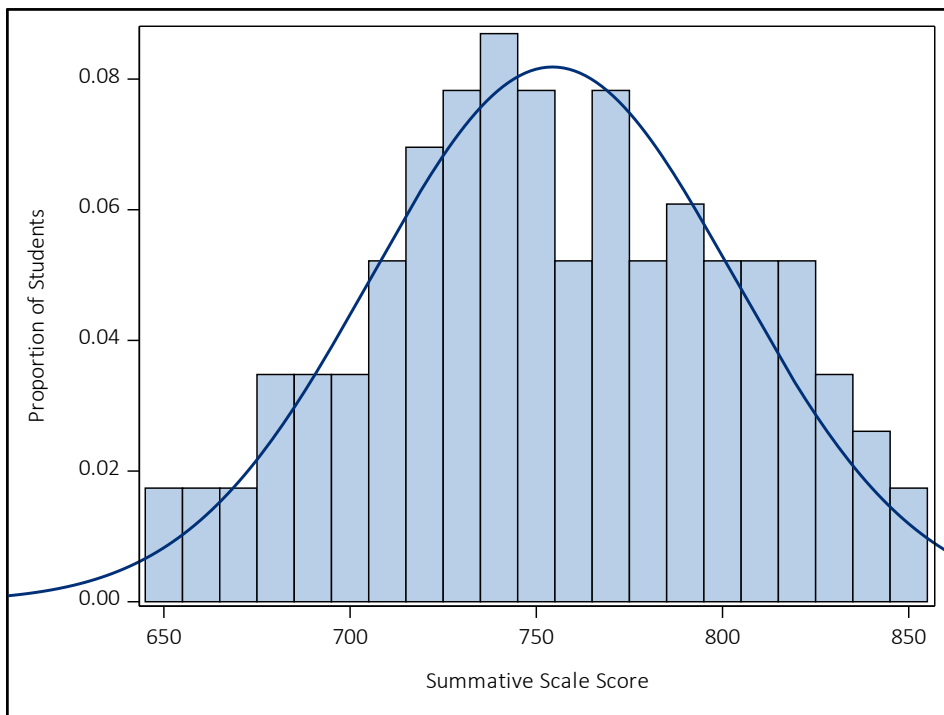


Figure A.7.3.4 Summative Scale Score Distribution for ELA Grade 7

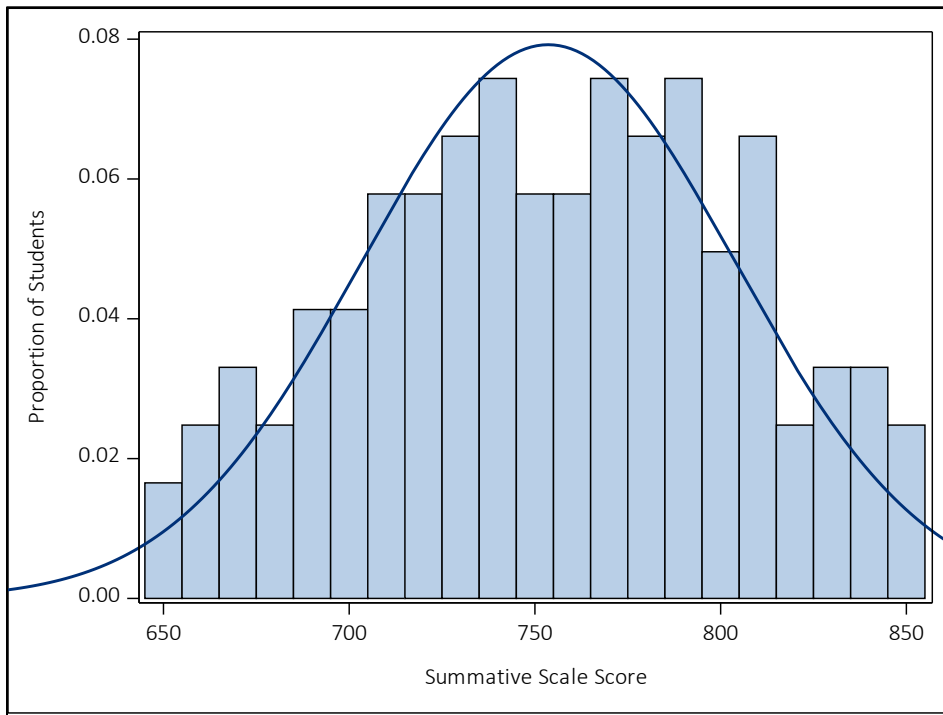


Figure A.7.3.5 Summative Scale Score Distribution for ELA Grade 8

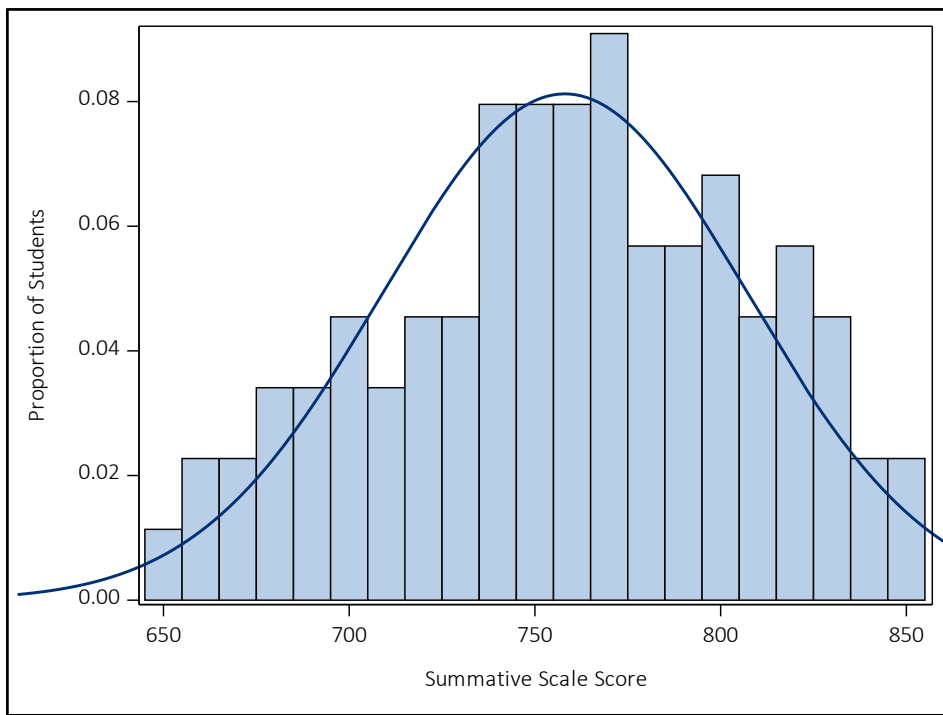


Figure A.7.3.6 Summative Scale Score Distribution for ELA Grade 9

Table A.7.3.1 Scale Score Cumulative Frequencies: ELA Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	697	0.75	697	0.75
655-659	273	0.29	970	1.04
660-664	584	0.63	1,554	1.67
665-669	455	0.49	2,009	2.16
670-674	828	0.89	2,837	3.05
675-679	1,605	1.72	4,442	4.77
680-684	1,826	1.95	6,268	6.72
685-689	1,927	2.06	8,195	8.78
690-694	1,883	2.02	10,078	10.80
695-699	1,801	1.93	11,879	12.73
700-704	1,786	1.91	13,665	14.64
705-709	3,315	3.55	16,980	18.19
710-714	1,661	1.78	18,641	19.97
715-719	3,426	3.67	22,067	23.64
720-724	3,456	3.70	25,523	27.34
725-729	3,610	3.86	29,133	31.20
730-734	3,831	4.10	32,964	35.30
735-739	3,964	4.24	36,928	39.54
740-744	4,209	4.51	41,137	44.05
745-749	4,319	4.62	45,456	48.67
750-754	5,451	5.84	50,907	54.51
755-759	4,412	4.72	55,319	59.23
760-764	4,417	4.73	59,736	63.96
765-769	4,414	4.73	64,150	68.69
770-774	4,258	4.56	68,408	73.25
775-779	4,005	4.29	72,413	77.54
780-784	3,762	4.03	76,175	81.57
785-789	3,434	3.68	79,609	85.25
790-794	2,408	2.58	82,017	87.83
795-799	2,791	2.99	84,808	90.82
800-804	2,270	2.43	87,078	93.25
805-809	913	0.98	87,991	94.23

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	1,517	1.62	89,508	95.85
815-819	856	0.92	90,364	96.77
820-824	1,029	1.10	91,393	97.87
825-829	572	0.61	91,965	98.48
830-834	316	0.34	92,281	98.82
835-839	348	0.37	92,629	99.19
840-844	308	0.33	92,937	99.52
845-849	72	0.08	93,009	99.60
850	403	0.43	93,412	100.00

Table A.7.3.2 Scale Score Cumulative Frequencies: ELA Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	553	0.58	553	0.58
655-659	9	0.01	562	0.59
660-664	676	0.71	1,238	1.30
665-669	12	0.01	1,250	1.31
670-674	1,151	1.21	2,401	2.52
675-679	1,453	1.53	3,854	4.05
680-684	1,815	1.91	5,669	5.96
685-689	1,283	1.35	6,952	7.31
690-694	2,752	2.90	9,704	10.21
695-699	1,895	2.00	11,599	12.21
700-704	1,903	2.01	13,502	14.22
705-709	2,856	3.01	16,358	17.23
710-714	2,419	2.55	18,777	19.78
715-719	3,490	3.68	22,267	23.46
720-724	2,710	2.86	24,977	26.32
725-729	2,563	2.70	27,540	29.02
730-734	3,671	3.87	31,211	32.89
735-739	4,781	5.04	35,992	37.93
740-744	4,061	4.28	40,053	42.21
745-749	4,215	4.44	44,268	46.65
750-754	5,399	5.69	49,667	52.34
755-759	5,709	6.02	55,376	58.36
760-764	4,413	4.65	59,789	63.01
765-769	4,412	4.65	64,201	67.66
770-774	4,354	4.59	68,555	72.25
775-779	4,960	5.23	73,515	77.48
780-784	3,880	4.09	77,395	81.57
785-789	3,504	3.69	80,899	85.26
790-794	1,629	1.72	82,528	86.98
795-799	2,888	3.05	85,416	90.03
800-804	2,421	2.55	87,837	92.58
805-809	1,028	1.08	88,865	93.66

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	1,281	1.35	90,146	95.01
815-819	1,485	1.57	91,631	96.58
820-824	592	0.62	92,223	97.20
825-829	510	0.54	92,733	97.74
830-834	567	0.60	93,300	98.34
835-839	346	0.36	93,646	98.70
840-844	259	0.27	93,905	98.97
845-849	258	0.27	94,163	99.24
850	678	0.71	94,841	100.00

Table A.7.3.3 Scale Score Cumulative Frequencies: ELA Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	274	0.28	274	0.28
655-659	505	0.53	779	0.81
660-664	5	0.01	784	0.82
665-669	587	0.61	1,371	1.43
670-674	1,297	1.35	2,668	2.78
675-679	455	0.47	3,123	3.25
680-684	1,724	1.79	4,847	5.04
685-689	1,244	1.29	6,091	6.33
690-694	1,805	1.88	7,896	8.21
695-699	3,540	3.68	11,436	11.89
700-704	1,651	1.72	13,087	13.61
705-709	2,531	2.63	15,618	16.24
710-714	2,220	2.31	17,838	18.55
715-719	3,254	3.38	21,092	21.93
720-724	4,200	4.37	25,292	26.30
725-729	3,484	3.62	28,776	29.92
730-734	4,557	4.74	33,333	34.66
735-739	4,748	4.94	38,081	39.60
740-744	4,753	4.94	42,834	44.54
745-749	6,099	6.34	48,933	50.88
750-754	4,098	4.26	53,031	55.14
755-759	6,425	6.68	59,456	61.82
760-764	5,178	5.39	64,634	67.21
765-769	4,224	4.39	68,858	71.60
770-774	5,817	6.05	74,675	77.65
775-779	3,524	3.67	78,199	81.32
780-784	3,273	3.40	81,472	84.72
785-789	3,667	3.81	85,139	88.53
790-794	2,484	2.58	87,623	91.11
795-799	1,650	1.72	89,273	92.83
800-804	1,901	1.98	91,174	94.81
805-809	1,115	1.16	92,289	95.97

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	1,344	1.40	93,633	97.37
815-819	520	0.54	94,153	97.91
820-824	655	0.68	94,808	98.59
825-829	386	0.40	95,194	98.99
830-834	127	0.13	95,321	99.12
835-839	271	0.28	95,592	99.40
840-844	237	0.25	95,829	99.65
845-849	55	0.06	95,884	99.71
850	259	0.27	96,143	100.00

Table A.7.3.4 Scale Score Cumulative Frequencies: ELA Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,338	1.38	1,338	1.38
655-659	1,126	1.16	2,464	2.54
660-664	—	—	2,464	2.54
665-669	1,389	1.43	3,853	3.97
670-674	—	—	3,853	3.97
675-679	1,521	1.57	5,374	5.54
680-684	1,530	1.58	6,904	7.12
685-689	1,423	1.47	8,327	8.59
690-694	1,484	1.53	9,811	10.12
695-699	1,371	1.41	11,182	11.53
700-704	2,654	2.74	13,836	14.27
705-709	1,927	1.99	15,763	16.26
710-714	2,207	2.27	17,970	18.53
715-719	2,685	2.77	20,655	21.30
720-724	2,822	2.91	23,477	24.21
725-729	3,744	3.86	27,221	28.07
730-734	3,345	3.45	30,566	31.52
735-739	4,336	4.47	34,902	35.99
740-744	4,304	4.44	39,206	40.43
745-749	3,620	3.73	42,826	44.16
750-754	5,509	5.68	48,335	49.84
755-759	3,792	3.91	52,127	53.75
760-764	5,664	5.84	57,791	59.59
765-769	3,925	4.05	61,716	63.64
770-774	5,530	5.70	67,246	69.34
775-779	3,650	3.76	70,896	73.10
780-784	3,568	3.68	74,464	76.78
785-789	4,224	4.35	78,688	81.13
790-794	3,902	4.02	82,590	85.15
795-799	2,167	2.23	84,757	87.38
800-804	2,588	2.67	87,345	90.05
805-809	1,785	1.84	89,130	91.89

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	1,611	1.66	90,741	93.55
815-819	1,756	1.81	92,497	95.36
820-824	769	0.79	93,266	96.15
825-829	709	0.73	93,975	96.88
830-834	963	0.99	94,938	97.87
835-839	275	0.28	95,213	98.15
840-844	496	0.51	95,709	98.66
845-849	411	0.42	96,120	99.08
850	897	0.92	97,017	100.00

Table A.7.3.5 Scale Score Cumulative Frequencies: ELA Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,160	1.16	1,160	1.16
655-659	613	0.61	1,773	1.77
660-664	388	0.39	2,161	2.16
665-669	860	0.86	3,021	3.02
670-674	536	0.54	3,557	3.56
675-679	1,594	1.59	5,151	5.15
680-684	1,024	1.02	6,175	6.17
685-689	1,757	1.75	7,932	7.92
690-694	1,675	1.67	9,607	9.59
695-699	3,132	3.13	12,739	12.72
700-704	1,541	1.54	14,280	14.26
705-709	2,843	2.84	17,123	17.10
710-714	1,384	1.38	18,507	18.48
715-719	2,923	2.92	21,430	21.40
720-724	2,943	2.94	24,373	24.34
725-729	3,043	3.04	27,416	27.38
730-734	4,062	4.06	31,478	31.44
735-739	4,179	4.17	35,657	35.61
740-744	3,489	3.48	39,146	39.09
745-749	5,414	5.41	44,560	44.50
750-754	3,743	3.74	48,303	48.24
755-759	3,745	3.74	52,048	51.98
760-764	5,869	5.86	57,917	57.84
765-769	4,009	4.00	61,926	61.84
770-774	4,073	4.07	65,999	65.91
775-779	4,057	4.05	70,056	69.96
780-784	3,999	3.99	74,055	73.95
785-789	3,788	3.78	77,843	77.73
790-794	3,639	3.63	81,482	81.36
795-799	2,626	2.62	84,108	83.98
800-804	3,186	3.18	87,294	87.16
805-809	2,249	2.25	89,543	89.41

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	2,120	2.12	91,663	91.53
815-819	1,972	1.97	93,635	93.50
820-824	1,179	1.18	94,814	94.68
825-829	1,019	1.02	95,833	95.70
830-834	941	0.94	96,774	96.64
835-839	1,141	1.14	97,915	97.78
840-844	241	0.24	98,156	98.02
845-849	545	0.54	98,701	98.56
850	1,414	1.41	100,115	100.00

Table A.7.3.6 Scale Score Cumulative Frequencies: ELA Grade 9

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	904	0.93	904	0.93
655-659	973	1.00	1,877	1.93
660-664	—	—	1,877	1.93
665-669	1,521	1.57	3,398	3.50
670-674	23	0.02	3,421	3.52
675-679	2,062	2.12	5,483	5.64
680-684	2,240	2.31	7,723	7.95
685-689	2,210	2.27	9,933	10.22
690-694	2,065	2.13	11,998	12.35
695-699	2,002	2.06	14,000	14.41
700-704	1,899	1.95	15,899	16.36
705-709	3,568	3.67	19,467	20.03
710-714	1,781	1.83	21,248	21.86
715-719	3,469	3.57	24,717	25.43
720-724	3,709	3.82	28,426	29.25
725-729	3,693	3.80	32,119	33.05
730-734	1,924	1.98	34,043	35.03
735-739	4,062	4.18	38,105	39.21
740-744	3,975	4.09	42,080	43.30
745-749	4,172	4.29	46,252	47.59
750-754	4,460	4.59	50,712	52.18
755-759	6,640	6.83	57,352	59.01
760-764	4,303	4.43	61,655	63.44
765-769	4,187	4.31	65,842	67.75
770-774	4,063	4.18	69,905	71.93
775-779	3,873	3.99	73,778	75.92
780-784	3,603	3.71	77,381	79.63
785-789	4,846	4.99	82,227	84.62
790-794	2,907	2.99	85,134	87.61
795-799	2,402	2.47	87,536	90.08
800-804	2,202	2.27	89,738	92.35
805-809	1,894	1.95	91,632	94.30

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
810-814	824	0.85	92,456	95.15
815-819	1,457	1.50	93,913	96.65
820-824	624	0.64	94,537	97.29
825-829	1,062	1.09	95,599	98.38
830-834	402	0.41	96,001	98.79
835-839	348	0.36	96,349	99.15
840-844	284	0.29	96,633	99.44
845-849	194	0.20	96,827	99.64
850	347	0.36	97,174	100.00

A.7.4. Math Score Distributions

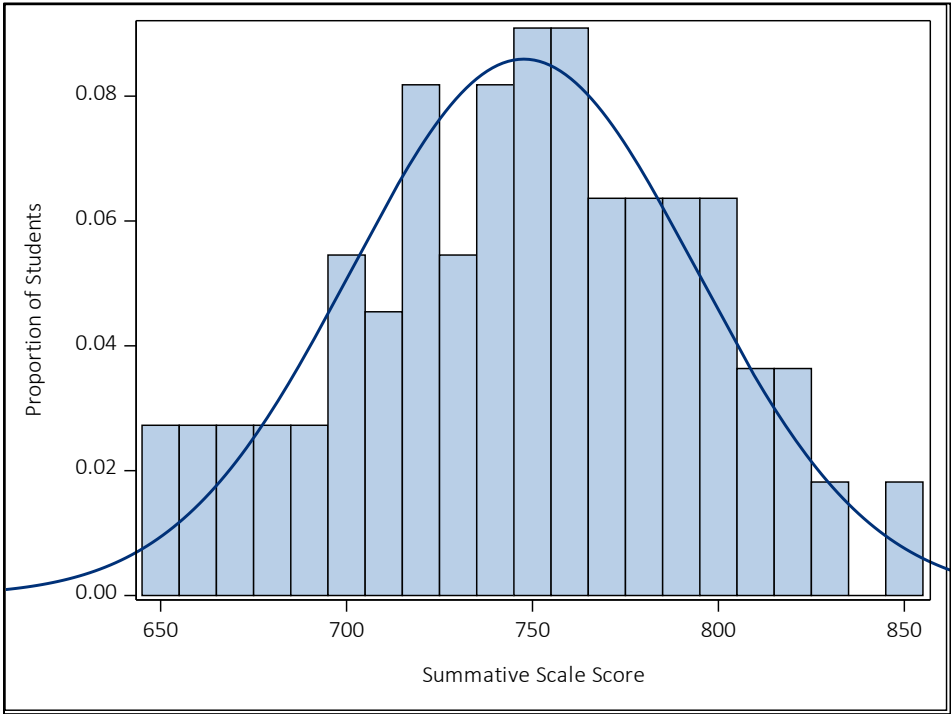


Figure A.7.4.1 Summative Scale Score Distribution for Mathematics Grade 4

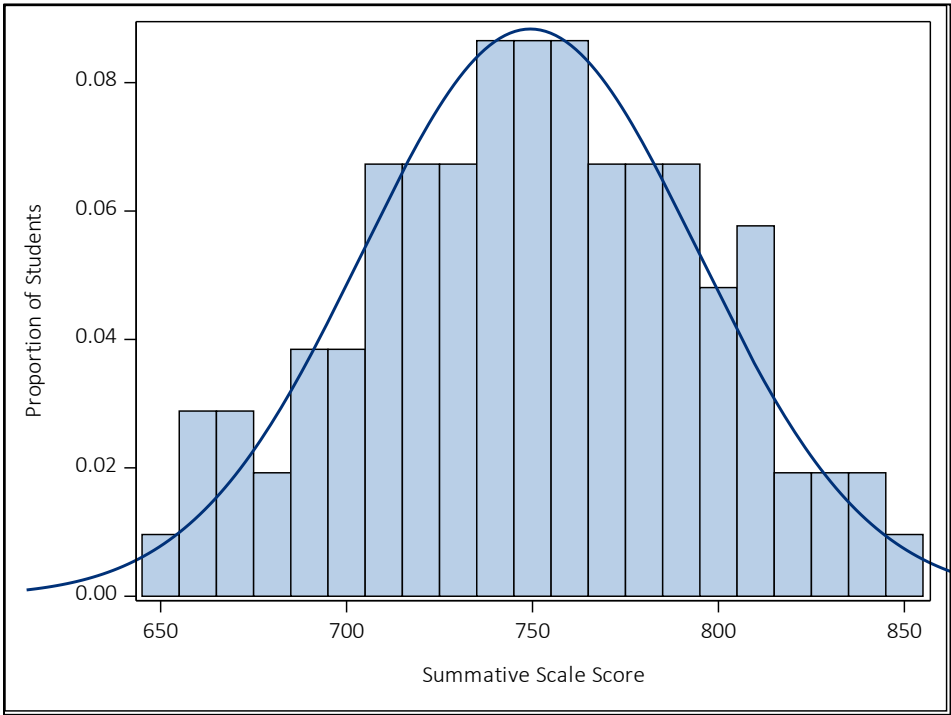


Figure A.7.4.2 Summative Scale Score Distribution for Mathematics Grade 5

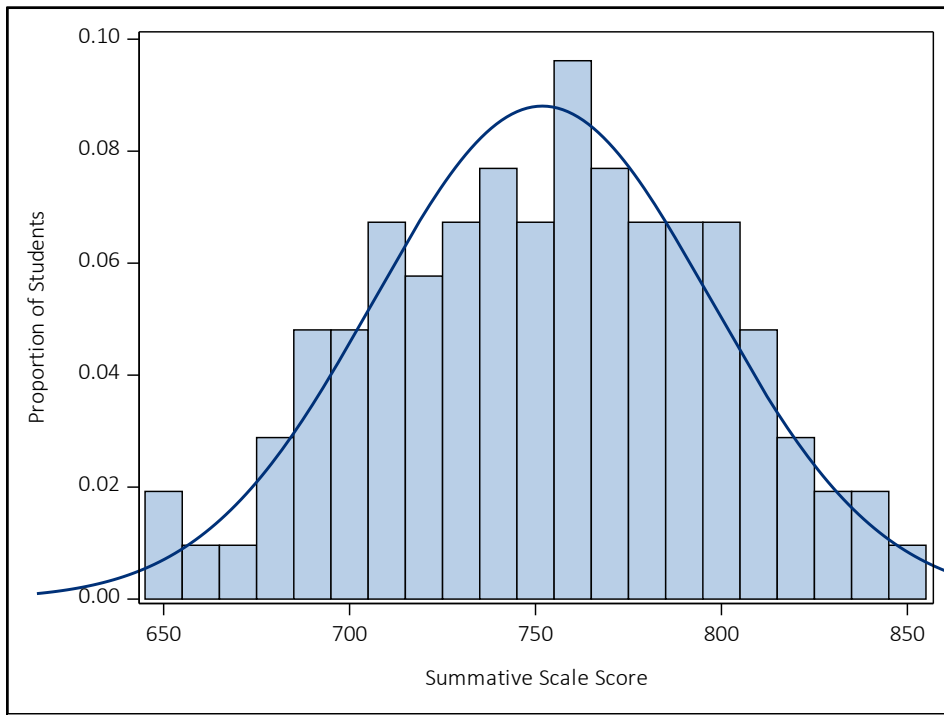


Figure A.7.4.3 Summative Scale Score Distribution for Mathematics Grade 6

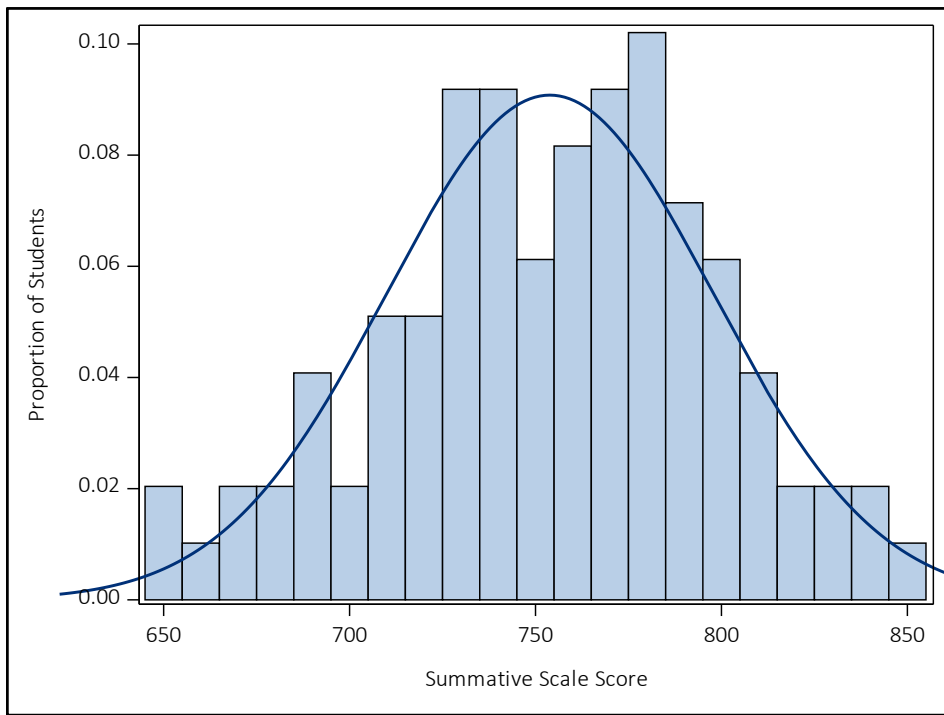


Figure A.7.4.4 Summative Scale Score Distribution for Mathematics Grade 7

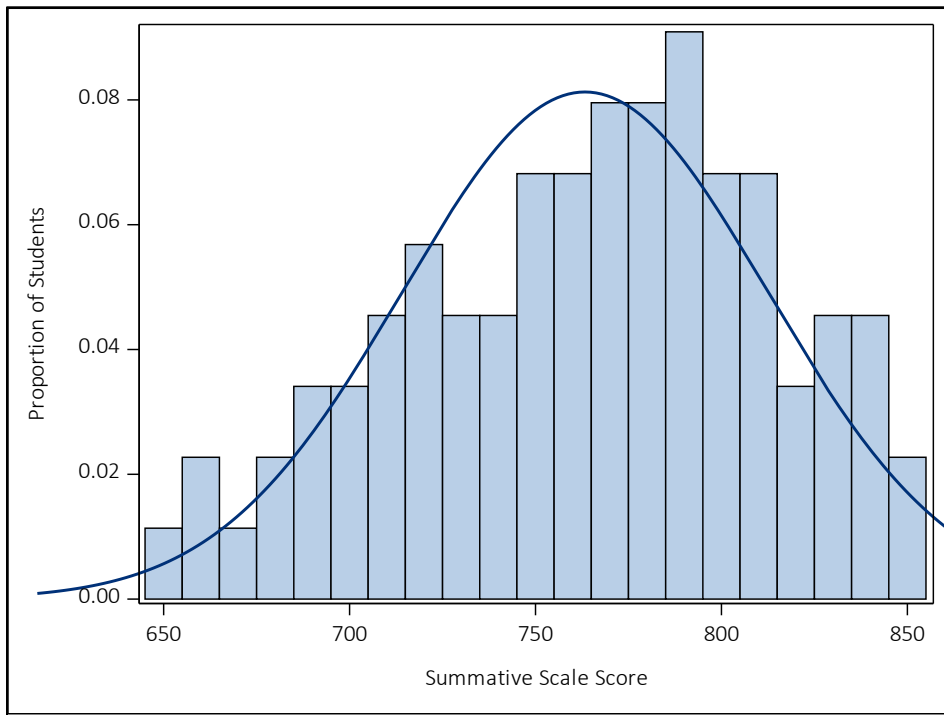


Figure A.7.4.5 Summative Scale Score Distribution for Mathematics Grade 8

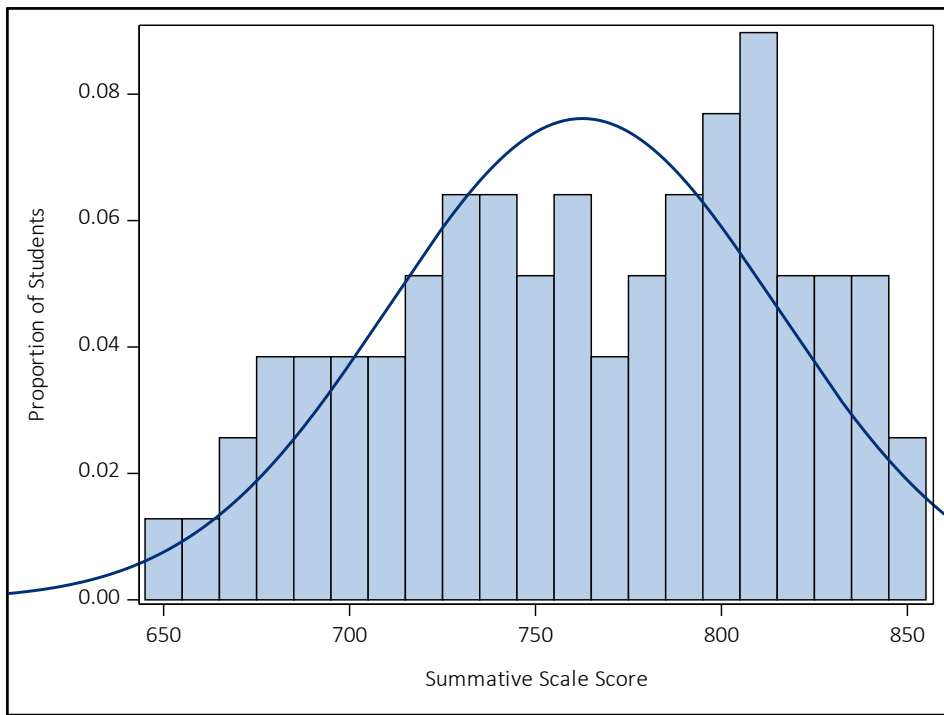


Figure A.7.4.6 Summative Scale Score Distribution for Algebra I

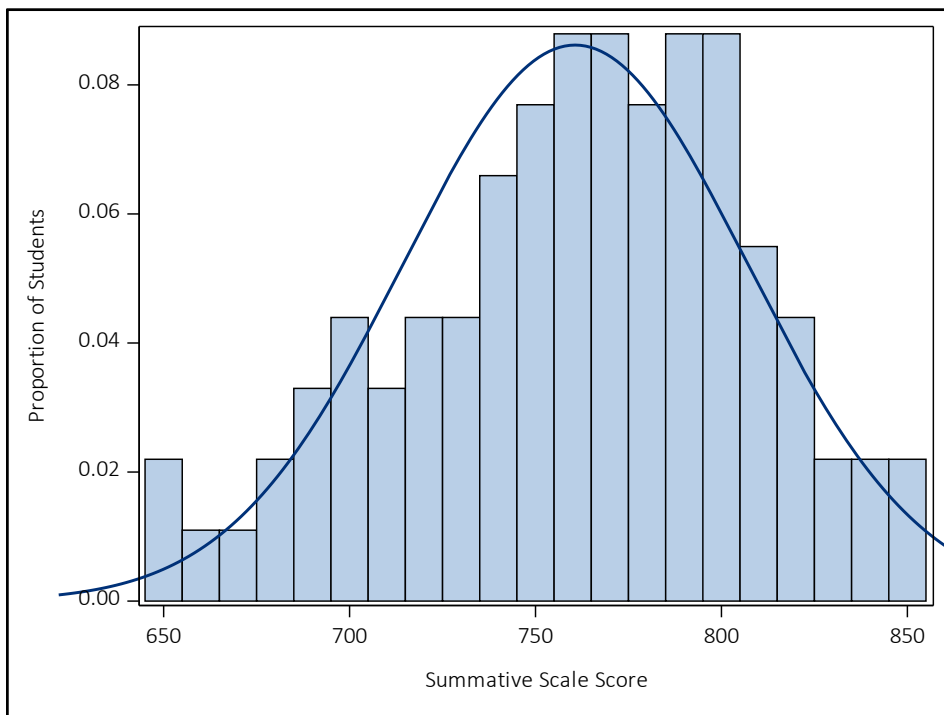


Figure A.7.4.7 Summative Scale Score Distribution for Algebra II

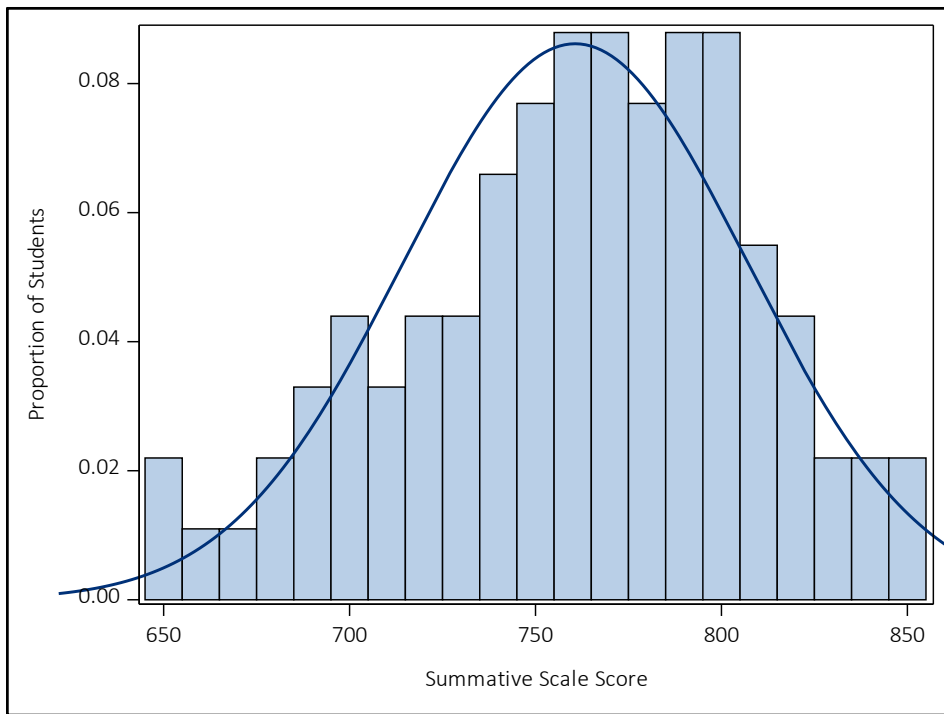


Figure A.7.4.8 Summative Scale Score Distribution for Geometry

Table A.7.4.1 Scale Score Cumulative Frequencies: Mathematics Grade 4

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1,278	1.35	1,278	1.35
655-659	—	—	1,278	1.35
660-664	1,077	1.13	2,355	2.48
665-669	—	—	2,355	2.48
670-674	1,331	1.40	3,686	3.88
675-679	651	0.69	4,337	4.57
680-684	1,696	1.79	6,033	6.36
685-689	990	1.04	7,023	7.40
690-694	2,042	2.15	9,065	9.55
695-699	3,302	3.48	12,367	13.03
700-704	2,541	2.67	14,908	15.70
705-709	3,760	3.96	18,668	19.66
710-714	2,652	2.79	21,320	22.45
715-719	3,889	4.09	25,209	26.54
720-724	4,062	4.28	29,271	30.82
725-729	3,797	4.00	33,068	34.82
730-734	4,105	4.32	37,173	39.14
735-739	5,224	5.50	42,397	44.64
740-744	5,211	5.49	47,608	50.13
745-749	5,179	5.45	52,787	55.58
750-754	4,916	5.17	57,703	60.75
755-759	5,809	6.11	63,512	66.86
760-764	4,468	4.70	67,980	71.56
765-769	4,216	4.44	72,196	76.00
770-774	4,002	4.21	76,198	80.21
775-779	3,887	4.09	80,085	84.30
780-784	2,561	2.70	82,646	87.00
785-789	2,614	2.75	85,260	89.75
790-794	2,164	2.28	87,424	92.03

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	2,146	2.26	89,570	94.29
800-804	1,214	1.28	90,784	95.57
805-809	1,129	1.19	91,913	96.76
810-814	464	0.49	92,377	97.25
815-819	849	0.89	93,226	98.14
820-824	689	0.73	93,915	98.87
825-829	—	—	93,915	98.87
830-834	501	0.53	94,416	99.40
835-839	—	—	94,416	99.40
840-844	—	—	94,416	99.40
845-849	338	0.36	94,754	99.76
850	246	0.26	95,000	100.00

Table A.7.4.2 Scale Score Cumulative Frequencies: Mathematics Grade 5

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	781	0.81	781	0.81
655-659	489	0.51	1,270	1.32
660-664	423	0.44	1,693	1.76
665-669	—	—	1,693	1.76
670-674	1,444	1.50	3,137	3.26
675-679	1,836	1.91	4,973	5.17
680-684	11	0.01	4,984	5.18
685-689	2,239	2.32	7,223	7.50
690-694	2,460	2.55	9,683	10.05
695-699	2,846	2.95	12,529	13.00
700-704	2,944	3.05	15,473	16.05
705-709	3,120	3.24	18,593	19.29
710-714	4,859	5.04	23,452	24.33
715-719	4,874	5.06	28,326	29.39
720-724	4,815	5.00	33,141	34.39
725-729	4,782	4.96	37,923	39.35
730-734	4,535	4.71	42,458	44.06
735-739	4,474	4.64	46,932	48.70
740-744	5,510	5.72	52,442	54.42
745-749	5,292	5.49	57,734	59.91
750-754	4,978	5.17	62,712	65.08
755-759	4,678	4.85	67,390	69.93
760-764	3,260	3.38	70,650	73.31
765-769	4,173	4.33	74,823	77.64
770-774	4,650	4.83	79,473	82.47
775-779	3,352	3.48	82,825	85.95
780-784	3,051	3.17	85,876	89.12
785-789	2,116	2.20	87,992	91.32
790-794	2,480	2.57	90,472	93.89

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	1,523	1.58	91,995	95.47
800-804	914	0.95	92,909	96.42
805-809	1,192	1.24	94,101	97.66
810-814	623	0.65	94,724	98.31
815-819	293	0.30	95,017	98.61
820-824	247	0.26	95,264	98.87
825-829	410	0.43	95,674	99.30
830-834	—	—	95,674	99.30
835-839	299	0.31	95,973	99.61
840-844	—	—	95,973	99.61
845-849	—	—	95,973	99.61
850	399	0.41	96,372	100.00

Table A.7.4.3 Scale Score Cumulative Frequencies: Mathematics Grade 6

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	1484	1.52	1484	1.52
655-659	—	—	1484	1.52
660-664	2	0.00	1486	1.52
665-669	2020	2.07	3506	3.59
670-674	—	—	3506	3.59
675-679	2816	2.88	6322	6.47
680-684	—	—	6322	6.47
685-689	3497	3.58	9819	10.05
690-694	1948	1.99	11767	12.04
695-699	1942	1.99	13709	14.03
700-704	3980	4.08	17689	18.11
705-709	3916	4.01	21605	22.12
710-714	5702	5.84	27307	27.96
715-719	5615	5.75	32922	33.71
720-724	3452	3.53	36374	37.24
725-729	6813	6.98	43187	44.22
730-734	3263	3.34	46450	47.56
735-739	6053	6.20	52503	53.76
740-744	5523	5.66	58026	59.42
745-749	6019	6.16	64045	65.58
750-754	4434	4.54	68479	70.12
755-759	4883	5.00	73362	75.12
760-764	4393	4.50	77755	79.62
765-769	3262	3.34	81017	82.96
770-774	2905	2.97	83922	85.93
775-779	3359	3.44	87281	89.37
780-784	2849	2.92	90130	92.29
785-789	1495	1.53	91625	93.82
790-794	1800	1.84	93425	95.66

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	1143	1.17	94568	96.83
800-804	979	1.00	95547	97.83
805-809	768	0.79	96315	98.62
810-814	419	0.43	96734	99.05
815-819	327	0.33	97061	99.38
820-824	130	0.13	97191	99.51
825-829	116	0.12	97307	99.63
830-834	92	0.09	97399	99.72
835-839	81	0.08	97480	99.80
840-844	59	0.06	97539	99.86
845-849	—	—	97539	99.86
850	114	0.12	97653	100.00

Table A.7.4.4 Scale Score Cumulative Frequencies: Mathematics Grade 7

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	912	0.98	912	0.98
655-659	—	—	912	0.98
660-664	455	0.49	1,367	1.47
665-669	7	0.01	1,374	1.48
670-674	978	1.05	2,352	2.53
675-679	1,034	1.11	3,386	3.64
680-684	1,402	1.50	4,788	5.14
685-689	1,599	1.71	6,387	6.85
690-694	3,567	3.82	9,954	10.67
695-699	1,801	1.93	11,755	12.60
700-704	2,129	2.28	13,884	14.88
705-709	3,999	4.28	17,883	19.16
710-714	5,880	6.29	23,763	25.45
715-719	5,521	5.91	29,284	31.36
720-724	3,717	3.98	33,001	35.34
725-729	5,612	6.00	38,613	41.34
730-734	5,260	5.63	43,873	46.97
735-739	6,631	7.09	50,504	54.06
740-744	5,982	6.40	56,486	60.46
745-749	5,295	5.67	61,781	66.13
750-754	4,612	4.93	66,393	71.06
755-759	5,883	6.29	72,276	77.35
760-764	4,167	4.46	76,443	81.81
765-769	3,358	3.59	79,801	85.40
770-774	3,632	3.89	83,433	89.29
775-779	3,007	3.22	86,440	92.51
780-784	2,176	2.33	88,616	94.84
785-789	1,401	1.50	90,017	96.34
790-794	912	0.98	90,929	97.32

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	986	1.05	91,915	98.37
800-804	396	0.42	92,311	98.79
805-809	312	0.33	92,623	99.12
810-814	281	0.30	92,904	99.42
815-819	142	0.15	93,046	99.57
820-824	84	0.09	93,130	99.66
825-829	155	0.17	93,285	99.83
830-834	—	—	93,285	99.83
835-839	—	—	93,285	99.83
840-844	111	0.12	93,396	99.95
845-849	—	—	93,396	99.95
850	67	0.07	93,463	100.00

Table A.7.4.5 Scale Score Cumulative Frequencies: Mathematics Grade 8

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,556	3.83	2,556	3.83
655-659	12	0.02	2,568	3.85
660-664	3,022	4.53	5,590	8.38
665-669	—	—	5,590	8.38
670-674	20	0.03	5,610	8.41
675-679	4,319	6.48	9,929	14.89
680-684	2,189	3.28	12,118	18.17
685-689	2,753	4.13	14,871	22.30
690-694	5,114	7.67	19,985	29.97
695-699	2,417	3.62	22,402	33.59
700-704	2,487	3.73	24,889	37.32
705-709	4,467	6.70	29,356	44.02
710-714	4,069	6.10	33,425	50.12
715-719	3,709	5.56	37,134	55.68
720-724	3,285	4.93	40,419	60.61
725-729	3,008	4.51	43,427	65.12
730-734	2,653	3.98	46,080	69.10
735-739	3,366	5.05	49,446	74.15
740-744	2,879	4.32	52,325	78.47
745-749	2,435	3.65	54,760	82.12
750-754	2,188	3.28	56,948	85.40
755-759	1,718	2.58	58,666	87.98
760-764	1,507	2.26	60,173	90.24
765-769	1,640	2.46	61,813	92.70
770-774	1,272	1.91	63,085	94.61
775-779	977	1.46	64,062	96.07
780-784	603	0.90	64,665	96.97
785-789	485	0.73	65,150	97.70
790-794	503	0.75	65,653	98.45

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	309	0.46	65,962	98.91
800-804	218	0.33	66,180	99.24
805-809	172	0.26	66,352	99.50
810-814	96	0.14	66,448	99.64
815-819	64	0.10	66,512	99.74
820-824	51	0.08	66,563	99.82
825-829	43	0.06	66,606	99.88
830-834	33	0.05	66,639	99.93
835-839	11	0.02	66,650	99.95
840-844	13	0.02	66,663	99.97
845-849	—	—	66,663	99.97
850	30	0.04	66,693	100.00

Table A.7.4.6 Scale Score Cumulative Frequencies: Algebra I

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	678	0.64	678	0.64
655-659	1,461	1.38	2,139	2.02
660-664	—	—	2,139	2.02
665-669	—	—	2,139	2.02
670-674	1,568	1.49	3,707	3.51
675-679	1,396	1.32	5,103	4.83
680-684	—	—	5,103	4.83
685-689	4,763	4.51	9,866	9.34
690-694	—	—	9,866	9.34
695-699	6,360	6.03	16,226	15.37
700-704	3,249	3.08	19,475	18.45
705-709	3,893	3.69	23,368	22.14
710-714	7,326	6.94	30,694	29.08
715-719	6,859	6.50	37,553	35.58
720-724	6,197	5.87	43,750	41.45
725-729	5,332	5.05	49,082	46.50
730-734	4,715	4.47	53,797	50.97
735-739	4,277	4.05	58,074	55.02
740-744	3,797	3.60	61,871	58.62
745-749	6,458	6.12	68,329	64.74
750-754	2,919	2.77	71,248	67.51
755-759	5,149	4.88	76,397	72.39
760-764	4,364	4.14	80,761	76.53
765-769	3,885	3.68	84,646	80.21
770-774	3,391	3.21	88,037	83.42
775-779	2,256	2.14	90,293	85.56
780-784	2,719	2.58	93,012	88.14
785-789	2,335	2.21	95,347	90.35
790-794	2,056	1.95	97,403	92.30
795-799	1,310	1.24	98,713	93.54

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
800-804	1,175	1.11	99,888	94.65
805-809	1,354	1.28	101,242	95.93
810-814	559	0.53	101,801	96.46
815-819	1,001	0.95	102,802	97.41
820-824	457	0.43	103,259	97.84
825-829	387	0.37	103,646	98.21
830-834	328	0.31	103,974	98.52
835-839	289	0.27	104,263	98.79
840-844	265	0.25	104,528	99.04
845-849	141	0.13	104,669	99.17
850	866	0.82	105,535	100.00

Table A.7.4.7 Scale Score Cumulative Frequencies: Algebra II

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	217	2.15	217	2.15
655-659	—	—	217	2.15
660-664	98	0.97	315	3.12
665-669	139	1.38	454	4.50
670-674	133	1.32	587	5.82
675-679	206	2.04	793	7.86
680-684	156	1.55	949	9.41
685-689	225	2.23	1,174	11.64
690-694	168	1.67	1,342	13.31
695-699	245	2.43	1,587	15.74
700-704	372	3.69	1,959	19.43
705-709	136	1.35	2,095	20.78
710-714	342	3.39	2,437	24.17
715-719	286	2.83	2,723	27.00
720-724	276	2.74	2,999	29.74
725-729	243	2.41	3,242	32.15
730-734	361	3.58	3,603	35.73
735-739	256	2.54	3,859	38.27
740-744	246	2.44	4,105	40.71
745-749	518	5.13	4,623	45.84
750-754	264	2.62	4,887	48.46
755-759	532	5.27	5,419	53.73
760-764	502	4.98	5,921	58.71
765-769	240	2.38	6,161	61.09
770-774	471	4.67	6,632	65.76
775-779	507	5.03	7,139	70.79
780-784	486	4.82	7,625	75.61
785-789	419	4.15	8,044	79.76
790-794	393	3.90	8,437	83.66

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	287	2.84	8,724	86.50
800-804	242	2.40	8,966	88.90
805-809	283	2.81	9,249	91.71
810-814	208	2.06	9,457	93.77
815-819	87	0.86	9,544	94.63
820-824	105	1.04	9,649	95.67
825-829	84	0.83	9,733	96.50
830-834	86	0.85	9,819	97.35
835-839	72	0.71	9,891	98.06
840-844	50	0.50	9,941	98.56
845-849	15	0.15	9,956	98.71
850	133	1.32	10,089	100.00

Table A.7.4.8 Scale Score Cumulative Frequencies: Geometry

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	94	0.31	94	0.31
655-659	—	—	94	0.31
660-664	40	0.13	134	0.44
665-669	—	—	134	0.44
670-674	154	0.51	288	0.95
675-679	97	0.32	385	1.27
680-684	—	—	385	1.27
685-689	450	1.49	835	2.76
690-694	—	—	835	2.76
695-699	788	2.61	1623	5.37
700-704	923	3.05	2546	8.42
705-709	1	0.00	2547	8.42
710-714	1089	3.60	3636	12.02
715-719	1109	3.67	4745	15.69
720-724	1200	3.97	5945	19.66
725-729	1191	3.94	7136	23.60
730-734	2251	7.44	9387	31.04
735-739	1131	3.74	10518	34.78
740-744	2254	7.45	12772	42.23
745-749	2133	7.05	14905	49.28
750-754	2046	6.77	16951	56.05
755-759	1931	6.39	18882	62.44
760-764	2549	8.43	21431	70.87
765-769	1745	5.77	23176	76.64
770-774	1875	6.20	25051	82.84
775-779	1561	5.16	26612	88.00
780-784	1267	4.19	27879	92.19
785-789	799	2.64	28678	94.83
790-794	674	2.23	29352	97.06

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
795-799	291	0.96	29643	98.02
800-804	238	0.79	29881	98.81
805-809	129	0.43	30010	99.24
810-814	63	0.21	30073	99.45
815-819	55	0.18	30128	99.63
820-824	33	0.11	30161	99.74
825-829	20	0.07	30181	99.81
830-834	10	0.03	30191	99.84
835-839	15	0.05	30206	99.89
840-844	8	0.03	30214	99.92
845-849	9	0.03	30223	99.95
850	17	0.06	30240	100.00

A.7.5. ELA Major Claim Score Distributions

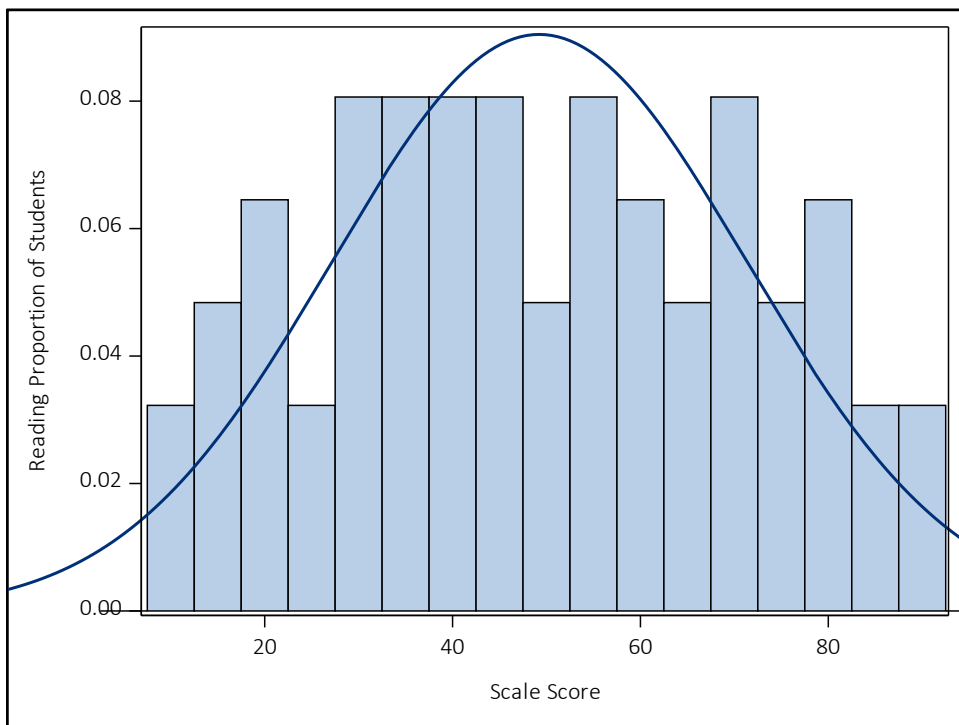


Figure A.7.5.1 Reading Major Claim Scale Score Distribution for ELA Grade 4

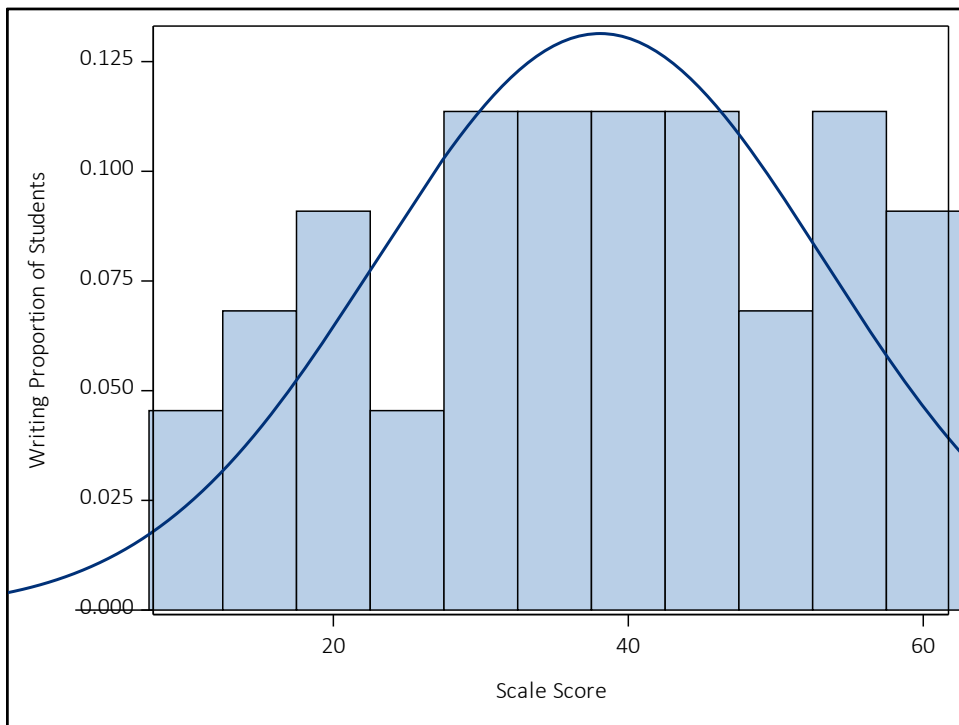


Figure A.7.5.2 Writing Major Claim Scale Score Distribution for ELA Grade 4

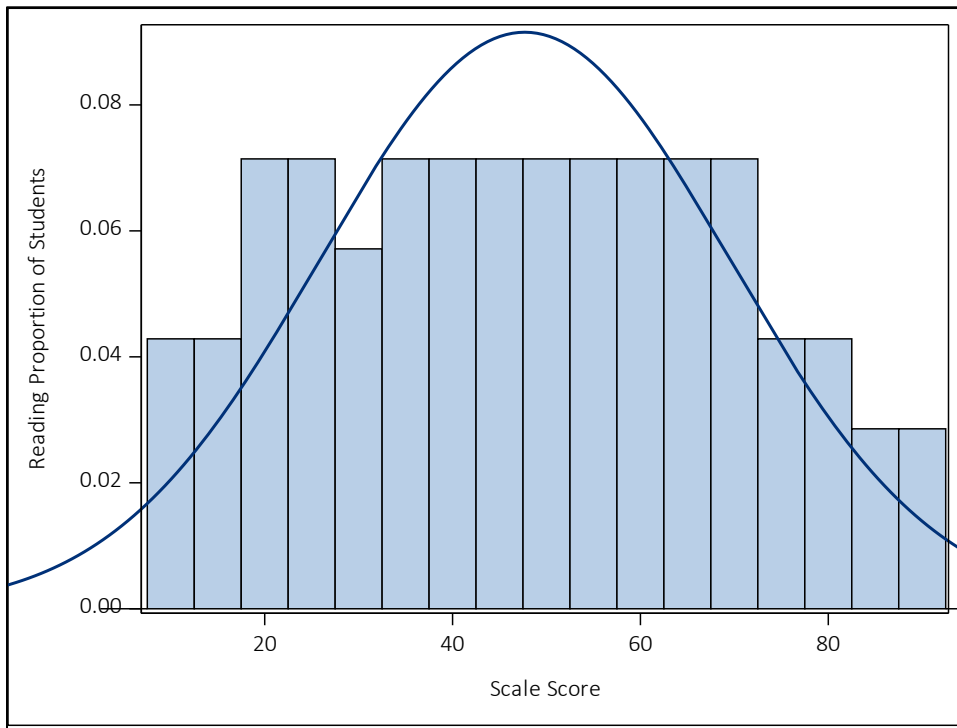
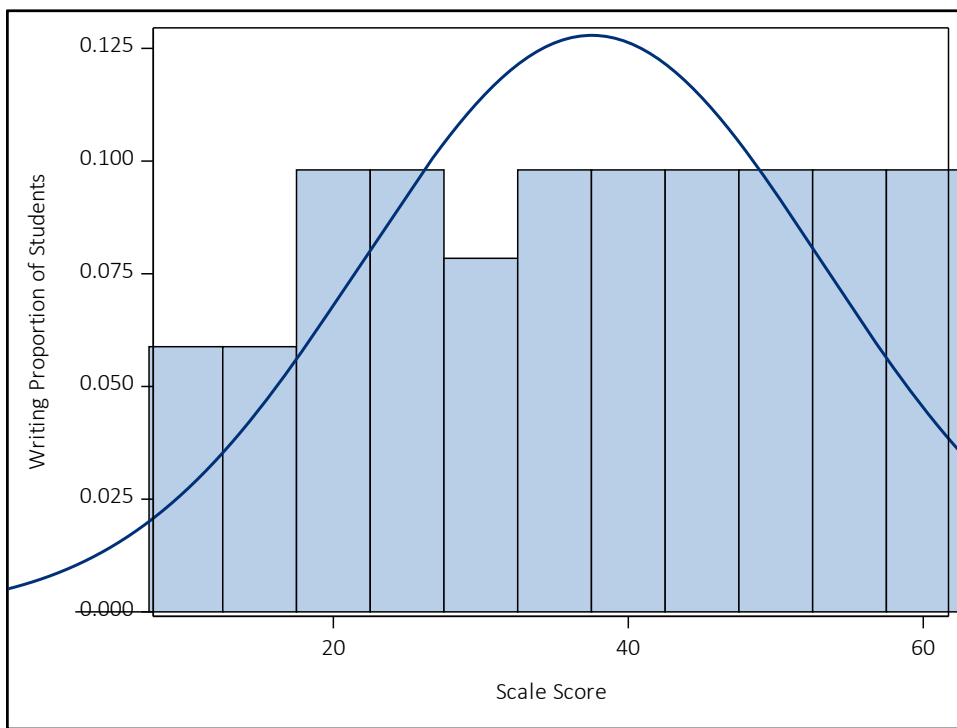


Figure A.7.5.3 Reading Major Claim Scale Score Distribution for ELA Grade 5



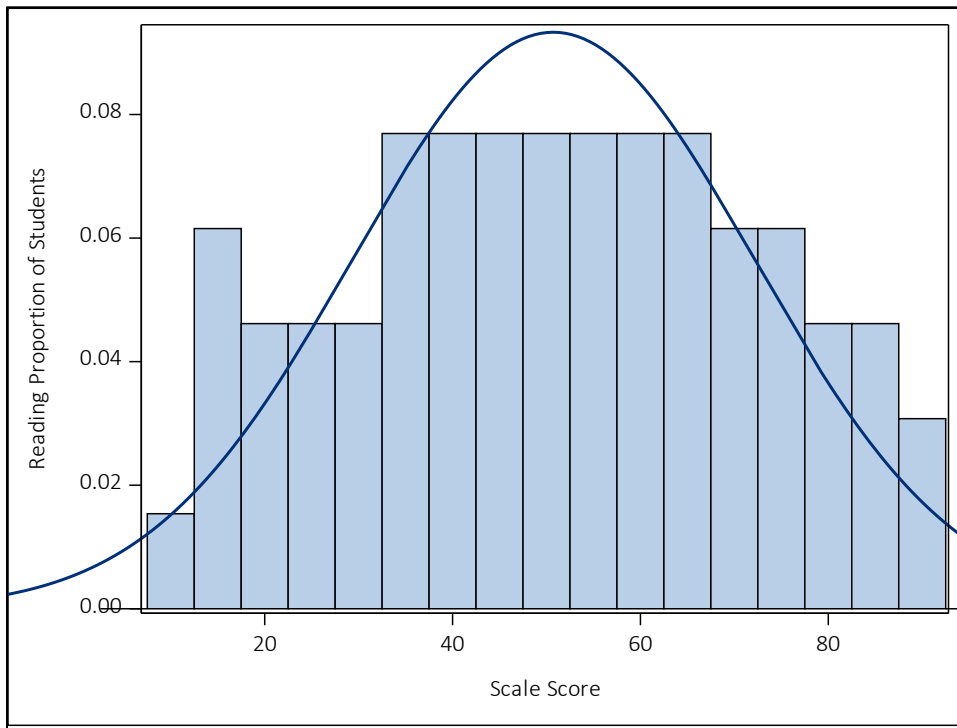


Figure A.7.5.5 Reading Major Claim Scale Score Distribution for ELA Grade 6

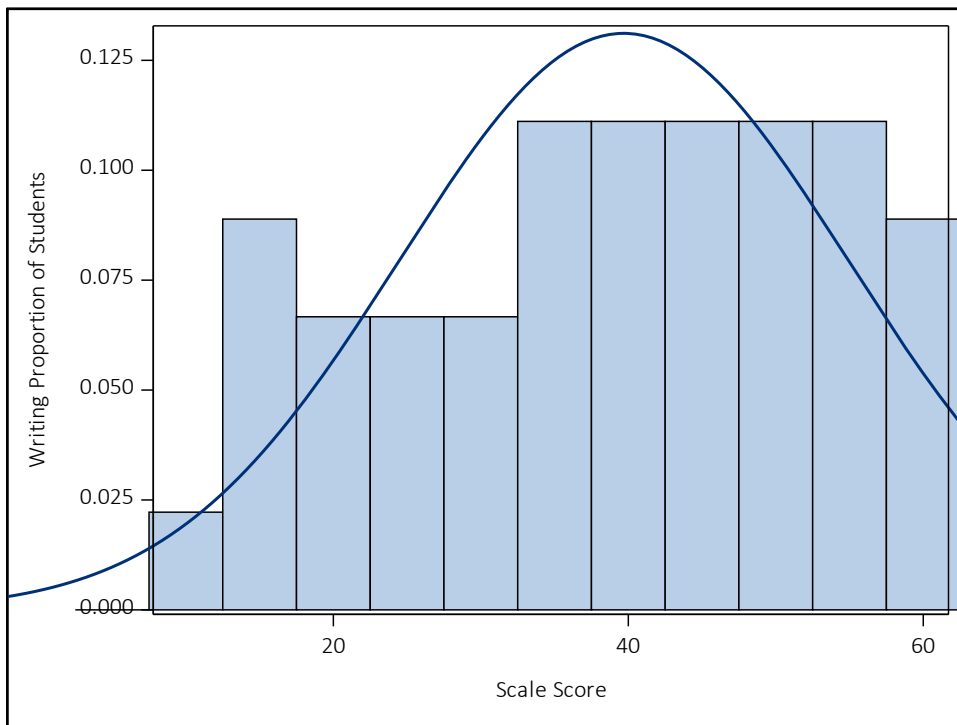


Figure A.7.5.6 Writing Major Claim Scale Score Distribution for ELA Grade 6

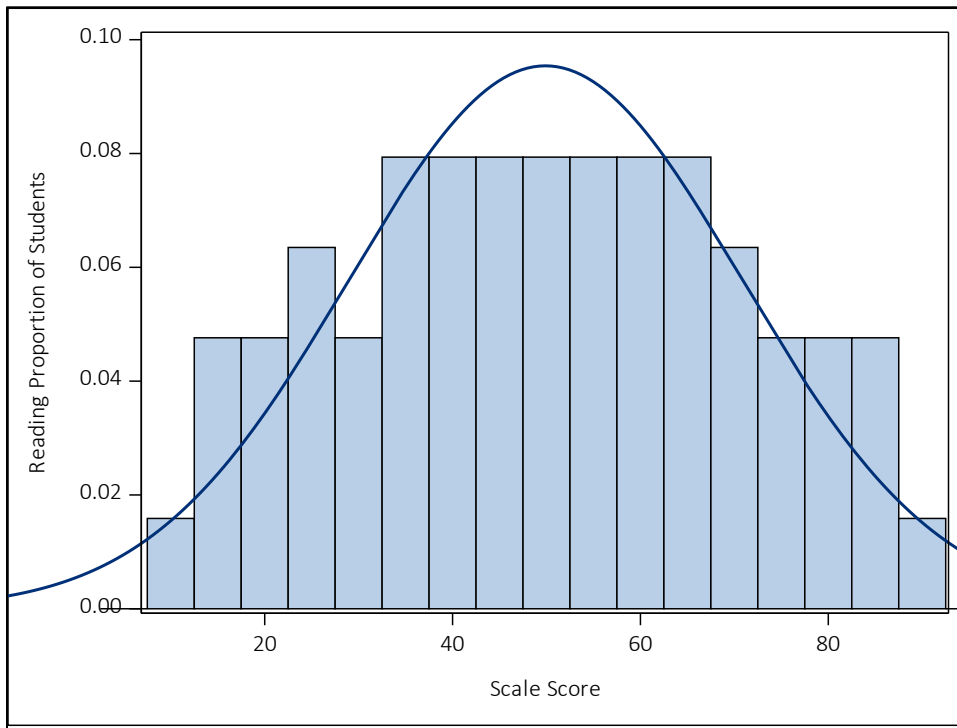


Figure A.7.5.7 Reading Major Claim Scale Score Distribution for ELA Grade 7

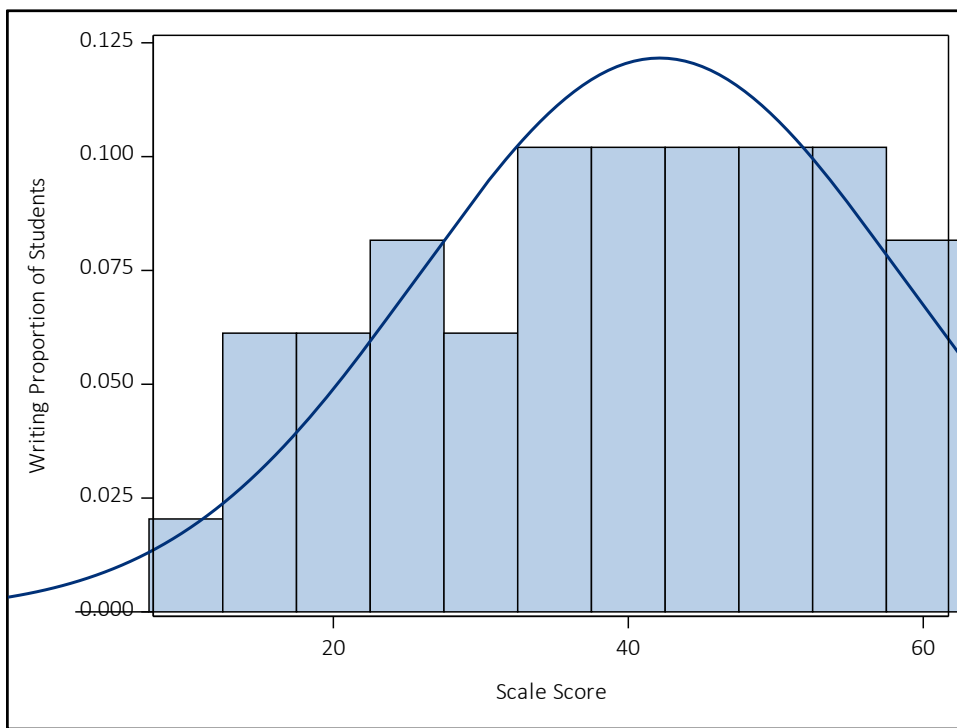


Figure A.7.5.8 Writing Major Claim Scale Score Distribution for ELA Grade 7

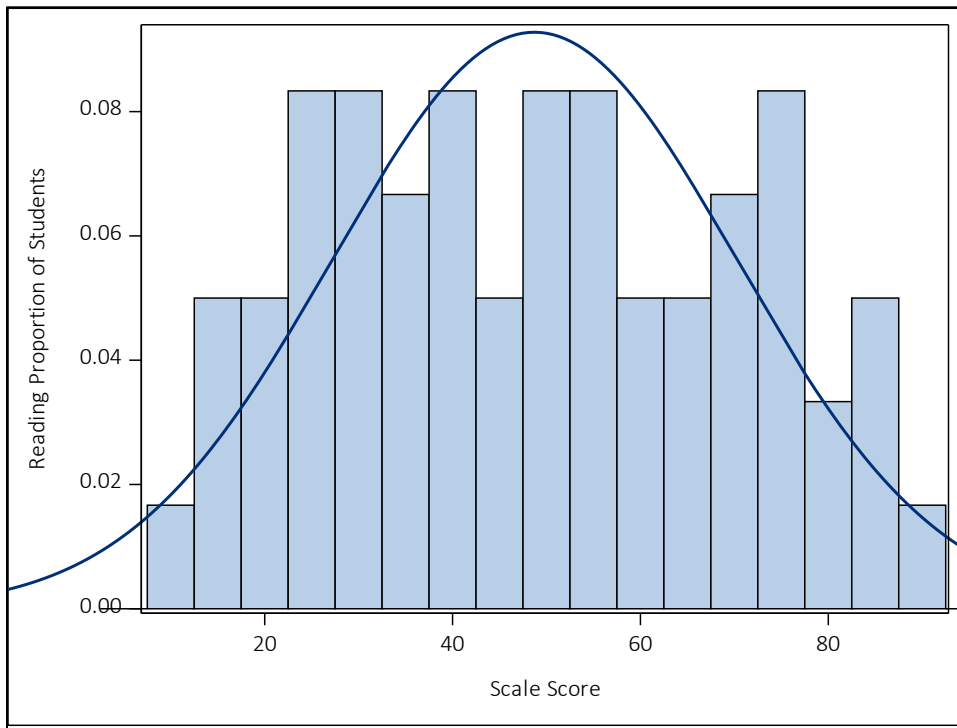


Figure A.7.5.9 Reading Major Claim Scale Score Distribution for ELA Grade 8

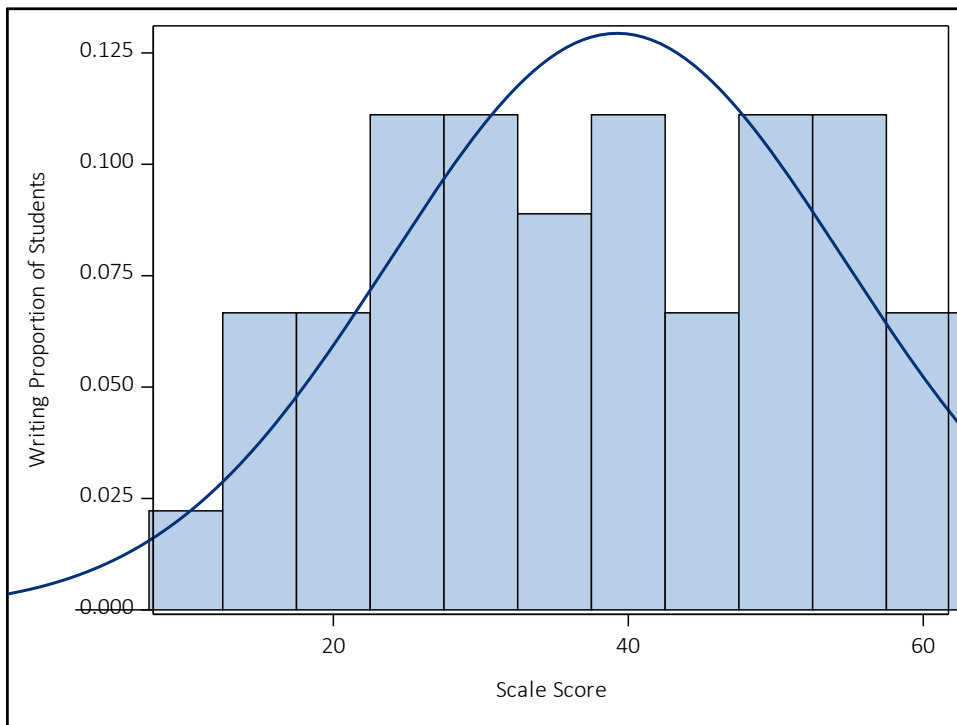


Figure A.7.5.10 Writing Major Claim Scale Score Distribution for ELA Grade 8

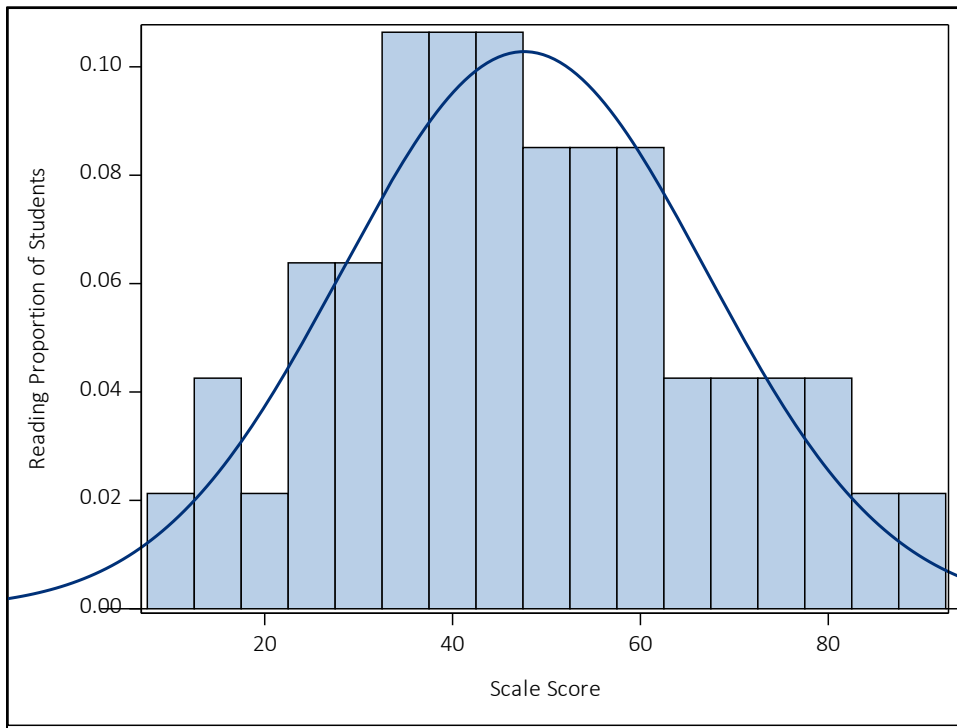


Figure A.7.5.11 Reading Major Claim Scale Score Distribution for ELA Grade 9

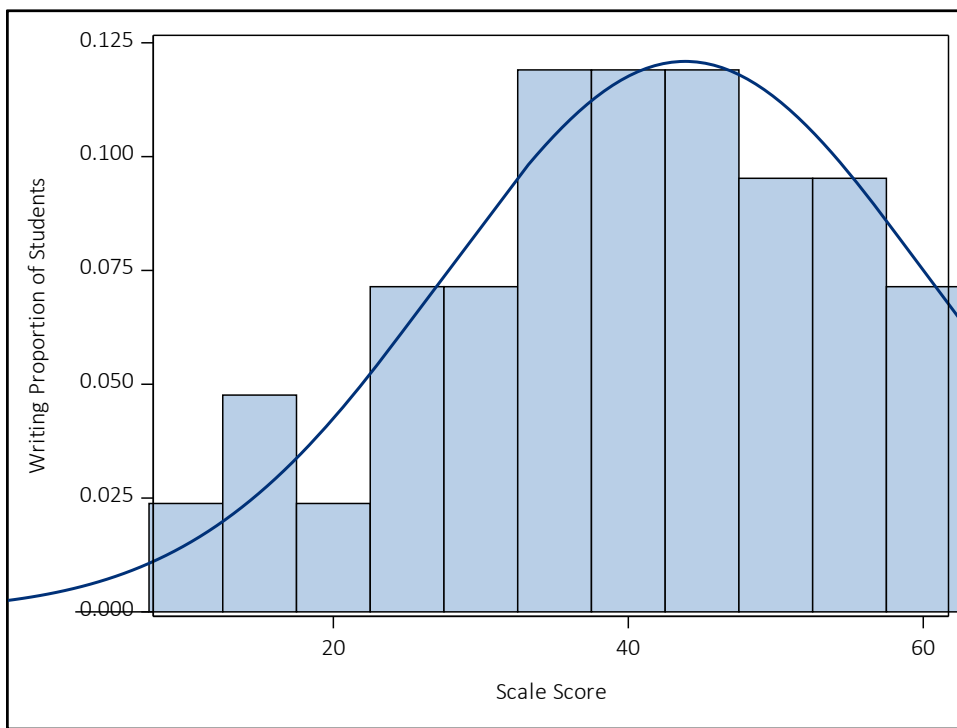


Figure A.7.5.12 Writing Major Claim Scale Score Distribution for ELA Grade 9

A.7.6. Demographic Scale Score Reporting

Table A.7.6.1 Subgroup Performance for ELA Grade 4 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		85,903	752.09	38.20	650	850
Gender	Female	43,204	754.84	38.10	650	850
	Male	42,694	749.30	38.10	650	850
Ethnicity	American Indian/Alaska Native	178	756.94	37.37	651	844
	Asian	9,933	777.72	33.80	650	850
	Black or African American	12,134	736.44	36.50	650	850
	Hispanic/Latino	27,576	737.41	36.68	650	850
	Native Hawaiian or Pacific Islander	145	759.10	38.03	655	846
	Two or more races	2,955	760.99	36.98	650	850
	White	32,965	761.55	33.75	650	850
Economic Status*	Not Economically Disadvantaged	54,957	762.03	35.84	650	850
	Economically Disadvantaged	30,946	734.43	35.80	650	850
English Learner Status	Non-English Learner	77,125	756.41	36.40	650	850
	English Learner	8,778	714.09	32.06	650	832
Disabilities	Students without Disabilities	73,570	754.96	37.54	650	850
	Student with Disability (SWD)	12,333	734.95	37.62	650	850
Reading Summative Score		85,903	50.46	15.64	10	90
Gender	Female	43,204	50.78	15.40	10	90
	Male	42,694	50.13	15.86	10	90
Ethnicity	American Indian/Alaska Native	178	52.01	15.42	11	90
	Asian	9,933	60.35	14.23	10	90
	Black or African American	12,134	44.53	14.81	10	90
	Hispanic/Latino	27,576	44.38	14.71	10	90
	Native Hawaiian or Pacific Islander	145	53.12	15.82	13	90
	Two or more races	2,955	54.37	15.25	10	90
	White	32,965	54.37	14.18	10	90
Economic Status*	Not Economically Disadvantaged	54,957	54.45	14.89	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
	Economically Disadvantaged	30,946	43.36	14.36	10	90
English Learner Status	Non-English Learner	77,125	52.21	15.01	10	90
	English Learner	8,778	35.09	12.30	10	85
Disabilities	Students without Disabilities	73,570	51.44	15.43	10	90
	Student with Disability (SWD)	12,333	44.61	15.61	10	90
Writing Summative Score		85,903	34.00	11.64	10	60
Gender	Female	43,204	35.38	11.20	10	60
	Male	42,694	32.59	11.90	10	60
Ethnicity	American Indian/Alaska Native	178	35.34	11.52	10	60
	Asian	9,933	40.68	8.87	10	60
	Black or African American	12,134	29.50	12.22	10	60
	Hispanic/Latino	27,576	30.34	12.12	10	60
	Native Hawaiian or Pacific Islander	145	35.37	12.06	10	55
	Two or more races	2,955	36.00	10.97	10	60
	White	32,965	36.50	10.02	10	60
Economic Status	Not Economically Disadvantaged	54,957	36.60	10.48	10	60
	Economically Disadvantaged	30,946	29.37	12.14	10	60
English Learner Status	Non-English Learner	77,125	35.16	10.99	10	60
	English Learner	8,778	23.77	12.16	10	60
Disabilities	Students without Disabilities	73,570	34.96	11.18	10	60
	Student with Disability (SWD)	12,333	28.23	12.62	10	60

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

Table A.7.6.2 Subgroup Performance for ELA Grade 5 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		87,320	753.76	37.97	650	850
Gender	Female	43,705	758.12	37.91	650	850
	Male	43,605	749.38	37.52	650	850
Ethnicity	American Indian/Alaska Native	156	754.96	39.78	651	850
	Asian	10,404	778.73	33.27	650	850
	Black or African American	12,591	736.86	35.69	650	850
	Hispanic/Latino	27,562	739.89	36.50	650	850
	Native Hawaiian or Pacific Islander	163	761.82	36.97	678	850
	Two or more races	2,866	762.47	36.51	650	850
	White	33,564	762.95	34.02	650	850
Economic Status*	Not Economically Disadvantaged	55,811	763.65	35.87	650	850
	Economically Disadvantaged	31,509	736.23	35.15	650	850
English Learner Status	Non-English Learner	80,601	757.48	36.19	650	850
	English Learner	6,719	709.10	29.43	650	850
Disabilities	Students without Disabilities	74,562	756.95	37.29	650	850
	Student with Disability (SWD)	12,758	735.07	36.48	650	850
Reading Summative Score		87,320	50.37	15.13	10	90
Gender	Female	43,705	51.35	15.18	10	90
	Male	43,605	49.37	15.01	10	90
Ethnicity	American Indian/Alaska Native	156	50.72	16.34	12	90
	Asian	10,404	60.45	13.84	10	90
	Black or African American	12,591	43.87	13.91	10	90
	Hispanic/Latino	27,562	44.55	14.04	10	90
	Native Hawaiian or Pacific Islander	163	52.29	14.86	22	90
	Two or more races	2,866	54.38	14.78	10	90
	White	33,564	54.10	13.78	10	90
Economic Status*	Not Economically Disadvantaged	55,811	54.39	14.51	10	90
	Economically Disadvantaged	31,509	43.23	13.48	10	90
English Learner Status	Non-English Learner	80,601	51.81	14.54	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
	English Learner	6,719	33.03	10.67	10	90
Disabilities	Students without Disabilities	74,562	51.45	14.96	10	90
	Student with Disability (SWD)	12,758	44.04	14.53	10	90
Writing Summative Score		87,320	34.58	12.51	10	60
Gender	Female	43,705	36.55	11.79	10	60
	Male	43,605	32.61	12.90	10	60
Ethnicity	American Indian/Alaska Native	156	35.23	12.29	10	60
	Asian	10,404	41.08	9.41	10	60
	Black or African American	12,591	29.72	13.09	10	60
	Hispanic/Latino	27,562	31.13	13.09	10	60
	Native Hawaiian or Pacific Islander	163	38.25	10.60	10	60
	Two or more races	2,866	36.39	11.82	10	60
	White	33,564	37.05	11.12	10	60
Economic Status	Not Economically Disadvantaged	55,811	37.20	11.35	10	60
	Economically Disadvantaged	31,509	29.94	13.11	10	60
English Learner Status	Non-English Learner	80,601	35.66	11.90	10	60
	English Learner	6,719	21.64	12.50	10	60
Disabilities	Students without Disabilities	74,562	35.70	11.99	10	60
	Student with Disability (SWD)	12,758	28.01	13.44	10	60

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

Table A.7.6.3 Subgroup Performance for ELA Grade 6 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		88,738	750.75	35.01	650	850
Gender	Female	44,345	754.55	34.67	650	850
	Male	44,371	746.95	34.95	650	850
Ethnicity	American Indian/Alaska Native	150	751.65	35.45	656	835
	Asian	10,239	774.69	31.41	650	850
	Black or African American	12,382	735.57	32.41	650	850
	Hispanic/Latino	28,201	738.15	33.89	650	850
	Native Hawaiian or Pacific Islander	176	758.79	32.88	656	841
	Two or more races	2,684	756.27	34.31	650	850
	White	34,875	758.85	31.28	650	850
Economic Status*	Not Economically Disadvantaged	57,591	759.10	33.32	650	850
	Economically Disadvantaged	31,147	735.31	32.74	650	850
English Learner Status	Non-English Learner	83,184	753.77	33.38	650	850
	English Learner	5,554	705.56	26.92	650	814
Disabilities	Students without Disabilities	75,819	753.97	34.12	650	850
	Student with Disability (SWD)	12,919	731.84	34.19	650	850
Reading Summative Score		88,738	49.90	14.09	10	90
Gender	Female	44,345	50.55	13.93	10	90
	Male	44,371	49.25	14.22	10	90
Ethnicity	American Indian/Alaska Native	150	50.04	14.03	13	90
	Asian	10,239	59.52	13.11	10	90
	Black or African American	12,382	44.29	13.06	10	90
	Hispanic/Latino	28,201	44.50	13.25	10	90
	Native Hawaiian or Pacific Islander	176	52.43	13.43	13	90
	Two or more races	2,684	52.72	13.99	10	90
	White	34,875	53.21	12.71	10	90
Economic Status*	Not Economically Disadvantaged	57,591	53.36	13.53	10	90
	Economically Disadvantaged	31,147	43.50	12.81	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
English Learner Status	Non-English Learner	83,184	51.10	13.50	10	90
	English Learner	5,554	31.86	9.68	10	76
Disabilities	Students without Disabilities	75,819	51.00	13.82	10	90
	Student with Disability (SWD)	12,919	43.42	13.92	10	90
Writing Summative Score		88,738	33.32	11.97	10	60
Gender	Female	44,345	35.20	11.27	10	60
	Male	44,371	31.45	12.35	10	60
Ethnicity	American Indian/Alaska Native	150	33.49	12.26	10	55
	Asian	10,239	39.87	9.17	10	60
	Black or African American	12,382	28.57	12.33	10	60
	Hispanic/Latino	28,201	30.01	12.55	10	60
	Native Hawaiian or Pacific Islander	176	36.49	10.29	10	55
	Two or more races	2,684	34.37	11.57	10	60
	White	34,875	35.68	10.58	10	60
Economic Status	Not Economically Disadvantaged	57,591	35.63	11.01	10	60
	Economically Disadvantaged	31,147	29.05	12.49	10	60
English Learner Status	Non-English Learner	83,184	34.23	11.42	10	60
	English Learner	5,554	19.71	11.79	10	55
Disabilities	Students without Disabilities	75,819	34.48	11.40	10	60
	Student with Disability (SWD)	12,919	26.51	12.93	10	60

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

Table A.7.6.4 Subgroup Performance for ELA Grade 7 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		89,798	756.14	40.75	650	850
Gender	Female	44,633	761.65	40.07	650	850
	Male	45,119	750.69	40.68	650	850
Ethnicity	American Indian/Alaska Native	149	759.07	43.68	650	850
	Asian	10,332	785.66	35.78	650	850
	Black or African American	12,826	739.96	38.44	650	850
	Hispanic/Latino	28,146	741.37	40.48	650	850
	Native Hawaiian or Pacific Islander	173	760.69	40.52	650	842
	Two or more races	2,609	762.50	39.92	650	850
	White	35,549	764.60	35.52	650	850
Economic Status*	Not Economically Disadvantaged	58,824	765.09	38.59	650	850
	Economically Disadvantaged	30,974	739.15	39.31	650	850
English Learner Status	Non-English Learner	84,232	759.73	38.57	650	850
	English Learner	5,566	701.80	33.47	650	842
Disabilities	Students without Disabilities	76,780	760.07	39.66	650	850
	Student with Disability (SWD)	13,018	732.99	39.33	650	850
Reading Summative Score		89,798	51.16	16.25	10	90
Gender	Female	44,633	52.46	16.06	10	90
	Male	45,119	49.86	16.34	10	90
Ethnicity	American Indian/Alaska Native	149	52.19	17.29	10	90
	Asian	10,332	62.75	14.87	10	90
	Black or African American	12,826	45.05	15.11	10	90
	Hispanic/Latino	28,146	45.04	15.73	10	90
	Native Hawaiian or Pacific Islander	173	52.63	16.23	10	90
	Two or more races	2,609	54.56	15.99	10	90
	White	35,549	54.58	14.43	10	90
Economic Status*	Not Economically Disadvantaged	58,824	54.80	15.53	10	90
	Economically Disadvantaged	30,974	44.23	15.31	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
English Learner Status	Non-English Learner	84,232	52.56	15.49	10	90
	English Learner	5,566	29.87	12.09	10	85
Disabilities	Students without Disabilities	76,780	52.51	15.93	10	90
	Student with Disability (SWD)	13,018	43.19	15.83	10	90
Writing Summative Score		89,798	35.87	11.86	10	60
Gender	Female	44,633	37.92	11.23	10	60
	Male	45,119	33.84	12.12	10	60
Ethnicity	American Indian/Alaska Native	149	36.64	12.51	10	60
	Asian	10,332	43.16	9.44	10	60
	Black or African American	12,826	31.48	12.20	10	60
	Hispanic/Latino	28,146	32.42	12.40	10	60
	Native Hawaiian or Pacific Islander	173	37.46	10.71	10	60
	Two or more races	2,609	36.74	11.70	10	60
	White	35,549	38.00	10.31	10	60
Economic Status	Not Economically Disadvantaged	58,824	38.04	11.01	10	60
	Economically Disadvantaged	30,974	31.76	12.32	10	60
English Learner Status	Non-English Learner	84,232	36.81	11.23	10	60
	English Learner	5,566	21.61	12.05	10	60
Disabilities	Students without Disabilities	76,780	37.07	11.30	10	60
	Student with Disability (SWD)	13,018	28.82	12.63	10	60

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

Table A.7.6.5 Subgroup Performance for ELA Grade 8 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		93,300	758.08	42.37	650	850
Gender	Female	46,050	765.49	41.42	650	850
	Male	47,187	750.84	42.05	650	850
Ethnicity	American Indian/Alaska Native	139	759.31	44.39	664	850
	Asian	10,431	788.74	36.81	650	850
	Black or African American	13,623	741.79	39.67	650	850
	Hispanic/Latino	29,111	744.19	41.68	650	850
	Native Hawaiian or Pacific Islander	195	768.88	38.81	672	850
	Two or more races	2,513	763.29	41.68	650	850
	White	37,254	765.91	38.43	650	850
Economic Status*	Not Economically Disadvantaged	61,662	766.41	40.80	650	850
	Economically Disadvantaged	31,638	741.85	40.65	650	850
English Learner Status	Non-English Learner	88,133	761.31	40.68	650	850
	English Learner	5,167	703.04	31.41	650	850
Disabilities	Students without Disabilities	79,358	762.41	41.22	650	850
	Student with Disability (SWD)	13,942	733.44	40.39	650	850
Reading Summative Score		93,300	51.99	16.84	10	90
Gender	Female	46,050	54.10	16.61	10	90
	Male	47,187	49.91	16.80	10	90
Ethnicity	American Indian/Alaska Native	139	52.22	17.61	10	90
	Asian	10,431	64.04	15.26	10	90
	Black or African American	13,623	46.04	15.76	10	90
	Hispanic/Latino	29,111	46.05	16.10	10	90
	Native Hawaiian or Pacific Islander	195	55.69	15.36	17	90
	Two or more races	2,513	55.01	16.81	10	90
	White	37,254	55.20	15.42	10	90
Economic Status*	Not Economically Disadvantaged	61,662	55.43	16.35	10	90
	Economically Disadvantaged	31,638	45.27	15.71	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
English Learner Status	Non-English Learner	88,133	53.23	16.27	10	90
	English Learner	5,167	30.78	11.17	10	90
Disabilities	Students without Disabilities	79,358	53.51	16.51	10	90
	Student with Disability (SWD)	13,942	43.34	16.02	10	90
Writing Summative Score		93,300	36.23	12.50	10	60
Gender	Female	46,050	38.69	11.67	10	60
	Male	47,187	33.83	12.81	10	60
Ethnicity	American Indian/Alaska Native	139	36.63	12.84	10	60
	Asian	10,431	43.79	9.64	10	60
	Black or African American	13,623	31.71	12.71	10	60
	Hispanic/Latino	29,111	33.00	13.10	10	60
	Native Hawaiian or Pacific Islander	195	39.29	11.70	10	60
	Two or more races	2,513	36.94	12.01	10	60
	White	37,254	38.24	11.20	10	60
Economic Status	Not Economically Disadvantaged	61,662	38.26	11.72	10	60
	Economically Disadvantaged	31,638	32.28	13.02	10	60
English Learner Status	Non-English Learner	88,133	37.13	11.93	10	60
	English Learner	5,167	20.93	12.11	10	60
Disabilities	Students without Disabilities	79,358	37.53	11.89	10	60
	Student with Disability (SWD)	13,942	28.84	13.28	10	60

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

Table A.7.6.6 Subgroup Performance for ELA Grade 9 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		92,037	750.33	40.68	650	850
Gender	Female	45,519	757.00	39.98	650	850
	Male	46,392	743.75	40.30	650	850
Ethnicity	American Indian/Alaska Native	158	751.18	41.91	668	850
	Asian	10,267	782.01	34.47	650	850
	Black or African American	13,128	734.43	37.43	650	850
	Hispanic/Latino	29,682	736.91	40.12	650	850
	Native Hawaiian or Pacific Islander	165	758.13	42.06	659	842
	Two or more races	2,315	756.52	40.15	650	850
	White	36,284	757.69	36.58	650	850
Economic Status*	Not Economically Disadvantaged	61,540	758.06	39.28	650	850
	Economically Disadvantaged	30,497	734.73	38.96	650	850
English Learner Status	Non-English Learner	86,794	753.51	39.10	650	850
	English Learner	5,243	697.72	28.55	650	833
Disabilities	Students without Disabilities	77,562	754.92	39.50	650	850
	Student with Disability (SWD)	14,475	725.74	38.02	650	850
Reading Summative Score		92,037	51.60	16.76	10	90
Gender	Female	45,519	53.60	16.51	10	90
	Male	46,392	49.62	16.78	10	90
Ethnicity	American Indian/Alaska Native	158	51.72	16.56	20	90
	Asian	10,267	64.43	14.78	10	90
	Black or African American	13,128	45.67	15.48	10	90
	Hispanic/Latino	29,682	46.13	16.23	10	90
	Native Hawaiian or Pacific Islander	165	54.16	17.37	16	90
	Two or more races	2,315	54.62	16.74	10	90
	White	36,284	54.38	15.36	10	90
Economic Status*	Not Economically Disadvantaged	61,540	54.71	16.37	10	90
	Economically Disadvantaged	30,497	45.31	15.74	10	90

Group Type	Subgroup	N	Mean	SD	Min.	Max.
English Learner Status	Non-English Learner	86,794	52.83	16.23	10	90
	English Learner	5,243	31.22	11.45	10	79
Disabilities	Students without Disabilities	77,562	53.27	16.40	10	90
	Student with Disability (SWD)	14,475	42.61	15.78	10	90
Writing Summative Score		92,037	32.38	12.46	10	60
Gender	Female	45,519	34.67	11.82	10	60
	Male	46,392	30.13	12.67	10	60
Ethnicity	American Indian/Alaska Native	158	32.40	13.11	10	60
	Asian	10,267	40.55	9.28	10	60
	Black or African American	13,128	27.67	12.52	10	60
	Hispanic/Latino	29,682	28.75	12.94	10	60
	Native Hawaiian or Pacific Islander	165	34.22	12.97	10	60
	Two or more races	2,315	33.61	12.22	10	60
	White	36,284	34.66	11.03	10	60
Economic Status	Not Economically Disadvantaged	61,540	34.51	11.69	10	60
	Economically Disadvantaged	30,497	28.07	12.86	10	60
English Learner Status	Non-English Learner	86,794	33.33	11.94	10	60
	English Learner	5,243	16.67	10.25	10	60
Disabilities	Students without Disabilities	77,562	33.82	11.86	10	60
	Student with Disability (SWD)	14,475	24.68	12.79	10	60

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

Table A.7.6.7 Subgroup Performance for Mathematics Grade 4 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		71,192	748.49	35.67	650	850
Gender	Female	36,027	746.01	34.46	650	850
	Male	35,159	751.02	36.70	650	850
Ethnicity	American Indian/Alaska Native	137	754.84	34.43	664	850
	Asian	8,494	776.38	31.86	650	850
	Black or African American	9,638	727.12	33.44	650	850
	Hispanic/Latino	22,329	733.35	32.51	650	850
	Native Hawaiian or Pacific Islander	110	754.94	34.59	663	834
	Two or more races	2,436	757.82	34.67	650	850
	White	28,028	758.59	29.95	650	850
Economic Status*	Not Economically Disadvantaged	46,845	758.09	33.48	650	850
	Economically Disadvantaged	24,347	730.02	32.35	650	850
English Learner Status	Non-English Learner	64,419	751.67	34.57	650	850
	English Learner	6,773	718.20	31.60	650	834
Disabilities	Students without Disabilities	62,454	749.85	35.31	650	850
	Student with Disability (SWD)	8,738	738.75	36.70	650	850
Language Form	Spanish	1,740	705.93	30.25	650	810

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

Table A.7.6.8 Subgroup Performance for Mathematics Grade 5 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		73,543	745.05	35.16	650	850
Gender	Female	37,069	742.96	33.78	650	850
	Male	36,467	747.18	36.39	650	850
Ethnicity	American Indian/Alaska Native	129	745.35	39.64	670	850
	Asian	9,107	774.83	31.97	650	850
	Black or African American	10,041	723.82	31.66	650	850
	Hispanic/Latino	22,844	729.84	31.48	650	850
	Native Hawaiian or Pacific Islander	144	753.38	35.74	679	850
	Two or more races	2,408	753.81	33.80	650	850
	White	28,857	754.31	29.55	650	850
Economic Status*	Not Economically Disadvantaged	48,231	754.82	33.18	650	850
	Economically Disadvantaged	25,312	726.43	31.07	650	850
English Learner Status	Non-English Learner	68,061	747.81	34.16	650	850
	English Learner	5,482	710.74	28.72	650	850
Disabilities	Students without Disabilities	64,643	746.54	35.04	650	850
	Student with Disability (SWD)	8,900	734.26	34.16	650	850
Language Form	Spanish	1,691	707.04	27.93	650	803

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.7.6.9 Subgroup Performance for Mathematics Grade 6 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		76,834	740.11	33.66	650	850
Gender	Female	38,593	738.37	32.79	650	850
	Male	38,224	741.87	34.43	650	850
Ethnicity	American Indian/Alaska Native	130	742.29	31.38	667	831
	Asian	9,346	768.96	30.89	650	850
	Black or African American	10,199	719.98	29.64	650	850
	Hispanic/Latino	24,156	725.49	30.14	650	850
	Native Hawaiian or Pacific Islander	149	748.06	31.03	667	814
	Two or more races	2,293	746.88	32.97	650	850
	White	30,534	749.03	28.51	650	850
Economic Status*	Not Economically Disadvantaged	51,076	748.85	32.11	650	850
	Economically Disadvantaged	25,758	722.78	29.72	650	850
English Learner Status	Non-English Learner	71,800	742.58	32.62	650	850
	English Learner	5,034	704.87	28.01	650	850
Disabilities	Students without Disabilities	67,306	741.94	33.30	650	850
	Student with Disability (SWD)	9,528	727.14	33.37	650	850
Language Form	Spanish	1,868	699.86	27.65	650	795

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.7.6.10 Subgroup Performance for Mathematics Grade 7 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		74,165	739.88	30.45	650	850
Gender	Female	37,249	738.03	30.15	650	850
	Male	36,876	741.75	30.64	650	850
Ethnicity	American Indian/Alaska Native	122	738.73	26.69	683	803
	Asian	7,336	762.74	29.44	650	850
	Black or African American	10,549	723.83	28.17	650	850
	Hispanic/Latino	23,922	727.90	27.80	650	850
	Native Hawaiian or Pacific Islander	149	743.97	30.11	653	796
	Two or more races	2,064	745.89	29.78	650	850
	White	30,007	749.07	26.38	650	850
Economic Status*	Not Economically Disadvantaged	48,535	747.23	29.19	650	850
	Economically Disadvantaged	25,630	725.97	27.82	650	850
English Learner Status	Non-English Learner	68,857	742.16	29.59	650	850
	English Learner	5,308	710.32	25.66	650	842
Disabilities	Students without Disabilities	64,604	741.71	29.98	650	850
	Student with Disability (SWD)	9,561	727.54	30.77	650	850
Language Form	Spanish	1,898	709.00	23.27	650	820

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.7.6.11 Subgroup Performance for Mathematics Grade 8 Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		51,196	719.79	34.72	650	850
Gender	Female	25,292	719.71	33.98	650	850
	Male	25,862	719.84	35.43	650	850
Ethnicity	American Indian/Alaska Native	75	711.01	34.02	650	825
	Asian	2,747	742.00	35.90	650	850
	Black or African American	8,742	705.72	31.46	650	850
	Hispanic/Latino	19,288	712.20	33.29	650	850
	Native Hawaiian or Pacific Islander	107	724.61	34.21	650	812
	Two or more races	1,246	723.21	34.84	650	850
	White	18,968	730.56	32.45	650	850
Economic Status*	Not Economically Disadvantaged	30,412	726.43	34.35	650	850
	Economically Disadvantaged	20,784	710.07	32.92	650	850
English Learner Status	Non-English Learner	46,775	722.12	34.27	650	850
	English Learner	4,421	695.08	29.42	650	836
Disabilities	Students without Disabilities	42,611	721.58	34.61	650	850
	Student with Disability (SWD)	8,585	710.86	33.87	650	850
Language Form	Spanish	1,783	693.84	30.07	650	825

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.7.6.12 Subgroup Performance for Algebra I Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		90,834	739.93	38.09	650	850
Gender	Female	44,664	738.35	36.31	650	850
	Male	46,074	741.46	39.69	650	850
Ethnicity	American Indian/Alaska Native	152	735.13	37.97	650	838
	Asian	9,964	774.84	38.93	650	850
	Black or African American	13,010	721.68	31.08	650	850
	Hispanic/Latino	29,969	724.91	32.33	650	850
	Native Hawaiian or Pacific Islander	190	747.52	36.63	659	850
	Two or more races	2,275	748.50	39.33	650	850
	White	35,237	749.02	34.70	650	850
Economic Status*	Not Economically Disadvantaged	60,714	747.90	38.40	650	850
	Economically Disadvantaged	30,120	723.84	31.88	650	850
English Learner Status	Non-English Learner	83,886	742.72	37.59	650	850
	English Learner	6,948	706.24	26.10	650	850
Disabilities	Students without Disabilities	77,808	742.74	38.04	650	850
	Student with Disability (SWD)	13,026	723.12	33.82	650	850
Language Form	Spanish	2,876	703.95	23.62	650	847

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

Table A.7.6.13 Subgroup Performance for Algebra II Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		8,793	756.12	43.26	650	850
Gender	Female	4,256	751.99	41.77	650	850
	Male	4,525	760.03	44.30	650	850
Ethnicity	American Indian/Alaska Native	11	761.73	46.75	696	850
	Asian	3,484	776.99	36.32	650	850
	Black or African American	644	712.89	37.90	650	850
	Hispanic/Latino	1,235	721.64	39.47	650	850
	Native Hawaiian or Pacific Islander	16	734.75	37.71	678	794
	Two or more races	274	755.03	42.61	650	850
	White	3,126	755.64	36.93	650	850
Economic Status*	Not Economically Disadvantaged	7,558	762.62	40.24	650	850
	Economically Disadvantaged	1,235	716.36	39.67	650	850
English Learner Status	Non-English Learner	8,604	757.43	42.52	650	850
	English Learner	189	696.71	33.80	650	850
Disabilities	Students without Disabilities	8,120	758.63	42.07	650	850
	Student with Disability (SWD)	673	725.91	45.94	650	850
Language Form	Spanish	116	688.46	22.84	650	745

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.7.6.14 Subgroup Performance for Geometry Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		27,057	749.43	27.52	650	850
Gender	Female	13,621	747.86	26.46	650	850
	Male	13,402	751.02	28.48	650	850
Ethnicity	American Indian/Alaska Native	48	756.17	31.55	695	835
	Asian	5,669	766.29	24.16	654	850
	Black or African American	2,301	729.57	26.24	650	815
	Hispanic/Latino	5,343	732.64	27.09	650	835
	Native Hawaiian or Pacific Islander	52	752.12	27.64	685	801
	Two or more races	823	754.37	26.48	650	850
	White	12,806	752.21	23.13	650	850
Economic Status*	Not Economically Disadvantaged	22,006	753.48	26.04	650	850
	Economically Disadvantaged	5,051	731.76	26.79	650	850
English Learner Status	Non-English Learner	26,487	750.24	26.99	650	850
	English Learner	570	711.85	25.67	650	811
Disabilities	Students without Disabilities	24,813	750.53	26.91	650	850
	Student with Disability (SWD)	2,244	737.27	31.02	650	846
Language Form	Spanish	280	705.21	22.97	650	775

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

Appendix 8

A.8.1. Student Demographic Information

Table A.8.1.1 Demographic Information: ELA Grade 4 Test Takers

Demographic	N	Percent
Economically Disadvantaged	34,669	37.1
Student with Disability (SWD)	19,838	21.2
English Learner	9,664	10.3
Male	47,380	50.7
Female	46,025	49.3
American Indian/ Alaska Native	191	0.2
Asian	10,180	10.9
Black or African American	13,409	14.4
Hispanic/Latino	30,587	32.7
White/Caucasian	35,658	38.2
Native Hawaiian or Pacific Islander	161	0.2
Two or more races	3,206	3.4
Unknown Ethnicity	20	0.0

Table A.8.1.2 Demographic Information: ELA Grade 5 Test Takers

Demographic	N	Percent
Economically Disadvantaged	35,249	37.2
Student with Disability (SWD)	20,258	21.4
English Learner	7,514	7.9
Male	48,288	50.9
Female	46,543	49.1
American Indian/ Alaska Native	165	0.2
Asian	10,626	11.2
Black or African American	14,039	14.8
Hispanic/Latino	30,515	32.2
White/Caucasian	36,257	38.2
Native Hawaiian or Pacific Islander	174	0.2
Two or more races	3,046	3.2
Unknown Ethnicity	19	0.0

Table A.8.1.3 Demographic Information: ELA Grade 6 Test Takers

Demographic	N	Percent
Economically Disadvantaged	34,751	36.1
Student with Disability (SWD)	20,322	21.1
English Learner	6,199	6.4
Male	49,115	51.1
Female	47,005	48.9
American Indian/ Alaska Native	165	0.2
Asian	10,452	10.9
Black or African American	13,808	14.4
Hispanic/Latino	31,083	32.3
White/Caucasian	37,554	39.1
Native Hawaiian or Pacific Islander	183	0.2
Two or more races	2,864	3.0
Unknown Ethnicity	34	0.0

Table A.8.1.4 Demographic Information: ELA Grade 7 Test Takers

Demographic	N	Percent
Economically Disadvantaged	34,405	35.5
Student with Disability (SWD)	20,236	20.9
English Learner	6,097	6.3
Male	49,648	51.2
Female	47,320	48.8
American Indian/ Alaska Native	164	0.2
Asian	10,565	10.9
Black or African American	14,233	14.7
Hispanic/Latino	30,908	31.9
White/Caucasian	38,152	39.3
Native Hawaiian or Pacific Islander	178	0.2
Two or more races	2,797	2.9
Unknown Ethnicity	20	0.0

Table A.8.1.5 Demographic Information: ELA Grade 8 Test Takers

Demographic	N	Percent
Economically Disadvantaged	34,907	34.9
Student with Disability (SWD)	20,755	20.7
English Learner	5,535	5.5
Male	51,495	51.4
Female	48,555	48.5
American Indian/ Alaska Native	152	0.2
Asian	10,628	10.6
Black or African American	15,013	15.0
Hispanic/Latino	31,702	31.7
White/Caucasian	39,715	39.7
Native Hawaiian or Pacific Islander	202	0.2
Two or more races	2,662	2.7
Unknown Ethnicity	41	0.0

Table A.8.1.6 Demographic Information: ELA Grade 9 Test Takers

Demographic	N	Percent
Economically Disadvantaged	32,645	33.6
Student with Disability (SWD)	19,536	20.1
English Learner	5,437	5.6
Male	49,630	51.1
Female	47,412	48.8
American Indian/ Alaska Native	165	0.2
Asian	10,412	10.7
Black or African American	14,149	14.6
Hispanic/Latino	31,513	32.4
White/Caucasian	38,304	39.4
Native Hawaiian or Pacific Islander	178	0.2
Two or more races	2,409	2.5
Unknown Ethnicity	44	0.0

Table A.8.1.7 Demographic Information: Mathematics Grade 4 Test Takers

Demographic	N	Percent
Economically Disadvantaged	35,517	37.4
Student with Disability (SWD)	19,797	20.8
English Learner	11,440	12.0
Male	48,195	50.7
Female	46,796	49.3
American Indian/ Alaska Native	191	0.2
Asian	10,337	10.9
Black or African American	13,441	14.1
Hispanic/Latino	31,816	33.5
White/Caucasian	35,826	37.7
Native Hawaiian or Pacific Islander	162	0.2
Two or more races	3,205	3.4
Unknown Ethnicity	22	0.0

Table A.8.1.8 Demographic Information: Mathematics Grade 5 Test Takers

Demographic	N	Percent
Economically Disadvantaged	36,105	37.5
Student with Disability (SWD)	20,192	21.0
English Learner	9,230	9.6
Male	49,079	50.9
Female	47,283	49.1
American Indian/ Alaska Native	168	0.2
Asian	10,778	11.2
Black or African American	14,022	14.5
Hispanic/Latino	31,727	32.9
White/Caucasian	36,422	37.8
Native Hawaiian or Pacific Islander	176	0.2
Two or more races	3,056	3.2
Unknown Ethnicity	23	0.0

Table A.8.1.9 Demographic Information: Mathematics Grade 6 Test Takers

Demographic	N	Percent
Economically Disadvantaged	35,639	36.5
Student with Disability (SWD)	20,248	20.7
English Learner	7,940	8.1
Male	49,860	51.1
Female	47,769	48.9
American Indian/ Alaska Native	166	0.2
Asian	10,572	10.8
Black or African American	13,831	14.2
Hispanic/Latino	32,324	33.1
White/Caucasian	37,682	38.6
Native Hawaiian or Pacific Islander	184	0.2
Two or more races	2,858	2.9
Unknown Ethnicity	36	0.0

Table A.8.1.10 Demographic Information: Mathematics Grade 7 Test Takers

Demographic	N	Percent
Economically Disadvantaged	34,717	37.1
Student with Disability (SWD)	19,929	21.3
English Learner	7,765	8.3
Male	47,632	51.0
Female	45,786	49.0
American Indian/ Alaska Native	156	0.2
Asian	8,333	8.9
Black or African American	14,019	15.0
Hispanic/Latino	31,548	33.8
White/Caucasian	36,623	39.2
Native Hawaiian or Pacific Islander	165	0.2
Two or more races	2,598	2.8
Unknown Ethnicity	21	0.0

Table A.8.1.11 Demographic Information: Mathematics Grade 8 Test Takers

Demographic	N	Percent
Economically Disadvantaged	28,126	42.2
Student with Disability (SWD)	17,685	26.5
English Learner	6,018	9.0
Male	34,644	51.9
Female	32,002	48.0
American Indian/ Alaska Native	109	0.2
Asian	3,385	5.1
Black or African American	11,975	18.0
Hispanic/Latino	25,162	37.7
White/Caucasian	24,272	36.4
Native Hawaiian or Pacific Islander	123	0.2
Two or more races	1,632	2.4
Unknown Ethnicity	35	0.1

Table A.8.1.12 Demographic Information: Algebra I Test Takers

Demographic	N	Percent
Economically Disadvantaged	36,440	34.5
Student with Disability (SWD)	20,167	19.1
English Learner	8,685	8.2
Male	54,475	51.6
Female	50,949	48.3
American Indian/ Alaska Native	177	0.2
Asian	10,971	10.4
Black or African American	15,337	14.5
Hispanic/Latino	35,907	34.0
White/Caucasian	40,297	38.2
Native Hawaiian or Pacific Islander	218	0.2
Two or more races	2,584	2.4
Unknown Ethnicity	44	0.0

Table A.8.1.13 Demographic Information: Algebra II Test Takers

Demographic	N	Percent
Economically Disadvantaged	2,110	20.9
Student with Disability (SWD)	949	9.4
English Learner	451	4.5
Male	5,224	51.8
Female	4,853	48.1
American Indian/ Alaska Native	15	0.1
Asian	3,656	36.2
Black or African American	717	7.1
Hispanic/Latino	2,079	20.6
White/Caucasian	3,302	32.7
Native Hawaiian or Pacific Islander	18	0.2
Two or more races	298	3.0
Unknown Ethnicity	4	0.0

Table A.8.1.14 Demographic Information: Geometry Test Takers

Demographic	N	Percent
Economically Disadvantaged	6,182	20.4
Student with Disability (SWD)	3,095	10.2
English Learner	769	2.5
Male	15,075	49.9
Female	15,128	50.0
American Indian/ Alaska Native	52	0.2
Asian	6,125	20.3
Black or African American	2,555	8.4
Hispanic/Latino	6,524	21.6
White/Caucasian	13,997	46.3
Native Hawaiian or Pacific Islander	59	0.2
Two or more races	911	3.0
Unknown Ethnicity	17	0.1

A.8.2. Post-Administration Differential Item Functioning Results

Table A.8.2.1 Post-Administration Differential Item Functioning for ELA Grade 4

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	31			1	3	29	94	1	3		
White vs. Black/ African American	31					31	100				
White vs. Hispanic/Latino	31					31	100				
White vs. Asian	31					31	100				
White vs. American Indian/ Alaska Native	31					31	100				
White vs. Native Hawaiian or Pacific Islander	31					31	100				
White vs. Two or more races	31					31	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	31					31	100				
Non-English Learner vs. English Learner	31	2	6	3	10	26	84				
Student without Disability vs. Student with Disability	31					31	100				

Table A.8.2.2 Post-Administration Differential Item Functioning for ELA Grade 5

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	31					30	97	1	3		
White vs. Black/ African American	31			1	3	30	97				
White vs. Hispanic/Latino	31			1	3	30	97				
White vs. Asian	31					31	100				
White vs. American Indian/ Alaska Native	31					31	100				
White vs. Native Hawaiian or Pacific Islander	31	1	3			30	97				
White vs. Two or more races	31					31	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	31					31	100				
Non-English Learner vs. English Learner	31	3	10	2	6	26	84				
Student without Disability vs. Student with Disability	31					31	100				

Table A.8.2.3 Post-Administration Differential Item Functioning for ELA Grade 6

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	31			2	6	27	87	2	6		
White vs. Black/ African American	31					31	100				
White vs. Hispanic/Latino	31					31	100				
White vs. Asian	31					30	97	1	3		
White vs. American Indian/ Alaska Native	31					31	100				
White vs. Native Hawaiian or Pacific Islander	31			1	3	30	97				
White vs. Two or more races	31					31	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	31					31	100				
Non-English Learner vs. English Learner	31	12	39	4	13	15	48				
Student without Disability vs. Student with Disability	31					31	100				

Table A.8.2.4 Post-Administration Differential Item Functioning for ELA Grade 7

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	31			1	3	29	94	1	3		
White vs. Black/ African American	31					31	100				
White vs. Hispanic/Latino	31			1	3	30	97				
White vs. Asian	31					31	100				
White vs. American Indian/ Alaska Native	31					31	100				
White vs. Native Hawaiian or Pacific Islander	31			1	3	30	97				
White vs. Two or more races	31					31	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	31			1	3	30	97				
Non-English Learner vs. English Learner	31	7	23	5	16	19	61				
Student without Disability vs. Student with Disability	31					31	100				

Table A.8.2.5 Post-Administration Differential Item Functioning for ELA Grade 8

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	31					30	97	1	3		
White vs. Black/ African American	31					31	100				
White vs. Hispanic/Latino	31					31	100				
White vs. Asian	31					31	100				
White vs. American Indian/ Alaska Native	31					31	100				
White vs. Native Hawaiian or Pacific Islander	31			1	3	30	97				
White vs. Two or more races	31					31	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	31					31	100				
Non-English Learner vs. English Learner	31	4	13	6	19	21	68				
Student without Disability vs. Student with Disability	31					31	100				

Table A.8.2.6 Post-Administration Differential Item Functioning for ELA Grade 9

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	20			1	5	18	90	1	5		
White vs. Black/ African American	20					20	100				
White vs. Hispanic/Latino	20					20	100				
White vs. Asian	20					20	100				
White vs. American Indian/ Alaska Native	20			1	5	18	90	1	5		
White vs. Native Hawaiian or Pacific Islander	20			1	5	19	95				
White vs. Two or more races	20					20	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	20					20	100				
Non-English Learner vs. English Learner	20	1	5	2	10	17	85				
Student without Disability vs. Student with Disability	20					20	100				

Table A.8.2.7 Post-Administration Differential Item Functioning for Mathematics Grade 4

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	54			2	4	51	94	1	2		
White vs. Black/ African American	54					53	98	1	2		
White vs. Hispanic/Latino	54					54	100				
White vs. Asian	54					54	100				
White vs. American Indian/ Alaska Native	54					53	98	1	2		
White vs. Native Hawaiian or Pacific Islander	54					54	100				
White vs. Two or more races	54					54	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	54					54	100				
Non-English Learner vs. English Learner	54			1	2	53	98				
Student without Disability vs. Student with Disability	54					54	100				

Table A.8.2.8 Post-Administration Differential Item Functioning for Mathematics Grade 5

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	56			1	2	55	98				
White vs. Black/ African American	56			2	4	54	96				
White vs. Hispanic/Latino	56					56	100				
White vs. Asian	56					56	100				
White vs. American Indian/ Alaska Native	56					56	100				
White vs. Native Hawaiian or Pacific Islander	56					56	100				
White vs. Two or more races	56					56	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	56					56	100				
Non-English Learner vs. English Learner	56			4	7	52	93				
Student without Disability vs. Student with Disability	56					55	98			1	2

Table A.8.2.9 Post-Administration Differential Item Functioning for Mathematics Grade 6

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	51			5	10	46	90				
White vs. Black/ African American	51					51	100				
White vs. Hispanic/Latino	51					51	100				
White vs. Asian	51					48	94	2	4	1	2
White vs. American Indian/ Alaska Native	51					50	98	1	2		
White vs. Native Hawaiian or Pacific Islander	51			1	2	50	98				
White vs. Two or more races	51					51	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	51					51	100				
Non-English Learner vs. English Learner	51			1	2	49	96	1	2		
Student without Disability vs. Student with Disability	51					50	98	1	2		

Table A.8.2.10 Post-Administration Differential Item Functioning for Mathematics Grade 7

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	55			1	2	53	96	1	2		
White vs. Black/ African American	55			1	2	49	89	5	9		
White vs. Hispanic/Latino	55					55	100				
White vs. Asian	55			1	2	49	89	5	9		
White vs. American Indian/ Alaska Native	55					53	96	2	4		
White vs. Native Hawaiian or Pacific Islander	55					55	100				
White vs. Two or more races	55					55	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	55					55	100				
Non-English Learner vs. English Learner	55			1	2	53	96	1	2		
Student without Disability vs. Student with Disability	55					55	100				

Table A.8.2.11 Post-Administration Differential Item Functioning for Mathematics Grade 8

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	53					53	100				
White vs. Black/ African American	53			1	2	52	98				
White vs. Hispanic/Latino	53					53	100				
White vs. Asian	53					53	100				
White vs. American Indian/ Alaska Native	53					53	100				
White vs. Native Hawaiian or Pacific Islander	53					53	100				
White vs. Two or more races	53					53	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	53					53	100				
Non-English Learner vs. English Learner	53					51	96	2	4		
Student without Disability vs. Student with Disability	53					52	98	1	2		

Table A.8.2.12 Post-Administration Differential Item Functioning for Algebra I

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	49			1	2	48	98				
White vs. Black/ African American	49			1	2	48	98				
White vs. Hispanic/Latino	49					49	100				
White vs. Asian	49					44	90	5	10		
White vs. American Indian/ Alaska Native	49					48	98	1	2		
White vs. Native Hawaiian or Pacific Islander	49					49	100				
White vs. Two or more races	49					49	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	49					49	100				
Non-English Learner vs. English Learner	49			3	6	44	90	2	4		
Student without Disability vs. Student with Disability	49					49	100				

Table A.8.2.13 Post-Administration Differential Item Functioning for Algebra II

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	51			2	4	49	96				
White vs. Black/ African American	51	2	4	2	4	46	90	1	2		
White vs. Hispanic/Latino	51			2	4	46	90	3	6		
White vs. Asian	51			1	2	41	80	8	16	1	2
White vs. American Indian/ Alaska Native	51					51	100				
White vs. Native Hawaiian or Pacific Islander	51					51	100				
White vs. Two or more races	51			1	2	48	94	2	4		
Not Economically Disadvantaged vs. Economically Disadvantaged	51			3	6	47	92	1	2		
Non-English Learner vs. English Learner	51					51	100				
Student without Disability vs. Student with Disability	51			3	6	48	94				

Table A.8.2.14 Post-Administration Differential Item Functioning for Geometry

DIF Comparison	N Items	C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
		N	%	N	%	N	%	N	%	N	%
Female vs. Male	53			1	2	52	98				
White vs. Black/ African American	53			1	2	52	98				
White vs. Hispanic/Latino	53					53	100				
White vs. Asian	53					51	96	2	4		
White vs. American Indian/ Alaska Native	53					53	100				
White vs. Native Hawaiian or Pacific Islander	53					53	100				
White vs. Two or more races	53					53	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	53					52	98	1	2		
Non-English Learner vs. English Learner	53	1	2	6	11	42	79	3	6	1	2
Student without Disability vs. Student with Disability	53					53	100				

Appendix 9

A.9.1. Reliability of Overall Scores for Demographic Subgroups

Table A.9.1.1 Summary of Test Reliability Estimates for Subgroups: ELA Grade 4

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	70	4.53	0.87	368	0.85	42848	0.89
Gender							
Male	70	4.45	0.88	227	0.86	21204	0.90
Female	70	4.61	0.87	141	0.83	21640	0.89
Ethnicity							
Black/African American	70	4.28	0.86	1236	0.81	5980	0.89
Asian/Pacific Islander	70	4.67	0.85	4990	0.84	258	0.86
Hispanic/Latino	70	4.34	0.86	180	0.82	13851	0.89
American Indian/Alaska Native	70	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	70	4.6	0.87	243	0.85	1462	0.89
White	70	4.6	0.85	16597	0.84	16368	0.87
Special Instruction Needs							
Economically Disadvantaged	70	4.26	0.86	184	0.8	15524	0.89
Not Economically Disadvantaged	70	4.62	0.86	27633	0.85	27324	0.88
English Learner	70	3.89	0.83	823	0.79	4429	0.86
Non-English Learner	70	4.56	0.87	305	0.85	38419	0.88
Students with Disabilities (SWD)	70	4.29	0.87	364	0.85	6077	0.90
Students without Disabilities	70	4.85	0.87	36799	0.86	36771	0.89
Students Taking Accommodated Forms							
ASL	67	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	67	4.14	0.82	176	0.82	176	0.82
Non-Screen Reader	67	4.28	0.86	119	0.86	119	0.86
Screen Reader	67	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.55	0.85	7102	0.85	7102	0.85

Note: n/r = not reported..

Table A.9.1.2 Summary of Test Reliability Estimates for Subgroups: ELA Grade 5

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	72	4.64	0.87	7116	0.84	43955	0.89
Gender							
Male	72	4.52	0.87	4428	0.83	22080	0.89
Female	72	4.72	0.87	150	0.84	21868	0.89
Ethnicity							
Black/African American	72	4.34	0.85	1391	0.78	6355	0.88
Asian/Pacific Islander	72	4.63	0.85	221	0.84	5333	0.86
Hispanic/Latino	72	4.49	0.86	2769	0.81	13802	0.88
American Indian/Alaska Native	72	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	72	4.66	0.87	172	0.85	1431	0.88
White	72	4.73	0.86	16608	0.85	144	0.87
Special Instruction Needs							
Economically Disadvantaged	72	4.41	0.85	3515	0.79	15827	0.87
Not Economically Disadvantaged	72	4.72	0.86	3601	0.85	180	0.87
English Learner	72	3.78	0.8	731	0.74	3354	0.82
Non-English Learner	72	4.66	0.86	6385	0.84	40601	0.88
Students with Disabilities (SWD)	72	4.40	0.87	7116	0.84	6421	0.88
Students without Disabilities	72	4.89	0.88	37028	0.87	37534	0.88
Students Taking Accommodated Forms							
ASL	74	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	74	4.33	0.85	191	0.85	191	0.85
Non-Screen Reader	74	4.55	0.85	114	0.85	114	0.85
Screen Reader	74	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.63	0.84	7084	0.84	7084	0.84

Note: n/r = not reported..

Table A.9.1.3 Summary of Test Reliability Estimates for Subgroups: ELA Grade 6

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	72	4.90	0.87	7161	0.85	44571	0.88
Gender							
Male	72	4.75	0.87	4602	0.85	22274	0.89
Female	72	5.00	0.87	2559	0.85	101	0.89
Ethnicity							
Black/African American	72	4.56	0.86	1386	0.83	6281	0.87
Asian/Pacific Islander	72	4.98	0.86	5190	0.84	211	0.87
Hispanic/Latino	72	4.75	0.86	2779	0.83	14073	0.88
American Indian/Alaska Native	72	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	72	4.89	0.86	170	0.81	1330	0.89
White	72	4.97	0.85	17303	0.84	17572	0.86
Special Instruction Needs							
Economically Disadvantaged	72	4.63	0.85	3475	0.82	15559	0.87
Not Economically Disadvantaged	72	4.98	0.87	28579	0.85	115	0.89
English Learner	72	3.80	0.81	615	0.77	2796	0.83
Non-English Learner	72	4.92	0.87	6546	0.85	214	0.88
Students with Disabilities (SWD)	72	4.58	0.88	7161	0.85	6542	0.89
Students without Disabilities	72	5.24	0.87	37790	0.86	38029	0.88
Students Taking Accommodated Forms							
ASL	74	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	74	4.30	0.85	106	0.85	106	0.85
Non-Screen Reader	74	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	74	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	3.82	0.85	7118	0.85	7118	0.85

Note: n/r = not reported..

Table A.9.1.4 Summary of Test Reliability Estimates for Subgroups: ELA Grade 7

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	72	5.30	0.88	250	0.83	44673	0.90
Gender							
Male	72	5.15	0.88	155	0.83	22307	0.90
Female	72	5.32	0.88	2592	0.87	22338	0.90
Ethnicity							
Black/African American	72	4.99	0.86	1356	0.83	6399	0.89
Asian/Pacific Islander	72	5.21	0.87	5347	0.85	233	0.90
Hispanic/Latino	72	5.14	0.88	2668	0.86	13989	0.90
American Indian/Alaska Native	72	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	72	5.27	0.88	1329	0.87	1280	0.90
White	72	5.26	0.86	17780	0.85	17769	0.88
Special Instruction Needs							
Economically Disadvantaged	72	5.02	0.86	121	0.79	15413	0.89
Not Economically Disadvantaged	72	5.40	0.87	129	0.82	29260	0.89
English Learner	72	4.17	0.83	2775	0.81	2791	0.86
Non-English Learner	72	5.31	0.87	227	0.83	41882	0.89
Students with Disabilities (SWD)	72	5.01	0.88	249	0.83	6429	0.90
Students without Disabilities	72	5.59	0.88	38536	0.87	38244	0.90
Students Taking Accommodated Forms							
ASL	70	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	70	4.91	0.84	108	0.84	108	0.84
Non-Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	4.06	0.87	6903	0.87	6903	0.87

Note: n/r = not reported..

Table A.9.1.5 Summary of Test Reliability Estimates for Subgroups: ELA Grade 8

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	72	5.35	0.87	185	0.82	46549	0.89
Gender							
Male	72	5.18	0.87	124	0.81	23559	0.89
Female	72	5.37	0.87	2444	0.85	22955	0.89
Ethnicity							
Black/African American	72	5.07	0.87	1349	0.84	6851	0.89
Asian/Pacific Islander	72	5.07	0.85	5327	0.84	5299	0.87
Hispanic/Latino	72	5.27	0.87	2527	0.84	14598	0.89
American Indian/Alaska Native	72	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	72	5.20	0.88	147	0.87	1247	0.89
White	72	5.30	0.86	18786	0.85	18468	0.88
Special Instruction Needs							
Economically Disadvantaged	72	5.20	0.86	3194	0.83	15844	0.88
Not Economically Disadvantaged	72	5.35	0.87	110	0.84	30705	0.89
English Learner	72	4.21	0.81	361	0.80	2589	0.84
Non-English Learner	72	5.34	0.87	178	0.82	43960	0.89
Students with Disabilities (SWD)	72	5.11	0.87	183	0.82	6916	0.89
Students without Disabilities	72	5.56	0.88	39725	0.87	39633	0.89
Students Taking Accommodated Forms							
ASL	70	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	4.38	0.85	6577	0.85	6577	0.85

Note: n/r = not reported..

Table A.9.1.6 Summary of Test Reliability Estimates for Subgroups: ELA Grade 9

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	70	5.16	0.86	4911	0.81	226	0.87
Gender							
Male	70	4.84	0.87	3101	0.82	137	0.90
Female	70	5.28	0.86	1805	0.81	45519	0.87
Ethnicity							
Black/African American	70	4.85	0.83	973	0.74	6772	0.87
Asian/Pacific Islander	70	5.82	0.79	137	0.64	10432	0.84
Hispanic/Latino	70	4.87	0.86	1758	0.82	29682	0.87
American Indian/Alaska Native	70	4.93	0.90	158	0.90	158	0.90
Two or more races	70	5.58	0.87	2315	0.86	1266	0.87
White	70	5.05	0.85	1941	0.85	1941	0.85
Special Instruction Needs							
Economically Disadvantaged	70	4.87	0.84	2065	0.78	30497	0.87
Not Economically Disadvantaged	70	5.23	0.85	2846	0.83	143	0.87
English Learner	70	3.70	0.80	177	0.78	5243	0.80
Non-English Learner	70	5.17	0.85	4734	0.81	209	0.87
Students with Disabilities (SWD)	70	4.66	0.86	4911	0.81	14475	0.88
Students without Disabilities	70	5.51	0.87	77562	0.86	39725	0.87
Students Taking Accommodated Forms							
ASL	70	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	70	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	70	4.02	0.81	4839	0.81	4839	0.81

Note: n/r = not reported..

Table A.9.1.7 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 4

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	3.59	0.89	470	0.86	13822	0.91
Gender							
Male	52	3.59	0.90	263	0.88	7505	0.92
Female	52	3.57	0.89	207	0.82	6315	0.91
Ethnicity							
Black/African American	52	3.29	0.89	1897	0.89	1875	0.90
Asian/Pacific Islander	52	3.73	0.88	4293	0.86	962	0.92
Hispanic/Latino	52	3.39	0.88	327	0.85	6776	0.89
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.69	0.90	1194	0.88	381	0.92
White	52	3.69	0.88	13942	0.86	3839	0.91
Special Instruction Needs							
Economically Disadvantaged	52	3.33	0.88	275	0.85	6874	0.89
Not Economically Disadvantaged	52	3.68	0.89	195	0.87	6948	0.92
English Learner	52	3.19	0.87	216	0.85	2582	0.88
Non-English Learner	52	3.63	0.90	254	0.87	9381	0.92
Students with Disabilities (SWD)	52	3.42	0.89	303	0.87	4259	0.90
Students without Disabilities	52	3.64	0.90	167	0.86	8416	0.92
Students Taking Accommodated Forms							
ASL	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	3.23	0.85	248	0.85	248	0.85
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.55	0.90	22351	0.90	22351	0.90
Students Taking Translated Forms							
Spanish Language	52	2.76	0.87	1740	0.87	987	0.87

Note: n/r = not reported..

Table A.9.1.8 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 5

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	3.58	0.89	430	0.85	13344	0.90
Gender							
Male	52	3.57	0.89	267	0.86	7335	0.91
Female	52	3.59	0.88	163	0.83	6005	0.89
Ethnicity							
Black/African American	52	3.24	0.86	1940	0.85	1992	0.87
Asian/Pacific Islander	52	3.88	0.88	4557	0.85	820	0.92
Hispanic/Latino	52	3.37	0.86	268	0.81	6478	0.86
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.74	0.89	1257	0.88	308	0.91
White	52	3.71	0.87	14304	0.85	3717	0.90
Special Instruction Needs							
Economically Disadvantaged	52	3.30	0.86	256	0.83	6696	0.86
Not Economically Disadvantaged	52	3.70	0.88	174	0.87	6648	0.91
English Learner	52	3.04	0.83	173	0.76	1853	0.85
Non-English Learner	52	3.63	0.88	257	0.87	9400	0.90
Students with Disabilities (SWD)	52	3.36	0.88	302	0.86	4397	0.89
Students without Disabilities	52	3.64	0.88	128	0.74	7841	0.91
Students Taking Accommodated Forms							
ASL	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	3.01	0.83	230	0.83	230	0.83
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.47	0.89	21526	0.89	21526	0.89
Students Taking Translated Forms							
Spanish Language	52	2.83	0.83	873	0.82	1691	0.84

Note: n/r = not reported..

Table A.9.1.9 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 6

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	3.49	0.89	257	0.84	9938	0.90
Gender							
Male	52	3.50	0.89	147	0.82	6859	0.90
Female	52	3.48	0.89	109	0.87	5628	0.89
Ethnicity							
Black/African American	52	3.02	0.86	1796	0.86	5101	0.87
Asian/Pacific Islander	52	3.94	0.89	4719	0.87	655	0.92
Hispanic/Latino	52	3.19	0.86	142	0.82	11082	0.87
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.65	0.90	272	0.89	288	0.90
White	52	3.65	0.88	15229	0.86	3531	0.89
Special Instruction Needs							
Economically Disadvantaged	52	3.11	0.86	157	0.80	12236	0.87
Not Economically Disadvantaged	52	3.66	0.89	100	0.88	6148	0.90
English Learner	52	2.73	0.83	1068	0.80	1559	0.85
Non-English Learner	52	3.54	0.89	162	0.87	8881	0.90
Students with Disabilities (SWD)	52	3.18	0.88	5300	0.85	4728	0.90
Students without Disabilities	52	3.62	0.90	32736	0.89	7192	0.90
Students Taking Accommodated Forms							
ASL	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	2.84	0.84	121	0.84	121	0.84
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.36	0.88	19785	0.88	19785	0.88
Students Taking Translated Forms							
Spanish Language	52	2.58	0.81	777	0.78	1868	0.82

Note: n/r = not reported..

Table A.9.1.10 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 7

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	3.29	0.89	264	0.88	9093	0.90
Gender							
Male	52	3.29	0.89	159	0.85	5104	0.91
Female	52	3.29	0.89	5344	0.89	104	0.91
Ethnicity							
Black/African American	52	2.88	0.88	1753	0.87	1679	0.88
Asian/Pacific Islander	52	3.90	0.89	3794	0.88	501	0.92
Hispanic/Latino	52	2.99	0.86	142	0.81	10897	0.88
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.48	0.90	1040	0.89	255	0.92
White	52	3.50	0.88	15036	0.87	3296	0.90
Special Instruction Needs							
Economically Disadvantaged	52	2.94	0.87	150	0.84	12105	0.88
Not Economically Disadvantaged	52	3.47	0.89	24046	0.88	4925	0.91
English Learner	52	2.61	0.84	3480	0.81	1685	0.87
Non-English Learner	52	3.35	0.89	34266	0.89	189	0.90
Students with Disabilities (SWD)	52	2.96	0.88	5106	0.85	4787	0.90
Students without Disabilities	52	3.43	0.89	31164	0.89	6733	0.90
Students Taking Accommodated Forms							
ASL	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	2.72	0.88	124	0.88	124	0.88
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	3.06	0.89	18278	0.89	18278	0.89
Students Taking Translated Forms							
Spanish Language	52	2.48	0.81	756	0.77	1898	0.83

Note: n/r = not reported..

Table A.9.1.11 Summary of Test Reliability Estimates for Subgroups: Mathematics Grade 8

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability N	Alpha	Maximum Reliability N	Alpha
Total Group	52	2.98	0.83	7381	0.82	213	0.84
Gender							
Male	52	2.97	0.84	121	0.81	12475	0.85
Female	52	2.98	0.83	3203	0.82	4258	0.84
Ethnicity							
Black/African American	52	2.67	0.80	1585	0.78	4356	0.81
Asian/Pacific Islander	52	3.47	0.86	1416	0.85	349	0.90
Hispanic/Latino	52	2.82	0.81	109	0.71	8770	0.83
American Indian/Alaska Native	52	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	52	3.06	0.84	202	0.80	183	0.87
White	52	3.15	0.82	9462	0.81	2622	0.85
Special Instruction Needs							
Economically Disadvantaged	52	2.78	0.82	3434	0.80	9835	0.83
Not Economically Disadvantaged	52	3.12	0.84	3947	0.82	118	0.86
English Learner	52	2.46	0.78	1347	0.75	1291	0.81
Non-English Learner	52	3.03	0.84	6871	0.82	156	0.85
Students with Disabilities (SWD)	52	2.76	0.81	128	0.76	4312	0.83
Students without Disabilities	52	3.07	0.84	2932	0.83	5144	0.86
Students Taking Accommodated Forms							
ASL	52	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	52	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	52	2.78	0.83	14784	0.83	14784	0.83
Students Taking Translated Forms							
Spanish Language	52	2.39	0.75	500	0.64	1783	0.81

Note: n/r = not reported..

Table A.9.1.12 Summary of Test Reliability Estimates for Subgroups: Algebra I

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Total Group	56	3.32	0.87	259	0.78	44010	0.88
Gender							
Male	56	3.36	0.88	146	0.77	22398	0.89
Female	56	3.27	0.87	112	0.80	21566	0.88
Ethnicity							
Black/African American	56	2.86	0.81	1137	0.79	6436	0.83
Asian/Pacific Islander	56	4.12	0.88	5142	0.87	507	0.91
Hispanic/Latino	56	2.93	0.80	109	0.45	13471	0.85
American Indian/Alaska Native	56	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	56	3.63	0.89	1115	0.87	151	0.90
White	56	3.54	0.86	17504	0.85	2507	0.88
Special Instruction Needs							
Economically Disadvantaged	56	2.90	0.80	123	0.49	14140	0.84
Not Economically Disadvantaged	56	3.51	0.88	136	0.82	5312	0.89
English Learner	56	2.50	0.73	4152	0.62	2047	0.78
Non-English Learner	56	3.40	0.87	224	0.79	6295	0.89
Students with Disabilities (SWD)	56	2.93	0.83	3489	0.77	6445	0.86
Students without Disabilities	56	3.40	0.86	132	0.59	6882	0.89
Students Taking Accommodated Forms							
ASL	55	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	3.25	0.87	13618	0.87	13618	0.87
Students Taking Translated Forms							
Spanish Language	55	2.34	0.58	824	0.50	2876	0.62

Note: n/r = not reported..

Table A.9.1.13 Summary of Test Reliability Estimates for Subgroups: Algebra II

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Total Group	56	3.76	0.88	4414	0.87	541	0.92
Gender							
Male	56	3.83	0.89	2277	0.88	296	0.93
Female	56	3.67	0.88	2132	0.86	245	0.91
Ethnicity							
Black/African American	56	2.96	0.86	325	0.86	319	0.87
Asian/Pacific Islander	56	4.55	0.82	1767	0.82	1733	0.83
Hispanic/Latino	56	2.99	0.82	642	0.67	573	0.87
American Indian/Alaska Native	56	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	56	4.11	0.87	136	0.86	138	0.89
White	56	4.05	0.85	1605	0.85	1605	0.85
Special Instruction Needs							
Economically Disadvantaged	56	2.91	0.81	635	0.62	557	0.88
Not Economically Disadvantaged	56	4.08	0.87	3830	0.85	233	0.90
English Learner	56	2.09	0.33	341	0.33	341	0.33
Non-English Learner	56	3.81	0.89	4373	0.87	504	0.92
Students with Disabilities (SWD)	56	3.15	0.89	139	0.88	134	0.90
Students without Disabilities	56	3.83	0.88	4071	0.86	407	0.92
Students Taking Accommodated Forms							
ASL	55	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	3.42	0.90	1107	0.90	1107	0.90
Students Taking Translated Forms							
Spanish Language	55	2.01	0.33	186	0.25	116	0.49

Note: n/r = not reported..

Table A.9.1.14 Summary of Test Reliability Estimates for Subgroups: Geometry

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Total Group	56	3.33	0.89	13316	0.89	1887	0.90
Gender							
Male	56	3.38	0.90	6571	0.89	984	0.91
Female	56	3.28	0.88	6730	0.88	901	0.89
Ethnicity							
Black/African American	56	2.78	0.88	1137	0.86	121	0.90
Asian/Pacific Islander	56	3.81	0.87	224	0.86	230	0.88
Hispanic/Latino	56	2.84	0.86	911	0.83	2548	0.87
American Indian/Alaska Native	56	n/r	n/r	n/r	n/r	n/r	n/r
Two or more races	56	3.65	0.89	390	0.89	390	0.89
White	56	3.38	0.87	6360	0.86	596	0.89
Special Instruction Needs							
Economically Disadvantaged	56	2.78	0.85	522	0.82	2418	0.87
Not Economically Disadvantaged	56	3.47	0.89	10852	0.88	1023	0.90
English Learner	56	2.50	0.83	426	0.72	152	0.87
Non-English Learner	56	3.37	0.89	13178	0.88	1492	0.90
Students with Disabilities (SWD)	56	2.98	0.89	418	0.86	1090	0.90
Students without Disabilities	56	3.39	0.89	12226	0.88	1469	0.90
Students Taking Accommodated Forms							
ASL	55	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	70	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Non-Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	3.27	0.88	3062	0.88	3062	0.88
Students Taking Translated Forms							
Spanish Language	55	2.16	0.73	280	0.73	280	0.73

Note: n/r = not reported..

Appendix 10

A.10.1. Intercorrelations of Subclaims

Table A.10.1.1 Average Intercorrelations between ELA Grade 4 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	93,412	1						
RL	93,412	0.91	1					
RI	93,412	0.87	0.69	1				
RV	93,412	0.76	0.54	0.49	1			
WR	93,412	0.74	0.69	0.73	0.42	1		
WE	93,412	0.74	0.68	0.72	0.42	0.99	1	
WKL	93,412	0.70	0.66	0.69	0.39	0.94	0.89	1

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

Table A.10.1.2 Average Intercorrelations between ELA Grade 5 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	94,841	1						
RL	94,841	0.89	1					
RI	94,841	0.88	0.68	1				
RV	94,841	0.86	0.63	0.64	1			
WR	94,841	0.71	0.61	0.75	0.52	1		
WE	94,841	0.70	0.60	0.74	0.51	0.99	1	
WKL	94,841	0.70	0.60	0.72	0.51	0.95	0.92	1

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

Table A.10.1.3 Average Intercorrelations between ELA Grade 6 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	96,143	1						
RL	96,143	0.91	1					
RI	96,143	0.86	0.65	1				
RV	96,143	0.80	0.61	0.55	1			
WR	96,143	0.75	0.68	0.71	0.53	1		
WE	96,143	0.74	0.67	0.70	0.52	1.00	1	
WKL	96,143	0.74	0.67	0.70	0.52	0.97	0.95	1

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

Table A.10.1.4 Average Intercorrelations between ELA Grade 7 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	97,017	1						
RL	97,017	0.91	1					
RI	97,017	0.90	0.73	1				
RV	97,017	0.82	0.61	0.62	1			
WR	97,017	0.76	0.69	0.77	0.50	1		
WE	97,017	0.74	0.68	0.76	0.48	1.00	1	
WKL	97,017	0.76	0.69	0.76	0.52	0.96	0.93	1

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

Table A.10.1.5 Average Intercorrelations between ELA Grade 8 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	100,115	1						
RL	100,115	0.91	1					
RI	100,115	0.90	0.72	1				
RV	100,115	0.78	0.58	0.59	1			
WR	100,115	0.79	0.71	0.77	0.54	1		
WE	100,115	0.78	0.71	0.76	0.53	1.00	1	
WKL	100,115	0.79	0.71	0.76	0.55	0.97	0.95	1

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

Table A.10.1.6 Average Intercorrelations between ELA Grade 9 Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	97,174	1						
RL	97,174	0.88	1					
RI	97,174	0.91	0.68	1				
RV	97,174	0.79	0.55	0.62	1			
WR	97,174	0.77	0.61	0.80	0.53	1		
WE	97,174	0.76	0.60	0.79	0.52	1.00	1	
WKL	97,174	0.77	0.62	0.79	0.54	0.97	0.95	1

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

Table A.10.1.7 Average Intercorrelations and Reliability between Mathematics Grade 4 Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	95,000	1			
ASC	95,000	0.77	1		
MR	95,000	0.74	0.75	1	
MP	95,000	0.69	0.62	0.64	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.8 Average Intercorrelations and Reliability between Mathematics Grade 5 Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	96,372	1			
ASC	96,372	0.74	1		
MR	96,372	0.68	0.66	1	
MP	96,372	0.77	0.67	0.64	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.9 Average Intercorrelations and Reliability between Mathematics Grade 6 Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	97,653	1			
ASC	97,653	0.78	1		
MR	97,653	0.77	0.75	1	
MP	97,653	0.73	0.67	0.66	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.10 Average Intercorrelations and Reliability between Mathematics Grade 7 Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	93,463	1			
ASC	93,463	0.71	1		
MR	93,463	0.75	0.66	1	
MP	93,463	0.74	0.64	0.66	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.11 Average Intercorrelations and Reliability between Mathematics Grade 8 Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	66,693	1			
ASC	66,693	0.66	1		
MR	66,693	0.64	0.60	1	
MP	66,693	0.60	0.52	0.55	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.12 Average Intercorrelations and Reliability between Algebra I Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	105,535	1			
ASC	105,535	0.71	1		
MR	105,535	0.70	0.66	1	
MP	105,535	0.69	0.61	0.67	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.13 Average Intercorrelations and Reliability between Algebra II Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	10,089	1			
ASC	10,089	0.59	1		
MR	10,089	0.66	0.71	1	
MP	10,089	0.76	0.66	0.70	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

Table A.10.1.14 Average Intercorrelations and Reliability between Geometry Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	30,240	1			
ASC	30,240	0.72	1		
MR	30,240	0.69	0.67	1	
MP	30,240	0.71	0.67	0.65	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice

A.10.2. Quality Testing Standards

Table A.10.2.1 ELA Grade 6 Form 1 Matching Results

ELA Grade 6 Form 1	Unmatched		DIFF*	Matched		DIFF*
	Current Form 1	Original Form 1		Current Form 1	Original Form 1	
Sample Size	119,838	31,031		30,667	30,667	
American Indian/Alaska Native	1.3	0.3	1	0.3	0.3	0
Asian	6.8	6.7	0.1	6.7	6.7	0
Black/African American	14.1	32.8	-18.6	32.2	32.2	0
Hispanic/Latino Ethnicity	31.4	18.9	12.5	19.1	19.1	0
Hawaiian/Pacific Islander	0.2	0.2	0	0.1	0.1	0
White	43.4	36.5	6.9	37	37	0
Two or More Races	2.9	4.7	-1.8	4.7	4.7	0
Female	49.7	49.4	0.3	49.4	49.4	0
Economic Disadvantage	48.3	44.1	4.2	44.5	44.5	0
English Learner	7.2	5.7	1.4	5.6	5.6	0
Students with Disabilities	14.4	13.9	0.5	13.7	13.7	0
Grade 6	100	100	0	100	100	0
Prior Year Scale Score	745	742.3	2.7	742.7	742.7	0
Prior Performance Level 1	10.2	11.7	-1.5	11.4	11.4	0
Prior Performance Level 2	18	19	-1	18.8	18.8	0
Prior Performance Level 3	26.4	26.3	0.1	26.4	26.4	0
Prior Performance Level 4	39.3	38.5	0.9	38.8	38.8	0
Prior Performance Level 5	6.1	4.6	1.5	4.6	4.6	0

*DIFF = Current Percent – Original Percent

Table A.10.2.2 Mathematics Grade 6 Form 1 Matching Results

Mathematics Grade 6 Form 1	Unmatched		DIFF*	Matched		DIFF*
	Current Form 1	Original Form 1		Current Form 1	Original Form 1	
Sample Size	95,174	28,514		27,677	27,677	
American Indian/Alaska Native	1.1	0.2	0.9	0.2	0.2	0
Asian	7.6	7	0.6	7.1	7.1	0
Black/African American	11.5	33.4	-21.9	31.6	31.6	0
Hispanic/Latino Ethnicity	28	17.9	10.1	18.5	18.5	0
Hawaiian/Pacific Islander	0.1	0.2	0	0.1	0.1	0
White	48.4	36.5	11.9	37.6	37.6	0
Two or More Races	3.2	4.8	-1.6	4.9	4.9	0
Female	50.2	50	0.2	50.1	50.1	0
Economic Disadvantage	42.6	42.4	0.3	43.2	43.2	0
English Learner	4.6	3.7	0.9	3.5	3.5	0
Students with Disabilities	9.8	11	-1.2	10.6	10.6	0
Grade 6	100	100	0	100	100	0
Prior Year Scale Score	743.9	741.1	2.8	741.7	741.7	0
Prior Performance Level 1	9	12.6	-3.6	12	12	0
Prior Performance Level 2	18.9	20.3	-1.4	20	20	0
Prior Performance Level 3	28.6	25.6	3	25.8	25.8	0
Prior Performance Level 4	35.7	33.8	1.9	34.3	34.3	0
Prior Performance Level 5	7.8	7.8	0	7.8	7.8	0

*DIFF = Current Percent – Original Percent

Table A.10.2.3 ELA Grade 10 Form 1 Matching Results

ELA Grade 10 Form 1	Unmatched		DIFF*	Matched		DIFF*
	Current Form 1	Original Form 1		Current Form 1	Original Form 1	
Sample Size	55,046	27,951		22,970	22,970	
American Indian/Alaska Native	2	0.3	1.7	0.3	0.3	0
Asian	9.3	7.5	1.8	8.6	8.6	0
Black/African American	11.1	33.2	-22	24.1	24.1	0
Hispanic/Latino Ethnicity	32.1	14.9	17.2	17.5	17.5	0
Hawaiian/Pacific Islander	0.2	0.1	0.1	0.1	0.1	0
White	44	39.5	4.5	46.9	46.9	0
Two or More Races	1.3	4.6	-3.3	2.6	2.6	0
Female	50.2	50.5	-0.2	50.5	50.5	0
Economic Disadvantage	35.8	35	0.9	32.6	32.6	0
English Learner	3.2	3.2	0	2.9	2.9	0
Students with Disabilities	15.6	14.7	0.9	14.4	14.4	0
Grade 9	1.3	3.5	-2.2	1.8	1.8	0
Grade 10	98.6	96.5	2.2	98.2	98.2	0
2017 Scale Score	755.5	740	15.5	746.3	746.2	0.1
2017 Performance Level 1	8.8	15.8	-7	11.3	11.3	0
2017 Performance Level 2	13	18.8	-5.8	15.9	15.9	0
2017 Performance Level 3	21.4	23.7	-2.2	24.5	24.5	0
2017 Performance Level 4	39.6	34	5.6	39.1	39.1	0
2017 Performance Level 5	17.3	7.7	9.5	9.3	9.3	0

*DIFF = Current Percent – Original Percent

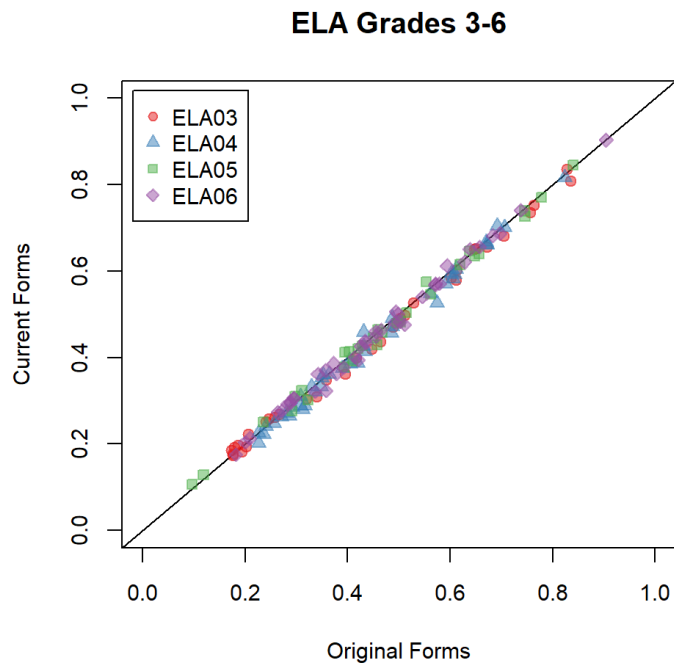


Figure A.10.2.1 ELA Grades 3–6 P-Values

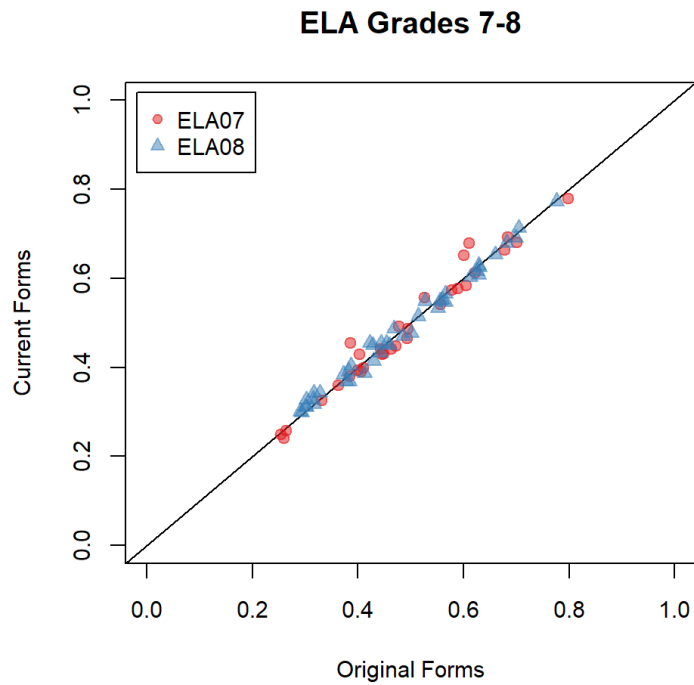


Figure A.10.2.2 ELA Grades 7–8 P-Values

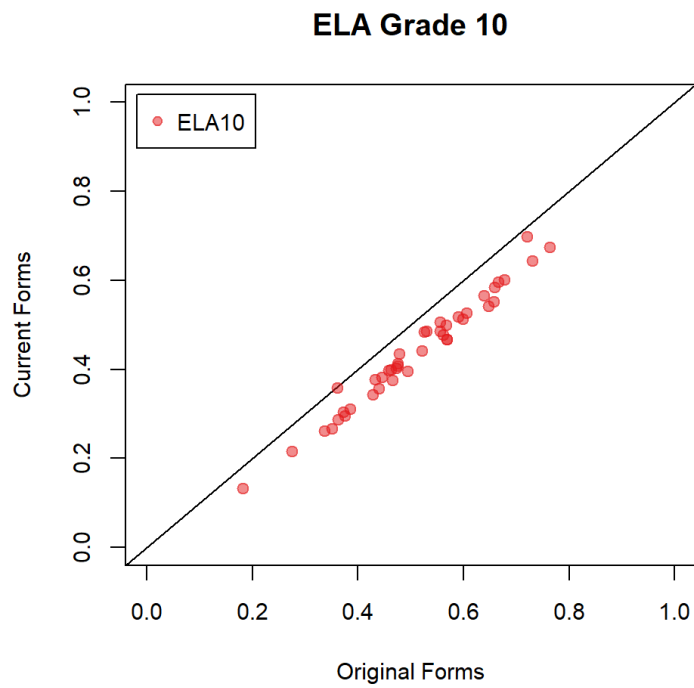


Figure A.10.2.3 ELA Grade 10 P-Values

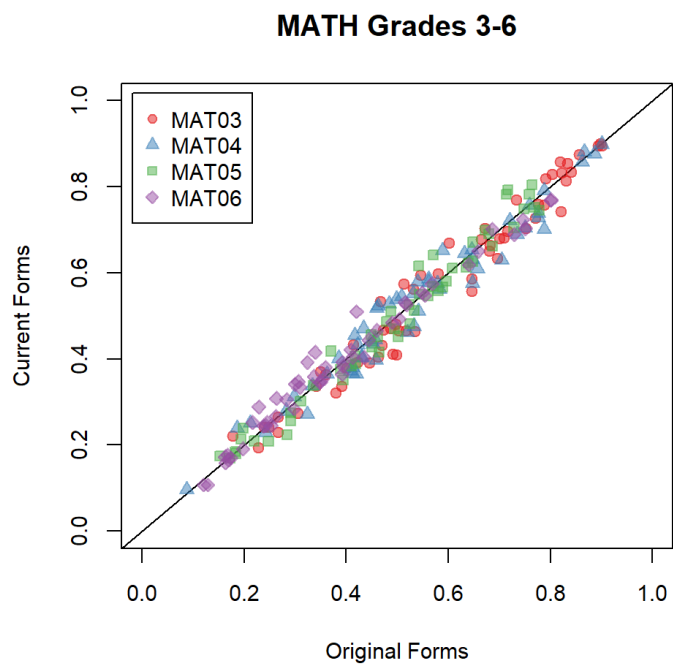


Figure A.10.2.4 Mathematics Grades 3–6 P-Values

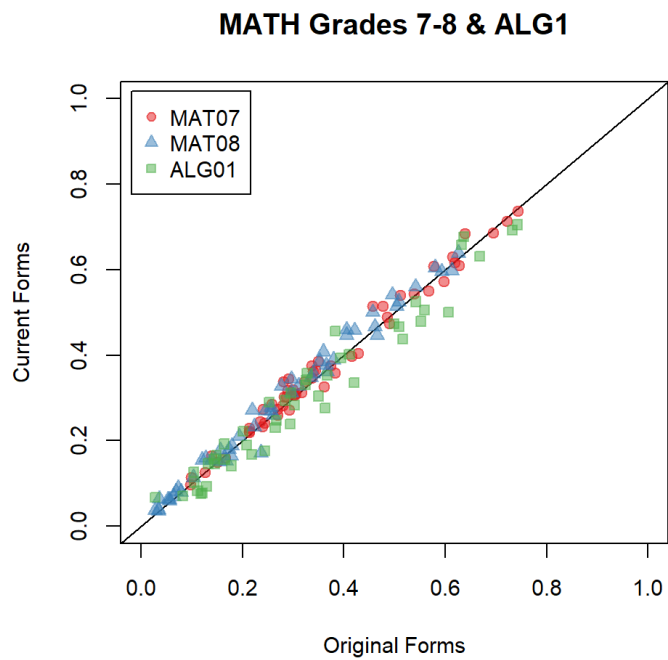


Figure A.10.2.5 Mathematics Grade 7–8 and Algebra I P-Values

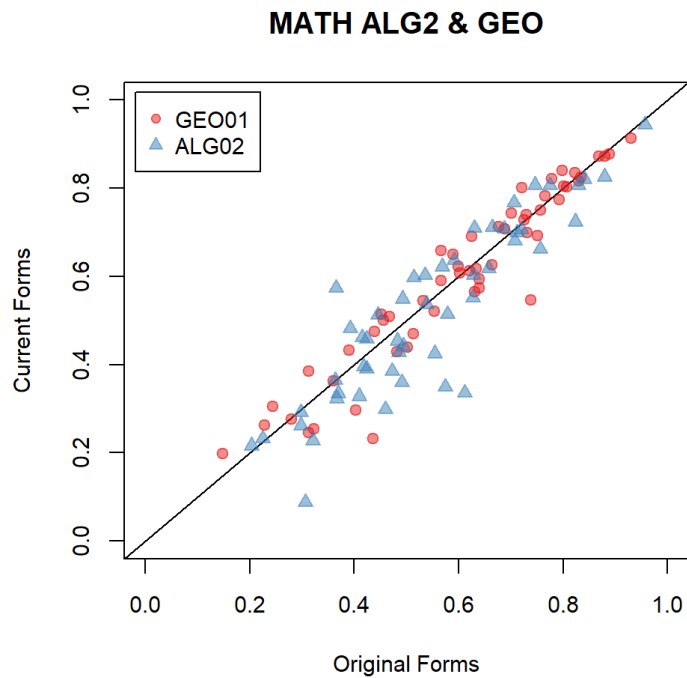


Figure A.10.2.6 Algebra II and Geometry P-Values

Table A.10.2.4 Distributions of P-Value Differences* for ELA

Grade	N	Min	25%	Median	75%	Max
3	34	-0.034	-0.017	-0.01	0.004	0.016
4	42	-0.049	-0.019	-0.01	-0.004	0.028
5	31	-0.029	-0.016	-0.006	0.009	0.021
6	42	-0.035	-0.008	-0.001	0.008	0.02
7	31	-0.026	-0.016	-0.006	0	0.07
8	42	-0.025	-0.01	0	0.011	0.032
10	42	-0.106	-0.085	-0.073	-0.062	-0.003

*Difference = Current p-value – Original p-value

Table A.10.2.5 Distributions of P-Value Differences* for Mathematics

Grade	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
A1	48	-0.105	-0.042	-0.019	0.014	0.073
GO	55	-0.204	-0.031	0.004	0.04	0.094
A2	51	-0.275	-0.062	-0.022	0.04	0.209

*Difference = Current p-value – Original p-value

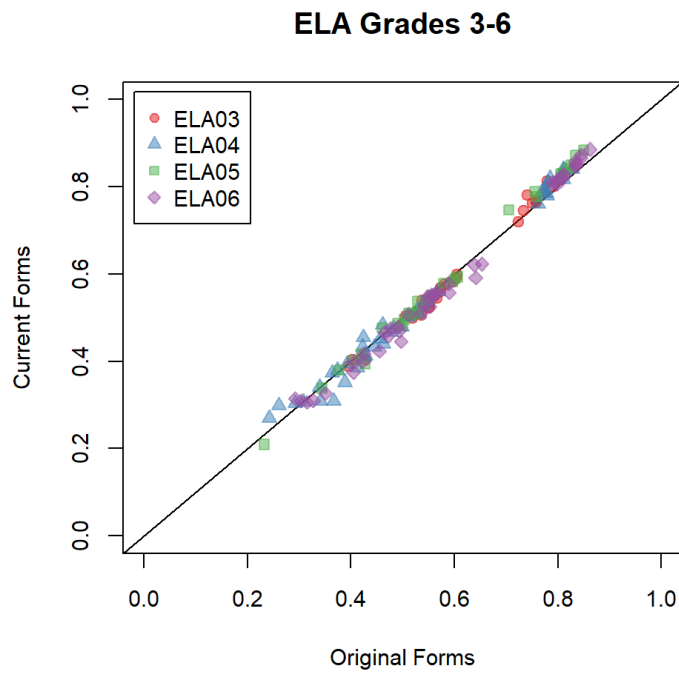


Figure A.10.2.7 Polyserial Correlations ELA Grades 3–6

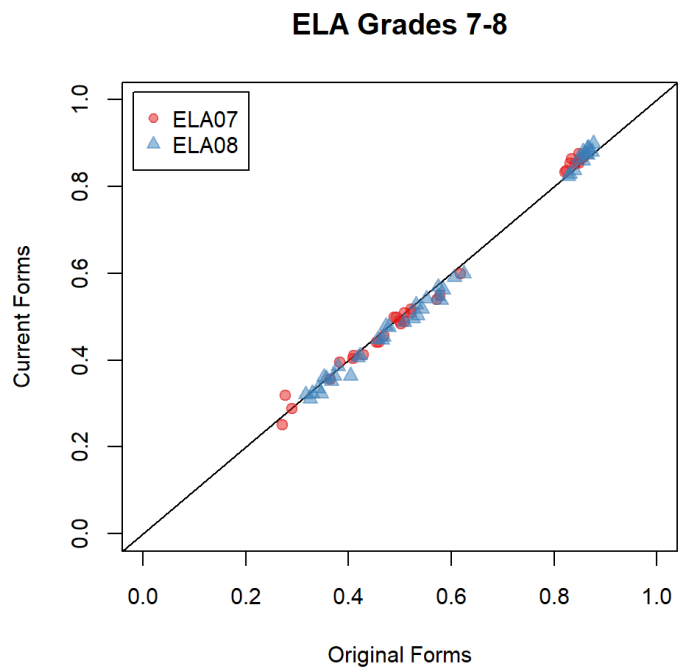


Figure A.10.2.8 Polyserial Correlations ELA Grades 7–8

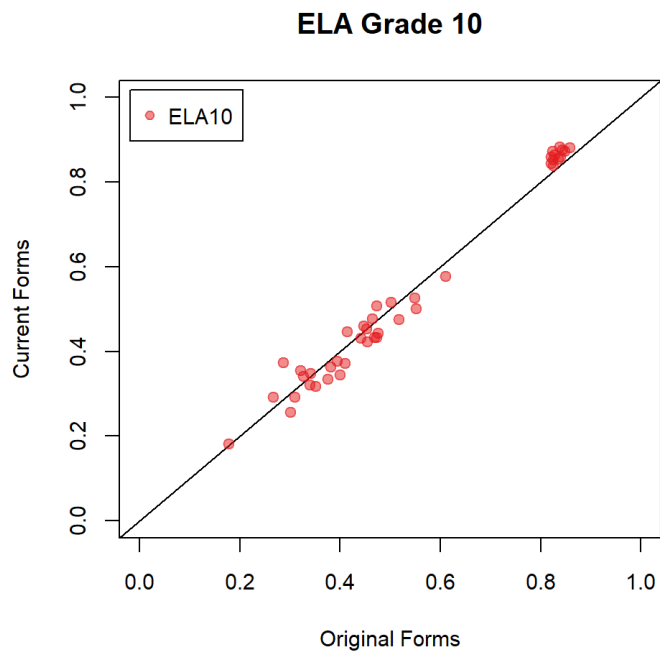


Figure A.10.2.9 Polyserial Correlations ELA Grade 10

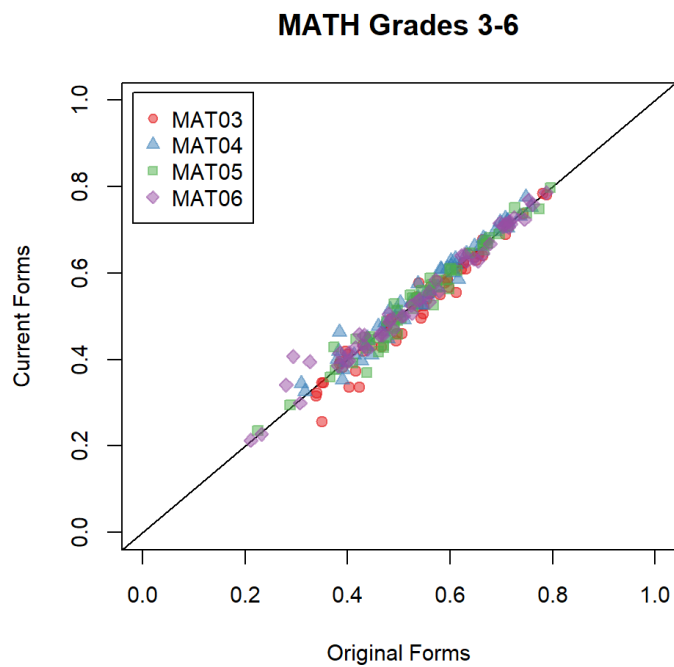


Figure A.10.2.10 Polyserial Correlations Mathematics Grades 3–6

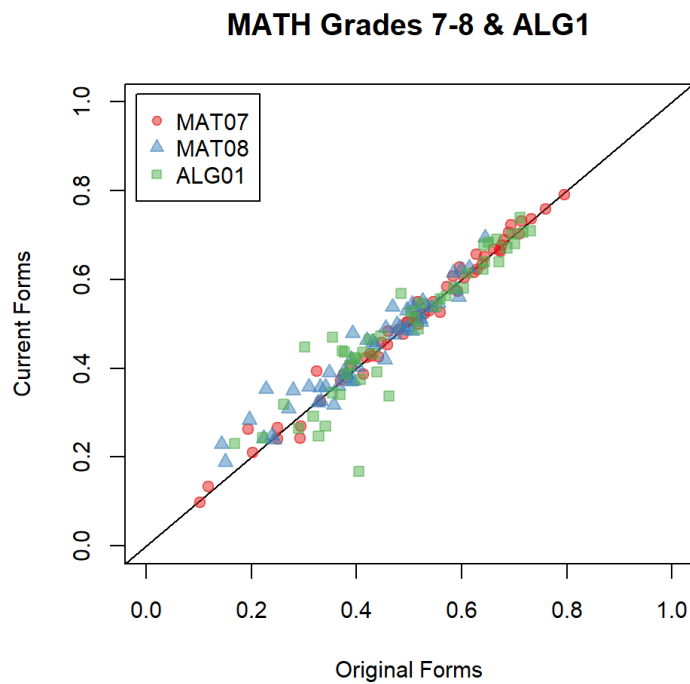


Figure A.10.2.11 Polyserial Correlations Mathematics Grades 7–8 and Algebra I

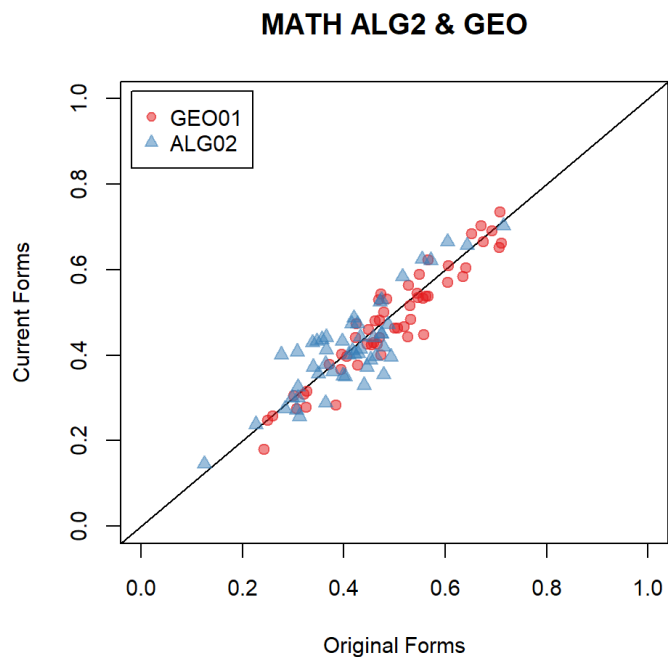


Figure A.10.2.12 Polyserial Correlations Algebra II and Geometry

Table A.10.2.6 Distributions of Polyserial Differences* for ELA

Grade	N	Min	25%	Median	75%	Max
3	34	-0.029	-0.015	-0.004	0.012	0.041
4	42	-0.058	-0.011	0	0.017	0.037
5	31	-0.034	-0.013	-0.003	0.020	0.042
6	42	-0.052	-0.022	-0.008	0.013	0.028
7	31	-0.031	-0.015	0	0.012	0.043
8	42	-0.042	-0.017	-0.007	0.005	0.023
10	42	-0.055	-0.032	0.010	0.026	0.088

*Difference = Current Polyserial – Original Polyserial

Table A.10.2.7 Distributions of Polyserial Differences* for Mathematics

Grade	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
A1	48	-0.238	-0.022	0.001	0.025	0.145
G0	55	-0.108	-0.037	-0.011	0.012	0.072
A2	51	-0.125	-0.025	0.002	0.052	0.125

*Difference = Current Polyserial – Original Polyserial

Table A.10.2.8 DIF Category Crosstabulations for ELA

ELA Grades 3-8 & 10	Percent of DIF Calculations		
	None	B DIF (Current)	C DIF (Current)
None	89.9% – 96.7%	0% – 2.7%	0% – 0.4%
B DIF (Original)	0.6% – 4.8%	1.2% – 2.4%	0%
C DIF (Original)	0% – 0.4%	0% – 1.8%	0% – 1.6%

Table A.10.2.9 DIF Category Crosstabulations for Mathematics Grades 3–8 and Algebra I

Mathematics Grades 3 – 8 & Algebra I	Percent of DIF Calculations		
	None	B DIF (Current)	C DIF (Current)
None	94.5% – 97.3%	0.2% – 2.1%	0% – 0.3%
B DIF (Original)	1.4% – 2.5%	0.2% – 2.2%	0% – 0.5%
C DIF (Original)	0% – 0.5%	0% – 0.5%	0% – 0.2%

Table A.10.2.10 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.10.2.11 ELA Reliability

Grade	Original		Current Form 1				Current Form 2			
	Pts	Alpha**	Pts	Alpha	SB	Diff*	Pts	Alpha	SB	Diff*
3	82	0.92	54	0.90	0.89	0.01	55	0.89	0.89	0
4	106	0.92	74	0.89	0.89	0	67	0.88	0.88	0
5	106	0.93	74	0.89	0.89	0	67	0.88	0.89	-0.01
6	109	0.94	74	0.92	0.92	0	70	0.90	0.90	0
7	109	0.94	74	0.91	0.91	0	70	0.90	0.91	-0.01
8	109	0.94	74	0.92	0.92	0	70	0.90	0.91	-0.01
10	109	0.93	74	0.90	0.89	0.01	70	0.88	0.89	-0.01

*DIFF = Current Alpha – Spearman Brown (SB) Prophecy

**Alpha = Weighted average of the stratified alphas from Original form 1 and Original form 2

Table A.10.2.12 ELA Raw Score Standard Error of Measurement

Grade	Original			Current Form 1			Current Form 2		
	RS Points	RS SEM	SEM/ Points	RS Points	RS SEM	SEM/ Points	RS Points	RS SEM	SEM/ Points
3	82	4.42	0.054	54	3.54	0.066	55	3.58	0.065
4	106	5.41	0.051	74	4.46	0.06	67	4.51	0.067
5	106	5.46	0.052	74	4.48	0.061	67	4.48	0.067
6	109	5.53	0.051	74	4.50	0.061	70	4.49	0.064
7	109	5.93	0.054	74	4.71	0.064	70	5.06	0.072
8	109	5.63	0.052	74	4.52	0.061	70	4.69	0.067
10	109	5.95	0.044	74	4.71	0.05	70	5.20	0.06

Table A.10.2.13 ELA 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

Table A.10.2.14 Mathematics Reliability

Grade	Original		Current Form 1 and Form 2		SB	Diff*
	Points	Alpha**	Points	Alpha**		
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
A1	81	0.93	55	0.90	0.90	0
GO	81	0.93	55	0.89	0.90	-0.01
A2	81	0.89	55	0.84	0.85	-0.01

**Alpha = Weighted average of the stratified alphas from form 1 and form 2

Table A.10.2.15 Mathematics Raw Score Standard Error of Measurement

Grade	Original			Current		
	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
A1	81	3.61	0.045	55	2.88	0.052
GO	81	4.21	0.052	55	3.51	0.064
A2	81	4.25	0.052	55	3.50	0.064

Table A.10.2.16 Mathematics Scale Score Standard Error of Measurement

Grade	Original				Current			
	Form 1		Form 2		Form 1		Form 2	
	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
A1	80	8.9	81	8.7	55	10.8	55	10.4
GO	81	6.4	81	6.4	55	7.9	55	8.0
A2	81	9.7	81	9.8	55	11.4	55	12.2

Table A.10.2.17 ELA Scale Score Descriptive Statistics

Grade	N	Current			Original			Diff*	D
		Mean	Median	SD	Mean	Median	SD		
3	62,753	737.6	739	41.9	739.2	740	42.3	-1.6	-0.04
4	61,139	742.3	742	38.5	744.7	746	37.3	-2.5	-0.06
5	62,463	744.3	743	36.2	744.6	745	35.0	-0.4	-0.01
6	61,173	743.2	744	33.9	742.6	744	32.7	0.6	0.02
7	59,137	746	747	40.8	747.4	749	39.2	-1.4	-0.04
8	58,210	746.6	748	41.5	745.1	746	40.5	1.5	0.04
10	40,163	749	752	46.9	767.1	770	42.7	-18.1	-0.40

*Diff = Current mean – Original mean

Table A.10.2.18 Mathematics Scale Score Descriptive Statistics

Grade	N	Current			Original			Diff*	D
		Mean	Median	SD	Mean	Median	SD		
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
A1	35,083	739.7	739	33.4	743.5	742	32.9	-3.82	-0.12
GO	3,054	773.4	776.5	24.9	772.6	775	24.7	0.81	0.03
A2	1,576	778.2	779	29.6	782.3	782	28.9	-4.09	-0.14

*Diff = Current mean – Original mean

Table A.10.2.19 ELA Writing Claim Score Descriptive Statistics

Grade	N	Current			Original			Diff*	D
		Mean	Median	SD	Mean	Median	SD		
3	62,753	45.3	45	16.8	46.7	47	17.3	-1.4	-0.08
4	61,139	47.2	47	15.5	48.2	48	15.1	-1	-0.07
5	62,463	47.7	47	14.6	48.3	49	14.3	-0.6	-0.04
6	61,173	47.5	47	13.4	47.5	47	13.3	0	0
7	59,137	48.6	49	16.3	49.3	50	16.0	-0.7	-0.04
8	58,210	48.9	48	16.8	48.8	49	16.4	0.1	0.01
10	40,163	49.3	49	18.6	57.2	57	17.8	-7.8	-0.43

*Diff = Current mean – Original mean

Table A.10.2.20 ELA Reading Claim Score Descriptive Statistics

Grade	N	Current			Original			Diff*	D
		Mean	Median	SD	Mean	Median	SD		
3	62,753	29	33	13.5	29.8	32	12.7	-0.8	-0.06
4	61,139	31.6	34	11.7	32.5	34	10.6	-0.9	-0.08
5	62,463	31.0	33	12.6	31.8	33	10.9	-0.8	-0.07
6	61,173	30.5	34	12.4	30.8	33	11.2	-0.3	-0.02
7	59,137	32.4	34	12.4	32.8	35	11.5	-0.4	-0.03
8	58,210	32.0	33	12.9	31.6	34	12.2	0.3	0.03
10	40,163	33.6	35	13.0	37.7	39	11.0	-4.1	-0.34

*Diff = Current mean – Original mean

Table A.10.2.21 ELA Subclaim Distributions

Form	Level	Percent of Students by Subclaim Performance Level				
		RL	RI	RV	WE	WKL
Current	1	45	42.2	44.9	39.5	38.2
	2	26.3	24.7	23.7	27.3	28.3
	3	28.7	33.1	31.4	33.1	33.4
Original	1	44.5	45.6	44.1	41.9	40
	2	25.2	22.4	24.7	25.4	26.1
	3	30.3	32.1	31.2	32.7	33.9
ES	-	0.02	0.04	0.01	0.03	0.03

Table A.10.2.22 Mathematics Subclaim Distributions

Form	Level	Percent of Students by Subclaim Performance Level			
		A (MC)	C (MR)	D (MP)	B (ASC)
Current	1	33.5	36.7	31	33.5
	2	30.5	27.1	26.4	33.9
	3	36	36.1	42.5	32.6
Original	1	32.6	37.5	32.1	33
	2	29	24.4	25.6	28.3
	3	38.4	38.1	42.2	38.7
ES	-	0.03	0.03	0.01	0.07

Table A.10.2.23 ELA Subclaim Distribution Comparison: Effect Size

Grade	Subclaim Distribution Effect Size				
	RL	RI	RV	WE	WKL
3	0.01	0.03	0.10	0.14	0.10
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	0.10	0.08
8	0.02	0.05	0.07	0.03	0.04
10	0.19	0.20	0.15	0.15	0.14

Table A.10.2.24 Mathematics Subclaim Distribution Comparison: Effect Size

Grade	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
A1	0.05	0.11	0.11	0.06
GO	0.03	0.05	0.04	0.02
A2	0.06	0.04	0.16	0.09

Table A.10.2.25 ELA 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

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.10.2.26 ELA Longitudinal Scale Score Comparison: Original to Original

Grade	2018 Original			2019 Original			2019-2018		
	N**	Mean	SD	N**	Mean	SD	DIFF*	SD	D
3	74,206	735.3	43.4	72,606	737.1	42.5	1.8	43	0.04
4	75,608	741.8	37.9	74,281	741.8	38.2	0	38.1	0
5	74,695	740.4	35.4	75,575	741.8	35.9	1.4	35.7	0.04
6	76,094	739.3	33	79,034	740.6	33.1	1.4	33.1	0.04
7	73,574	742.8	39.8	75,398	745.2	39.6	2.3	39.7	0.06
8	72,661	739.6	40.3	72,976	743	40.8	3.3	40.5	0.08
9	3,449	728.5	39.9	3,468	731.7	40.9	3.2	40.4	0.08
10	72,150	744.2	49.4	74,517	747.8	48.6	3.6	49	0.07

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.10.2.27 Mathematics Longitudinal Scale Score Comparison: Original to Current

Grade	2018 Original			2019 Current			2019-2018		
	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
GO	112,873	732.6	27.4	105,676	731.9	29.5	-0.7	28.4	-0.02
A2	20,658	714.8	33.2	21,414	712.4	34.8	-2.4	34.0	-0.07

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.10.2.28 Mathematics Longitudinal Scale Score Comparison: Original to Original

Grade	2018 Original			2019 Original			2019-2018		
	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
GO	7,110	726.1	24.6	6,967	727.5	27.2	1.5	25.9	0.06
A2	2,841	727.6	33.6	2,943	725.5	34.1	-2.2	33.9	-0.06

*DIFF = 2019 Current mean – 2018 Original mean

**All students (not matched samples)

Table A.10.2.29 ELA Longitudinal Regression

Grade (Prior Grade)	Sample Size			R2		
	Original-Current	Original-Original	All	Full	Reduced	Change
4 (3)	251,957	70,459	322,416	0.6486	0.648	0.0007
5 (4)	258,568	71,980	330,548	0.6948	0.6948	0
6 (5)	261,213	69,545	330,758	0.6967	0.6966	0.0001
7 (6)	255,849	70,466	326,315	0.7093	0.709	0.0004
8 (7)	253,432	68,542	321,974	0.7263	0.7261	0.0002
9 (8)	109,156	3,015	112,171	0.7306	0.7306	0.0001
10 (8)	103,001	53,963	156,964	0.6598	0.6338	0.026

Table A.10.2.30 Mathematics Longitudinal Regression

Grade (Prior Grade)	Sample Size			R2		
	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
GO (A1)	92,531	11,117	103,648	0.6769	0.6707	0.0062
A2 (A1, GO)	60,547	4,136	64,683	0.6793	0.6766	0.0027

Table A.10.2.31 ELA Grade 3 Performance Level Comparison

Level	N Count		Percent		DIFF
	Current	Original	Current	Original	
1	12,869	12,533	20.5	20	0.5
2	11,212	10,901	17.9	17.4	0.5
3	13,896	12,699	22.1	20.2	1.9
4	21,847	23,625	34.8	37.6	-2.8
5	2,929	2,995	4.7	4.8	-0.1
Cramer's V Effect Size = .03					

Table A.10.2.32 Mathematics Grade 3 Performance Level Comparison

Level	N Count		Percent		DIFF
	Current	Original	Current	Original	
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.10.2.33 Performance Level Comparison Summary: Effect Sizes

ELA		Mathematics	
Grade	Cramer's V Effect Size	Grade	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	A1	0.09
—	—	GO	0.04
—	—	A2	0.07

Table A.10.2.34 College and Career Readiness Comparison Summary: Effect Sizes

Proportion of Students at or Above the CCR Cut							
ELA				Mathematics			
Grade	Current	Original	Cohen's h^{**}	Grade	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	A1	0.38	0.42	-0.09
—	—	—	—	GO	0.87	0.86	0.03
—	—	—	—	A2	0.86	0.89	-0.09

**Computed as Current proportion – Original proportion

Table A.10.2.35 ELA Classification Accuracy

Grade	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	Cohen's h	Current	Original	Cohen's h
3	0.71	0.75	-0.10	0.90	0.92	-0.05
4	0.68	0.74	-0.13	0.89	0.91	-0.06
5	0.72	0.78	-0.15	0.90	0.92	-0.08
6	0.74	0.79	-0.13	0.91	0.92	-0.06
7	0.71	0.77	-0.13	0.91	0.93	-0.06
8	0.71	0.77	-0.13	0.91	0.93	-0.07
10	0.67	0.77	-0.23	0.90	0.93	-0.10

Table A.10.2.36 ELA Classification Consistency

Grade	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	Cohen's h	Current	Original	Cohen's h
3	0.61	0.66	-0.10	0.86	0.88	-0.06
4	0.57	0.64	-0.15	0.85	0.88	-0.07
5	0.62	0.70	-0.17	0.86	0.89	-0.09
6	0.64	0.71	-0.15	0.87	0.89	-0.08
7	0.60	0.67	-0.15	0.87	0.90	-0.07
8	0.62	0.69	-0.15	0.87	0.90	-0.08
10	0.57	0.69	-0.25	0.86	0.90	-0.12

Table A.10.2.37 Mathematics Classification Accuracy

Grade	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	Cohen's <i>h</i>	Current	Original	Cohen's <i>h</i>
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
A1	0.74	0.79	-0.11	0.91	0.92	-0.06
GO	0.81	0.85	-0.11	0.96	0.96	-0.03
A2	0.82	0.86	-0.1	0.92	0.95	-0.10

Table A.10.2.38 Mathematics Classification Consistency

Grade	Performance Level Classification			College and Career Readiness* Classification		
	Current	Original	<i>h</i>	Current	Original	<i>h</i>
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
A1	0.65	0.70	-0.13	0.87	0.89	-0.07
GO	0.73	0.78	-0.13	0.94	0.94	-0.04
A2	0.74	0.79	-0.12	0.89	0.92	-0.12

Table A.10.2.39 ELA Grade 6 Performance Level Comparison

Level	Original to Current			Original to Original		
	Current States 2018	Current States 2019	DIFF	Original States 2018	Original States 2019	DIFF
1	10.2	11.3	1.1	12.4	12.6	0.2
2	20.1	17.9	-2.2	21.3	18.8	-2.5
3	28	28.5	0.5	27.7	27.5	-0.2
4	33.3	33.8	0.5	32.1	34.3	2.2
5	8.3	8.4	0.1	6.6	6.8	0.2
Cramer's V Effect Size = .03				Cramer's V Effect Size = .03		

Table A.10.2.40 Mathematics Grade 6 Performance Level Comparison

Level	Original to Current			Original to Original		
	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.10.2.41 Performance Level Comparison Summary: Effect Sizes

ELA			Mathematics		
Grade	Original to Current	Original to Original	Grade	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	A1	0.10	0.05
10	0.09	0.04	GO	0.07	0.06
			A2	0.05	0.05

Table A.10.2.42 ELA Reading Claim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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.10.2.43 ELA 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

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.44 ELA 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

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.45 ELA 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

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.46 ELA Reading Vocabulary (RV) Subclaim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
3	10	0.68	8	0.61	0.63	-0.02
4	12	0.61	9	0.56	0.54	0.02
5	15	0.75	11	0.67	0.69	-0.02
6	14	0.72	8	0.58	0.56	-0.02
7	15	0.66	11	0.62	0.59	0.03
8	17	0.69	11	0.53	0.59	-0.06
10	14	0.6	10	0.47	0.52	-0.05

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.47 ELA Writing Knowledge and Conventions (WKL) Subclaim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
3	9	0.87	6	0.82	0.82	0
4	9	0.88	6	0.84	0.83	0.01
5	9	0.88	6	0.84	0.83	0.01
6	9	0.89	6	0.85	0.84	0.01
7	9	0.89	6	0.86	0.84	0.02
8	9	0.91	6	0.87	0.87	0
10	9	0.89	6	0.86	0.84	0.02

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.48 ELA Written Expression (WE) Subclaim Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
3	27	0.81	18	0.74	0.74	0
4	33	0.83	22	0.77	0.76	0.01
5	33	0.81	23	0.72	0.75	-0.03
6	36	0.86	24	0.81	0.8	0.01
7	36	0.88	24	0.85	0.83	0.02
8	36	0.9	24	0.86	0.86	0
10	36	0.88	24	0.85	0.83	0.02

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.49 Mathematics Subclaim A Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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
A1	26	0.79	17	0.72	0.71	0.01
GO	30	0.84	18	0.79	0.76	0.03
A2	25	0.74	16	0.66	0.65	0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.50 Mathematics Subclaim B Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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
A1	17	0.73	9	0.64	0.59	0.05
GO	19	0.79	12	0.65	0.7	-0.05
A2	20	0.7	12	0.55	0.58	-0.03

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.51 Mathematics Subclaim C Reliability

Grade	Original		Current		SB	Diff*
	Points	Alpha	Points	Alpha		
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
A1	14	0.75	10	0.7	0.68	0.02
GO	14	0.64	10	0.6	0.56	0.04
A2	14	0.55	10	0.44	0.47	-0.03

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Table A.10.2.52 Mathematics Subclaim D Reliability

Grade	Original		Current		SB	Diff*
	Pts.	Alpha	Pts.	Alpha		
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	—	—
A1	18	0.75	15	0.69	0.71	-0.02
GO	18	0.7	15	0.64	0.66	-0.02
A2	18	0.59	15	0.56	0.55	0.01

*Diff: Current Alpha – Spearman Brown (SB) Prophecy

Appendix 11

Appendix 11 provides the summary growth results for subgroups for grades 4 through 9 ELA and mathematics 4 through 8 and high school.

Table A.11.1 Summary of Student Growth Percentile Estimates for Subgroups: Grade 4 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	44,622	50.4	13.5	51
Female	43,404	49.5	13.4	49
Ethnicity				
White	34,374	52.0	13.4	53
African American	12,521	45.4	13.7	44
Asian	9,461	58.2	13.3	61
Pacific Islander	149	50.0	13.5	50
American Indian/Alaska Native	177	52.9	13.6	55
Hispanic	28,301	46.4	13.6	45
Multiple	3,038	52.3	13.5	54
Special Instruction Needs				
Economically Disadvantaged	32,481	45.5	13.6	44
Not-economically Disadvantaged	55,549	52.5	13.4	54
English Learner (EL)	7,995	44.2	14.0	42
Non-English Learner	80,036	50.5	13.4	51
Students with Disabilities (SWD)	15,203	44.0	14.0	42
Students without Disabilities	69,183	51.3	13.4	52

Table A.11.2 Summary of Student Growth Percentile Estimates for Subgroups: Grade 5 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,516	47.6	14.1	47
Female	44,115	52.4	13.8	53
Ethnicity				
White	34,982	51.3	13.8	52
African American	13,190	45.3	14.2	43
Asian	9,919	56.4	13.7	59
Pacific Islander	164	50.4	13.7	48.5
American Indian/Alaska Native	147	52.6	13.1	57
Hispanic	28,323	48.1	14.1	47
Multiple	2,902	51.1	13.7	52
Special Instruction Needs				
Economically Disadvantaged	33,088	46.7	14.1	45
Not-economically Disadvantaged	56,546	51.9	13.8	53
English Learner (EL)	5,943	44.7	14.9	42
Non-English Learner	83,691	50.3	13.9	50
Students with Disabilities (SWD)	15,445	43.3	14.8	41
Students without Disabilities	70,245	51.6	13.8	52

Table A.11.3 Summary of Student Growth Percentile Estimates for Subgroups: Grade 6 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,385	48.9	13.2	48
Female	44,444	51.1	13.2	52
Ethnicity				
White	36,204	51.0	13.2	51
African American	12,920	45.7	13.2	44
Asian	9,833	57.9	13.4	61
Pacific Islander	172	50.5	13.4	50.5
American Indian/Alaska Native	153	51.0	13.2	49
Hispanic	28,830	47.8	13.1	47
Multiple	2,708	49.9	13.2	50
Special Instruction Needs				
Economically Disadvantaged	32,582	47.2	13.2	46
Not-economically Disadvantaged	58,262	51.5	13.2	52
English Learner (EL)	4,698	44.2	13.9	42
Non-English Learner	86,148	50.3	13.2	50
Students with Disabilities (SWD)	15,226	41.4	13.5	38
Students without Disabilities	71,413	51.9	13.2	53

Table A.11.4 Summary of Student Growth Percentile Estimates for Subgroups: Grade 7 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,986	49.1	13.0	49
Female	44,904	50.9	13.1	51
Ethnicity				
White	36,929	50.2	13.1	50
African American	13,399	46.9	13.0	46
Asian	10,038	59.4	13.5	63
Pacific Islander	171	46.5	13.7	47
American Indian/Alaska Native	144	52.3	12.3	55.5
Hispanic	28,581	47.6	12.9	47
Multiple	2,660	51.3	12.9	53
Special Instruction Needs				
Economically Disadvantaged	32,288	47.3	12.9	46
Not-economically Disadvantaged	59,644	51.4	13.2	52
English Learner (EL)	4,503	43.7	13.1	41
Non-English Learner	87,431	50.3	13.1	50
Students with Disabilities (SWD)	15,049	42.3	13.1	39
Students without Disabilities	72,533	51.5	13.1	52

Table A.11.5 Summary of Student Growth Percentile Estimates for Subgroups: Grade 8 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	48,934	47.4	13.5	46
Female	46,039	52.6	13.6	54
Ethnicity				
White	38,472	49.7	13.6	49
African American	14,055	48.2	13.3	48
Asian	10,169	56.6	14.3	59
Pacific Islander	195	50.7	13.4	49
American Indian/Alaska Native	139	53.1	14.3	54
Hispanic	29,436	48.8	13.4	49
Multiple	2,535	50.4	13.5	50
Special Instruction Needs				
Economically Disadvantaged	32,784	48.6	13.3	48
Not-economically Disadvantaged	62,249	50.7	13.7	51
English Learner (EL)	3,988	46.8	14.0	45
Non-English Learner	91,045	50.1	13.5	50
Students with Disabilities (SWD)	14,875	42.5	13.5	39
Students without Disabilities	75,207	51.6	13.6	52

Table A.11.6 Summary of Student Growth Percentile Estimates for Subgroups: Grade 9 ELA

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,310	48.5	13.4	48
Female	43,359	51.4	13.5	52
Ethnicity				
White	35,925	49.7	13.4	50
African American	12,710	47.3	13.2	46.5
Asian	9,759	58.9	14.2	62
Pacific Islander	155	52.9	14.7	57
American Indian/Alaska Native	140	50.6	14.0	50.5
Hispanic	27,843	48.2	13.2	47
Multiple	2,219	50.3	13.4	51
Special Instruction Needs				
Economically Disadvantaged	29,459	47.4	13.2	46
Not-economically Disadvantaged	59,319	51.2	13.5	52
English Learner (EL)	3,353	42.4	14.1	39
Non-English Learner	85,425	50.2	13.4	50
Students with Disabilities (SWD)	13,497	40.7	13.5	37
Students without Disabilities	70,818	51.8	13.4	53

Table A.11.7 Summary of Student Growth Percentile Estimates for Subgroups: Grade 4 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	45,198	50.2	13.7	50
Female	43,951	49.7	13.7	49
Ethnicity				
White	34,485	51.4	13.7	52
African American	12,534	45.5	13.7	44
Asian	9,525	57.5	14.2	61
Pacific Islander	152	54.9	14.3	52.5
American Indian/Alaska Native	177	51.0	13.4	48
Hispanic	29,231	47.5	13.6	46
Multiple	3,041	51.6	13.9	52
Special Instruction Needs				
Economically Disadvantaged	33,167	46.7	13.7	45
Not-economically Disadvantaged	55,987	51.9	13.8	53
English Learner (EL)	9,161	46.5	13.7	45
Non-English Learner	79,994	50.3	13.7	50
Students with Disabilities (SWD)	15,209	44.4	13.7	42
Students without Disabilities	70,310	51.1	13.8	52

Table A.11.8 Summary of Student Growth Percentile Estimates for Subgroups: Grade 5 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,103	50.1	13.8	50
Female	44,604	49.7	13.9	50
Ethnicity				
White	35,081	50.7	13.7	51
African American	13,176	44.9	14.3	43
Asian	10,003	58.5	13.5	62
Pacific Islander	166	56.5	14.1	57
American Indian/Alaska Native	151	51.6	13.1	54
Hispanic	29,226	48.2	14.0	48
Multiple	2,904	50.7	13.7	51
Special Instruction Needs				
Economically Disadvantaged	33,761	47.1	14.1	46
Not-economically Disadvantaged	56,953	51.6	13.7	52
English Learner (EL)	7,092	47.9	14.4	47
Non-English Learner	83,622	50.1	13.8	50
Students with Disabilities (SWD)	15,415	47.1	14.4	46
Students without Disabilities	71,362	50.6	13.8	51

Table A.11.9 Summary of Student Growth Percentile Estimates for Subgroups: Grade 6 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	46,860	50.0	14.5	50
Female	44,948	49.8	14.6	50
Ethnicity				
White	36,253	50.6	14.1	51
African American	12,926	45.6	15.5	44
Asian	9,886	57.1	13.7	60
Pacific Islander	173	51.5	14.1	52
American Indian/Alaska Native	152	49.7	14.6	49.5
Hispanic	29,704	48.6	15.0	48
Multiple	2,707	49.2	14.4	48
Special Instruction Needs				
Economically Disadvantaged	33,263	48.1	15.2	47
Not-economically Disadvantaged	58,561	50.9	14.2	51
English Learner (EL)	5,835	45.2	15.9	43
Non-English Learner	85,991	50.2	14.5	50
Students with Disabilities (SWD)	15,175	43.2	15.7	40
Students without Disabilities	72,459	51.3	14.4	52

Table A.11.10 Summary of Student Growth Percentile Estimates for Subgroups: Grade 7 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	44,730	52.0	14.3	53
Female	43,130	47.7	14.4	47
Ethnicity				
White	35,303	51.5	13.9	52
African American	13,135	46.8	15.2	46
Asian	7,801	57.2	13.5	61
Pacific Islander	157	50.8	14.1	52
American Indian/Alaska Native	137	49.0	14.5	43
Hispanic	28,895	47.2	14.8	46
Multiple	2,458	51.9	14.1	54
Special Instruction Needs				
Economically Disadvantaged	32,346	47.0	14.9	46
Not-economically Disadvantaged	55,552	51.6	14.0	52
English Learner (EL)	5,660	45.3	15.7	43
Non-English Learner	82,240	50.2	14.3	50
Students with Disabilities (SWD)	14,897	42.4	15.4	39
Students without Disabilities	68,819	51.5	14.2	52

Table A.11.11 Summary of Student Growth Percentile Estimates for Subgroups: Grade 8 Mathematics

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	32,280	48.2	16.1	48
Female	29,734	51.6	16.1	52
Ethnicity				
White	23,254	52.7	15.2	54
African American	11,053	45.8	17.4	44
Asian	3,068	57.0	14.3	60
Pacific Islander	117	57.0	16.0	62
American Indian/Alaska Native	98	46.6	17.1	43.5
Hispanic	22,904	47.8	16.7	47
Multiple	1,534	51.3	15.6	53
Special Instruction Needs				
Economically Disadvantaged	26,004	47.1	16.9	46
Not-economically Disadvantaged	36,052	51.8	15.5	53
English Learner (EL)	4,257	45.4	18.4	43
Non-English Learner	57,799	50.2	16.0	50
Students with Disabilities (SWD)	13,473	46.5	17.9	45
Students without Disabilities	45,256	50.7	15.6	51

Table A.11.12 Summary of Student Growth Percentile Estimates for Subgroups: Algebra I

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	44,898	50.6	15.9	51
Female	42,446	49.1	16.3	49
Ethnicity				
White	34,565	51.5	15.4	52
African American	12,378	46.6	17.8	45
Asian	9,864	56.6	13.6	59
Pacific Islander	179	51.0	16.2	49
American Indian/Alaska Native	130	49.8	17.0	52.5
Hispanic	28,084	46.7	17.2	45
Multiple	2,206	51.0	15.3	51.5
Special Instruction Needs				
Economically Disadvantaged	29,372	46.7	17.5	45
Not-economically Disadvantaged	58,052	51.4	15.4	52
English Learner (EL)	4,652	42.1	18.8	40
Non-English Learner	82,772	50.3	16.0	50
Students with Disabilities (SWD)	11,707	43.8	18.5	42
Students without Disabilities	71,421	50.8	15.7	51

Table A.11.13 Summary of Student Growth Percentile Estimates for Subgroups: Geometry

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	13,052	50.0	13.6	50
Female	13,142	49.8	13.9	50
Ethnicity				
White	12,408	49.6	13.3	49
African American	2,226	43.8	15.6	42
Asian	5,367	57.8	12.8	61
Pacific Islander	—	—	—	—
American Indian/Alaska Native	—	—	—	—
Hispanic	5,361	44.9	15.0	43
Multiple	765	51.9	13.3	53
Special Instruction Needs				
Economically Disadvantaged	5,087	44.7	15.2	43
Not-economically Disadvantaged	21,139	51.2	13.4	52
English Learner (EL)	514	45.6	17.5	44
Non-English Learner	25,712	50.0	13.7	50
Students with Disabilities (SWD)	1,228	42.0	16.5	39
Students without Disabilities	23,738	50.4	13.6	51

Note. “—” indicates insufficient sample for student growth percentile calculation for these tests.

Table A.11.14 Summary of Student Growth Percentile Estimates for Subgroups: Algebra II

	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Gender				
Male	4,376	51.5	14.2	52
Female	4,136	48.0	14.6	47
Ethnicity				
White	2,804	46.7	13.8	45
African American	487	41.7	15.6	39
Asian	3,354	55.6	13.6	58
Pacific Islander	—	—	—	—
American Indian/Alaska Native	—	—	—	—
Hispanic	1,597	45.5	16.6	44
Multiple	255	49.9	14.0	50
Special Instruction Needs				
Economically Disadvantaged	1,593	44.4	16.8	42
Not-economically Disadvantaged	6,928	51.1	13.8	52
English Learner (EL)	263	42.6	19.4	39
Non-English Learner	8,258	50.0	14.2	50
Students with Disabilities (SWD)	294	41.3	17.0	37.5
Students without Disabilities	7,915	50.4	14.3	51

Note. “—” indicates insufficient sample for student growth percentile calculation for these tests.