



Illinois Assessment of Readiness (IAR)

Technical Report

2024–2025

Prepared by Pearson for the Illinois State Board of Education (ISBE)
December 1, 2025

Table of Contents

Section 1: Introduction	9
1.1. Assessment Overview	9
1.2. Background	9
1.3. Student Participation	10
1.4. Organizations and Groups Involved.....	10
Section 2: Test Design	12
2.1. Claims and Subclaims.....	12
2.2. Test Blueprints	13
2.3. Item Types	14
2.4. Test Sections and Testing Times	14
2.5. Test Structure	18
Section 3: Test Development	20
3.1. Asset Development Plan	20
3.2. Passage Selection.....	20
3.3. Item Development	20
3.4. Form Construction	21
3.4.1. Operational Forms	21
3.4.2. Field Test Forms	21
Accommodated Forms	22
3.5. Data Review	22
Section 4: Test Administration.....	24
4.1. Accessibility Features and Accommodations	24
4.2. Test Security	26
4.3. Testing Irregularities and Security Breaches	27
Section 5: Scoring.....	29
5.1. Machine Scoring	29
5.2. Human Scoring of Constructed-Response Items	30
5.2.1. Scorer Training	31
5.2.2. Scorer Qualification.....	32
5.2.3. Scorer Monitoring	33
5.2.3.1. Second Scoring	33
5.2.3.2. Backreading	33
5.2.3.3. Validity Responses.....	33
5.2.3.4. Calibration Sets.....	34
5.3. Automated Scoring of PCR Items.....	34
5.3.1. Sampling Responses Used for Training IEA.....	35
5.3.2. Quality Criteria for Evaluating IEA Performance.....	35
5.4. Inter-Rater Agreement	36
5.5. Hierarchy of Assigned Scores for Reporting	38
Section 6: Reporting.....	39
6.1. Available Reports	39
6.2. Interpretation of Test Scores	39
6.2.1. Total Scale Scores and Performance Levels	39
6.2.2. Claim and Subclaim Scores.....	40
6.2.3. Additional Measures	40
Section 7: Standard Setting.....	41

7.1. Standard Setting Process	41
7.2. Cut Scores	43
Section 8: Student Characteristics and Test Results	44
8.1. Student Participation	44
8.2. Scale Score Distributions	46
Section 9: Classical Item Analysis.....	51
9.1. Data Preparation.....	51
9.2. Item Analyses.....	51
9.2.1. Item Difficulty (P-value)	52
9.2.2. Item Discrimination (Item-Total Correlation)	52
9.2.3. Percentage of Students Choosing Each Answer Option	53
9.2.4. Percentage of Students Omitting or Not Reaching Each Item	53
9.2.5. Distribution of Item Scores	54
9.3. Flagging Criteria	54
Section 10: Differential Item Functioning (DIF)	55
10.1. DIF Methods	55
10.2. Classification	56
10.3. Comparisons	56
10.4. Results.....	57
Section 11: Calibration, Equating, and Scaling.....	58
11.1. IRT Model.....	58
11.2. IRT Analysis Results.....	58
11.3. Establishing the Reporting Scale.....	59
11.3.1. Summative Score Scale and Performance Levels.....	59
11.3.2. Reading and Writing Claim Scale	61
11.3.3. Subclaims Scale	61
11.4. Types of Scores on the IAR Individual Student Report	61
11.4.1. Scale Score	61
11.4.2. Performance Level	62
11.4.3. Subclaim Performance Indicators	62
11.4.4. Conversion Tables	62
11.4.5. Scaling Constants	64
Section 12: Quality Control Procedures.....	66
12.1. Quality Control of the Item Bank.....	66
12.2. Quality Co	66
12.3. Control of Test Form Development	66
12.4. Quality Control of Test Materials	66
12.5. Quality Control of Scoring.....	67
12.5.1. Quality Control of Scanning	67
12.5.2. Quality Control of Image Editing.....	68
12.5.3. Quality Control of Answer Document and Data	68
12.6. Quality Control of Psychometric Processes	69
Section 13: Reliability.....	71
13.1. Internal Consistency and SEM	71
13.1.1. Raw Score Estimation.....	72
13.1.2. Scale Score Estimation	72
13.1.3. Results	73
13.2. Decision Accuracy and Consistency.....	76

Section 14: Validity	78
14.1. Evidence Based on Test Content	78
14.2. Evidence Based on Internal Structure	79
14.2.1. Intercorrelations	79
14.2.2. Reliability.....	82
14.2.3. Local Item Independence.....	82
14.3. Evidence Based on Relationships to Other Variables	82
14.4. Evidence from Special Studies	83
14.4.1. Content Alignment Studies	83
14.4.2. Mode and Device Comparability Studies	85
14.4.3. Alternate Blueprint Study	86
14.5. Evidence Based on Response Processes.....	88
14.6. Evidence Based on the Consequences to Testing.....	89
14.7. Summary.....	89
Section 15: Student Growth Measures.....	90
15.1. Norm Groups	90
15.2. SGP Estimation.....	91
15.3. SGP Results	92
15.3.1. Summary for Total Group	92
15.3.2. Subgroups of Interest.....	92
References	94
Appendix A: Scale Score Cumulative Frequencies	98
Appendix B: Scale Score Performance by Demographic Subgroup	115
Appendix C: Differential Item Functioning (DIF) Results	127
Appendix D: TCCs, CSEM Curves, and TIF Curves	134
Appendix E: Reliability by Subgroup	146
Appendix F: Decision Accuracy and Consistency by Performance Level	158
Appendix G: Student Growth Percentile (SGP) Estimates by Subgroup	161

List of Tables

Table 1.1. Organizations and Groups Involved	11
Table 2.1. Claims and Subclaims—ELA/L	12
Table 2.2. Claims and Subclaims—Mathematics	12
Table 2.3. High-Level Blueprint—ELA/L	13
Table 2.4. High-Level Blueprint—Mathematics	13
Table 2.5. Comparison between State Population and the Sample	16
Table 2.6. Test Sections and Testing Times	17
Table 2.7. Contribution of PCR Items in ELA/L: Number of Possible Points by Task	18
Table 2.8. Contribution in Item Types in Math: Number of items by Type	19
Table 3.1. Number of Test Forms Constructed in Spring 2025	21
Table 3.2. Number of Items and Field Test Forms for ELA	22
Table 3.3. Supported Accommodations.....	22
Table 3.4. 2025 Data Review Results	23
Table 4.1. Test Administration Activities	24
Table 4.2. Test Irregularity and Security Breach Examples.....	27
Table 5.1. Scoring Training Materials.....	31
Table 5.2. Mathematics Scorer Qualification Requirements.....	33

Table 5.3. Scoring Validity Agreement Requirements	34
Table 5.4. Inter-Rater Agreement Expectations and Spring 2025 Results	37
Table 5.5. ELA/L PCR Average Agreement Indices	37
Table 5.8. Scoring Hierarchy Rules	38
Table 7.1. Scale Score Ranges and Cut Scores	43
Table 8.1. Student Participation by Administration Mode	44
Table 8.2. Student Participation by Demographic Subgroup—ELA/L Grades 3-5	44
Table 8.3. Student Participation by Demographic Subgroup—ELA/L Grades 6-8	45
Table 8.4. Student Participation by Demographic Subgroup— Mathematics Grades 3-5	45
Table 8.5. Student Participation by Demographic Subgroup— Mathematics Grades 6-8	45
Table 9.1. Summary of <i>p</i> -Values	52
Table 9.2. Summary of Item-Total Correlations.....	53
Table 10.1. DIF Categories	56
Table 10.2. DIF Comparison Groups	57
Table 11.1. Pre-Equated IRT Parameter Estimates Summary—ELA/L.....	58
Table 11.2. Pre-Equated IRT Parameter Estimates Summary—Mathematics.....	59
Table 11.3. Calculating Scaling Constants for Reading and Writing Claim Scores	61
Table 11.4. Cut Scores and Scaling Constants—ELA/L.....	64
Table 11.5. Cut Scores and Scaling Constants—Reading and Writing	64
Table 11.6. Cut Scores and Scaling Constants—Mathematics.....	65
Table 13.1. Summary of Raw Score Test Reliability for Total Group	74
Table 13.2. Summary of Scale Score Test Reliability for Total Group.....	74
Table 13.3. Average Reliability Estimates by Subclaim—ELA/L	75
Table 13.4. Average Reliability Estimates by Subclaim—Mathematics.....	76
Table 13.5. Decision Accuracy and Consistency Summary	77
Table 14.1. Average Interrelations and Reliability between Subclaims—ELA/L.....	80
Table 14.2. Average Interrelations and Reliability between Subclaims—Mathematics.....	81
Table 14.3. Correlations between ELA/L and Mathematics	83
Table 14.4. 2019 Alternate Blueprint Study: Prior Grades used in Matching.....	87
Table 14.5. 2019 Alternate Blueprint Study: Matching Sample Size Results.....	87
Table 15.1. SGP Grade-Level Progressions for One-Year Prior Scores.....	90
Table 15.2. Summary of SGP Estimates for Total Group	92
Table A.1. Scale Score Cumulative Frequencies—ELA/L Grade 3	98
Table A.2. Scale Score Cumulative Frequencies—ELA/L Grade 3 Reading	99
Table A.3. Scale Score Cumulative Frequencies—ELA/L Grade 3 Writing	99
Table A.4. Scale Score Cumulative Frequencies—ELA/L Grade 4	99
Table A.5. Scale Score Cumulative Frequencies—ELA/L Grade 4 Reading	100
Table A.6. Scale Score Cumulative Frequencies—ELA/L Grade 4 Writing	101
Table A.7. Scale Score Cumulative Frequencies—ELA/L Grade 5	101
Table A.8. Scale Score Cumulative Frequencies—ELA/L Grade 5 Reading	102
Table A.9. Scale Score Cumulative Frequencies—ELA/L Grade 5 Writing	102
Table A.10. Scale Score Cumulative Frequencies—ELA/L Grade 6	103
Table A.11. Scale Score Cumulative Frequencies—ELA/L Grade 6 Reading	104
Table A.12. Scale Score Cumulative Frequencies—ELA/L Grade 6 Writing	104
Table A.13. Scale Score Cumulative Frequencies—ELA/L Grade 7	104
Table A.14. Scale Score Cumulative Frequencies—ELA/L Grade 7 Reading	105
Table A.15. Scale Score Cumulative Frequencies—ELA/L Grade 7 Writing	106
Table A.16. Scale Score Cumulative Frequencies—ELA/L Grade 8	106

Table A.17. Scale Score Cumulative Frequencies—ELA/L Grade 8 Reading	107
Table A.18. Scale Score Cumulative Frequencies—ELA/L Grade 8 Writing	108
Table A.19. Scale Score Cumulative Frequencies—Mathematics Grade 3	108
Table A.20. Scale Score Cumulative Frequencies—Mathematics Grade 4	109
Table A.21. Scale Score Cumulative Frequencies—Mathematics Grade 5	110
Table A.22. Scale Score Cumulative Frequencies—Mathematics Grade 6	111
Table A.23. Scale Score Cumulative Frequencies—Mathematics Grade 7	112
Table A.24. Scale Score Cumulative Frequencies—Mathematics Grade 8	113
Table B.1. Scale Score Performance by Demographic Subgroup—ELA/L Grade 3	115
Table B.2. Scale Score Performance by Demographic Subgroup—ELA/L Grade 4	116
Table B.3. Scale Score Performance by Demographic Subgroup—ELA/L Grade 5	118
Table B.4. Scale Score Performance by Demographic Subgroup—ELA/L Grade 6	119
Table B.5. Scale Score Performance by Demographic Subgroup—ELA/L Grade 7	121
Table B.6. Scale Score Performance by Demographic Subgroup—ELA/L Grade 8	122
Table B.7. Scale Score Performance by Demographic Subgroup—Mathematics Grade 3	124
Table B.8. Scale Score Performance by Demographic Subgroup—Mathematics Grade 4	124
Table B.9. Scale Score Performance by Demographic Subgroup—Mathematics Grade 5	125
Table B.10. Scale Score Performance by Demographic Subgroup—Mathematics Grade 6	125
Table B.11. Scale Score Performance by Demographic Subgroup—Mathematics Grade 7	126
Table B.12. Scale Score Performance by Demographic Subgroup—Mathematics Grade 8	126
Table C.1. Pre-Administration DIF Results—ELA/L Grade 3	127
Table C.2. Pre-Administration DIF Results—ELA/L Grade 4	128
Table C.3. Pre-Administration DIF Results—ELA/L Grade 5	128
Table C.4. Pre-Administration DIF Results—ELA/L Grade 6	129
Table C.5. Pre-Administration DIF Results—ELA/L Grade 7	129
Table C.6. Pre-Administration DIF Results—ELA/L Grade 8	130
Table C.7. Pre-Administration DIF Results—Mathematics Grade 3	130
Table C.8. Pre-Administration DIF Results—Mathematics Grade 4	131
Table C.9. Pre-Administration DIF Results—Mathematics Grade 5	131
Table C.10. Pre-Administration DIF Results—Mathematics Grade 6	132
Table C.11. Pre-Administration DIF Results—Mathematics Grade 7	132
Table C.12. Pre-Administration DIF Results—Mathematics Grade 8	133
Table E.1. Test Reliability Estimates by Subgroup—ELA/L Grade 3	146
Table E.2. Test Reliability Estimates by Subgroup—ELA/L Grade 4	147
Table E.3. Test Reliability Estimates by Subgroup—ELA/L Grade 5	148
Table E.4. Test Reliability Estimates by Subgroup—ELA/L Grade 6	149
Table E.5. Test Reliability Estimates by Subgroup—ELA/L Grade 7	150
Table E.6. Test Reliability Estimates by Subgroup—ELA/L Grade 8	151
Table E.7. Test Reliability Estimates by Subgroup—Mathematics Grade 3	152
Table E.8. Test Reliability Estimates by Subgroup—Mathematics Grade 4	153
Table E.9. Test Reliability Estimates by Subgroup—Mathematics Grade 5	154
Table E.10. Test Reliability Estimates by Subgroup—Mathematics Grade 6	155
Table E.11. Test Reliability Estimates by Subgroup—Mathematics Grade 7	156
Table E.12. Test Reliability Estimates by Subgroup—Mathematics Grade 8	157
Table F.1. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 3	158
Table F.2. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 4	158
Table F.3. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 5	158
Table F.4. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 6	158

Table F.5. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 7	159
Table F.6. Decision Accuracy and Consistency by Performance Level—ELA/L Grade 8	159
Table F.7. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 3	159
Table F.8. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 4	159
Table F.9. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 5	160
Table F.10. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 6	160
Table F.11. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 7	160
Table F.12. Decision Accuracy and Consistency by Performance Level—Mathematics Grade 8	160
Table G.1. SGP Estimates by Subgroup—ELA/L Grade 4	161
Table G.2. SGP Estimates by Subgroup—ELA/L Grade 5	161
Table G.3. SGP Estimates by Subgroup—ELA/L Grade 6	162
Table G.4. SGP Estimates by Subgroup—ELA/L Grade 7	162
Table G.5. SGP Estimates by Subgroup—ELA/L Grade 8	163
Table G.6. SGP Estimates by Subgroup—Mathematics Grade 4	163
Table G.7. SGP Estimates by Subgroup—Mathematics Grade 5	164
Table G.8. SGP Estimates by Subgroup—Mathematics Grade 6	164
Table G.9. SGP Estimates by Subgroup—Mathematics Grade 7	165
Table G.10. SGP Estimates by Subgroup—Mathematics Grade 8	165

List of Figures

Figure 8.1. Scale Score Distributions—ELA/L.....	47
Figure 8.2. Scale Score Distributions—Reading.....	48
Figure 8.3. Scale Score Distributions—Writing.....	49
Figure 8.4. Scale Score Distributions—Mathematics.....	50
Figure D.1. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 3.....	134
Figure D.2. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 4.....	135
Figure D.3. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 5.....	136
Figure D.4. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 6.....	137
Figure D.5. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 7.....	138
Figure D.6. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 8.....	139
Figure D.7. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 3.....	140
Figure D.8. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 4.....	141
Figure D.9. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 5.....	142
Figure D.10. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 6.....	143
Figure D.11. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 7.....	144
Figure D.12. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 8.....	145

Section 1: Introduction

This technical report documents the evidence of reliability and validity to support test users in evaluating the intended purposes, uses, and interpretations of the test scores for the spring 2025 administration of the Illinois Assessment of Readiness (IAR) assessments in English language arts/literacy (ELA/L) and mathematics in grades 3–8. The evidence includes descriptions of the test design, development, and administration procedures; the student test results; and psychometric analyses including calibration, equating, and scaling to ensure that the test results are comparable across different test forms and administrations. The information provided herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

1.1. Assessment Overview

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

The assessments are administered online with paper accommodated forms available as needed, along with a wide range of accessibility features for all students and accommodations for students with disabilities, including screen readers, braille, large print, and American Sign Language (ASL). Student results are reported as an overall scale score and performance level with subclaim performance indicators. The four performance levels are Level 4: *Above Proficient*, Level 3: *Proficient*, Level 2: *Approaching Proficient*, and Level 1: *Below Proficient*. Students performing at Levels 3 and 4 are proficient or above proficient and have demonstrated readiness for the next grade level.

1.2. Background

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

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

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

In 2025, Illinois transitioned to the ACT® as the high school accountability assessment for ELA/L mathematics, and science beginning in the 2024–2025 school year. The performance level cut scores were established in 2025 during a standard setting to unify the performance levels across the ACT, IAR, and Illinois Science Assessment (ISA).

1.3. Student Participation

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

1.4. Organizations and Groups Involved

Table 1.1 presents the organizations and groups involved in ensuring the success of the IARs, with each contributor playing a vital role in making sure the assessments yield valid and reliable test results. Input from Illinois educators in the test development and review process is vital to ensure that the assessments reflect the Illinois student population, and feedback from the Technical Advisory Committee (TAC) has been reviewed, addressed, and incorporated into the IAR program. Pearson also uses several experts from various vendors to better support ISBE's goals; for example, Human Resources Research Organization (HumRRO) provides third-party replication, MetaMetrics provides Lexile® and Quantile® licensing and professional development services, the Center for Assessment calculates the student growth percentiles, and edCount conducts expert accommodation reviews and facilitates bias committees for Math, and provides resources, and training support.

Table 1.1. Organizations and Groups Involved

Organization/Group	Roles and Responsibilities
Illinois State Board of Education (ISBE)	<ul style="list-style-type: none">• Carries out the state and federal requirements for the implementation of the statewide assessments• Oversees the planning, scheduling, and implementation of all major assessment activities and supervises the current contract with partnering testing organizations• Conducts quality control activities for every aspect of the test development, administration, scoring, and reporting processes and monitors the security provisions of the assessment program
Pearson	<ul style="list-style-type: none">• Responsible for all test development, administration, and psychometric analyses, including scoring all item responses and providing score reports• Develops future content for the assessment program, including locating and developing appropriate stimulus materials, creating standards-aligned items, coordinating and facilitating item review workshops, and selecting items for the operational and field test forms during form construction
Technical Advisory Committee (TAC)	<ul style="list-style-type: none">• A group of psychometric, assessment design, and administration experts who provide consulting and advice for the assessment system• Provides input on the assessment design, equating and scaling, peer review, standard setting, and reliability and validity issues
Illinois Educators	<ul style="list-style-type: none">• Reviews test items and associated stimuli to ensure appropriate grade-level content, alignment to the content standards, and consistency with classroom instruction• Participates in bias and sensitivity reviews to ensure that items are fair and appropriate for all students• Reviews field tested items during data review to determine their eligibility to be included in the operational item pool• Participates in performance level descriptor (PLD) development and standard setting as needed

Section 2: Test Design

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

2.1. Claims and Subclaims

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

Table 2.1. Claims and Subclaims—ELA/L

Type	Description
Master Claim	Students must demonstrate that they are college and career ready or on track to readiness as demonstrated through reading and comprehending of grade-level texts of appropriate complexity and writing effectively when using and/or analyzing sources.
Major Claims	(1) Reading and comprehending a range of sufficiently complex texts independently (2) Writing effectively when using and/or analyzing sources
Subclaims	The claims and evidence are grouped into the following categories: <ul style="list-style-type: none"> • Reading: Literary Text • Reading: Informational Text • Reading: Vocabulary • Writing: Written Expression • Writing: Knowledge of Language and Conventions

Table 2.2. Claims and Subclaims—Mathematics

Type	Description
Master Claim	Students solve grade-level problems aligned to the Standards for Mathematical Content with connections to the Standards for Mathematical Practice to determine the degree to which a student is college or career ready or on track to being ready in mathematics.
Subclaims	The claims and evidence are grouped into the following categories: <ul style="list-style-type: none"> • Subclaim A: Major Content with Connections to Practices • Subclaim B: Additional and Supporting Content with Connections to Practices • Subclaim C: Highlighted Practices with Connections to Content: Expressing mathematical reasoning by constructing viable arguments, critiquing the reasoning of others, and/or attending to precision when making mathematical statements • Subclaim D: Highlighted Practice with Connections to Content: Modeling/Application by solving real-world problems by applying knowledge and skills articulated in the standards

2.2. Test Blueprints

Each IAR assessment is based on the content area and grade-level test blueprint that outlines the range and distribution of content and the distribution of points across the subclaims and item types to guide test construction. Table 2.3 and *Note*. Reading accounts for 55% of the total score points for grade 3 and 60% for grades 4–8. Writing accounts for 45% of the total score points for grade 3 and 40% for grades 4–8. While the blueprint emphasizes that writing tasks (WE and WKL) account for 40% of the total points (30 out of 74) in grades 4-8, with Written Expression alone contributes 32%, it is also important to consider that up to 8 additional points contribute to a Reading sub-claim based on performance on the two PCR items. Therefore, the full weight of the PCR items is 38 points (out of 74) on test blueprints comprised of the RST & LAT task model options in grades 4-8. 4 points per PCR align to a Reading subclaim, but the points are earned from the PCR response, so those points are tied to how the student responds to the writing prompt.

Table 2.4 present a high-level overview of the IAR blueprints that show the percentage of points for each subclaim. Public-facing blueprints can be found online at <https://www.isbe.net/iar> under “Test Design.” Content developers use additional documents with more detailed blueprint information and sequencing guides when building test forms to ensure consistency in content and psychometric properties.

Table 2.3. High-Level Blueprint—ELA/L

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

Note. Reading accounts for 55% of the total score points for grade 3 and 60% for grades 4–8. Writing accounts for 45% of the total score points for grade 3 and 40% for grades 4–8. While the blueprint emphasizes that writing tasks (WE and WKL) account for 40% of the total points (30 out of 74) in grades 4-8, with Written Expression alone contributes 32%, it is also important to consider that up to 8 additional points contribute to a Reading sub-claim based on performance on the two PCR items. Therefore, the full weight of the PCR items is 38 points (out of 74) on test blueprints comprised of the RST & LAT task model options in grades 4-8. 4 points per PCR align to a Reading subclaim, but the points are earned from the PCR response, so those points are tied to how the student responds to the writing prompt.

Table 2.4. High-Level Blueprint—Mathematics

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

2.3. Item Types

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

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

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

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

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

2.4. Test Sections and Testing Times

Each assessment includes several sections, as detailed in Table 2.5. For mathematics tests, field test items are integrated within the operational sections, so students cannot distinguish which questions are field test items and which are operational items. Each student answers field test items totaling between 6 and 8 points. Because math items can be worth 1, 2, 3, 4, or 6 points, clusters of field test items are assembled automatically using test assembly software. These clusters, or testlets, are pre-compiled to optimize their placement in available field test slots. There are 52 field test forms that contain testlets distributed across the two online math forms to ensure broad deployment of the field test items.

The ELA/L assessments consist of two operational sections and one field test section, while the mathematics assessments consist of three operational sections with embedded field testing. A field test sampling plan determines the total number of ELA/L students required to take the field test, with only those students within schools that are selected in the sampling plan participating in the third field test section.

Illinois policy requires that schools participate in the field test no more than once every three years. However, some historical context is needed to describe how the 2025 sampling plan was derived in relationship to previous years. In 2022, approximately 50% of schools from Illinois participated in the field test since Illinois was still part of the New Meridian contract that shared items across states, while they were also developing and field testing Illinois-only items for the IAR. Therefore, Illinois schools supported the equivalent of two field test administrations in 2022. In 2023, about 31% of schools participated in the field test. Therefore, in 2024, only the remaining 19% of schools were included in field testing. For 2025, the sampling plan allowed for a new rotation, with approximately one-third of schools selected to participate.

The ELA sampling plan was constructed in December 2024 and aimed to reflect the demographic composition of the state using prior-year summative data. Schools selected for the field test were screened to exclude those participating in NAEP testing (based on a list provided by ISBE), as well as any private schools or homeschooled students that appeared in the school lists. The finalized sample was reviewed by the state, and adjustments were made to account for closed schools. For comparison purposes, the table below shows percentages of the main demographic groups in the state population and the selected ELA sample.

Table 2.5. Comparison between State Population and the Sample

Percentage	G3		G4		G5		G6		G7		G8	
	State	Sample	State	Sample	State	Sample	State	Sample	State	Sample	State	Sample
Economic Disadvantage Status	52.23	52.37	51.61	51.88	51.28	50.95	50.94	49.40	50.14	48.97	49.76	48.58
English Language Learner	21.37	22.16	21.44	22.16	16.76	17.10	14.70	14.11	15.60	15.02	15.91	15.57
American Indian/Alaska Native	0.22	0.23	0.26	0.27	0.24	0.25	0.23	0.24	0.21	0.24	0.21	0.23
Asian	5.78	6.38	5.99	6.54	5.89	6.50	5.96	7.13	5.93	6.86	5.78	6.79
Black/African American	16.09	16.24	16.11	16.16	16.06	16.27	15.89	14.92	15.80	14.95	15.91	15.02
Hispanic/Latino	27.85	26.90	28.32	27.76	28.14	27.11	28.11	27.24	28.55	28.26	28.90	28.31
Middle Eastern/North African	0.24	0.35	0.23	0.29	0.21	0.26	0.22	0.21	0.22	0.22	0.23	0.22
Native Hawaiian/Pacific Islander	0.07	0.07	0.08	0.09	0.07	0.08	0.08	0.07	0.09	0.08	0.09	0.11
Two or more races	4.90	4.83	4.73	4.75	4.58	4.49	4.52	4.48	4.32	4.16	4.15	3.93
White	44.85	45.00	44.29	44.14	44.80	45.05	44.99	45.71	44.88	45.23	44.73	45.39
Female	49.23	49.70	49.06	49.15	48.88	49.11	48.93	48.87	48.93	48.88	48.80	48.72
Male	50.76	50.29	50.93	50.83	51.10	50.87	51.04	51.09	51.05	51.10	51.16	51.24
Student With Disabilities	18.95	18.42	19.40	18.97	19.78	19.14	19.54	18.80	18.99	18.54	18.68	18.12

Due to the number of forms field tested (up to 18 per grade) for the ELA assessments, the required student n-count per form was reduced to approximately 2,000 to ensure feasibility and adequate sample size for calibration. This sample size was determined based on several limiting factors, including the policy that students may only be sampled once within a three-year period. For mathematics assessments, where up to 52 test forms per grade were administered to all students, the target sample size was considerably larger than 2,000 due to its embedded FT design in which all students participated in field testing and were randomly assigned one of the FT test forms. Psychometricians used the total number of students in the sample to determine how many forms could be field tested. Final approval included confirmation from the state.

The IAR is a timed assessment, with the testing time limited to the section testing times presented in Table 2.6 (except for an extended time accommodation). The section testing time is the amount of time that must be provided to any student who needs it to complete the section, and the total testing time reflects the operational testing time only. A new section cannot be started until all students in the testing environment are finished or until the section testing time has expired. If all students have completed testing before the end of the section testing time, the section may end. Once the section testing time has elapsed, the section must end (except for students with an extended time accommodation).

Table 2.6. Test Sections and Testing Times

Assessment(s)	Section(s)	Testing Time per Section	Total Testing Time
ELA/L 3	Sections 1–2	75 minutes	150 minutes or 225 minutes (for schools assigned to the field test)
	Section 3 (field test) – only given to students in the field test sample. Schools are eligible to be in the field test sample once every three years.	75 minutes	
ELA/L 4–8	Sections 1–2	90 minutes	180 minutes or 270 minutes (for schools assigned to the field test)
	Section 3 (field test) – only given to students in the field test sample. Schools are eligible to be in the field test sample once every three years.	90 minutes	
Mathematics 3–5	Sections 1–3 (non-calculator)	60 minutes	180 minutes
Mathematics 6–7	Section 1 (majority is a calculator subsection followed by a shorter non-calculator subsection)	60 minutes	180 minutes
	Sections 2–3 (calculator)	60 minutes	
Mathematics 8	Section 1 (non-calculator)	60 minutes	180 minutes
	Sections 2–3 (calculator)	60 minutes	

2.5. Test Structure

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

All ELA/L assessments include an RST and either the LAT or the NWT. The LAT and the RST are scored for three traits: Reading Comprehension, Written Expression, and Knowledge of Conventions. The NWT is scored for two traits: Written Expression and Knowledge of Conventions. All traits are initially scored as either 0–3 or 0–4 points, with the Written Expression traits then multiplied by 3 (or weighted) to increase their contribution to the total score, making possible subclaim scores 0, 3, 6, and 9 or 0, 3, 6, 9, and 12. Table 2.7 presents the maximum possible points for the PCR items.

Table 2.7. Contribution of PCR Items in ELA/L: Number of Possible Points by Task

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

The mathematics assessments, depicted in Table 2.8 tasks that measure a combination of conceptual understanding, applications, skills, and procedures. Each grade-level assessment includes both short- and extended-response items focused on applying skills and concepts to solve problems that require demonstration of the mathematical practices from the Illinois Learning Standards with a focus on modeling and reasoning with precision. Mathematics constructed-response items consist of tasks designed to assess a student’s ability to use mathematics to solve real-life problems. Some of the tasks require students to describe how they solved a problem, while other tasks measure conceptual understanding and the ability to apply concepts by means of selected-response or technology-enhanced items. Students are also required to demonstrate their skills and knowledge by answering innovative selected-response and short-answer items that measure concepts and skills.

Table 2.8. Contribution in Item Types in Math: Number of items by Type

Items	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Type 1 Items						
1 point	24	20	20	20	20	20
2 points	3	5	5	5	5	5
Type 1 Totals	27	25	25	25	25	25
Type 2 Items						
3 points	2	2	2	2	2	2
4 points	1	1	1	1	1	1
Type 2 Totals	3	3	3	3	3	3
Type 3 Items						
3 points	2	2	2	2	2	2
6 points	1	1	1	1	1	1
Type 3 Totals	3	3	3	3	3	3

Section 3: Test Development

This section describes the process for developing new items for the IAR assessments to be embedded as field test items in the operational test administration, along with the process for constructing the test forms. Pearson constructed the spring 2025 test forms with custom-developed IAR items from the operationally ready item pool.

3.1. Asset Development Plan

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

3.2. Passage Selection

The ELA/L assessments are based on selected texts, including multimedia stimuli. Using the *Passage Selection Guidelines* with the text complexity framework and guidance on selecting a variety of text types and passages, ELA/L subject matter experts were trained to search for and commission appropriate texts to support an annual pool of passages for consideration. Content experts then reviewed the texts for adherence to the *Text Selection Guidelines* and the annual asset development plan in the number and distribution of genres and topics. Next, the Text Review Committee provided feedback about the grade-level appropriateness, content, and potential bias concerns and reached consensus about which texts would move forward for development. ELA/L item development was not conducted until after the texts were approved.

3.3. Item Development

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

Next, every newly developed item was reviewed during content and bias and sensitivity committees to ensure that they aligned to the standards and were fair for all student populations. The meetings were conducted virtually and included large group training on the expectations and processes of each meeting, followed by breakout meetings by content area and grade level where additional training was provided. Each item also underwent an editorial review for grammar, punctuation, clarity, and adherence to the style guide.

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

3.4. Form Construction

Test form construction is the process of selecting and sequencing a set of operational and field test items for administration. It is a complex, interactive task that requires both content and psychometric expertise aimed at ensuring that scores be comparable across forms and administrations. Table 3.1 presents the number of test forms constructed for spring 2025. Both ELA/L and mathematics had two core operational forms and one accommodated operational form. The forms were constructed to reflect the test blueprint in terms of content, item types, test length, and expected difficulty and performance along the ability continuum. They were also constructed to adhere to the following goals outlined in the test construction specifications:

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

Table 3.1. Number of Test Forms Constructed in Spring 2025

Assessment	#Core OP Forms	#Accommodated OP Base Forms	#FT Forms
ELA/L 3	2	1	18
ELA/L 4	2	1	18
ELA/L 5	2	1	18
ELA/L 6	2	1	18
ELA/L 7	2	1	18
ELA/L 8	2	1	18
Mathematics 3	2	1	52
Mathematics 4	2	1	52
Mathematics 5	2	1	52
Mathematics 6	2	1	49
Mathematics 7	2	1	50
Mathematics 8	2	1	42

Note. OP = operational, FT = field test

3.4.1. Operational Forms

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

3.4.2. Field Test Forms

In spring 2025, all students received one of two core forms of operational items or an accommodated form of operational items. Different sets of field test items were either embedded in the mathematics sections or administered to a select sample of students in a separate third section for ELA/L (i.e., census field testing is conducted for mathematics, while a sampling plan is used for ELA/L). Mathematics forms include embedded items in Sections 2 and 3 only for grades 3–5 and in each section for all other grades.

Table 3.2. Number of Items and Field Test Forms for ELA

ELA	# of forms per grade	Total # of forms	# of items per forms* including duplicates	# of items per grade	Total # of items
Grades 3-8	18	108	10 -11	180-198	1080-1188

Accommodated Forms

Table 3.3 presents the accommodated forms constructed based on the one accommodated operational form developed for each content area and grade, as well as the accommodations available on the operational core form. The forms are accommodated to support braille, large print, human reader/human signers, and text-to-speech (TTS). Spanish forms are provided for mathematics only.

Table 3.3. Supported Accommodations

Test Form	ELA/L	Mathematics
Accommodated Base Form (ACC1)	<ul style="list-style-type: none"> • Paper-Based Form • Large Print • Read Aloud • Human Reader • Human Signer • ASL • Braille • Screen Reader • Non-Screen Reader • TTS 	<ul style="list-style-type: none"> • Paper-Based Form • Large Print • Read Aloud • Human Reader • Human Signer • ASL • Braille • Screen Reader • Non-Screen Reader • Spanish Paper • Spanish Large Print • Spanish Human Reader
Core Form (Online1)	N/A	<ul style="list-style-type: none"> • TTS • Spanish Online • Spanish TTS

3.5. Data Review

Following the spring 2025 test administration, an educator data review committee met in August 2025 to evaluate the field-tested items and associated performance data in terms of appropriateness, level of difficulty, and any potential differential item functioning (DIF) for groups of interest. The committee recommended acceptance or rejection of each field-tested item for inclusion in the operational item bank and, for mathematics, made recommendations for some items to be revised and re-field tested. Items approved by the committee became eligible for use on future operational assessments.

Table 3.4. 2025 Data Review Results

Math REVIEWED Items at Data Review				
Grade	Total	ACCEPT	% Accept	REJECT
3	173	148	86%	25
4	176	162	92%	14
5	167	163	98%	4
6	164	162	99%	2
7	165	160	97%	5
8	176	171	97%	5
	1021	966	95%	55
ELA REVIEWED Items at Data Review				
Grade	Total	ACCEPT	% Accept	REJECT
3	80	55	69%	25
4	66	52	79%	14
5	80	76	95%	4
6	80	78	98%	2
7	67	62	93%	5
8	86	81	94%	5
	459	404	88%	55

Section 4: Test Administration

Table 4.1 presents the spring 2025 test administration dates. The IAR assessments are administered online, with paper accommodated forms available as needed. The online administration takes place in TestNav, Pearson’s online testing platform. ADAM (Assessment Delivery and Management) is the student test management portal that Test Administrators use to manage student tests, registrations and order materials if needed.

Table 4.1. Test Administration Activities

Event	Dates
Administration Training Modules	October 2024 – March 2025
Receive Materials	February 20, 2025
Online Testing Window	March 3 – April 18, 2025
Paper Testing Window	March 3 – April 4, 2025
Return Materials	April 11, 2025

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

4.1. Accessibility Features and Accommodations

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

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

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

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

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

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

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

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

4.2. Test Security

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

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

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

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

4.3. Testing Irregularities and Security Breaches

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

Table 4.2. Test Irregularity and Security Breach Examples

Topic	Examples
Electronic Devices	<p>Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner) while secure test materials are still distributed, while students are testing, after a student turns in their test materials, or during a break</p> <p><i>Exceptions:</i></p> <ul style="list-style-type: none"> • School Test Coordinators, Technology Coordinators, and Test Administrators can use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed. • Certain electronic devices may be allowed for medical or audiological purposes during testing. If a student needs their device for medical reasons, a unique accommodation form must be submitted in advance of testing. The student should be seated near the Test Proctor during testing.
Test Supervision	<ul style="list-style-type: none"> • Coaching students during testing (e.g., giving students verbal or nonverbal cues, hints, suggestions, or paraphrasing or defining any part of the test) • Engaging in activities (e.g., grading papers, reading a book, newspaper, or magazine) that prevent proper student supervision while secure test materials are still distributed or while students are testing • Leaving students unattended while secure test materials are still distributed or while students are testing • Deviating from testing time procedures • Allowing cheating of any kind • Providing unauthorized persons with access to secure materials • Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore not appropriate • Allowing students to test before or after the test administration window
Test Materials	<ul style="list-style-type: none"> • Losing a student test booklet or answer document • Losing a student testing ticket • Leaving test materials unattended or failing to keep test materials secure at all times • Reading or viewing the passages or test items before, during, or after testing • Copying or reproducing (e.g., taking a picture of) any part of the passages or items or any secure test materials or online test forms • Revealing or discussing passages or test items with anyone, including students and school staff, through verbal exchange, email, social media, or any other form of communication • Removing secure test materials from the school’s campus or removing them from locked storage for any purpose other than administering the test <p><i>Exception:</i> Administration of a human reader/signer accessibility feature for mathematics or accommodation for ELA/L that requires a Test Administrator to access passages or items</p>

Topic	Examples
Testing Environment	<ul style="list-style-type: none"><li data-bbox="386 237 1029 268">• Allowing unauthorized visitors in the testing environment<li data-bbox="386 268 1349 331">• Failing to follow administration directions exactly as specified in the <i>Test Administrator Manual</i><li data-bbox="386 331 1300 394">• Displaying testing aids in the testing environment (e.g., a bulletin board containing relevant instructional materials) during testing

Section 5: Scoring

Selected-response, technology-enabled, and technology-enhanced items are machine scored; constructed-response items are human scored using Pearson’s scoring platform, OSCAR (Online Scoring and Reporting); and the ELA/L PCR items are primarily scored by Pearson’s automated scoring engine known as the Intelligent Essay Assessor (IEA), with a 10% reliability score and some outlier scoring (where the IEA score and human score differ by more than 1 point) by human scorers (i.e., 10% of the PCR item scores are also scored by humans in addition to IEA to compute the inter-rater agreement and monitor scoring).

5.1. Machine Scoring

Pearson performed a key check and adjudication near the end of the test administration and before reporting to verify that the answer keys were correct for each item. The keycheck process is a quality assurance step to ensure that multiple-choice or multi-select items are scoring correctly. This process involves a review of item statistics to confirm that the designated correct answer(s) are being awarded full credit and that the scoring logic is functioning as intended. Reviewers check the scoring data to verify that all correct responses receive full credit, and that incorrect or partially correct responses are scored appropriately according to the established rules.

If a test map contains an incorrect key, such as a misaligned correct answer or scoring rule, this issue is typically flagged during the statistical keycheck process, which uses item-level metrics and response patterns to detect anomalies. Once identified, the item is escalated for content review to confirm the error and determine its source. The test map is then corrected and re-published, ensuring alignment with scoring rules and platform logic. A re-analysis is conducted to validate the fix, and downstream systems are updated accordingly. If the correction affects scoring logic or item metadata, the Change Control Board (CCB) oversees the implementation to prevent disruptions. The corrected item is re-exported, rescored, and verified by psychometrics and content teams before finalization. This process ensures scoring integrity and prevents propagation of errors across systems.

The adjudication process is a specialized workflow designed to ensure the accuracy and fairness of scoring for technology-enhanced and free-response items, distinct from the standard key check. It provides a structured approach for addressing discrepancies and maintaining scoring integrity.

The adjudication workflow begins with the preparation phase. Relevant scoring materials—including item files, answer keys, scoring rubrics, and student responses—are gathered and made accessible within secure scoring platforms such as OSCAR. All reviewers involved in the adjudication process are granted appropriate access to these materials to facilitate a thorough review.

The initial review is conducted by a trained content specialist or designated scorer. This reviewer examines flagged items or student responses where scoring discrepancies have been identified, either through automated scoring algorithms or manual checks. The first reviewer documents the nature of the discrepancy, marks the issue in the scoring system, and provides an initial assessment or recommended action.

If the first reviewer is unable to resolve the discrepancy or if there is disagreement regarding the scoring decision, the issue is escalated to a senior content lead or adjudication manager. The second reviewer performs a detailed analysis of the flagged item or response, referencing scoring guidelines and consulting with subject matter experts as needed. This step ensures that all perspectives are considered and that the scoring aligns with established criteria.

For issues that remain unresolved after the second review, a third adjudication is conducted by the Test Development Manager (TDM) or a designated member of the Content Committee Board (CCB). This final review involves a comprehensive evaluation of all supporting documentation, prior reviewer notes, and scoring rubrics. The TDM or CCB approves the final scoring decision and oversees the implementation of any necessary changes to the scoring rules or answer keys.

After adjudication decisions are finalized, a series of post-adjudication checks are performed. These include rescoring affected responses, conducting a secondary review of item spreadsheets, and, if required, replacing or updating items within the scoring system. These steps confirm that adjudication outcomes have been correctly applied and that scoring reliability is maintained.

Throughout the adjudication workflow, all actions, findings, and decisions are logged in secure tracking systems. Clear communication is maintained among reviewers, scoring managers, and project stakeholders to ensure transparency. A summary of adjudication outcomes is provided to relevant parties, and procedural updates are incorporated into future scoring guidelines.

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

5.2. Human Scoring of Constructed-Response Items

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

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

5.2.1. Scorer Training

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

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

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

Table 5.1. Scoring Training Materials

Training Material	Description	Specifications
Anchor Sets	Anchor sets consist of responses that are clear examples of student performance at each score point and are the primary reference for scorers as they internalize the rubric. The responses selected are representative of typical approaches to the task and are arranged to reflect a continuum of performance. All scorers have access to the anchor set when they are training and scoring and are directed to refer to it regularly.	The mathematics prototype anchor set includes three annotated responses per score point, whereas the abbreviated anchor set includes 1–3 annotated responses per score point. The ELA/L prototype anchor sets include three annotated responses per score point, including separate complete anchor sets for each scoring trait (Reading Comprehension and Written Expression and Conventions for Research Simulation and Literary Analysis Tasks, Written Expression for Narrative Writing Tasks, and Knowledge of Language and Conventions for all task types).

Training Material	Description	Specifications
Practice Sets	Practice sets are used to help scorers practice applying the scoring guidelines. Scorers review the anchor sets, score the practice sets, and then compare their assigned scores for the practice sets to the actual assigned scores to help them learn. Some of these responses clearly reinforce the scoring guidelines presented in the anchor set, whereas others are more difficult to evaluate, fall near the boundary between two score categories, or represent unusual approaches to the task to provide guidance and practice in defining the line between score categories and applying the scoring criteria to a wider range of response types.	The mathematics prototype and abbreviated practice sets include 2–3 sets of 10 annotated responses. The ELA/L prototype practice sets include two sets of five annotated responses and two sets of 10 annotated responses, whereas the abbreviated practice sets include two sets of 10 annotated responses.
Qualification Sets	Qualification sets consist of student responses that are clear examples of score points to reinforce the application of the scoring criteria illustrated in the anchor set. These sets are used to confirm that scorers understand how to score the responses accurately. Scorers are required to meet specified agreement percentages on qualification sets to score student responses.	The mathematics and ELA/L prototype qualification sets include three sets of 10 responses each (not annotated). The subsequent abbreviated items do not include qualification sets.

5.2.2. Scorer Qualification

To demonstrate that they could accurately apply the scoring methodology, scorers applied scores to three qualification sets consisting of 10 responses each. ELA/L scorers applied a score for each trait on each response in the qualification sets¹, and mathematics scorers applied a score for each part of an item that was a constructed response ranging from 1–4 parts. Scorers were required to match the approved score at a certain percentage to qualify. For ELA/L qualification, scorers were required to meet the following conditions:

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

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

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

Table 5.2. Mathematics Scorer Qualification Requirements

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

5.2.3. Scorer Monitoring

Score monitoring consisted of second scoring of at least 10% of the responses, backreading, the use of validity responses and calibration sets, and inter-rater reliability (see Section 0).

5.2.3.1. Second Scoring

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

5.2.3.2. Backreading

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

5.2.3.3. Validity Responses

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

Table 5.3. Scoring Validity Agreement Requirements

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

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

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

5.2.3.4. Calibration Sets

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

5.3. Automated Scoring of PCR Items

Automated scoring performed by Pearson’s IEA automat For IAR constructed-response items, Pearson uses the Intelligent Essay Assessor (IEA) to generate automated scores based on human-scored training ed scoring engine was the default option for scoring the summative assessment’s online PCR tasks. Under the default option, it was assumed that operational scores for approximately 90% of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. Human scoring was applied to responses that were scored while IEA was being trained, as well as to additional responses routed to human scoring when there was uncertainty about the automated scores. For 10% of responses, a second reliability score was assigned to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. When IEA provided the first score of record, the second reliability score was a human score.

For each IAR constructed-response item and trait, the responses used to train and evaluate the IEA models are selected using simple random sampling from the pool of available field test human-scored responses. The selected responses are divided into training, validation, and holdout sets to support model development, tuning, and independent evaluation. In developing each model, Pearson applies text preprocessing and extracts a rich set of features designed to reflect the scoring rubric, including indicators of writing quality (e.g., mechanics, grammar, organization, and development) and content-based semantic features aligned with the expectations of the item. Multiple supervised machine-learning algorithms are trained separately for each trait, and the model demonstrating the strongest performance on validation and holdout data is selected for operational use. Pearson’s Continuous Flow procedures allow new operational responses to be incorporated into the models when appropriate, while maintaining human oversight to ensure consistency with IAR scoring standards. All sampling used in automated scoring model development for IAR is conducted using simple random sampling to ensure that the training data are representative of the full pool of scored responses.

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

When the engine is less confident in scoring a response, the response is marked with a low confidence flag that automatically routes it to human scorers (known as Smart Routing). Smart Routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores. Smart Routing can be applied prompt-by-prompt to the extent needed to meet scoring quality criteria for automated scoring. It was assumed for the spring administration that operational scores for approximately 95% of the online PCR responses would be assigned by IEA, while the operational scores for the remaining online responses were assigned by human scorers.

5.3.1. Sampling Responses Used for Training IEA

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

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

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

5.3.2. Quality Criteria for Evaluating IEA Performance

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

Additional measures of inter-rater agreement for evaluating automated scoring are used, including the Pearson correlation (r), kappa, quadratic weighted kappa (QWK), exact agreement, and standardized mean difference (SMD). These measures are computed between pairs of human scores and between IEA

and humans to evaluate how performance was the same or different. Criteria for evaluating the training of IEA given these measures include the following:

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

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

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

5.4. Inter-Rater Agreement

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

Pearson evaluates automated scoring performance for IAR within the same validity framework used for human scoring. Following guidance from Williamson, Xi, and Breyer (2012), agreement between human and automated scores is summarized using multiple indices, including quadratic weighted kappa (QWK), Pearson correlation, and standardized mean difference (SMD) between score distributions. For automated scoring models to be considered operationally comparable to human scoring, QWK values are expected to meet or exceed 0.70, to fall within approximately 0.10 of the corresponding human-human QWK values, and SMD values are expected to remain below 0.15 in absolute value. Percent exact agreement is also reported for context but is treated as an ancillary measure rather than a primary indicator of scoring validity. This approach ensures that the automated scoring models used for IAR are evaluated according to industry-standard criteria that appropriately reflect the ordinal nature and functional purpose of the scoring scales. The agreement results presented in Table 5.5 therefore demonstrate that the automated scoring models used for IAR meet the expected criteria for operational comparability to human scoring.

Table 5.4 presents the inter-rater agreement expectations and results for the constructed-response items from the spring 2025 administration across all grades based on human scoring, and Table 5.5. ELA/L PCR Average Agreement Indices

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

Table 5.4. Inter-Rater Agreement Expectations and Spring 2025 Results

Content Area	#Items	Score Point Range	Perfect Agreement Expectation	Perfect Agreement 2025 Result	Within 1 Point Expectation*	Within 1 Point 2025 Result
ELA/L	27	Multi-trait	65%	100%	96%	100%
Mathematics	15	0–2	80%	100%	96%	96%
	28	0–3	70%	100%	96%	100%
	11	0–4	65%	100%	95%	100%
	5	0–5	65%	100%	95%	100%
	6	0–6	65%	100%	95%	100%

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

Table 5.5. ELA/L PCR Average Agreement Indices

Trait	Assessment	#PCRs	#Tasks	Perfect	Kappa	QWK	<i>r</i>
Written Expression	ELA/L 3	5	3	82%	0.64	0.81	0.82
	ELA/L 4	4	3	81%	0.66	0.82	0.83
	ELA/L 5	5	3	78%	0.67	0.85	0.86
	ELA/L 6	4	3	80%	0.69	0.87	0.87
	ELA/L 7	5	3	81%	0.69	0.88	0.88
	ELA/L 8	4	3	77%	0.65	0.85	0.86
Writing Knowledge Language & Conventions	ELA/L 3	5	3	80%	0.62	0.80	0.81
	ELA/L 4	4	3	82%	0.69	0.84	0.84
	ELA/L 5	5	3	78%	0.66	0.84	0.84
	ELA/L 6	4	3	80%	0.68	0.86	0.86
	ELA/L 7	5	3	80%	0.68	0.88	0.88
	ELA/L 8	4	3	76%	0.63	0.84	0.84

5.5. Hierarchy of Assigned Scores for Reporting

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

Table 5.6. Scoring Hierarchy Rules

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

Section 6: Reporting

6.1. Available Reports

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

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

6.2. Interpretation of Test Scores

6.2.1. Total Scale Scores and Performance Levels

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

Based on a student's total score, an inference is drawn about how much knowledge and skill in the content area the student has acquired. The overall scale scores also determine a student's performance level that classifies a student's competency based on their test performance as reflected by their test results, as provided below. Each performance level is defined by a range of overall scale scores for the assessment established during the standard setting (see Section 7 for more details). Students performing at Levels 3 and 4 are considered proficient or above proficient and have demonstrated readiness for the next grade level/course and, ultimately, are likely on track for college and careers. The full PLDs are available online at <https://www.isbe.net/IAR>.

- Level 4: *Above Proficient*
- Level 3: *Proficient*
- Level 2: *Approaching Proficient*
- Level 1: *Below Proficient*

6.2.2. Claim and Subclaim Scores

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

Within each reporting category are specific skill sets (subclaims) students demonstrate on the IAR. Each subclaim category includes the header identifying the subclaim, an explanatory icon representing the student's performance, and an explanation of whether the student has met the expectations of the subclaim.

A student's subclaim category indicates how well the student performed within a specific skill set (subclaim) on the assessment. Similar to overall and reporting category scores, proficiency in each subclaim is estimated using a common measurement scale. Subclaim performance is reported using three readiness categories, rather than as scale scores or performance levels:

- Higher-level-readiness (H): The letter "H" for a given subclaim signifies that the student demonstrated a higher level of readiness, corresponding to proficiency consistent with Performance Level 3 or 4. Students in this category are generally academically well prepared for further study in that content area and may benefit from additional enrichment.
- Middle-level-readiness (M): The letter "M" indicates that the student demonstrated a middle level of readiness, which reflects proficiency consistent with Performance Level 2. These students may need academic support to progress successfully in further studies related to the subclaim content area.
- Lower-level readiness (L): The letter "L" indicates that the student demonstrated a lower level of readiness, corresponding to proficiency consistent with Performance Level 1. Students in this category are likely not yet academically prepared for additional study in the subclaim content area and will likely require targeted instructional interventions to improve achievement.

Student performance for each subclaim is visually marked with one of these indicators (H, M, or L) to clearly communicate readiness in specific skill areas.

6.2.3. Additional Measures

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

Section 7: Standard Setting

Student performance on state assessments is required to be reported in performance levels that reflect how well students demonstrate the knowledge and skills expected at their grade level, as defined by a state’s content standards. These performance levels are defined by cut scores, or points on the test’s score scale (i.e., the full range of possible scores) that mark the minimum score a student must earn to be classified into a given performance level. During a standard setting meeting, committees of educators review test items and apply their content expertise and professional judgment to recommend these cut scores. This section summarizes the 2025 standard setting that took place from July 14–18, 2025, in Springfield, Illinois, to establish the most recent cut scores for the ACT, IAR, and ISA assessments, with the full details about the process provided in the standard setting report (Gardner, T. & Moore, J, 2025).

Illinois transitioned in 2025 to the ACT as the high school accountability assessment for ELA/L, mathematics, and science. In response to findings that Illinois’ previous cut scores were among the highest in the nation, Illinois took advantage of the opportunity presented by the shift in high school assessment to unify the names, number, and definition of the performance levels across the ACT, IAR, and ISA assessments. This required a standard setting to recommend new cut scores that divide each assessment score scale into four levels: *Above Proficient*, *Proficient*, *Approaching Proficient*, and *Below Proficient*. Despite differences in test design, this unified approach aimed to maintain high expectations while better reflecting college and career readiness and establishing a unified, coherent reporting structure across the Illinois assessment system.

The standard setting used spring 2025 data and followed the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake et al., 2005) for all assessments in addition to the Modified Briefing Book approach (Camara et al., 2017; Haertel et al., 2012) for the ACT to establish the empirical link with college and career readiness benchmark data. Seventeen committees with 12 panelists each for grades 3 and 5-8 and two committees with 18 panelists each for the grade 4 ELA and mathematics IAR assessments were convened, for a total of 155² panelists covering 240 slots (14 subject and grade-level committees for IAR and ISA and five committees for each ACT subject: English, reading, writing, mathematics, and science). A vertical articulation committee was then convened on the last day with 34 participants to ensure an appropriate progression of cuts across grades and subject areas. See the standard setting report for full details of the process (Gardner, T. & Moore, J, 2025).

7.1. Standard Setting Process

Each standard setting committee recommended three cut scores to divide the score scale into the four performance levels: the *Approaching Proficient* cut (between *Below Proficient* and *Approaching Proficient*), the *Proficient* cut (between *Approaching Proficient* and *Proficient*), and the *Above Proficient* cut (between *Proficient* and *Above Proficient*). Following the Extended Modified Yes/No Angoff method, panelists reviewed each item on one form of the spring 2025 operational assessment in test administration order and judged whether most borderline students (i.e., students scoring near the lower end of a performance level) would likely answer the item correctly (for single-point items) or how many points they would likely earn (for multi-point items). Because constructed-response items have a large impact on the IAR and ISA test scores, these judgments were supplemented by student score profiles illustrating how actual students performed across the score scale to ground panelists’ decisions in actual student performance.

² Final participation was 147 of the 155 panelists recruited.

While the IAR and ISA are designed to measure mastery of the grade-level Illinois Learning Standards, the ACT has an additional purpose to predict college readiness. Therefore, in addition to the item-level Extended Modified Yes/No Angoff content judgments, ACT panelists received briefing books containing empirical data linking ACT scores to real-world student outcomes, such as the likelihood of earning a B or C in first-year college courses, high school GPA, and college enrollment rates, as well as performance on other assessments (e.g., IAR, ISA, NAEP). This information enabled panelists to interpret ACT scores in the context of college and career readiness and ensure that the recommended cut scores reflect meaningful readiness benchmarks in addition to content mastery.

Each standard setting meeting began with an overview of the purpose, test design, panelist role, and orientation to the materials, including the standard setting tool where the item judgments were made. Panelists first “experienced the assessment” by taking the spring 2025 operational assessment, followed by discussing the PLDs and exploring examples of performance that distinguishes students just entering the performance level from the full range of expected performance in the band. After additional system training and a practice activity, panelists confirmed their readiness before beginning the item-level judgments. Three rounds were conducted for IAR and ISA:

1. Round 1: Panelists reviewed each item on the test form and answered one of the following judgment questions for each performance level, beginning with the *Proficient* cut:
 - a. Single-point items: “*Considering a variety of students at the lower end of the performance level, would most students get this item correct?*” Panelists answered “yes” or “no.”
 - b. Multi-point items: “*Considering a variety of students, which score point most likely represents the most common response for students at the lower end of this performance level?*” Panelists chose between 0–6 score points depending on the item type and maximum score.
2. Round 2: Panelists received the score profiles and discussed the Round 1 results before revising the initial judgments as needed following the same steps as Round 1.
3. Round 3: Final revisions were made after discussing the Round 2 results, including the impact data that showed the percentage of students who would be classified into each performance level based on the recommended cuts and performance of students on the spring 2025 assessment.

Four rounds were conducted for the ACT assessments using a hybrid approach:

1. Round 1 (Extended Yes/No Modified Angoff): Panelists followed the same judgment process as IAR/ISA.
2. Round 2 (Extended Yes/No Modified Angoff): Panelists discussed the Round 1 results, followed by a discussion of the impact and outcome data. Equipped with both the Round 1 results and an understanding of the relationships between performance on the ACT and performance in first-year college courses, panelists revised their initial judgments following the same steps as Round 1.
3. Round 3 (Modified Briefing Book): The English, reading, and writing committees were combined to form a 12-panelist ACT ELA committee to support coherence across the full ELA domain, ensuring that the cut scores for the three subtests reflected a unified definition of college and career readiness. After discussing the Round 2 results, panelists focused the discussion on the briefing books and used the empirical data to recommend ACT scores that define minimal performance for *Approaching Proficient*, *Proficient*, and *Above Proficient*.

4. Round 4 (Modified Briefing Book): Final cut recommendations were submitted after further discussion.

For all assessments, each panelist’s judgments were summed across the items for each performance level to determine the test-level raw cut score, with “yes” = 1, “no” = 0, and score points used for constructed-response items. All raw scores were transformed to scale scores via a raw-to-scale score (RSS) conversion table, and all results were presented to the panelists on the scaled score metric. Final committee-level cut scores were the median of all the individual panelists’ cut scores from the final round. Select panelists then participated in a vertical articulation meeting to refine the cut scores across grades 3–11 to ensure logical progression, including the statistically linked cut scores for the PreACT 9 Secure and PreACT Secure.

Panelists also completed three evaluation surveys throughout the meeting to determine their understanding of the process and their confidence in the results: after the practice activity, after Round 3, and after the vertical articulation. The ACT panelists completed an additional survey after Round 2 to capture input from the panelists who were not retained for the ELA panel. Overall, results indicated a strong understanding of the process and high confidence in the recommended cut scores.

7.2. Cut Scores

Table 7.1 presents the resulting IAR scale score cut scores (i.e., the minimum score students must receive to be classified into a certain performance level), as shown in bold.

Table 7.1. Scale Score Ranges and Cut Scores

Assessment	Level 1: <i>Below Proficient</i>	Level 2: <i>Approaching Proficient</i>	Level 3: <i>Proficient</i>	Level 4: <i>Above Proficient</i>
ELA/L 3	650–684	685–734	735–779	780–850
ELA/L 4	650–694	695–736	737–779	780–850
ELA/L 5	650–699	700–738	739–779	780–850
ELA/L 6	650–704	705–740	741–779	780–850
ELA/L 7	650–709	710–742	743–784	785–850
ELA/L 8	650–709	710–744	745–794	795–850
Mathematics 3	650–704	705–731	732–780	781–850
Mathematics 4	650–707	708–739	740–783	784–850
Mathematics 5	650–708	709–739	740–781	782–850
Mathematics 6	650–704	705–741	742–772	773–850
Mathematics 7	650–711	712–744	745–780	781–850
Mathematics 8	650–704	705–744	745–790	791–850

Note. Cut scores used to identify Levels 2, 3, and 4 are shown in bold. Students with a score below the cut for Level 2 are placed in Level 1.

Section 8: Student Characteristics and Test Results

8.1. Student Participation

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

Table 8.1. Student Participation by Administration Mode

Assessment	#Valid Cases	Online N	%	Paper N	%
ELA/L 3	132,063	131,864	99.8	199	0.2
ELA/L 4	130,742	130,541	99.8	201	0.2
ELA/L 5	130,422	130,226	99.8	196	0.2
ELA/L 6	129,426	129,281	99.9	145	0.1
ELA/L 7	130,924	130,794	99.9	130	0.1
ELA/L 8	134,023	133,892	99.9	131	0.1
ELA/L Total	787,600	786,598	99.9	1,002	0.1
Mathematics 3	131,915	131,711	99.8	204	0.2
Mathematics 4	130,600	130,413	99.9	187	0.1
Mathematics 5	130,283	130,107	99.9	176	0.1
Mathematics 6	129,232	129,082	99.9	150	0.1
Mathematics 7	130,683	130,550	99.9	133	0.1
Mathematics 8	133,755	133,621	99.9	134	0.1
Mathematics Total	786,468	785,484	99.9	984	0.1

Table .2–Table 8.5 present the number of students with valid scores by demographic subgroup as captured in ADAM by means of a student data upload. The demographic data were verified by ISBE prior to score reporting. Students missing information on one or more of the demographic variables were omitted from the subgroup analyses.

Table .2. Student Participation by Demographic Subgroup—ELA/L Grades 3-5

Demographic	Grade 3		Grade 4		Grade 5	
	N	%	N	%	N	%
Economically Disadvantaged	69,077	52.3%	67,560	51.7%	66,942	51.3%
Students with Disabilities (SWD)	24,818	18.8%	25,169	19.3%	25,604	19.6%
English Learner (EL)	28,264	21.4%	28,071	21.5%	21,882	16.8%
Male	66,959	50.7%	66,540	50.9%	66,620	51.1%
Female	65,089	49.3%	64,183	49.1%	63,780	48.9%
American Indian/Alaska Native	291	0.2%	335	0.3%	318	0.2%
Asian	7,635	5.8%	7,839	6.0%	7,685	5.9%
Black/African American	21,294	16.1%	21,079	16.1%	20,922	16.0%
Hispanic/Latino	36,829	27.9%	37,074	28.4%	36,748	28.2%
Middle Eastern or North African	317	0.2%	304	0.2%	279	0.2%
Native Hawaiian or Other Pacific Islander	89	0.1%	102	0.1%	96	0.1%
Two or More Races Reported	6,433	4.9%	6,173	4.7%	5,967	4.6%
White/Caucasian	59,175	44.8%	57,836	44.2%	58,407	44.8%

Table 8.3. Student Participation by Demographic Subgroup—ELA/L Grades 6-8

Demographic	Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%
Economically Disadvantaged	65,969	51.0%	65,644	50.1%	66,698	49.8%
Students with Disabilities (SWD)	25,105	19.4%	24,702	18.9%	24,818	18.5%
English Learner (EL)	19,041	14.7%	20,440	15.6%	21,365	15.9%
Male	66,000	51.0%	66,824	51.0%	68,554	51.2%
Female	63,392	49.0%	64,064	48.9%	65,419	48.8%
American Indian/Alaska Native	303	0.2%	272	0.2%	286	0.2%
Asian	7,738	6.0%	7,784	5.9%	7,761	5.8%
Black/African American	20,570	15.9%	20,639	15.8%	21,296	15.9%
Hispanic/Latino	36,427	28.1%	37,437	28.6%	38,833	29.0%
Middle Eastern or North African	288	0.2%	288	0.2%	301	0.2%
Native Hawaiian or Other Pacific Islander	97	0.1%	124	0.1%	127	0.1%
Two or More Races Reported	5,842	4.5%	5,645	4.3%	5,548	4.1%
White/Caucasian	58,161	44.9%	58,735	44.9%	59,871	44.7%

Table 8.4. Student Participation by Demographic Subgroup— Mathematics Grades 3-5

Demographic	Grade 3		Grade 4		Grade 5	
	N	%	N	%	N	%
Economically Disadvantaged	68,981	52.3%	67,432	51.6%	66,853	51.3%
Students with Disabilities (SWD)	24,769	18.8%	25,116	19.2%	25,583	19.6%
English Learner (EL)	28,217	21.4%	28,032	21.5%	21,820	16.7%
Male	66,878	50.7%	66,454	50.9%	66,536	51.1%
Female	65,022	49.3%	64,127	49.1%	63,725	48.9%
American Indian/Alaska Native	291	0.2%	334	0.3%	316	0.2%
Asian	7,622	5.8%	7,836	6.0%	7,683	5.9%
Black/African American	21,253	16.1%	21,026	16.1%	20,908	16.0%
Hispanic/Latino	36,782	27.9%	37,038	28.4%	36,670	28.1%
Middle Eastern or North African	318	0.2%	304	0.2%	278	0.2%
Native Hawaiian or Other Pacific Islander	89	0.1%	102	0.1%	96	0.1%
Two or More Races Reported	6,432	4.9%	6,171	4.7%	5,959	4.6%
White/Caucasian	59,128	44.8%	57,789	44.2%	58,373	44.8%

Table 8.5. Student Participation by Demographic Subgroup— Mathematics Grades 6-8

Demographic	Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%
Economically Disadvantaged	65,824	50.9%	65,469	50.1%	66,524	49.7%
Students with Disabilities (SWD)	25,028	19.4%	24,635	18.9%	24,733	18.5%
English Learner (EL)	19,001	14.7%	20,383	15.6%	21,328	15.9%
Male	65,906	51.0%	66,720	51.1%	68,417	51.2%
Female	63,292	49.0%	63,927	48.9%	65,288	48.8%
American Indian/Alaska Native	302	0.2%	271	0.2%	285	0.2%
Asian	7,733	6.0%	7,772	5.9%	7,744	5.8%
Black/African American	20,511	15.9%	20,582	15.7%	21,237	15.9%
Hispanic/Latino	36,369	28.1%	37,349	28.6%	38,752	29.0%

Demographic	Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%
Middle Eastern or North African	288	0.2%	290	0.2%	301	0.2%
Native Hawaiian or Other Pacific Islander	96	0.1%	124	0.1%	128	0.1%
Two or More Races Reported	5,832	4.5%	5,640	4.3%	5,542	4.1%
White/Caucasian	58,101	45.0%	58,655	44.9%	59,766	44.7%

8.2. Scale Score Distributions

Figure 8.1 – Figure 8.4 present the spring 2025 IAR scale score distributions. The vertical y-axis labeled “Density” represents the proportion of students earning the scale score point indicated along the horizontal x-axis. The overall score scale ranges from 650 to 850, the Reading score scale ranges from 10 to 90, and the Writing score scale ranges from 10 to 60. Appendix A presents the cumulative frequency distribution for the overall scale scores, and Appendix B presents the subgroup statistics for the summative, Reading, and Writing scale scores.

Scale score distributions for mathematics peaked between approximately 700 and 750, and the distributions of the ELA/L overall scale scores were centered around 750. Reading scale scores tended to be centered around 40-55.

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

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

Figure 8.1. Scale Score Distributions—ELA/L

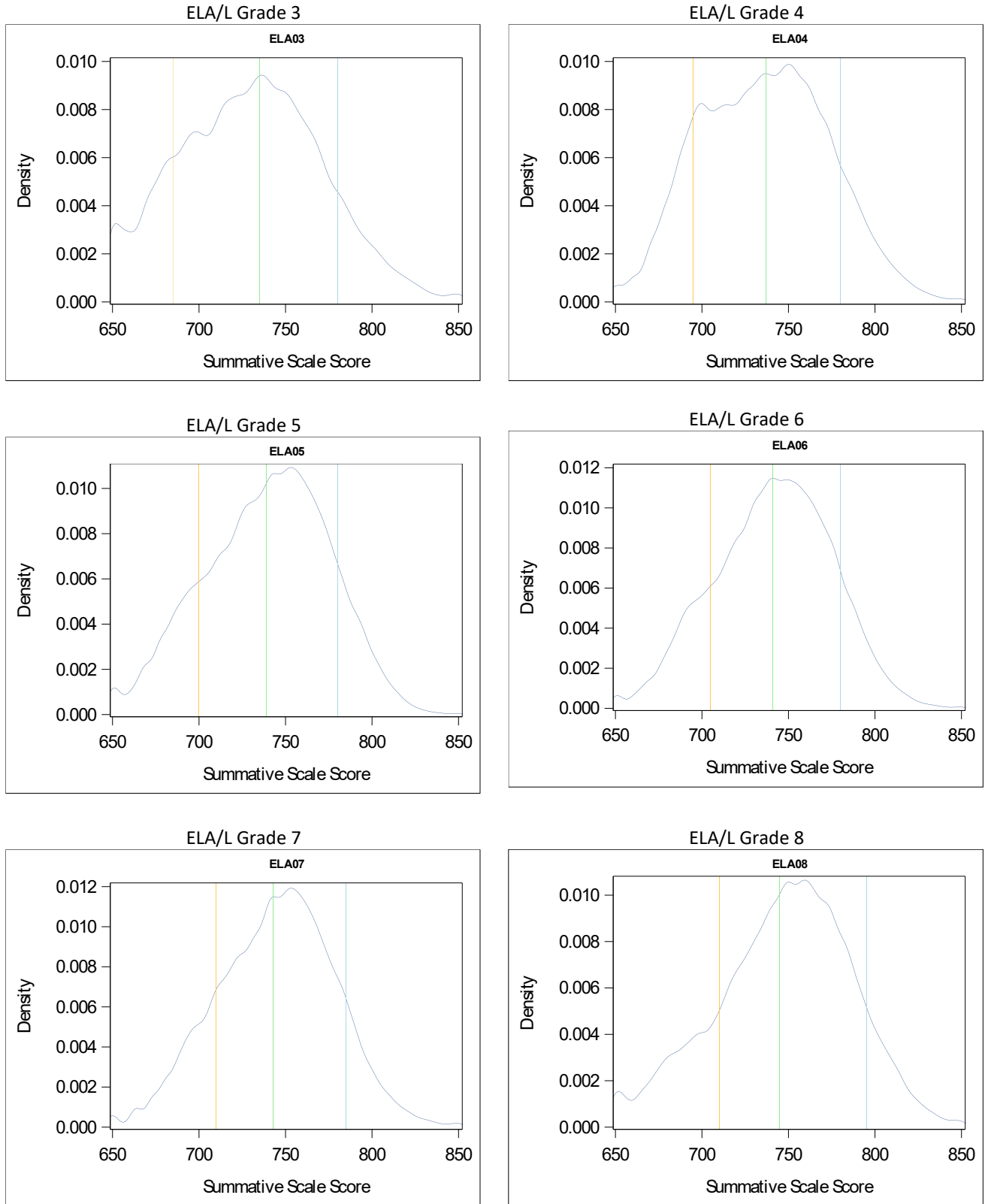


Figure 8.2. Scale Score Distributions—Reading

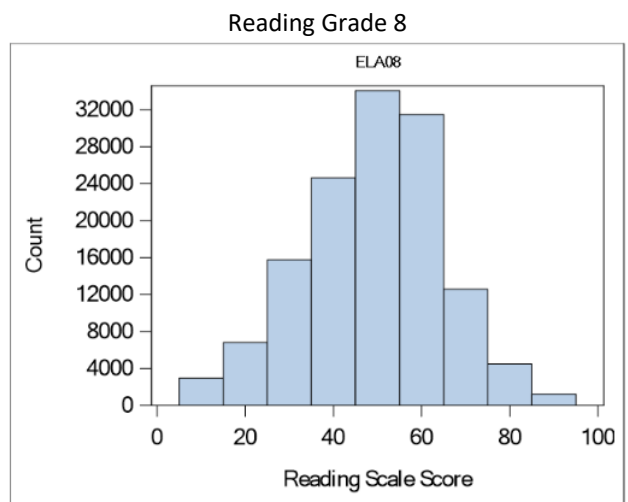
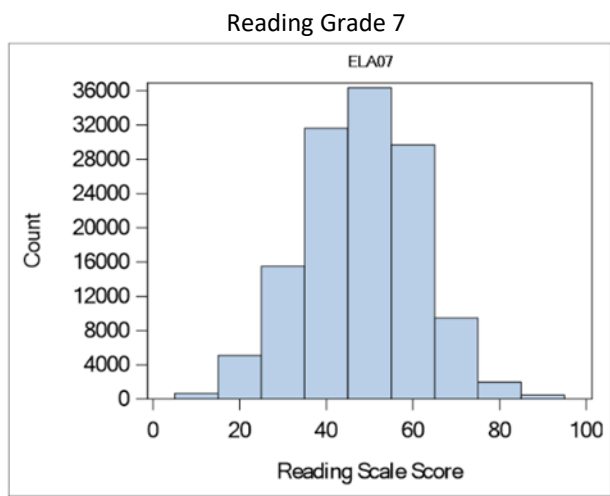
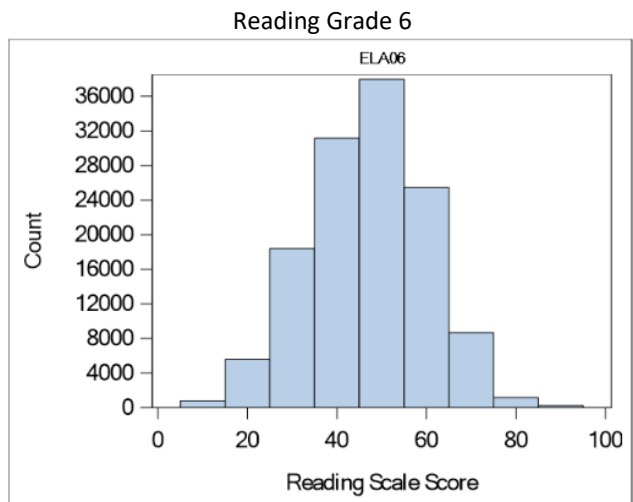
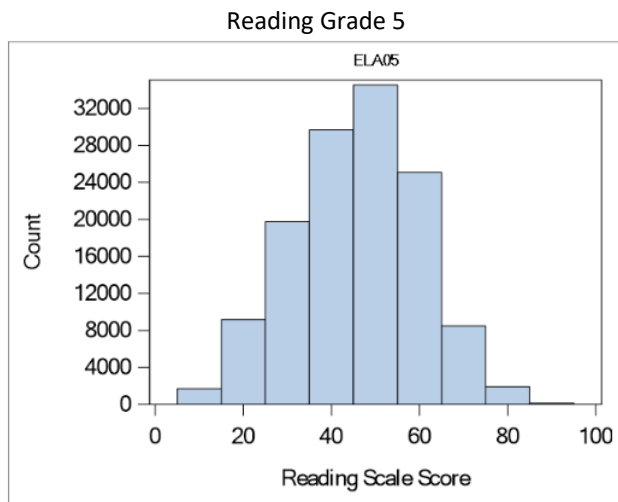
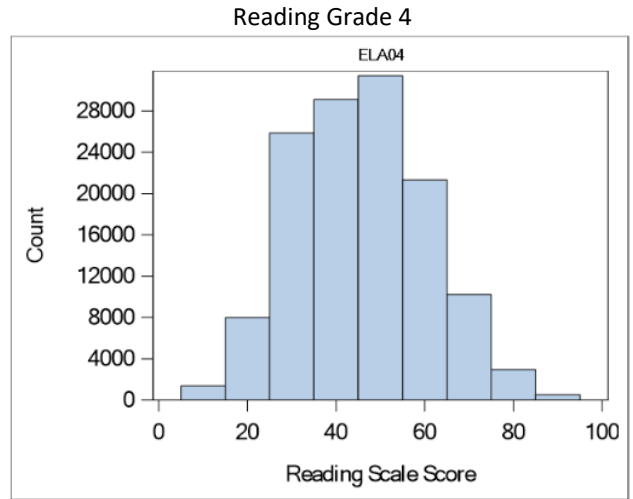
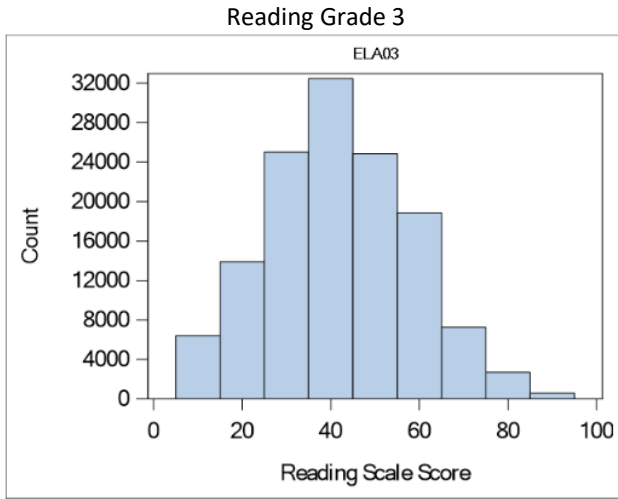


Figure 8.3. Scale Score Distributions—Writing

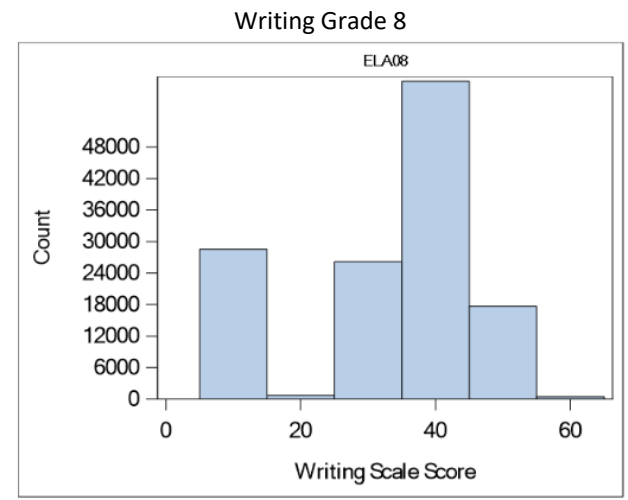
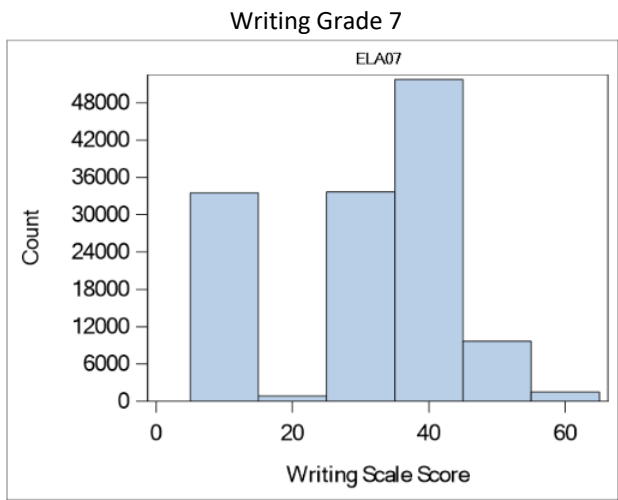
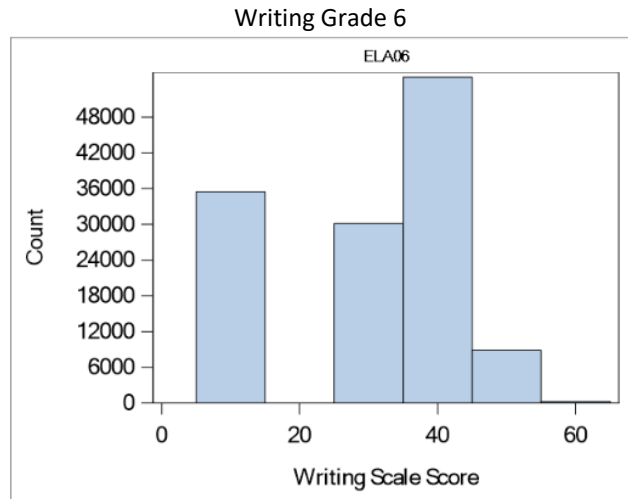
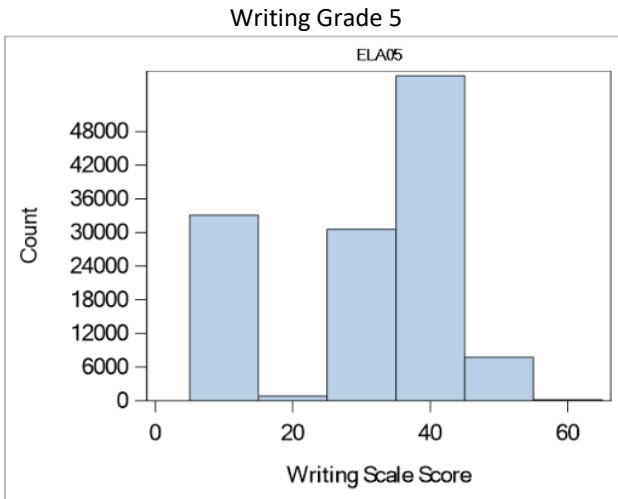
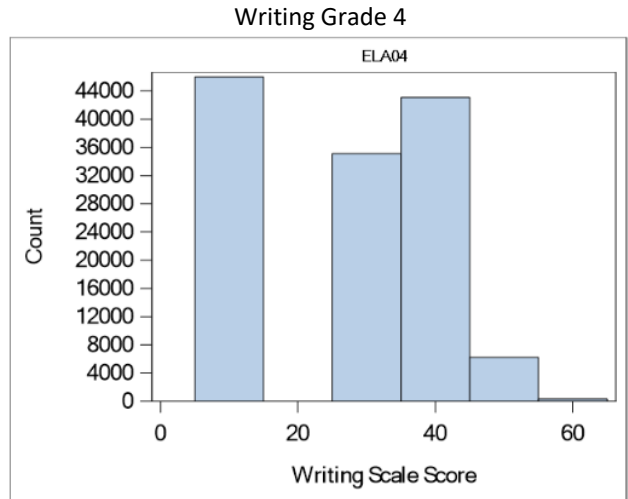
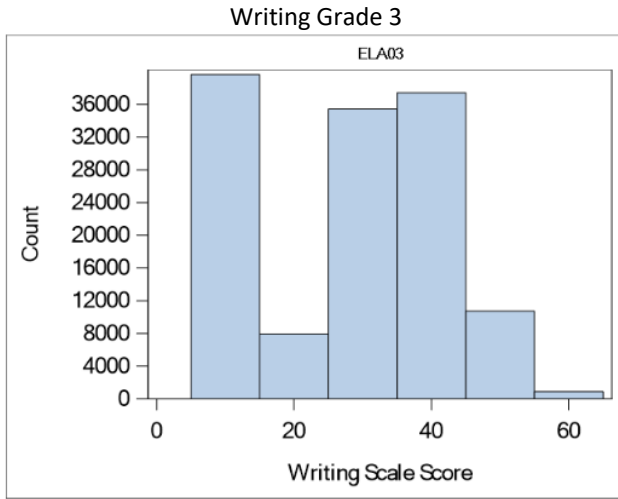
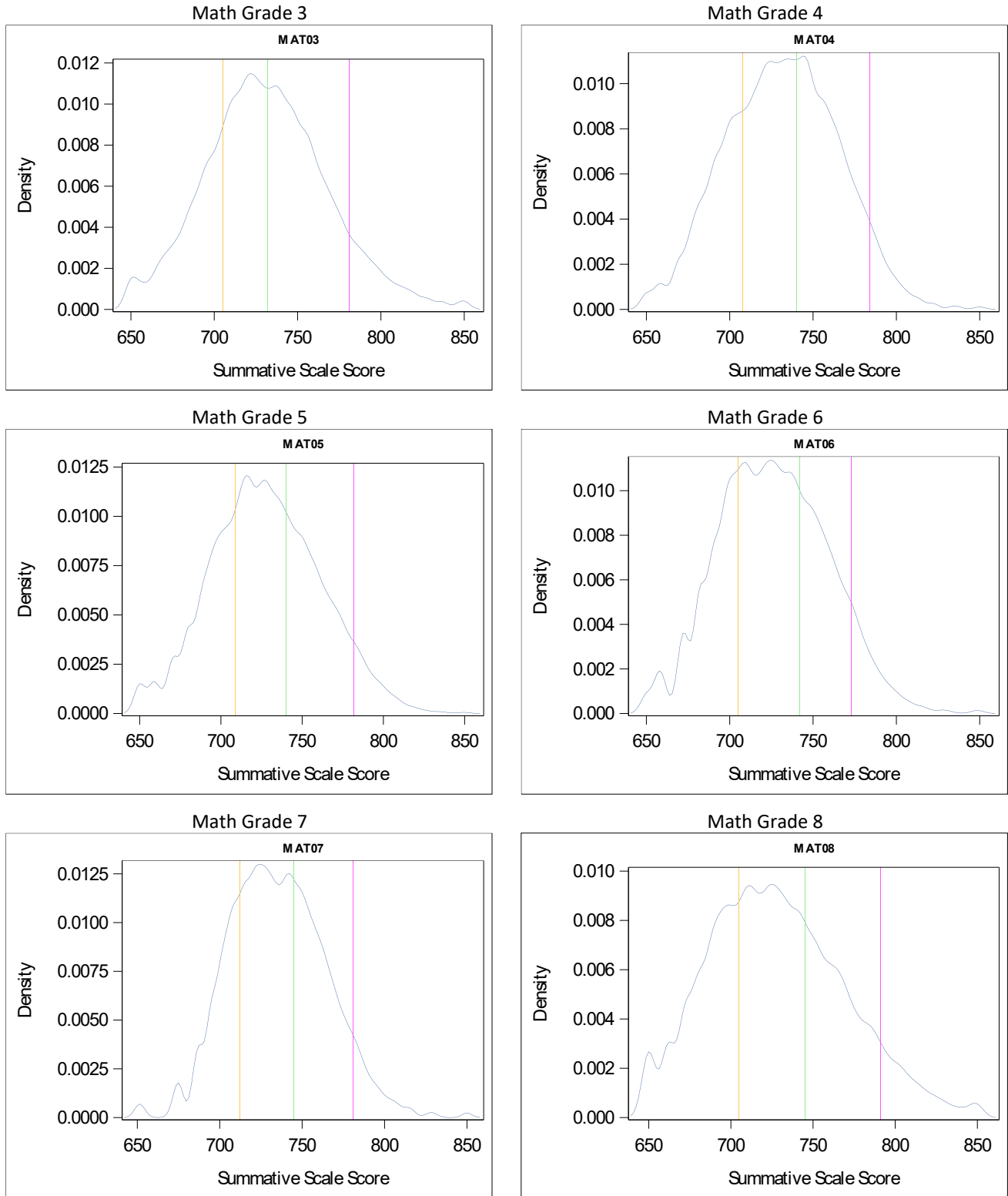


Figure 8.4. Scale Score Distributions—Mathematics



Section 9: Classical Item Analysis

This section presents item analysis results for the operational items included on the spring 2025 test forms. All assessments were pre-equated, meaning the scoring was based on item parameters estimated using data from earlier administrations. As a result, the item analysis results are from prior operational administrations that were used to make decisions during the test construction process and for score reporting.

9.1. Data Preparation

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

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

The following factors were also considered during the analyses:

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

9.2. Item Analyses

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

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

9.2.1. Item Difficulty (*P*-value)

When constructing tests, a wide range of item difficulties is desired (from easy to hard items) so that students of all ability levels can be assessed with precision. Item difficulty is measured by the *p*-value statistic bounded by 0 and 1 that indicates how easy or hard an item is for students. The *p*-value for dichotomous items is based on the proportion of students who answered an item correctly and is derived by dividing the number of students who got the item correct by the total number of students who answered it. For polytomous items, the *p*-value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item. A high *p*-value indicates that an item is easy (high proportion of students answered it correctly), whereas a low *p*-value indicates that an item is difficult. For example, a *p*-value of 0.79 indicates that 79% of students answered the item correctly. Items were flagged for review if the *p*-value was above 0.95 (i.e., too easy) or below 0.25 (i.e., too difficult).

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

Table 9.1. Summary of *p*-Values

Assessment	#Unique Items	Mean	SD	Min.	Max.	Median
ELA/L 3	40	0.42	0.15	0.17	0.68	0.42
ELA/L 4	48	0.46	0.15	0.17	0.73	0.48
ELA/L 5	50	0.43	0.15	0.15	0.71	0.41
ELA/L 6	44	0.47	0.16	0.12	0.76	0.5
ELA/L 7	53	0.48	0.17	0.16	0.84	0.44
ELA/L 8	43	0.5	0.16	0.23	0.79	0.53
Mathematics 3	69	0.55	0.22	0.26	0.9	0.5
Mathematics 4	73	0.51	0.21	0.25	0.89	0.5
Mathematics 5	77	0.44	0.2	0.05	0.9	0.41
Mathematics 6	70	0.41	0.21	0.07	0.89	0.37
Mathematics 7	69	0.44	0.17	0.17	0.82	0.41
Mathematics 8	70	0.38	0.19	0.03	0.8	0.37

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

9.2.2. Item Discrimination (*Item-Total Correlation*)

Item discrimination is represented by the item-total correlation bounded by -1 and 1 that describes the relationship between performance on a specific item and performance on the total test and indicates how well an item discriminates, or distinguishes, between low- and high-performing students. Students who do well on a test are expected to select the right answer to any given item, and students who do poorly are expected to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. An item with a high positive item-total correlation discriminates between low- and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that low-performing students performed better on an item than high-performing students, an indication that the item may be flawed. The item-total correlation was calculated for both dichotomous and polytomous items as an estimate of the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson et al., 1982). Item-total correlations below 0.15 were flagged for review.

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

Table 9.2. Summary of Item-Total Correlations

Assessment	#Unique Items	Mean	SD	Min.	Max.	Median
ELA/L 3	40	0.58	0.11	0.38	0.75	0.6
ELA/L 4	48	0.6	0.13	0.28	0.83	0.6
ELA/L 5	50	0.52	0.14	0.28	0.79	0.51
ELA/L 6	44	0.54	0.12	0.33	0.77	0.53
ELA/L 7	53	0.54	0.15	0.22	0.81	0.55
ELA/L 8	43	0.51	0.14	0.24	0.83	0.53
Mathematics 3	69	0.49	0.15	0.21	0.79	0.5
Mathematics 4	73	0.52	0.15	0.2	0.78	0.52
Mathematics 5	77	0.5	0.16	0.08	0.77	0.51
Mathematics 6	70	0.52	0.18	0.08	0.82	0.54
Mathematics 7	69	0.5	0.16	0.19	0.8	0.49
Mathematics 8	70	0.49	0.13	0.19	0.77	0.48

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

The item-total correlation was also calculated for the distractors of selected-response items to describe the relationship between selecting an incorrect response (i.e., a distractor) for an item and performance on the total test. Items with distractor-total correlations above 0.00 were flagged for review as these items may have multiple correct answers, be miskeyed, or have other content issues.

9.2.3. Percentage of Students Choosing Each Answer Option

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

9.2.4. Percentage of Students Omitting or Not Reaching Each Item

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

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

9.2.5. Distribution of Item Scores

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

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

9.3. Flagging Criteria

The review process for test items includes a systematic keycheck analysis of all operational items. During this process, operational items are evaluated for statistical flags that may indicate potential issues. If any operational items are flagged, they undergo a thorough internal review by Pearson’s psychometrics team, followed by further examination by the Illinois State Board of Education (ISBE). For the 2025 administration, no operational items were flagged based on the keycheck analysis. Field test items are also subjected to statistical flagging, and any identified items are subsequently reviewed in a formal data review process involving Illinois educators. All flagged items, whether operational or field test, must be evaluated by Illinois educators to determine their suitability for inclusion in the item bank. These items are either accepted for future use, rejected, or, when appropriate, revised and refield tested.

- *P*-values below 0.15 that indicates too difficult items
- Item-total correlations below 0.20 indicate items that do not effectively distinguish between higher- and lower-performing students. Without approval from ISBE, items with item-total correlations near or below zero will not be used operationally.
- Distractor-total correlations above 0.05 as these items may have multiple correct answers, be miskeyed, or have other content issues
- 40% or more of students choosing a distractor over the keyed response, which indicates that the item may have multiple correct answers or is miskeyed
- High omit and not-reached rates above 5%, which may indicate that students did not have adequate time if patterns of high omit or not-reached rates for items are located near the end of a test section

Section 10: Differential Item Functioning (DIF)

Differential item functioning (DIF) is a statistical procedure used to flag items for potential bias when students from different demographic groups with the same overall ability have a different probability of getting an item correct (e.g., an item that seems easy for female students but not for male students). This section presents DIF results for the operational items included on the spring 2025 test forms. All assessments were pre-equated, meaning that the scoring was based on item parameters estimated using data from earlier administrations. As a result, the DIF results are from prior operational administrations.

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

10.1. DIF Methods

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

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

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

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

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

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

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

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

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

10.2. Classification

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

Table 10.1. DIF Categories

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

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

10.3. Comparisons

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

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

Table 10.2. DIF Comparison Groups

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

10.4. Results

Appendix C presents the DIF results for the operational items included on the spring 2025 test forms (i.e., the DIF results are from a previous year’s bank). Spoiled or “do not score” items were excluded from the total test score for each form in the DIF analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, multiple correct answers, or no correct answers. However, the tables may include items for certain grade levels that were excluded from scoring based on later analyses.

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

Section 11: Calibration, Equating, and Scaling

This section describes the item response theory (IRT) model used in this assessment program, provides descriptive statistics of the item parameters, and describes how the reporting scale was established. All IAR assessments in spring 2025 were pre-equated.

11.1. IRT Model

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

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

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

11.2. IRT Analysis Results

Table 11.1 and Table 11.2 present the pre-equated IRT b - and a -parameter estimates for the operational items administered in spring 2025. The tables present the statistics for the Reading and Writing items for ELA/L and by item type for mathematics (see Section 2.3 for a description of the item types), including the total number of items and score points, mean, standard deviation (SD), minimum, and maximum.

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

Grade	Item Grouping	Points #	Points %	#Items	b				a			
					Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
3	All Items	100	100.00	45	0.34	0.80	-1.00	1.74	0.62	0.20	0.30	1.05
	Reading	70	70.00	35	0.03	0.62	-1.00	1.28	0.54	0.15	0.30	0.84
	Writing	30	30.00	10	1.40	0.32	0.78	1.74	0.87	0.12	0.65	1.05
4	All Items	115	100.00	52	0.19	0.79	-1.71	1.86	0.60	0.23	0.21	1.14
	Reading	88	76.52	44	-0.01	0.67	-1.71	1.45	0.53	0.17	0.21	1.02
	Writing	27	23.48	8	1.33	0.27	1.03	1.86	0.98	0.11	0.81	1.14
5	All Items	123	100.00	55	0.41	0.86	-1.55	2.33	0.52	0.24	0.19	1.11
	Reading	90	73.17	45	0.21	0.81	-1.55	2.33	0.43	0.14	0.19	0.91
	Writing	33	26.83	10	1.31	0.36	0.91	1.97	0.94	0.10	0.77	1.11
6	All Items	108	100.00	48	0.22	0.80	-1.44	1.86	0.55	0.23	0.25	1.16
	Reading	80	74.07	40	0.02	0.69	-1.44	1.60	0.46	0.13	0.25	0.75
	Writing	28	25.93	8	1.25	0.41	0.72	1.86	0.98	0.11	0.87	1.16
7	All Items	131	100.00	58	0.13	0.72	-1.54	1.70	0.64	0.27	0.21	1.26
	Reading	96	73.28	48	-0.01	0.70	-1.54	1.70	0.56	0.23	0.21	1.26
	Writing	35	26.72	10	0.82	0.31	0.27	1.33	1.02	0.10	0.85	1.15
8	All Items	106	100.00	47	0.09	0.85	-1.26	2.37	0.55	0.31	0.16	1.34
	Reading	78	73.58	39	-0.09	0.82	-1.26	2.37	0.44	0.17	0.16	0.86
	Writing	28	26.42	8	0.95	0.31	0.45	1.33	1.14	0.10	1.03	1.34

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

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

Grade	Item Grouping	Points #	Points %	#Items	<i>b</i> Mean	<i>b</i> SD	<i>b</i> Min.	<i>b</i> Max.	<i>a</i> Mean	<i>a</i> SD	<i>a</i> Min.	<i>a</i> Max.
3	All Items	99	100.00	69	-0.37	1.21	-2.63	2.02	0.78	0.30	0.25	1.41
	Type I	64	64.65	59	-0.57	1.19	-2.63	2.02	0.82	0.31	0.25	1.41
	Type II	17	17.17	5	0.83	0.59	-0.13	1.49	0.55	0.15	0.38	0.73
	Type III	18	18.18	5	0.77	0.24	0.47	1.01	0.62	0.16	0.42	0.84
4	All Items	114	100.00	73	-0.23	0.97	-2.48	1.63	0.83	0.28	0.23	1.46
	Type I	73	64.04	62	-0.39	0.96	-2.48	1.63	0.86	0.29	0.23	1.46
	Type II	14	12.28	4	0.71	0.35	0.20	0.99	0.67	0.13	0.52	0.83
	Type III	27	23.68	7	0.63	0.45	0.25	1.50	0.69	0.18	0.46	0.91
5	All Items	119	100.00	77	-0.05	1.34	-6.36	2.41	0.74	0.30	0.11	1.47
	Type I	78	65.55	66	-0.21	1.38	-6.36	2.41	0.76	0.32	0.11	1.47
	Type II	17	14.29	5	0.81	0.33	0.32	1.27	0.63	0.19	0.37	0.85
	Type III	24	20.17	6	0.97	0.27	0.52	1.19	0.63	0.16	0.51	0.86
6	All Items	113	100.00	70	0.23	1.16	-2.65	4.01	0.80	0.28	0.10	1.33
	Type I	69	61.06	58	0.12	1.21	-2.65	4.01	0.82	0.30	0.10	1.33
	Type II	20	17.70	6	0.99	0.70	0.40	2.29	0.69	0.15	0.44	0.89
	Type III	24	21.24	6	0.53	0.54	-0.50	0.92	0.73	0.16	0.56	0.92
7	All Items	103	100.00	69	0.33	0.82	-1.26	2.62	0.76	0.30	0.23	1.43
	Type I	71	68.93	60	0.25	0.83	-1.26	2.62	0.77	0.31	0.23	1.43
	Type II	14	13.59	4	0.76	0.52	0.07	1.28	0.61	0.22	0.33	0.79
	Type III	18	17.48	5	1.02	0.35	0.52	1.49	0.71	0.08	0.62	0.78
8	All Items	111	100.00	70	0.67	1.05	-1.36	3.24	0.67	0.24	0.20	1.34
	Type I	70	63.06	59	0.47	1.01	-1.36	3.24	0.69	0.26	0.20	1.34
	Type II	17	15.32	5	1.79	0.48	1.24	2.36	0.55	0.13	0.42	0.77
	Type III	24	21.62	6	1.78	0.40	1.10	2.21	0.55	0.10	0.39	0.65

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

11.3. Establishing the Reporting Scale

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

11.3.1. Summative Score Scale and Performance Levels

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

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

$$ScaleScore_{2015} = A_{2015} \times \theta_{2015} + B_{2015} \quad (11-2)$$

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

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

and

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

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

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

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

The slope ($slope_{2015_to_2016}$) and intercept ($intercept_{2015_to_2016}$) generated during the year-to-year linking defined the linear relationship between the 2015 theta scale (θ_{2015}) and the 2016 theta scale (θ_{2016}). These values were included in the scale score formula, and the formulas were used to solve for the slope (SA_{2016}) and (SB_{2016}) intercept for 2016. The slope (A_{2016}) was updated using the following formula:

$$SA_{2016} = \frac{A_{2015}}{slope_{2015_to_2016}} \quad (11-6)$$

where A_{2015} is the current scale score multiplicative constant, $slope_{2015_to_2016}$ is the multiplicative coefficient from the year-to-year linking, and SA_{2016} is the scale score slope constant for 2016 and beyond. The intercept (B_{2016}) was updated using the following formula:

$$SB_{2016} = B_{2015} - A_{2016} \times intercept_{2015_to_2016} \quad (11-7)$$

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

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

11.3.2. Reading and Writing Claim Scale

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

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

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

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

11.3.3. Subclaims Scale

11.4. Types of Scores on the IAR Individual Student Report

Student performance on the IAR is described on the individual student report using scale scores, performance levels, and subclaim performance indicators. State average results are included in relevant sections of the report to help parents understand how their child’s performance compares to that of other students. The reader can find more information about the score reports in the [Illinois Assessment of Readiness Score Report Interpretation Guide For Parents](#).

11.4.1. Scale Score

A scale score is a numerical value that summarizes student performance. Not all students respond to the same set of test items, so each student’s raw score (actual points earned on test items) is adjusted for the slight differences in difficulty among the various forms and administrations of the test. The resulting scale score allows for an accurate comparison across test forms and administration years within a grade or course and content area. IAR reports provide overall scale scores for English language arts/literacy and mathematics, which determine a student’s performance level. IAR scale scores range from 650 to 850 for all tests. Additionally, IAR English language arts/literacy reports provide separate scale scores for both Reading and Writing. IAR Reading scale scores range from 10 to 90, and IAR Writing scale scores range from 10 to 60. For example, a student who earns an overall scale score of 800 on one form of the grade 8 mathematics assessment would be expected to earn an overall scale score of 800 on any other form of the grade 8 mathematics assessment. Furthermore, the student’s overall scale score and level of mastery of concepts and skills would be comparable to a student who took the same assessment the previous year or following year.

11.4.2. Performance Level

Starting in 2025, the IAR performance level structure has been updated to include four performance levels: Below Proficient, Approaching Proficient, Proficient, and Above Proficient. Each performance level is a broad, categorical level defined by a student’s overall scale score and is used to report overall student performance by describing how well students met the expectations for their grade level or course. Each performance level is defined by a range of overall scale scores for the assessment. The new performance levels for the Illinois Assessment of Readiness are as follows:

- Level 4: Above Proficient
- Level 3: Proficient
- Level 2: Approaching Proficient
- Level 1: Below Proficient

Students performing at Levels 3 (Proficient) and 4 (Above Proficient) have demonstrated readiness for the next grade level or course. The updated performance level labels, descriptors, and cut scores align with this revised structure. Performance Level Descriptors (PLDs) describe the knowledge, skills, and practices that students should know and be able to demonstrate at each performance level in each content area (ELA/L and mathematics), and at each grade level or course.

11.4.3. Subclaim Performance Indicators

Subclaim performance indicators for the IAR are reported using graphical representations that indicate how the student performed relative to the overall performance of students who were Proficient or Approaching Proficient for the content area. Beginning in 2025, the subclaim reporting structure aligns with the revised four-level performance framework, but subclaims themselves are summarized in three categories:

- High (H)
- Medium (M)
- Low (L)

It is important to note that the readiness indicators previously labeled as High (H), Medium (M), and Low (L) are no longer directly comparable to prior years. The shift to four performance levels redefines thresholds and category descriptions, which means H, M, and L cannot be mapped one-to-one with the new levels. Historic interpretations of these indicators should not be used for direct year-over-year comparisons.

The fundamental scoring process for subclaims remains unchanged: subclaim scores are determined using the IRT theta (θ) scale, with cut scores for each performance level set according to updated 2025 definitions. Though the underlying scale and calculation methodology are consistent, the way student readiness is interpreted and reported now reflects the new performance levels.

11.4.4. Conversion Tables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$ScaleScore = A \times \theta + B \quad (11-14)$$

$$CSEM = CSEM_\theta \times A \quad (11-15)$$

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

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

11.4.5. Scaling Constants

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

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

Assessment	Cut	Theta	Scale Score	A	B
ELA/L 3	Level 2	-1.3733	685	36.7227	735.4297
	Level 3	-0.0117	735		
	Level 4	1.2137	780		
ELA/L 4	Level 2	-1.4589	695	31.5462	741.0214
	Level 3	-0.1275	737		
	Level 4	1.2356	780		
ELA/L 5	Level 2	-1.3411	700	29.4580	739.5050
	Level 3	-0.0171	739		
	Level 4	1.3747	780		
ELA/L 6	Level 2	-1.1890	705	28.3160	738.6673
	Level 3	0.0824	741		
	Level 4	1.4597	780		
ELA/L 7	Level 2	-0.9539	710	33.9161	742.3542
	Level 3	0.0190	743		
	Level 4	1.2574	785		
ELA/L 8	Level 2	-0.9799	710	34.1183	743.4330
	Level 3	0.0459	745		
	Level 4	1.5114	795		

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

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

Table 11.6. Cut Scores and Scaling Constants—Mathematics

Assessment	Cut	Theta	Scale Score	A	B
Mathematics 3	Level 2	-1.2584	705	32.1135	745.4119
	Level 3	-0.4176	732		
	Level 4	1.1082	781		
Mathematics 4	Level 2	-1.1166	708	29.9167	741.4049
	Level 3	-0.0470	740		
	Level 4	1.4238	784		
Mathematics 5	Level 2	-1.1471	709	29.0301	742.2997
	Level 3	-0.0792	740		
	Level 4	1.3676	782		
Mathematics 6	Level 2	-1.2053	705	28.1465	738.9252
	Level 3	0.1092	742		
	Level 4	1.2106	773		
Mathematics 7	Level 2	-0.9684	712	25.1033	736.3102
	Level 3	0.3462	745		
	Level 4	1.7802	781		
Mathematics 8	Level 2	-0.7333	705	32.9505	729.1640
	Level 3	0.4806	745		
	Level 4	1.8766	791		

Section 12: Quality Control Procedures

Quality control in a testing program is a comprehensive and ongoing process. This section describes procedures put into place to monitor the quality of the item bank, test form, and ancillary material development. Additional quality information can be found in the Program Quality Plan document.

12.1. Quality Control of the Item Bank

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

12.2. Quality Control of Test Form Development

The operational test forms were built based on targets and the established blueprints set, and items were pulled into forms based on the criteria approved in the test specifications. The forms then went through an internal review process to ensure content accuracy, completeness, style guide conformity, and tools function. Revisions were incorporated into the forms before final review and approval. The forms quality assurance was performed by Pearson’s Assessment and Information Quality (AIQ) organization. AIQ completed a comprehensive review of all online forms for the administration cycle. This group is part of Pearson’s larger Organizational Quality group and operates exclusively to validate form operability. The group verifies that the functionality of every online form is working to specifications. The overall functionality and maneuverability of each form is checked, and the behavior of each item within the form is verified. The items within each form were tested to verify that they operated as expected for students. As a further aspect of the testing process, AIQ confirmed that forms were loaded correctly and that the audio was correct when compared to text. Sections and overviews were reviewed. Technology-enhanced items also were tested as an additional measure. As enumerated in the Technology Guidelines for Assessments, user interfaces were compatible with a range of common computer devices, operating systems, and browsers.

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

12.3. Quality Control of Test Materials

Pearson provided high-quality materials in a timely and efficient manner to meet the test administration needs. Finline printed the non-scoreable materials for grade 4–8, while all scoreable materials were

printed at Pearson. Strict security requirements were employed to protect secure materials production. Materials were produced according to the style guide and to the detailed specifications supplied in the materials list.

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

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

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

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

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

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

12.4. Quality Control of Scoring

12.4.1. Quality Control of Scanning

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

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

12.4.2. Quality Control of Image Editing

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

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

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

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

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

12.4.3. Quality Control of Answer Document and Data

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

- Raw score validation (e.g., score key validation; evidence statement, field test nonscore; double-grid combinations; possible correct combination, if applicable; out-of-range/negative test cases)

- Matching (e.g., validation of high-confidence criteria, low-confidence criteria, cross document, external or forced matching by customer; prior to and after data updates; extract file of matched and unmatched documents)
- Demographic update tests (e.g., verification of data extract against corresponding layout; valid values for updatable fields; invalid values for updatable/nonupdatable fields; negative test for nonexisting record or empty file)

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

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

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

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

12.5. Quality Control of Psychometric Processes

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

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

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

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

Finally, conversion tables are used to generate reported scores for students and must be accurate. Comprehensive records were maintained on item-level decisions, and thorough checks were made to ensure that the correct items were included in the final score. Pre-equated conversion tables were developed independently by two psychometricians and completely matched. A reasonableness check was also conducted by psychometricians for each content and grade level to make sure the results were in alignment with observations during the analyses prior to conversion table creation. Refer to Section 11.4.4 for the procedure to create the conversion tables.

Section 13: Reliability

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

- Internal consistency
- Standard error of measurement (SEM)
- Decision accuracy and consistency
- Inter-rater agreement (see Section 0)

13.1. Internal Consistency and SEM

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

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

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

13.1.1. Raw Score Estimation

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

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

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

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

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

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

The formula for the SEM is as follows:

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

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

13.1.2. Scale Score Estimation

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

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

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

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

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

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

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

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

13.1.3. Results

Reliability results are presented at the overall, subgroup, and subclaim levels. Table 13.1 and Table 13.2 present the raw and scale score test reliability estimates for the total testing group, including the average reliability that is estimated by averaging the internal consistency estimates computed for all the individual forms of the test. The spring 2025 administration had three forms: two online core forms (Online1 and Online2) and one accommodated form (ACC1) taken by a small number of students. The tables present the average reliability across all forms and by form.

The average raw score reliability estimates for ELA/L range from 0.87 to 0.91, and the average raw score SEM is consistently between 3 and 4 points. The average reliability estimates for mathematics range from 0.89 to 0.93, and the raw score SEM was consistently about 3 points. Average scale score reliabilities for ELA/L range from 0.87 to 0.91, and the average SEM ranges from 9.94 to 13.68. Average scale score reliability estimates range from 0.88 to 0.92 for mathematics, and the average scale score SEM ranges from 8.57 to 12.19.

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

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

Assessment	#Forms	Max. Possible Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
ELA/L 3	3	55	3.46	0.87	65,523	0.90	65,211	0.88	931	0.82
ELA/L 4	3	74	3.82	0.91	64,883	0.93	64,675	0.91	869	0.88
ELA/L 5	3	74	3.82	0.89	64,800	0.92	64,433	0.90	887	0.84
ELA/L 6	3	74	4.09	0.90	64,314	0.93	64,019	0.91	662	0.87
ELA/L 7	3	74	4.37	0.91	65,139	0.93	64,682	0.91	559	0.89
ELA/L 8	3	74	4.25	0.91	66,565	0.92	66,249	0.91	543	0.90
Mathematics 3	3	52	2.96	0.91	87,873	0.92	42,477	0.91	1,118	0.90
Mathematics 4	3	52	2.93	0.93	87,096	0.94	42,087	0.93	992	0.92
Mathematics 5	3	52	2.90	0.89	85,694	0.93	43,199	0.92	999	0.82
Mathematics 6	3	52	2.88	0.92	85,421	0.94	42,417	0.93	712	0.88
Mathematics 7	3	52	2.97	0.92	85,011	0.94	44,259	0.93	556	0.90
Mathematics 8	3	52	2.79	0.89	94,594	0.92	37,840	0.92	557	0.83

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

Assessment	#Forms	Avg. Scale Score SEM	Avg. Reliability	Online1 Reliability	Online2 Reliability	ACC1 Reliability
ELA/L 3	3	13.68	0.88	0.88	0.88	0.87
ELA/L 4	3	10.54	0.90	0.90	0.90	0.91
ELA/L 5	3	11.25	0.87	0.88	0.88	0.86
ELA/L 6	3	9.94	0.89	0.90	0.89	0.89
ELA/L 7	3	10.47	0.91	0.92	0.91	0.91
ELA/L 8	3	11.17	0.90	0.90	0.90	0.90
Mathematics 3	3	10.45	0.91	0.91	0.90	0.91
Mathematics 4	3	8.70	0.92	0.92	0.92	0.92
Mathematics 5	3	9.33	0.91	0.91	0.91	0.90
Mathematics 6	3	8.69	0.92	0.91	0.92	0.92
Mathematics 7	3	8.57	0.90	0.90	0.90	0.90
Mathematics 8	3	12.19	0.88	0.88	0.88	0.87

Table 13.3 and Table 13.4 present the reliability estimates for each major claim and subclaim. Subclaims with greater numbers of points tend to have greater reliability estimates. Across grades, the average reliabilities range from 0.55 to 0.88 for Reading and 0.65 to 0.85 for Writing. The average reliabilities across all subclaims for mathematics range from 0.55 to 0.86 across grades.

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

Subclaim	Assessment	RS Range	Avg. Reliability
Reading: Total	ELA/L 3	30–31	0.82
	ELA/L 4	40–44	0.88
	ELA/L 5	40–44	0.84
	ELA/L 6	40–44	0.85
	ELA/L 7	40–44	0.86
	ELA/L 8	40–44	0.85
Reading: Literature	ELA/L 3	11–14	0.70
	ELA/L 4	16–18	0.75
	ELA/L 5	16–18	0.69
	ELA/L 6	16–18	0.72
	ELA/L 7	16–18	0.77
	ELA/L 8	16–18	0.72
Reading: Information	ELA/L 3	11–11	0.59
	ELA/L 4	16–16	0.70
	ELA/L 5	16–16	0.67
	ELA/L 6	14–16	0.65
	ELA/L 7	16–16	0.68
	ELA/L 8	16–16	0.61
Reading: Vocabulary	ELA/L 3	6–8	0.55
	ELA/L 4	8–10	0.68
	ELA/L 5	8–10	0.55
	ELA/L 6	8–10	0.62
	ELA/L 7	8–10	0.57
	ELA/L 8	8–12	0.67
Writing: Total	ELA/L 3	24–24	0.73
	ELA/L 4	27–30	0.79
	ELA/L 5	27–30	0.79
	ELA/L 6	30–30	0.79
	ELA/L 7	30–30	0.82
	ELA/L 8	30–30	0.84
Writing Expression	ELA/L 3	18–18	0.65
	ELA/L 4	21–24	0.74
	ELA/L 5	21–24	0.75
	ELA/L 6	24–24	0.76
	ELA/L 7	24–24	0.83
	ELA/L 8	24–24	0.84
Writing: Knowledge Language & Conventions	ELA/L 3	6–6	0.76
	ELA/L 4	6–6	0.81
	ELA/L 5	6–6	0.80
	ELA/L 6	6–6	0.77
	ELA/L 7	6–6	0.84
	ELA/L 8	6–6	0.85

Note. RS = raw score, Avg. = average

Table 13.4. Average Reliability Estimates by Subclaim—Mathematics

Subclaim	Assessment	RS Range	Avg. Reliability
Major Content	Mathematics 3	20–20	0.85
	Mathematics 4	21–21	0.86
	Mathematics 5	20–20	0.75
	Mathematics 6	20–20	0.83
	Mathematics 7	20–20	0.83
	Mathematics 8	20–22	0.78
Additional & Supporting Content	Mathematics 3	10–10	0.61
	Mathematics 4	9–9	0.65
	Mathematics 5	10–10	0.65
	Mathematics 6	10–10	0.59
	Mathematics 7	10–10	0.65
	Mathematics 8	8–10	0.61
Mathematics Reasoning	Mathematics 3	10–10	0.62
	Mathematics 4	10–10	0.71
	Mathematics 5	10–10	0.59
	Mathematics 6	10–10	0.67
	Mathematics 7	10–10	0.73
	Mathematics 8	10–10	0.55
Modeling Practice	Mathematics 3	12–12	0.67
	Mathematics 4	12–12	0.68
	Mathematics 5	12–12	0.68
	Mathematics 6	12–12	0.74
	Mathematics 7	12–12	0.76
	Mathematics 8	12–12	0.67

Note. RS = raw score, Avg. = average

13.2. Decision Accuracy and Consistency

The reliability of the classifications for the students was calculated using the computer program BB-CLASS (Brennan, 2004), which operationalizes a statistical method developed by Livingston and Lewis (1993, 1995). As Livingston and Lewis (1993, 1995) explain, this method uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient) to estimate two kinds of statistics, decision accuracy and decision consistency. Decision accuracy refers to the extent to which the classifications of students based on their scores on the test form agree with the classifications made based on the classifications that would be made if the test scores were perfectly reliable. Decision consistency refers to the agreement between these classifications based on two nonoverlapping, equally difficult forms of the test. In the case when no parallel test forms exist (which is the case here), BB-CLASS computes the decision consistency by comparing the actual observed score distribution with observed score distribution based on a hypothetical test form predicted from the model.

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

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

Table 13.5. Decision Accuracy and Consistency Summary

Statistic	Assessment	Exact Level	Level 3 or Higher vs. 2 or Lower	
Accuracy	ELA/L 3	0.76	0.89	
	ELA/L 4	0.80	0.91	
	ELA/L 5	0.77	0.90	
	ELA/L 6	0.78	0.90	
	ELA/L 7	0.79	0.91	
	ELA/L 8	0.79	0.91	
	Mathematics 3	0.79	0.91	
	Mathematics 4	0.83	0.92	
	Mathematics 5	0.78	0.90	
	Mathematics 6	0.81	0.92	
	Mathematics 7	0.82	0.92	
	Mathematics 8	0.77	0.91	
	Consistency	ELA/L 3	0.66	0.85
		ELA/L 4	0.72	0.87
ELA/L 5		0.68	0.86	
ELA/L 6		0.70	0.87	
ELA/L 7		0.71	0.87	
ELA/L 8		0.70	0.88	
Mathematics 3		0.70	0.87	
Mathematics 4		0.75	0.89	
Mathematics 5		0.69	0.86	
Mathematics 6		0.73	0.89	
Mathematics 7		0.74	0.88	
Mathematics 8		0.68	0.87	

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

Section 14: Validity

As stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations” (p. 11). The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, and psychometric characteristics. This chapter summarizes the evidence based on test content and the internal structure of the tests.

14.1. Evidence Based on Test Content

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

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

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

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

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

14.2. Evidence Based on Internal Structure

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

14.2.1. Intercorrelations

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

Table 14.1 and Table 14.2 present the weighted average Pearson intercorrelations between subclaims by averaging the intercorrelations computed for all the core operational forms of each assessment. The shaded values along the diagonal are the reliabilities from Section 13.1.3. The average intercorrelations are provided in the lower portion of the tables, and the total sample sizes are provided in the upper portion of the tables. Results are as follows:

- For ELA/L, the WR, WE, and WKL scores show strong intercorrelations, indicating a high degree of association among the writing-related subclaims. The RL, RI, and RV subclaims, which all pertain to Reading, display moderate to high correlations with each other. Additionally, the WR claim and the WE and WKL subclaims are moderately correlated with the RD subclaims (RL, RI, and RV). These moderate to high intercorrelations among ELA/L subclaims provide evidence for the unidimensionality of the ELA/L assessments. Moreover, the observed correlations support the sufficiency of the claim scores for individual student reporting, as the subclaims and claims are adequately related to one another.
- In mathematics, the intercorrelations are moderate overall. The MC subclaim consistently exhibits somewhat higher correlations with the ASC, MR, and MP subclaims, while the intercorrelations among ASC, MR, and MP are generally slightly lower. These patterns suggest that, while the mathematics assessments are likely unidimensional, there may be minor secondary dimensions present. Nonetheless, the level of intercorrelation is sufficient to support the structure of the mathematics tests as primarily measuring a single construct.

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

Grade	Subclaim	RD	RL	RI	RV	WR	WE	WKL
3	RD	0.82	131,665	131,665	131,665	131,665	131,665	131,665
	RL	0.90	0.70	131,665	131,665	131,665	131,665	131,665
	RI	0.85	0.62	0.59	131,665	131,665	131,665	131,665
	RV	0.85	0.66	0.60	0.55	131,665	131,665	131,665
	WR	0.73	0.65	0.67	0.56	0.73	131,665	131,665
	WE	0.70	0.62	0.66	0.54	0.99	0.65	131,665
	WKL	0.67	0.61	0.60	0.54	0.87	0.77	0.76
	4	RD	0.88	130,427	130,427	130,427	130,427	130,427
RL	0.93	0.75	130,427	130,427	130,427	130,427	130,427	
RI	0.88	0.72	0.70	130,427	130,427	130,427	130,427	
RV	0.87	0.71	0.67	0.68	130,427	130,427	130,427	
WR	0.74	0.66	0.73	0.58	0.79	130,427	130,427	
WE	0.72	0.64	0.72	0.57	0.99	0.74	130,427	
WKL	0.72	0.65	0.71	0.57	0.93	0.89	0.81	
5	RD	0.84	130,120	130,120	130,120	130,120	130,120	130,120
	RL	0.90	0.69	130,120	130,120	130,120	130,120	130,120
	RI	0.89	0.71	0.67	130,120	130,120	130,120	130,120
	RV	0.80	0.58	0.57	0.55	130,120	130,120	130,120
	WR	0.72	0.67	0.72	0.47	0.79	130,120	130,120
	WE	0.71	0.65	0.71	0.46	0.99	0.75	130,120
	WKL	0.71	0.66	0.70	0.45	0.94	0.90	0.80
	6	RD	0.85	128,995	128,995	128,995	128,995	128,995
RL		0.92	0.72	128,995	128,995	128,995	128,995	128,995
RI		0.90	0.71	0.65	128,995	128,995	128,995	128,995
RV		0.82	0.64	0.63	0.62	128,995	128,995	128,995
WR		0.76	0.67	0.76	0.54	0.79	128,995	128,995
WE		0.75	0.66	0.75	0.53	1.00	0.76	128,995
WKL		0.75	0.66	0.75	0.54	0.96	0.94	0.77

Grade	Subclaim	RD	RL	RI	RV	WR	WE	WKL
7	RD	0.86	130,380	130,380	130,380	130,380	130,380	130,380
	RL	0.93	0.77	130,380	130,380	130,380	130,380	130,380
	RI	0.90	0.75	0.68	130,380	130,380	130,380	130,380
	RV	0.81	0.65	0.61	0.57	130,380	130,380	130,380
	WR	0.74	0.66	0.77	0.48	0.82	130,380	130,380
	WE	0.73	0.65	0.76	0.47	1.00	0.83	130,380
	WKL	0.74	0.67	0.77	0.49	0.97	0.95	0.84
8	RD	0.85	133,357	133,357	133,357	133,357	133,357	133,357
	RL	0.90	0.72	133,357	133,357	133,357	133,357	133,357
	RI	0.85	0.63	0.61	133,357	133,357	133,357	133,357
	RV	0.82	0.62	0.57	0.67	133,357	133,357	133,357
	WR	0.74	0.65	0.72	0.54	0.84	133,357	133,357
	WE	0.74	0.64	0.72	0.53	1.00	0.84	133,357
	WKL	0.74	0.64	0.70	0.54	0.97	0.95	0.85

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

Table 14.2. Average Interrelations and Reliability between Subclaims—Mathematics

Grade	Subclaim	MC	ASC	MR	MP
3	MC	0.85	131,468	131,468	131,468
	ASC	0.77	0.61	131,468	131,468
	MR	0.73	0.65	0.62	131,468
	MP	0.72	0.64	0.66	0.67
4	MC	0.86	130,175	130,175	130,175
	ASC	0.71	0.65	130,175	130,175
	MR	0.79	0.66	0.71	130,175
	MP	0.76	0.70	0.72	0.68
5	MC	0.75	129,892	129,892	129,892
	ASC	0.75	0.65	129,892	129,892
	MR	0.80	0.70	0.59	129,892
	MP	0.74	0.66	0.71	0.68
6	MC	0.83	128,550	128,550	128,550
	ASC	0.74	0.59	128,550	128,550
	MR	0.75	0.69	0.67	128,550
	MP	0.82	0.70	0.76	0.74
7	MC	0.83	129,826	129,826	129,826
	ASC	0.73	0.65	129,826	129,826
	MR	0.79	0.67	0.73	129,826
	MP	0.80	0.74	0.77	0.76
8	MC	0.78	132,991	132,991	132,991
	ASC	0.78	0.61	132,991	132,991
	MR	0.71	0.65	0.55	132,991
	MP	0.74	0.68	0.69	0.67

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

14.2.2. Reliability

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

14.2.3. Local Item Independence

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

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

14.3. Evidence Based on Relationships to Other Variables

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

The relationship of the scores across the IAR assessments were evaluated Pearson correlations between the ELA/L and the mathematics scores and between the ELA/L subclaims, as shown in Table 14.3. Students must have a valid test score from spring 2025 for both ELA/L and mathematics at the same grade level to be included in the tables, and only correlations for pairings with total sample sizes of at least 100 are shown in the tables (blank cells indicate pairings with sample sizes less than 100). The correlation is presented in the lower triangle, and the sample size is presented in the upper triangle.

ELA/L, Reading, and Writing are low to moderately correlated with mathematics, with correlations ranging from 0.64 to 0.79. These correlations suggest that the ELA/L and mathematics tests are assessing different content. The higher intercorrelations between the ELA/L, Reading, and Writing scores suggest stronger internal relationships when compared to the correlations with mathematics.

Table 14.3. Correlations between ELA/L and Mathematics

Grade	Content Area	ELA/L	Reading	Writing	Mathematics
3	ELA/L	–	131,111	131,111	131,111
	Reading	0.95	–	131,111	131,111
	Writing	0.90	0.73	–	131,111
	Mathematics	0.76	0.75	0.65	–
4	ELA/L	–	129,877	129,877	129,877
	Reading	0.96	–	129,877	129,877
	Writing	0.90	0.74	–	129,877
	Mathematics	0.79	0.78	0.68	–
5	ELA/L	–	129,598	129,598	129,598
	Reading	0.95	–	129,598	129,598
	Writing	0.90	0.72	–	129,598
	Mathematics	0.75	0.73	0.65	–
6	ELA/L	–	128,153	128,153	128,153
	Reading	0.95	–	128,153	128,153
	Writing	0.92	0.76	–	128,153
	Mathematics	0.76	0.75	0.68	–
7	ELA/L	–	129,320	129,320	129,320
	Reading	0.94	–	129,320	129,320
	Writing	0.92	0.74	–	129,320
	Mathematics	0.76	0.76	0.64	–
8	ELA/L	–	132,387	132,387	132,387
	Reading	0.94	–	132,387	132,387
	Writing	0.92	0.74	–	132,387
	Mathematics	0.72	0.69	0.66	–

14.4. Evidence from Special Studies

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

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

14.4.1. Content Alignment Studies

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

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

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

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

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

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

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

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

14.4.2. Mode and Device Comparability Studies

A two-pronged study consisted of a mode comparability analysis and a device comparability analysis. The mode comparability analysis compared scores from the paper and online administrations, and the device comparability analysis compared the online scores from tests administered using a tablet and tests administered from any other type of electronic administration where a tablet was not present (i.e., laptops, desktops, Chromebooks). The goal of this study was threefold: (a) to investigate whether test items were of similar difficulty across the levels of conditions for each analysis (i.e., paper or online for the mode comparability analysis and tablet and non-tablet for the device comparability analysis), (b) to determine whether the psychometric properties of test scores were similar across the levels of conditions for each analysis, and (c) to determine whether overall test performance was similar across the levels of conditions for each analysis. This study examined performance on 12 assessments, split evenly between mathematics and ELA/L. Students were matched on demographic variables and on the score from the summative assessment in the same content area in the prior year, creating comparable samples that allowed for an unbiased comparison of performance across different conditions.

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

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

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

14.4.3. Alternate Blueprint Study

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

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

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

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

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

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

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

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

Assessment	Form	#Unmatched Forms	Original #Unmatched Forms	#Matched Forms	Original #Matched Forms
ELA/L 3	1	105,482	32,034	31,481	31,481
	2	105,309	31,861	31,272	31,272
ELA/L 4	1	105,826	28,153	27,695	27,695
	2	126,875	34,071	33,444	33,444
ELA/L 5	1	136,148	36,313	35,742	35,742
	2	101,869	27,272	26,721	26,721
ELA/L 6	1	119,838	31,031	30,667	30,667
	2	120,218	30,802	30,506	30,506
ELA/L 7	1	116,933	29,877	29,544	29,544
	2	117,757	29,835	29,593	29,593
ELA/L 8	1	118,198	29,638	29,312	29,312
	2	119,059	29,248	28,898	28,898
Mathematics 3	1	88,858	26,531	25,970	25,970
	2	88,919	26,595	25,987	25,987
Mathematics 4	1	87,291	25,941	25,070	25,070
	2	87,488	26,192	25,207	25,207

Assessment	Form	#Unmatched Forms	Original #Unmatched Forms	#Matched Forms	Original #Matched Forms
Mathematics 5	1	91,136	27,333	26,377	26,377
	2	91,739	27,611	26,754	26,754
Mathematics 6	1	95,174	28,514	27,677	27,677
	2	94,800	28,342	27,665	27,665
Mathematics 7	1	93,777	24,547	23,855	23,855
	2	93,265	24,141	23,485	23,485
Mathematics 8	1	83,289	15,293	14,962	14,962
	2	76,135	13,973	13,695	13,695

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

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

14.5. Evidence Based on Response Processes

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

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

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

14.6. Evidence Based on the Consequences to Testing

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

14.7. Summary

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

Psychometric analyses further provided evidence that the assessments measure what is intended. For example, the intercorrelations of the subclaims and the reliability analyses indicate that the summative assessments are both unidimensional, and the correlations between ELA/L and mathematics indicate that the two assessments are measuring different content. Several studies have also been conducted, including the content alignment studies, the benchmarking study conducted in support of the standard setting meeting, and the mode and device comparability studies. In addition to the validity information presented in this section of the technical report, other information in support of the uses and interpretations of the scores appear in the following sections:

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

Section 15: Student Growth Measures

Student growth percentiles (SGPs) are normative measures of annual progress that help answer questions like “How does my academic progress compare with the academic progress of my peers?” Unlike criterion-referenced measures that describe growth toward a specific goal, SGPs are norm-referenced—they describe a student’s growth relative to other students who started at the same place academically (Betebenner, 2009). SGPs track individual student progress by comparing a student’s scores from one year to the next, and then situate that student’s growth among their academic peers—defined as students who had similar scores on the same assessment in prior years.

SGPs indicate a student’s location within the distribution of current test scores for all students who performed similarly in the past, showing the percentage of academic peers above whom the student scored. The percentile ranges from 1 to 99, with higher numbers representing greater growth and lower numbers indicating less growth. For example, an SGP of 60 on grade 7 ELA/L means the student scored better than 60% of students who took the grade 7 ELA/L assessment in spring 2019 and who had achieved a similar score on the grade 6 ELA/L assessment in spring 2018 and the grade 5 ELA/L assessment in spring 2017. An SGP of 50 represents typical (median) student growth.

It’s important to note that SGPs require at least one prior year’s test score in the same content area to be calculated, and when available, up to two years of prior scores are used. For example, a 5th grade student’s SGP may be calculated using both their 3rd and 4th grade scores. SGPs do not measure the amount of growth, but rather how a student’s growth compares to their academically similar peers. SGPs are available for students in grades where prior year assessment data exists; for instance, 3rd graders will not have an SGP since there is no 2nd grade assessment.

SGPs can be interpreted as follows: scores above 60 indicate strong growth, scores between 40 and 60 indicate average growth, and scores below 40 indicate lower growth compared to academic peers. Schools and districts often look at the mean SGP to understand overall growth, but it’s important to examine student groups individually, as growth can vary widely.

15.1. Norm Groups

The norm groups consisted of students with the same prior scores based on grade progressions (academic peers). In the SGP analysis for the spring 2025 administration, SGPs were based on up to one year of prior test scores from the spring 2024 administration and two years of prior test scores from the spring 2023 administration, as shown in Table 15.1. Students who did not have a previous test score, which included any new students and all grade 3 students, did not receive an SGP. The sample size threshold of conducting SGP analysis was 1,000, so it was not conducted for the high school assessments due to low sample sizes.

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

Assessment	2025 Median	2025 N	2024 Median	2024 N	2023 Median	2023 N
ELA/L 4	50	124,435	—	—	—	—
ELA/L 5	50	124,405	50	118,272	—	—
ELA/L 6	50	123,245	50	117,503	50	111,841
ELA/L 7	50	124,910	50	118,959	50	113,584
ELA/L 8	50	128,047	50	122,372	50	116,976

Assessment	2025 Median	2025 N	2024 Median	2024 N	2023 Median	2023 N
Mathematics 4	50	119,968	–	–	–	–
Mathematics 5	50	120,882	50	115,887	–	–
Mathematics 6	50	120,249	50	115,659	50	110,280
Mathematics 7	50	122,461	50	117,569	48	112,406
Mathematics 8	50	125,795	50	121,020	50	115,897

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

15.2. SGP Estimation

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

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

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

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

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 an SGP of 60.

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

15.3. SGP Results

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

15.3.1. Summary for Total Group

Table 15.2 summarizes the SGP estimates for the total testing group from the spring 2025 IAR administration for ELA/L and mathematics. Median SGPs were all close to 50. If the model is a perfect fit, the median is expected to be 50 with norm-referenced data. The average standard error for the SGPs is within expectations for these models. In general, SGPs can be divided into three categories: (a) an SGP below 30, indicating that a student is not meeting a year’s worth of growth; (b) an SGP of 30–70, indicating that a student did achieve a year’s worth of growth; and (c) an SGP over 70, indicating that the student surpassed a year’s worth of growth. It is important to note that definitions such as these are not inherent to the SGP method but rather require expert judgment (Betebenner, 2009). The average standard errors, ranging from 13.00 to 15.90 across content areas and grades, support these interpretations (Betebenner et al., 2017).

Table 15.2. Summary of SGP Estimates for Total Group

Assessment	Sample Size	Average SGP	Average Standard Error	Median SGP
ELA/L 4	124,439	49.85	13.22	50
ELA/L 5	124,413	49.76	14.70	50
ELA/L 6	123,254	49.76	14.46	50
ELA/L 7	124,935	49.77	14.11	50
ELA/L 8	128,056	49.72	14.51	50
Mathematics 4	119,995	49.93	13.00	50
Mathematics 5	120,901	50.05	14.85	50
Mathematics 6	120,259	50.01	14.92	50
Mathematics 7	122,479	50.00	15.41	50
Mathematics 8	125,817	49.93	15.90	50

15.3.2. Subgroups of Interest

Appendix G presents the SGP results for subgroups of interest from the spring 2025 IAR administration. With norm-referenced data, the median of all SGPs is expected to be close to 50. Median subgroup SGPs below 50 represent growth lower than the median, and median SGPs above 50 represent growth higher than the median. As shown in the appendix, the median SGPs for subgroups of interest fell within the band of 30–70, which is considered adequate growth. Results by subgroup are as follows:

Gender:

- ELA/L: The median SGPs for females range from 50 to 54, while for males, the median SGPs range from 46 to 49. The standard errors for both groups are similar to those of the total group.
- Mathematics: The median SGPs for females generally range from 48 to 52, and for males, from 48 to 52. The standard errors for both groups are comparable to the total group.

Ethnicity:

- ELA/L: American Indian/Alaska Native students have median SGPs ranging from 40 to 48. For other ethnicity groups, median SGPs range from 49 to 52. The standard errors for all groups are similar to those of the total group.
- Mathematics: American Indian/Alaska Native students have median SGPs ranging from 46 to 52. Other ethnicity groups have median SGPs within the range of 49 to 52. The standard errors for all groups are under 20 points.

Special Instructional Needs:

- ELA/L: Students with disabilities have median SGPs ranging from 40 to 44, while students without disabilities have median SGPs ranging from 50 to 52. The standard errors for both subgroups are similar to those observed for the total group.
- Mathematics: Students with disabilities have median SGPs ranging from 40 to 46, while students without disabilities have median SGPs ranging from 51 to 52. The standard errors for both groups are similar to the total group.

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.
- Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012). *Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments*. Pearson.
- Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51.
- Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. National Center for the Improvement of Educational Assessment.
- Betebenner, D. W., Van Iwaarden, A., Domingue, B., & Shang, Y. (2017). *SGP: Student growth percentiles & percentile growth trajectories* (R package version, 1-7 [Computer software]).
- Boyd, A., Minchen, N., & McBride, M. (2018). *Alternative blueprinting options research report*. Pearson.
- Brandt, R., Bercovitz, E., McNally, S., & Zimmerman, L. (2015a). *Drawing response interaction usability study for PARCC*. Partnership for Assessment of Readiness for College and Careers. <https://files.eric.ed.gov/fulltext/ED599260.pdf>
- Brandt, R., Bercovitz, E., & Zimmerman, L. (2015b). *Drawing response interaction usability study for PARCC*. Pearson. <https://eric.ed.gov/?id=ED599261>
- Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0). (CASMA Research Report No. 9). Center for Advanced Studies in Measurement, University of Iowa.
- Camara, W. J., Allen, J. M., & Moore, J. L. (2017). Empirically-based college and career readiness cut scores and performance standards. In K. L. McClarty, K. D. Mattern & M. N. Gaertner (Eds.), *Preparing students for college and careers: Theory, measurement, and educational practice*. Routledge.
- Center for Assessment. (2018). *PARCC comparability review guidelines*.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

- Davis, L. L., & Moyer, E. L. (2015). *PARCC performance level setting technical report*. Partnership for Assessment of Readiness for College and Careers (PARCC).
- Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute.
- Dorans, N. J. (2013). *ETS contributions to the quantitative assessment of item, test and score fairness* (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04). Educational Testing Service.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. RR-91-47). Educational Testing Service.
- Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.
- Haertel, E. H., Beimers, J. N., & Miles, J. A. (2012). The briefing book method. In Cizek G. J. (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 283–299). Routledge.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum.
- Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1–24. <https://doi.org/10.1093/pan/mpr013>
- Koenker, R. (2005). *Quantile regression*. Cambridge University Press.
- Kolen, M. J. (2004). *POLYCEM windows console version* [Computer software]. The Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33(2), 129–140.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Livingston, S. A., & Lewis, C. (1993). *Estimating the consistency and accuracy of classifications based on test scores* (ETS Research Report No. RR-93-48). Educational Testing Service.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179–197.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8(4), 453–461.

- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303), 690–700.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.
- McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson.
- McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: establishing a validity framework for cut scores. *Educational Researcher*, 42(2), 78–88.
- Minchen, N., Boyd, A., & McBride, M. (2018a). *Alternative blueprinting options 2018 research report*. Pearson.
- Minchen, N., LaSalle, A., & Boyd, A. (2018b). *Operational study 4: Accessibility of new items/functionality component 4 report*. Pearson.
- New Meridian. (2023). *Technical report 2021–2022, alternate blueprint*. <https://www.isbe.net/Documents/PARCC-Spring-2023-Tech-Manual.pdf>
- Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Biometrika*, 47, 337–347.
- Pike, C. K., & Hudson, W. W. (1998). Reliability and measurement error in the presence of homogeneity. *Journal of Social Service Research*, 24(1–2), 149–163.
- Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting multiple performance standards using the Yes/No method: An alternative item mapping method*. Meeting of the NCME, Montreal, Canada.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353.
- Schultz, S. R., Norman Dvorak, R., & Chen, J. (2017). *Evaluating the quality and alignment of PARCC ELA/literacy and mathematics assessments: Grades 3, 4, 6, and 7*. (HumRRO Report 2017 No. 040). Human Resources Research Organization.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237–247.
- Steedle, J., & LaSalle, A. (2016). *Operational study 4: Accessibility of new items/functionality component 3 report*. Pearson.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach’s alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes.

- Wainer, H., & Thissen, D. (2001). *Test scoring*. Lawrence Erlbaum.
- Wei, Y., & He, X. (2006). Conditional growth charts. *Annals of Statistics*, 34(5), 2069–2097.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, 31(1), 2–13.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S.C. (2003). *Effects of local dependence on the validity of IRT item test, and ability statistics* (Technical Report). American College Admissions Test.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Lawrence Erlbaum
- Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (ETS Research Report RR-97-05). Educational Testing Service.

Appendix A: Scale Score Cumulative Frequencies

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

Score Band	N	%	Cumulative N	Cumulative %
650–654	3,881	2.94	3,881	2.94
655–659	1,934	1.46	5,815	4.40
660–664	1,914	1.45	7,729	5.85
665–669	2,400	1.82	10,129	7.67
670–674	2,541	1.92	12,670	9.59
675–679	5,055	3.83	17,725	13.42
680–684	2,682	2.03	20,407	15.45
685–689	2,610	1.98	23,017	17.43
690–694	4,749	3.60	27,766	21.02
695–699	4,511	3.42	32,277	24.44
700–704	6,176	4.68	38,453	29.12
705–709	5,785	4.38	44,238	33.50
710–714	3,949	2.99	48,187	36.49
715–719	3,954	2.99	52,141	39.48
720–724	7,702	5.83	59,843	45.31
725–729	3,744	2.84	63,587	48.15
730–734	7,572	5.73	71,159	53.88
735–739	7,658	5.80	78,817	59.68
740–744	3,784	2.87	82,601	62.55
745–749	7,321	5.54	89,922	68.09
750–754	5,224	3.96	95,146	72.05
755–759	5,034	3.81	100,180	75.86
760–764	4,772	3.61	104,952	79.47
765–769	5,858	4.44	110,810	83.91
770–774	2,558	1.94	113,368	85.84
775–779	3,369	2.55	116,737	88.39
780–784	3,148	2.38	119,885	90.78
785–789	2,473	1.87	122,358	92.65
790–794	1,613	1.22	123,971	93.87
795–799	1,827	1.38	125,798	95.26
800–804	1,173	0.89	126,971	96.14
805–809	1,533	1.16	128,504	97.31
810–814	776	0.59	129,280	97.89
815–819	679	0.51	129,959	98.41
820–824	688	0.52	130,647	98.93
825–829	396	0.30	131,043	99.23
830–834	199	0.15	131,242	99.38
835–839	206	0.16	131,448	99.53
840–844	182	0.14	131,630	99.67
845–849	149	0.11	131,779	99.78
850	284	0.22	132,063	100.00

Table A.2. Scale Score Cumulative Frequencies—ELA/L Grade 3 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	6,385	4.83	6,385	4.83
15-19	4,930	3.73	11,315	8.57
20-24	8,977	6.8	20,292	15.37
25-29	12,793	9.69	33,085	25.05
30-34	12,216	9.25	45,301	34.3
35-39	17,873	13.53	63,174	47.84
40-44	14,604	11.06	77,778	58.89
45-49	14,139	10.71	91,917	69.6
50-54	10,732	8.13	102,649	77.73
55-59	12,377	9.37	115,026	87.1
60-64	6,498	4.92	121,524	92.02
65-69	4,919	3.72	126,443	95.74
70-74	2,333	1.77	128,776	97.51
75-79	1,684	1.28	130,460	98.79
80-84	1,011	0.77	131,471	99.55
85-89	1	0	131,472	99.55
90	591	0.45	132,063	100

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

Score Band	N	%	Cumulative N	Cumulative %
10-14	39,639	30.02	39,639	30.02
15-19	-	-	-	-
20-24	7,957	6.03	47,596	36.04
25-29	3,369	2.55	50,965	38.59
30-34	32,063	24.28	83,028	62.87
35-39	25,456	19.28	108,484	82.15
40-44	11,970	9.06	120,454	91.21
45-49	7,968	6.03	128,422	97.24
50-54	2,748	2.08	131,170	99.32
55-59	685	0.52	131,855	99.84
60	208	0.16	132,063	100

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

Score Band	N	%	Cumulative N	Cumulative %
650–654	796	0.61	796	0.61
655–659	505	0.39	1,301	1.00
660–664	877	0.67	2,178	1.67
665–669	36	0.03	2,214	1.69
670–674	2,595	1.98	4,809	3.68
675–679	3,821	2.92	8,630	6.60
680–684	74	0.06	8,704	6.66
685–689	4,458	3.41	13,162	10.07
690–694	4,841	3.70	18,003	13.77
695–699	6,734	5.15	24,737	18.92

Score Band	N	%	Cumulative N	Cumulative %
700–704	4,089	3.13	28,826	22.05
705–709	7,158	5.47	35,984	27.52
710–714	3,152	2.41	39,136	29.93
715–719	5,902	4.51	45,038	34.45
720–724	5,560	4.25	50,598	38.70
725–729	5,376	4.11	55,974	42.81
730–734	6,842	5.23	62,816	48.05
735–739	6,592	5.04	69,408	53.09
740–744	5,406	4.13	74,814	57.22
745–749	6,758	5.17	81,572	62.39
750–754	6,696	5.12	88,268	67.51
755–759	5,161	3.95	93,429	71.46
760–764	7,432	5.68	100,861	77.15
765–769	4,740	3.63	105,601	80.77
770–774	4,403	3.37	110,004	84.14
775–779	3,854	2.95	113,858	87.09
780–784	3,381	2.59	117,239	89.67
785–789	3,651	2.79	120,890	92.46
790–794	2,463	1.88	123,353	94.35
795–799	1,526	1.17	124,879	95.52
800–804	1,662	1.27	126,541	96.79
805–809	1,301	1.00	127,842	97.78
810–814	1,007	0.77	128,849	98.55
815–819	597	0.46	129,446	99.01
820–824	415	0.32	129,861	99.33
825–829	355	0.27	130,216	99.60
830–834	158	0.12	130,374	99.72
835–839	111	0.08	130,485	99.80
840–844	85	0.07	130,570	99.87
845–849	87	0.07	130,657	99.93
850	85	0.07	130,742	100.00

Table A.5. Scale Score Cumulative Frequencies—ELA/L Grade 4 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	1,367	1.05	1,367	1.05
15-19	3,737	2.86	5,104	3.9
20-24	4,250	3.25	9,354	7.15
25-29	13,094	10.02	22,448	17.17
30-34	12,765	9.76	35,213	26.93
35-39	14,141	10.82	49,354	37.75
40-44	14,963	11.44	64,317	49.19
45-49	15,304	11.71	79,621	60.9
50-54	16,098	12.31	95,719	73.21
55-59	11,992	9.17	107,711	82.38
60-64	9,342	7.15	117,053	89.53
65-69	7,457	5.7	124,510	95.23

Score Band	N	%	Cumulative N	Cumulative %
70-74	2,750	2.1	127,260	97.34
75-79	2,225	1.7	129,485	99.04
80-84	729	0.56	130,214	99.6
85-89	371	0.28	130,585	99.88
90	157	0.12	130,742	100

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

Score Band	N	%	Cumulative N	Cumulative %
10-14	45,968	35.16	45,968	35.16
15-19	-	-	-	-
20-24	-	-	-	-
25-29	7,410	5.67	53,378	40.83
30-34	27,708	21.19	81,086	62.02
35-39	23,800	18.2	104,886	80.22
40-44	19,280	14.75	124,166	94.97
45-49	4,590	3.51	128,756	98.48
50-54	1,642	1.26	130,398	99.74
55-59	220	0.17	130,618	99.91
60	124	0.09	130,742	100

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

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,541	1.18	1,541	1.18
655–659	33	0.03	1,574	1.21
660–664	977	0.75	2,551	1.96
665–669	2,285	1.75	4,836	3.71
670–674	–	–	–	–
675–679	3,040	2.33	7,876	6.04
680–684	1,762	1.35	9,638	7.39
685–689	3,398	2.61	13,036	10.00
690–694	3,433	2.63	16,469	12.63
695–699	3,308	2.54	19,777	15.16
700–704	4,481	3.44	24,258	18.60
705–709	3,043	2.33	27,301	20.93
710–714	5,834	4.47	33,135	25.41
715–719	4,159	3.19	37,294	28.59
720–724	5,716	4.38	43,010	32.98
725–729	5,900	4.52	48,910	37.50
730–734	5,876	4.51	54,786	42.01
735–739	6,060	4.65	60,846	46.65
740–744	9,297	7.13	70,143	53.78
745–749	6,318	4.84	76,461	58.63
750–754	6,244	4.79	82,705	63.41
755–759	7,552	5.79	90,257	69.20
760–764	5,810	4.45	96,067	73.66

Score Band	N	%	Cumulative N	Cumulative %
765–769	6,694	5.13	102,761	78.79
770–774	6,005	4.60	108,766	83.40
775–779	5,277	4.05	114,043	87.44
780–784	3,523	2.70	117,566	90.14
785–789	2,957	2.27	120,523	92.41
790–794	3,096	2.37	123,619	94.78
795–799	1,917	1.47	125,536	96.25
800–804	1,878	1.44	127,414	97.69
805–809	1,023	0.78	128,437	98.48
810–814	822	0.63	129,259	99.11
815–819	542	0.42	129,801	99.52
820–824	261	0.20	130,062	99.72
825–829	127	0.10	130,189	99.82
830–834	77	0.06	130,266	99.88
835–839	65	0.05	130,331	99.93
840–844	45	0.03	130,376	99.96
845–849	17	0.01	130,393	99.98
850	29	0.02	130,422	100.00

Table A.8. Scale Score Cumulative Frequencies—ELA/L Grade 5 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	1,696	1.3	1,696	1.3
15-19	3,614	2.77	5,310	4.07
20-24	5,544	4.25	10,854	8.32
25-29	8,863	6.8	19,717	15.12
30-34	10,880	8.34	30,597	23.46
35-39	13,663	10.48	44,260	33.94
40-44	16,003	12.27	60,263	46.21
45-49	19,871	15.24	80,134	61.44
50-54	14,654	11.24	94,788	72.68
55-59	15,123	11.6	109,911	84.27
60-64	9,939	7.62	119,850	91.89
65-69	4,715	3.62	124,565	95.51
70-74	3,775	2.89	128,340	98.4
75-79	1,208	0.93	129,548	99.33
80-84	707	0.54	130,255	99.87
85-89	119	0.09	130,374	99.96
90	48	0.04	130,422	100

Table A.9. Scale Score Cumulative Frequencies—ELA/L Grade 5 Writing

Score Band	N	%	Cumulative N	Cumulative %
10-14	33,112	25.39	33,112	25.39
15-19	-	-	-	-
20-24	850	0.65	33,962	26.04
25-29	7,327	5.62	41,289	31.66

Score Band	N	%	Cumulative N	Cumulative %
30-34	23,254	17.83	64,543	49.49
35-39	30,340	23.26	94,883	72.75
40-44	27,563	21.13	122,446	93.88
45-49	6,321	4.85	128,767	98.73
50-54	1,451	1.11	130,218	99.84
55-59	88	0.07	130,306	99.91
60	116	0.09	130,422	100

Table A.10. Scale Score Cumulative Frequencies—ELA/L Grade 6

Score Band	N	%	Cumulative N	Cumulative %
650–654	775	0.60	775	0.60
655–659	9	0.01	784	0.61
660–664	539	0.42	1,323	1.02
665–669	586	0.45	1,909	1.47
670–674	944	0.73	2,853	2.20
675–679	2,091	1.62	4,944	3.82
680–684	2,609	2.02	7,553	5.84
685–689	2,816	2.18	10,369	8.01
690–694	2,985	2.31	13,354	10.32
695–699	2,810	2.17	16,164	12.49
700–704	3,871	2.99	20,035	15.48
705–709	3,830	2.96	23,865	18.44
710–714	4,858	3.75	28,723	22.19
715–719	4,967	3.84	33,690	26.03
720–724	5,221	4.03	38,911	30.06
725–729	6,656	5.14	45,567	35.21
730–734	6,866	5.30	52,433	40.51
735–739	7,039	5.44	59,472	45.95
740–744	7,084	5.47	66,556	51.42
745–749	8,380	6.47	74,936	57.90
750–754	6,785	5.24	81,721	63.14
755–759	8,085	6.25	89,806	69.39
760–764	6,182	4.78	95,988	74.16
765–769	6,161	4.76	102,149	78.92
770–774	5,251	4.06	107,400	82.98
775–779	5,874	4.54	113,274	87.52
780–784	4,088	3.16	117,362	90.68
785–789	2,802	2.16	120,164	92.84
790–794	2,808	2.17	122,972	95.01
795–799	2,251	1.74	125,223	96.75
800–804	1,379	1.07	126,602	97.82
805–809	999	0.77	127,601	98.59
810–814	590	0.46	128,191	99.05
815–819	443	0.34	128,634	99.39
820–824	330	0.25	128,964	99.64
825–829	196	0.15	129,160	99.79

Score Band	N	%	Cumulative N	Cumulative %
830–834	112	0.09	129,272	99.88
835–839	68	0.05	129,340	99.93
840–844	–	–	–	–
845–849	33	0.03	129,373	99.96
850	53	0.04	129,426	100.00

Table A.11. Scale Score Cumulative Frequencies—ELA/L Grade 6 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	810	0.63	810	0.63
15-19	2,157	1.67	2,967	2.29
20-24	3,404	2.63	6,371	4.92
25-29	8,298	6.41	14,669	11.33
30-34	10,120	7.82	24,789	19.15
35-39	14,305	11.05	39,094	30.21
40-44	16,844	13.01	55,938	43.22
45-49	18,783	14.51	74,721	57.73
50-54	19,161	14.8	93,882	72.54
55-59	14,891	11.51	108,773	84.04
60-64	10,548	8.15	119,321	92.19
65-69	5,797	4.48	125,118	96.67
70-74	2,870	2.22	127,988	98.89
75-79	724	0.56	128,712	99.45
80-84	447	0.35	129,159	99.79
85-89	76	0.06	129,235	99.85
90	191	0.15	129,426	100

Table A.12. Scale Score Cumulative Frequencies—ELA/L Grade 6 Writing

Score Band	N	%	Cumulative N	Cumulative %
10-14	35,494	27.42	35,494	27.42
15-19	-	-	-	-
20-24	-	-	-	-
25-29	1,879	1.45	37,373	28.88
30-34	28,247	21.82	65,620	50.7
35-39	37,097	28.66	102,717	79.36
40-44	17,593	13.59	120,310	92.96
45-49	7,575	5.85	127,885	98.81
50-54	1,272	0.98	129,157	99.79
55-59	18	0.01	129,175	99.81
60	251	0.19	129,426	100

Table A.13. Scale Score Cumulative Frequencies—ELA/L Grade 7

Score Band	N	%	Cumulative N	Cumulative %
650–654	663	0.51	663	0.51

Score Band	N	%	Cumulative N	Cumulative %
655–659	–	–	–	–
660–664	1,046	0.80	1,709	1.31
665–669	–	–	–	–
670–674	1,573	1.20	3,282	2.51
675–679	43	0.03	3,325	2.54
680–684	2,087	1.59	5,412	4.13
685–689	2,500	1.91	7,912	6.04
690–694	2,561	1.96	10,473	8.00
695–699	3,611	2.76	14,084	10.76
700–704	2,572	1.96	16,656	12.72
705–709	5,065	3.87	21,721	16.59
710–714	5,097	3.89	26,818	20.48
715–719	5,072	3.87	31,890	24.36
720–724	5,081	3.88	36,971	28.24
725–729	4,956	3.79	41,927	32.02
730–734	6,169	4.71	48,096	36.74
735–739	7,671	5.86	55,767	42.59
740–744	7,993	6.11	63,760	48.70
745–749	6,328	4.83	70,088	53.53
750–754	7,976	6.09	78,064	59.63
755–759	7,774	5.94	85,838	65.56
760–764	7,564	5.78	93,402	71.34
765–769	6,813	5.20	100,215	76.54
770–774	5,276	4.03	105,491	80.57
775–779	5,856	4.47	111,347	85.05
780–784	4,415	3.37	115,762	88.42
785–789	4,504	3.44	120,266	91.86
790–794	2,404	1.84	122,670	93.70
795–799	2,143	1.64	124,813	95.33
800–804	1,712	1.31	126,525	96.64
805–809	1,395	1.07	127,920	97.71
810–814	850	0.65	128,770	98.35
815–819	500	0.38	129,270	98.74
820–824	784	0.60	130,054	99.34
825–829	140	0.11	130,194	99.44
830–834	256	0.20	130,450	99.64
835–839	129	0.10	130,579	99.74
840–844	143	0.11	130,722	99.85
845–849	–	–	–	–
850	202	0.15	130,924	100.00

Table A.14. Scale Score Cumulative Frequencies—ELA/L Grade 7 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	680	0.52	680	0.52
15-19	1,094	0.84	1,774	1.35
20-24	4,010	3.06	5,784	4.42

Score Band	N	%	Cumulative N	Cumulative %
25-29	5,870	4.48	11,654	8.9
30-34	9,643	7.37	21,297	16.27
35-39	14,964	11.43	36,261	27.7
40-44	16,683	12.74	52,944	40.44
45-49	20,025	15.3	72,969	55.73
50-54	16,347	12.49	89,316	68.22
55-59	18,114	13.84	107,430	82.06
60-64	11,588	8.85	119,018	90.91
65-69	6,943	5.3	125,961	96.21
70-74	2,538	1.94	128,499	98.15
75-79	1,584	1.21	130,083	99.36
80-84	392	0.3	130,475	99.66
85-89	258	0.2	130,733	99.85
90	191	0.15	130,924	100

Table A.15. Scale Score Cumulative Frequencies—ELA/L Grade 7 Writing

Score Band	N	%	Cumulative N	Cumulative %
10-14	33,522	25.6	33,522	25.6
15-19	-	-	-	-
20-24	847	0.65	34,369	26.25
25-29	3,800	2.9	38,169	29.15
30-34	29,870	22.81	68,039	51.97
35-39	27,678	21.14	95,717	73.11
40-44	24,058	18.38	119,775	91.48
45-49	6,757	5.16	126,532	96.65
50-54	2,923	2.23	129,455	98.88
55-59	9	0.01	129,464	98.88
60	1,460	1.12	130,924	100

Table A.16. Scale Score Cumulative Frequencies—ELA/L Grade 8

Score Band	N	%	Cumulative N	Cumulative %
650-654	2,050	1.53	2,050	1.53
655-659	705	0.53	2,755	2.06
660-664	791	0.59	3,546	2.65
665-669	962	0.72	4,508	3.36
670-674	2,057	1.53	6,565	4.90
675-679	2,135	1.59	8,700	6.49
680-684	2,140	1.60	10,840	8.09
685-689	2,062	1.54	12,902	9.63
690-694	1,959	1.46	14,861	11.09
695-699	1,992	1.49	16,853	12.57
700-704	3,697	2.76	20,550	15.33
705-709	1,907	1.42	22,457	16.76
710-714	3,905	2.91	26,362	19.67
715-719	4,335	3.23	30,697	22.90

Score Band	N	%	Cumulative N	Cumulative %
720–724	4,650	3.47	35,347	26.37
725–729	4,872	3.64	40,219	30.01
730–734	5,171	3.86	45,390	33.87
735–739	6,880	5.13	52,270	39.00
740–744	7,229	5.39	59,499	44.39
745–749	6,227	4.65	65,726	49.04
750–754	6,369	4.75	72,095	53.79
755–759	7,979	5.95	80,074	59.75
760–764	7,613	5.68	87,687	65.43
765–769	6,185	4.61	93,872	70.04
770–774	5,914	4.41	99,786	74.45
775–779	5,544	4.14	105,330	78.59
780–784	6,340	4.73	111,670	83.32
785–789	4,584	3.42	116,254	86.74
790–794	3,923	2.93	120,177	89.67
795–799	3,295	2.46	123,472	92.13
800–804	2,039	1.52	125,511	93.65
805–809	2,514	1.88	128,025	95.52
810–814	2,016	1.50	130,041	97.03
815–819	1,164	0.87	131,205	97.90
820–824	928	0.69	132,133	98.59
825–829	697	0.52	132,830	99.11
830–834	382	0.29	133,212	99.39
835–839	251	0.19	133,463	99.58
840–844	199	0.15	133,662	99.73
845–849	134	0.10	133,796	99.83
850	227	0.17	134,023	100.00

Table A.17. Scale Score Cumulative Frequencies—ELA/L Grade 8 Reading

Score Band	N	%	Cumulative N	Cumulative %
10-14	2,960	2.21	2,960	2.21
15-19	1,919	1.43	4,879	3.64
20-24	4,881	3.64	9,760	7.28
25-29	7,642	5.7	17,402	12.98
30-34	8,107	6.05	25,509	19.03
35-39	11,447	8.54	36,956	27.57
40-44	13,185	9.84	50,141	37.41
45-49	15,816	11.8	65,957	49.21
50-54	18,256	13.62	84,213	62.83
55-59	18,326	13.67	102,539	76.51
60-64	13,167	9.82	115,706	86.33
65-69	8,332	6.22	124,038	92.55
70-74	4,240	3.16	128,278	95.71
75-79	2,873	2.14	131,151	97.86
80-84	1,629	1.22	132,780	99.07
85-89	684	0.51	133,464	99.58

Score Band	N	%	Cumulative N	Cumulative %
90	559	0.42	134,023	100

Table A.18. Scale Score Cumulative Frequencies—ELA/L Grade 8 Writing

Score Band	N	%	Cumulative N	Cumulative %
10-14	28,555	21.31	28,555	21.31
15-19	-	-	-	-
20-24	693	0.52	29,248	21.82
25-29	3,262	2.43	32,510	24.26
30-34	22,865	17.06	55,375	41.32
35-39	38,399	28.65	93,774	69.97
40-44	22,116	16.5	115,890	86.47
45-49	13,687	10.21	129,577	96.68
50-54	3,984	2.97	133,561	99.66
55-59	20	0.01	133,581	99.67
60	442	0.33	134,023	100

Table A.19. Scale Score Cumulative Frequencies—Mathematics Grade 3

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,664	1.26	1,664	1.26
655–659	1,086	0.82	2,750	2.08
660–664	293	0.22	3,043	2.31
665–669	1,764	1.34	4,807	3.64
670–674	1,884	1.43	6,691	5.07
675–679	2,630	1.99	9,321	7.07
680–684	3,128	2.37	12,449	9.44
685–689	3,541	2.68	15,990	12.12
690–694	4,009	3.04	19,999	15.16
695–699	4,350	3.30	24,349	18.46
700–704	4,534	3.44	28,883	21.90
705–709	6,084	4.61	34,967	26.51
710–714	8,351	6.33	43,318	32.84
715–719	6,213	4.71	49,531	37.55
720–724	7,876	5.97	57,407	43.52
725–729	8,768	6.65	66,175	50.16
730–734	5,514	4.18	71,689	54.34
735–739	7,940	6.02	79,629	60.36
740–744	5,975	4.53	85,604	64.89
745–749	7,007	5.31	92,611	70.21
750–754	6,286	4.77	98,897	74.97
755–759	6,015	4.56	104,912	79.53
760–764	3,760	2.85	108,672	82.38
765–769	3,936	2.98	112,608	85.36
770–774	4,396	3.33	117,004	88.70
775–779	1,976	1.50	118,980	90.19
780–784	1,833	1.39	120,813	91.58

Score Band	N	%	Cumulative N	Cumulative %
785–789	3,280	2.49	124,093	94.07
790–794	1,401	1.06	125,494	95.13
795–799	1,324	1.00	126,818	96.14
800–804	703	0.53	127,521	96.67
805–809	1,121	0.85	128,642	97.52
810–814	915	0.69	129,557	98.21
815–819	338	0.26	129,895	98.47
820–824	418	0.32	130,313	98.79
825–829	594	0.45	130,907	99.24
830–834	–	–	–	–
835–839	444	0.34	131,351	99.57
840–844	–	–	–	–
845–849	150	0.11	131,501	99.69
850	414	0.31	131,915	100.00

Table A.20. Scale Score Cumulative Frequencies—Mathematics Grade 4

Score Band	N	%	Cumulative N	Cumulative %
650–654	767	0.59	767	0.59
655–659	1,160	0.89	1,927	1.48
660–664	307	0.24	2,234	1.71
665–669	2,015	1.54	4,249	3.25
670–674	447	0.34	4,696	3.60
675–679	2,717	2.08	7,413	5.68
680–684	3,852	2.95	11,265	8.63
685–689	3,943	3.02	15,208	11.64
690–694	4,082	3.13	19,290	14.77
695–699	4,048	3.10	23,338	17.87
700–704	5,101	3.91	28,439	21.78
705–709	6,982	5.35	35,421	27.12
710–714	5,032	3.85	40,453	30.97
715–719	6,509	4.98	46,962	35.96
720–724	7,357	5.63	54,319	41.59
725–729	7,075	5.42	61,394	47.01
730–734	8,100	6.20	69,494	53.21
735–739	6,674	5.11	76,168	58.32
740–744	6,578	5.04	82,746	63.36
745–749	9,166	7.02	91,912	70.38
750–754	5,680	4.35	97,592	74.73
755–759	5,334	4.08	102,926	78.81
760–764	5,165	3.95	108,091	82.76
765–769	4,755	3.64	112,846	86.41
770–774	4,302	3.29	117,148	89.70
775–779	3,900	2.99	121,048	92.69
780–784	2,471	1.89	123,519	94.58
785–789	2,285	1.75	125,804	96.33
790–794	1,222	0.94	127,026	97.26

Score Band	N	%	Cumulative N	Cumulative %
795–799	1,004	0.77	128,030	98.03
800–804	831	0.64	128,861	98.67
805–809	627	0.48	129,488	99.15
810–814	503	0.39	129,991	99.53
815–819	–	–	–	–
820–824	291	0.22	130,282	99.76
825–829	–	–	–	–
830–834	80	0.06	130,362	99.82
835–839	112	0.09	130,474	99.90
840–844	–	–	–	–
845–849	–	–	–	–
850	126	0.10	130,600	100.00

Table A.21. Scale Score Cumulative Frequencies—Mathematics Grade 5

Score Band	N	%	Cumulative N	Cumulative %
650–654	1,657	1.27	1,657	1.27
655–659	1,773	1.36	3,430	2.63
660–664	–	–	–	–
665–669	545	0.42	3,975	3.05
670–674	2,826	2.17	6,801	5.22
675–679	843	0.65	7,644	5.87
680–684	3,660	2.81	11,304	8.68
685–689	5,335	4.09	16,639	12.77
690–694	5,766	4.43	22,405	17.20
695–699	6,077	4.66	28,482	21.86
700–704	5,951	4.57	34,433	26.43
705–709	5,646	4.33	40,079	30.76
710–714	5,485	4.21	45,564	34.97
715–719	10,096	7.75	55,660	42.72
720–724	6,353	4.88	62,013	47.60
725–729	7,002	5.37	69,015	52.97
730–734	7,696	5.91	76,711	58.88
735–739	6,932	5.32	83,643	64.20
740–744	7,400	5.68	91,043	69.88
745–749	5,695	4.37	96,738	74.25
750–754	5,130	3.94	101,868	78.19
755–759	4,681	3.59	106,549	81.78
760–764	4,300	3.30	110,849	85.08
765–769	3,877	2.98	114,726	88.06
770–774	3,485	2.67	118,211	90.73
775–779	2,989	2.29	121,200	93.03
780–784	2,615	2.01	123,815	95.04
785–789	1,787	1.37	125,602	96.41
790–794	1,278	0.98	126,880	97.39
795–799	788	0.60	127,668	97.99
800–804	1,040	0.80	128,708	98.79

Score Band	N	%	Cumulative N	Cumulative %
805–809	664	0.51	129,372	99.30
810–814	316	0.24	129,688	99.54
815–819	172	0.13	129,860	99.68
820–824	97	0.07	129,957	99.75
825–829	150	0.12	130,107	99.86
830–834	–	–	–	–
835–839	107	0.08	130,214	99.95
840–844	–	–	–	–
845–849	41	0.03	130,255	99.98
850	28	0.02	130,283	100.00

Table A.22. Scale Score Cumulative Frequencies—Mathematics Grade 6

Score Band	N	%	Cumulative N	Cumulative %
650–654	910	0.70	910	0.70
655–659	1,810	1.40	2,720	2.10
660–664	383	0.30	3,103	2.40
665–669	–	–	–	–
670–674	4,060	3.14	7,163	5.54
675–679	–	–	–	–
680–684	5,866	4.54	13,029	10.08
685–689	1,215	0.94	14,244	11.02
690–694	6,888	5.33	21,132	16.35
695–699	7,025	5.44	28,157	21.79
700–704	6,589	5.10	34,746	26.89
705–709	6,279	4.86	41,025	31.75
710–714	7,081	5.48	48,106	37.22
715–719	8,176	6.33	56,282	43.55
720–724	5,736	4.44	62,018	47.99
725–729	7,836	6.06	69,854	54.05
730–734	6,697	5.18	76,551	59.24
735–739	8,701	6.73	85,252	65.97
740–744	5,089	3.94	90,341	69.91
745–749	6,917	5.35	97,258	75.26
750–754	5,280	4.09	102,538	79.34
755–759	5,607	4.34	108,145	83.68
760–764	4,345	3.36	112,490	87.05
765–769	3,254	2.52	115,744	89.56
770–774	3,747	2.90	119,491	92.46
775–779	2,793	2.16	122,284	94.62
780–784	1,805	1.40	124,089	96.02
785–789	1,577	1.22	125,666	97.24
790–794	882	0.68	126,548	97.92
795–799	789	0.61	127,337	98.53
800–804	614	0.48	127,951	99.01
805–809	471	0.36	128,422	99.37
810–814	159	0.12	128,581	99.50

Score Band	N	%	Cumulative N	Cumulative %
815–819	190	0.15	128,771	99.64
820–824	90	0.07	128,861	99.71
825–829	147	0.11	129,008	99.83
830–834	53	0.04	129,061	99.87
835–839	–	–	–	–
840–844	–	–	–	–
845–849	82	0.06	129,143	99.93
850	89	0.07	129,232	100.00

Table A.23. Scale Score Cumulative Frequencies—Mathematics Grade 7

Score Band	N	%	Cumulative N	Cumulative %
650–654	716	0.55	716	0.55
655–659	–	–	–	–
660–664	–	–	–	–
665–669	–	–	–	–
670–674	246	0.19	962	0.74
675–679	1,583	1.21	2,545	1.95
680–684	492	0.38	3,037	2.32
685–689	2,991	2.29	6,028	4.61
690–694	952	0.73	6,980	5.34
695–699	5,297	4.05	12,277	9.39
700–704	6,087	4.66	18,364	14.05
705–709	6,342	4.85	24,706	18.91
710–714	9,614	7.36	34,320	26.26
715–719	5,010	3.83	39,330	30.10
720–724	9,050	6.93	48,380	37.02
725–729	10,078	7.71	58,458	44.73
730–734	6,902	5.28	65,360	50.01
735–739	9,223	7.06	74,583	57.07
740–744	8,634	6.61	83,217	63.68
745–749	7,777	5.95	90,994	69.63
750–754	7,092	5.43	98,086	75.06
755–759	6,381	4.88	104,467	79.94
760–764	5,747	4.40	110,214	84.34
765–769	4,332	3.31	114,546	87.65
770–774	4,183	3.20	118,729	90.85
775–779	3,025	2.31	121,754	93.17
780–784	2,733	2.09	124,487	95.26
785–789	1,751	1.34	126,238	96.60
790–794	1,030	0.79	127,268	97.39
795–799	929	0.71	128,197	98.10
800–804	790	0.60	128,987	98.70
805–809	660	0.51	129,647	99.21
810–814	306	0.23	129,953	99.44
815–819	201	0.15	130,154	99.60
820–824	–	–	–	–

Score Band	N	%	Cumulative N	Cumulative %
825–829	158	0.12	130,312	99.72
830–834	135	0.10	130,447	99.82
835–839	–	–	–	–
840–844	–	–	–	–
845–849	–	–	–	–
850	236	0.18	130,683	100.00

Table A.24. Scale Score Cumulative Frequencies—Mathematics Grade 8

Score Band	N	%	Cumulative N	Cumulative %
650–654	3,569	2.67	3,569	2.67
655–659	31	0.02	3,600	2.69
660–664	3,903	2.92	7,503	5.61
665–669	44	0.03	7,547	5.64
670–674	5,205	3.89	12,752	9.53
675–679	71	0.05	12,823	9.59
680–684	6,288	4.70	19,111	14.29
685–689	6,807	5.09	25,918	19.38
690–694	5,163	3.86	31,081	23.24
695–699	1,705	1.27	32,786	24.51
700–704	6,744	5.04	39,530	29.55
705–709	6,505	4.86	46,035	34.42
710–714	6,210	4.64	52,245	39.06
715–719	5,902	4.41	58,147	43.47
720–724	6,984	5.22	65,131	48.69
725–729	8,628	6.45	73,759	55.14
730–734	4,700	3.51	78,459	58.66
735–739	5,503	4.11	83,962	62.77
740–744	6,529	4.88	90,491	67.65
745–749	4,473	3.34	94,964	71.00
750–754	5,188	3.88	100,152	74.88
755–759	3,653	2.73	103,805	77.61
760–764	4,932	3.69	108,737	81.30
765–769	3,506	2.62	112,243	83.92
770–774	2,541	1.90	114,784	85.82
775–779	3,357	2.51	118,141	88.33
780–784	2,513	1.88	120,654	90.21
785–789	2,677	2.00	123,331	92.21
790–794	1,173	0.88	124,504	93.08
795–799	2,145	1.60	126,649	94.69
800–804	1,009	0.75	127,658	95.44
805–809	1,750	1.31	129,408	96.75
810–814	796	0.60	130,204	97.35
815–819	697	0.52	130,901	97.87
820–824	606	0.45	131,507	98.32
825–829	547	0.41	132,054	98.73
830–834	443	0.33	132,497	99.06

Appendix A: Scale Score Cumulative Frequencies

Score Band	N	%	Cumulative N	Cumulative %
835–839	373	0.28	132,870	99.34
840–844	–	–	–	–
845–849	305	0.23	133,175	99.57
850	580	0.43	133,755	100.00

Appendix B: Scale Score Performance by Demographic Subgroup

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	132,063	100.0%	729.69	41.23	650	850
Female	65,089	49.3%	733.47	41.87	650	850
Male	66,959	50.7%	726.01	40.26	650	850
American Indian/Alaska Native	291	0.2%	723.73	41.70	650	849
Asian	7,635	5.8%	750.02	42.76	650	850
Black or African American	21,294	16.1%	712.48	37.07	650	850
Hispanic/Latino	36,829	27.9%	715.98	39.20	650	850
Middle Eastern or North African	317	0.2%	729.66	46.24	650	850
Native Hawaiian or Pacific Islander	89	0.1%	734.82	41.00	650	850
Two or More Races	6,433	4.9%	735.22	41.43	650	850
White	59,175	44.8%	741.21	38.53	650	850
Not Economically Disadvantaged	62,986	47.7%	744.94	39.22	650	850
Economically Disadvantaged	69,077	52.3%	715.78	37.97	650	850
Non-English Learner (EL)	103,799	78.6%	735.59	40.42	650	850
English Learner (EL)	28,264	21.4%	708.01	36.66	650	850
Students without Disabilities	107,245	81.2%	734.84	40.35	650	850
Student with Disability (SWD)	24,818	18.8%	707.42	37.41	650	850
Reading Claim Score	132,063	100.0%	41.36	16.06	10	90
Female	65,089	49.3%	42.49	16.23	10	90
Male	66,959	50.7%	40.27	15.81	10	90
American Indian/Alaska Native	291	0.2%	38.64	15.99	10	90
Asian	7,635	5.8%	48.89	16.76	10	90
Black or African American	21,294	16.1%	34.89	14.27	10	90
Hispanic/Latino	36,829	27.9%	35.96	14.89	10	90
Middle Eastern or North African	317	0.2%	39.78	16.33	10	90
Native Hawaiian or Pacific Islander	89	0.1%	42.67	15.77	10	90
Two or More Races	6,433	4.9%	43.69	16.26	10	90
White	59,175	44.8%	45.85	15.32	10	90
Not Economically Disadvantaged	62,986	47.7%	47.27	15.60	10	90
Economically Disadvantaged	69,077	52.3%	35.98	14.50	10	90
Non-English Learner	103,799	78.6%	43.73	15.89	10	90
English Learner	28,264	21.4%	32.68	13.48	10	90
Students without Disabilities	107,245	81.2%	43.27	15.76	10	90
Student with Disability (SWD)	24,818	18.8%	33.14	14.69	10	90

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Writing Claim Score	132,063	100.0%	28.22	13.32	10	60
Female	65,089	49.3%	29.68	13.21	10	60
Male	66,959	50.7%	26.81	13.28	10	60
American Indian/Alaska Native	291	0.2%	27.10	13.58	10	56
Asian	7,635	5.8%	34.03	12.62	10	60
Black or African American	21,294	16.1%	23.20	12.86	10	60
Hispanic/Latino	36,829	27.9%	24.52	13.27	10	60
Middle Eastern or North African	317	0.2%	28.87	15.18	10	60
Native Hawaiian or Pacific Islander	89	0.1%	30.56	13.09	10	60
Two or More Races	6,433	4.9%	29.39	13.38	10	60
White	59,175	44.8%	31.46	12.33	10	60
Not Economically Disadvantaged	62,986	47.7%	32.46	12.28	10	60
Economically Disadvantaged	69,077	52.3%	24.36	13.06	10	60
Non-English Learner	103,799	78.6%	29.76	13.00	10	60
English Learner	28,264	21.4%	22.57	12.96	10	60
Students without Disabilities	107,245	81.2%	29.79	12.98	10	60
Student with Disability (SWD)	24,818	18.8%	21.47	12.69	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,742	100.0%	736.19	36.84	650	850
Female	64,183	49.1%	738.89	37.38	650	850
Male	66,540	50.9%	733.57	36.11	650	850
American Indian/Alaska Native	335	0.3%	730.26	35.51	650	837
Asian	7,839	6.0%	755.96	37.24	650	850
Black or African American	21,079	16.1%	718.78	32.60	650	850
Hispanic/Latino	37,074	28.4%	723.57	34.62	650	850
Middle Eastern or North African	304	0.2%	737.38	40.34	650	840
Native Hawaiian or Pacific Islander	102	0.1%	750.58	33.66	662	850
Two or More Races	6,173	4.7%	741.45	37.37	650	850
White	57,836	44.2%	747.38	34.10	650	850
Not Economically Disadvantaged	63,182	48.3%	750.65	34.86	650	850
Economically Disadvantaged	67,560	51.7%	722.66	33.34	650	850
Non-English Learner (EL)	102,671	78.5%	742.12	36.20	650	850
English Learner (EL)	28,071	21.5%	714.48	30.47	650	845
Students without Disabilities	105,573	80.7%	741.30	35.72	650	850
Student with Disability (SWD)	25,169	19.3%	714.74	33.59	650	850

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Reading Claim Score	130,742	100.0%	45.05	14.92	10	90
Female	64,183	49.1%	45.60	14.92	10	90
Male	66,540	50.9%	44.52	14.90	10	90
American Indian/Alaska Native	335	0.3%	42.49	13.96	10	83
Asian	7,839	6.0%	52.66	15.14	10	90
Black or African American	21,079	16.1%	38.37	13.12	10	90
Hispanic/Latino	37,074	28.4%	39.88	13.76	10	90
Middle Eastern or North African	304	0.2%	44.15	14.79	10	83
Native Hawaiian or Pacific Islander	102	0.1%	50.87	13.66	16	90
Two or More Races	6,173	4.7%	47.34	15.26	10	90
White	57,836	44.2%	49.54	14.09	10	90
Not Economically Disadvantaged	63,182	48.3%	50.84	14.36	10	90
Economically Disadvantaged	67,560	51.7%	39.64	13.31	10	90
Non-English Learner	102,671	78.5%	47.52	14.75	10	90
English Learner	28,071	21.5%	36.05	11.73	10	90
Students without Disabilities	105,573	80.7%	47.00	14.51	10	90
Student with Disability (SWD)	25,169	19.3%	36.90	13.81	10	90
Writing Claim Score	130,742	100.0%	27.16	13.56	10	60
Female	64,183	49.1%	28.75	13.49	10	60
Male	66,540	50.9%	25.63	13.45	10	60
American Indian/Alaska Native	335	0.3%	25.20	13.68	10	60
Asian	7,839	6.0%	33.58	12.53	10	60
Black or African American	21,079	16.1%	21.09	12.69	10	60
Hispanic/Latino	37,074	28.4%	23.44	13.27	10	60
Middle Eastern or North African	304	0.2%	28.43	14.58	10	58
Native Hawaiian or Pacific Islander	102	0.1%	31.39	12.63	10	57
Two or More Races	6,173	4.7%	28.43	13.58	10	60
White	57,836	44.2%	30.75	12.63	10	60
Not Economically Disadvantaged	63,182	48.3%	31.74	12.56	10	60
Economically Disadvantaged	67,560	51.7%	22.88	13.05	10	60
Non-English Learner	102,671	78.5%	28.85	13.33	10	60
English Learner	28,071	21.5%	20.99	12.54	10	60
Students without Disabilities	105,573	80.7%	28.93	13.20	10	60
Student with Disability (SWD)	25,169	19.3%	19.72	12.46	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,422	100.0%	739.26	35.95	650	850
Female	63,780	48.9%	743.37	35.84	650	850
Male	66,620	51.1%	735.31	35.62	650	850
American Indian/Alaska Native	318	0.2%	729.78	35.17	650	811
Asian	7,685	5.9%	759.07	35.92	650	850
Black or African American	20,922	16.0%	722.86	33.34	650	850
Hispanic/Latino	36,748	28.2%	728.02	34.92	650	850
Middle Eastern or North African	279	0.2%	736.94	40.28	650	848
Native Hawaiian or Pacific Islander	96	0.1%	741.75	36.09	650	813
Two or More Races	5,967	4.6%	744.21	36.02	650	846
White	58,407	44.8%	749.15	32.76	650	850
Not Economically Disadvantaged	63,480	48.7%	752.91	32.87	650	850
Economically Disadvantaged	66,942	51.3%	726.31	33.90	650	850
Non-English Learner (EL)	108,540	83.2%	744.86	34.43	650	850
English Learner (EL)	21,882	16.8%	711.45	29.87	650	831
Students without Disabilities	104,818	80.4%	745.04	33.78	650	850
Student with Disability (SWD)	25,604	19.6%	715.58	34.86	650	850
Reading Claim Score	130,422	100.0%	45.23	14.22	10	90
Female	63,780	48.9%	46.24	14.09	10	90
Male	66,620	51.1%	44.27	14.29	10	90
American Indian/Alaska Native	318	0.2%	41.54	13.58	10	74
Asian	7,685	5.9%	52.87	14.70	10	90
Black or African American	20,922	16.0%	39.14	13.03	10	90
Hispanic/Latino	36,748	28.2%	40.71	13.48	10	90
Middle Eastern or North African	279	0.2%	42.56	14.34	10	88
Native Hawaiian or Pacific Islander	96	0.1%	45.83	14.64	10	77
Two or More Races	5,967	4.6%	47.39	14.43	10	90
White	58,407	44.8%	49.06	13.25	10	90
Not Economically Disadvantaged	63,480	48.7%	50.60	13.39	10	90
Economically Disadvantaged	66,942	51.3%	40.15	13.08	10	90
Non-English Learner	108,540	83.2%	47.47	13.74	10	90
English Learner	21,882	16.8%	34.14	11.05	10	80
Students without Disabilities	104,818	80.4%	47.32	13.52	10	90
Student with Disability (SWD)	25,604	19.6%	36.71	13.86	10	90

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Writing Claim Score	130,422	100.0%	30.08	12.68	10	60
Female	63,780	48.9%	32.01	12.19	10	60
Male	66,620	51.1%	28.23	12.86	10	60
American Indian/Alaska Native	318	0.2%	27.39	12.86	10	51
Asian	7,685	5.9%	35.96	10.94	10	60
Black or African American	20,922	16.0%	24.77	12.79	10	60
Hispanic/Latino	36,748	28.2%	26.95	12.97	10	60
Middle Eastern or North African	279	0.2%	30.72	14.16	10	60
Native Hawaiian or Pacific Islander	96	0.1%	31.13	12.53	10	51
Two or More Races	5,967	4.6%	31.19	12.56	10	60
White	58,407	44.8%	33.07	11.43	10	60
Not Economically Disadvantaged	63,480	48.7%	34.11	11.10	10	60
Economically Disadvantaged	66,942	51.3%	26.25	12.90	10	60
Non-English Learner	108,540	83.2%	31.67	12.12	10	60
English Learner	21,882	16.8%	22.16	12.42	10	56
Students without Disabilities	104,818	80.4%	32.08	11.81	10	60
Student with Disability (SWD)	25,604	19.6%	21.87	12.80	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	129,426	100.0%	741.26	33.77	650	850
Female	63,392	49.0%	745.50	33.63	650	850
Male	66,000	51.0%	737.18	33.41	650	850
American Indian/Alaska Native	303	0.2%	735.52	34.78	650	833
Asian	7,738	6.0%	761.32	33.20	650	850
Black or African American	20,570	15.9%	726.84	30.91	650	833
Hispanic/Latino	36,427	28.1%	729.85	32.93	650	850
Middle Eastern or North African	288	0.2%	744.53	37.47	653	833
Native Hawaiian or Pacific Islander	97	0.1%	749.43	36.33	650	815
Two or More Races	5,842	4.5%	745.59	33.75	650	850
White	58,161	44.9%	750.40	30.83	650	850
Not Economically Disadvantaged	63,457	49.0%	753.94	31.09	650	850
Economically Disadvantaged	65,969	51.0%	729.06	31.70	650	850
Non-English Learner (EL)	110,385	85.3%	746.66	31.91	650	850
English Learner (EL)	19,041	14.7%	709.94	26.51	650	815
Students without Disabilities	104,321	80.6%	746.73	31.79	650	850
Student with Disability (SWD)	25,105	19.4%	718.50	32.25	650	850

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Reading Claim Score	129,426	100.0%	46.19	13.09	10	90
Female	63,392	49.0%	47.22	12.85	10	90
Male	66,000	51.0%	45.21	13.25	10	90
American Indian/Alaska Native	303	0.2%	44.21	13.70	10	84
Asian	7,738	6.0%	53.93	13.30	10	90
Black or African American	20,570	15.9%	41.23	12.14	10	90
Hispanic/Latino	36,427	28.1%	41.82	12.69	10	90
Middle Eastern or North African	288	0.2%	45.83	13.78	12	90
Native Hawaiian or Pacific Islander	97	0.1%	49.18	13.58	10	84
Two or More Races	5,842	4.5%	48.03	13.11	10	90
White	58,161	44.9%	49.48	12.07	10	90
Not Economically Disadvantaged	63,457	49.0%	50.92	12.25	10	90
Economically Disadvantaged	65,969	51.0%	41.65	12.24	10	90
Non-English Learner	110,385	85.3%	48.27	12.41	10	90
English Learner	19,041	14.7%	34.14	10.12	10	78
Students without Disabilities	104,321	80.6%	48.19	12.38	10	90
Student with Disability (SWD)	25,105	19.4%	37.90	12.72	10	90
Writing Claim Score	129,426	100.0%	29.52	12.86	10	60
Female	63,392	49.0%	31.59	12.38	10	60
Male	66,000	51.0%	27.54	13.00	10	60
American Indian/Alaska Native	303	0.2%	27.77	13.01	10	60
Asian	7,738	6.0%	35.90	10.84	10	60
Black or African American	20,570	15.9%	23.87	12.82	10	60
Hispanic/Latino	36,427	28.1%	26.09	12.94	10	60
Middle Eastern or North African	288	0.2%	32.36	12.99	10	60
Native Hawaiian or Pacific Islander	97	0.1%	32.41	12.51	10	52
Two or More Races	5,842	4.5%	30.40	12.93	10	60
White	58,161	44.9%	32.72	11.67	10	60
Not Economically Disadvantaged	63,457	49.0%	33.78	11.33	10	60
Economically Disadvantaged	65,969	51.0%	25.43	12.92	10	60
Non-English Learner	110,385	85.3%	31.23	12.28	10	60
English Learner	19,041	14.7%	19.63	11.62	10	51
Students without Disabilities	104,321	80.6%	31.52	12.06	10	60
Student with Disability (SWD)	25,105	19.4%	21.21	12.75	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,924	100.0%	744.41	34.04	650	850
Female	64,064	48.9%	749.88	34.02	650	850
Male	66,824	51.0%	739.15	33.21	650	850
American Indian/Alaska Native	272	0.2%	736.53	31.41	663	816
Asian	7,784	5.9%	765.33	33.67	650	850
Black or African American	20,639	15.8%	729.70	30.92	650	850
Hispanic/Latino	37,437	28.6%	733.31	32.86	650	850
Middle Eastern or North African	288	0.2%	743.88	38.13	650	829
Native Hawaiian or Pacific Islander	124	0.1%	751.75	33.71	650	850
Two or More Races	5,645	4.3%	748.00	34.79	650	850
White	58,735	44.9%	753.56	31.30	650	850
Not Economically Disadvantaged	65,280	49.9%	756.89	31.61	650	850
Economically Disadvantaged	65,644	50.1%	731.99	31.74	650	850
Non-English Learner (EL)	110,484	84.4%	749.71	32.41	650	850
English Learner (EL)	20,440	15.6%	715.72	27.70	650	820
Students without Disabilities	106,222	81.1%	749.57	32.23	650	850
Student with Disability (SWD)	24,702	18.9%	722.22	32.65	650	850
Reading Claim Score	130,924	100.0%	47.61	13.14	10	90
Female	64,064	48.9%	48.97	12.99	10	90
Male	66,824	51.0%	46.30	13.15	10	90
American Indian/Alaska Native	272	0.2%	44.72	12.65	16	84
Asian	7,784	5.9%	55.58	13.26	10	90
Black or African American	20,639	15.8%	42.34	12.06	10	90
Hispanic/Latino	37,437	28.6%	43.20	12.56	10	90
Middle Eastern or North African	288	0.2%	44.91	13.63	10	79
Native Hawaiian or Pacific Islander	124	0.1%	50.55	12.67	11	84
Two or More Races	5,645	4.3%	49.33	13.44	10	90
White	58,735	44.9%	51.06	12.17	10	90
Not Economically Disadvantaged	65,280	49.9%	52.40	12.33	10	90
Economically Disadvantaged	65,644	50.1%	42.84	12.16	10	90
Non-English Learner	110,484	84.4%	49.71	12.54	10	90
English Learner	20,440	15.6%	36.24	10.20	10	76
Students without Disabilities	106,222	81.1%	49.50	12.46	10	90
Student with Disability (SWD)	24,702	18.9%	39.46	12.88	10	90

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Writing Claim Score	130,924	100.0%	30.31	13.04	10	60
Female	64,064	48.9%	32.91	12.44	10	60
Male	66,824	51.0%	27.81	13.12	10	60
American Indian/Alaska Native	272	0.2%	28.03	12.69	10	60
Asian	7,784	5.9%	37.01	11.28	10	60
Black or African American	20,639	15.8%	25.04	12.89	10	60
Hispanic/Latino	37,437	28.6%	27.09	13.06	10	60
Middle Eastern or North African	288	0.2%	32.31	13.90	10	60
Native Hawaiian or Pacific Islander	124	0.1%	32.55	12.90	10	60
Two or More Races	5,645	4.3%	30.80	13.35	10	60
White	58,735	44.9%	33.27	12.03	10	60
Not Economically Disadvantaged	65,280	49.9%	34.34	11.74	10	60
Economically Disadvantaged	65,644	50.1%	26.30	13.03	10	60
Non-English Learner	110,484	84.4%	31.91	12.54	10	60
English Learner	20,440	15.6%	21.64	12.27	10	60
Students without Disabilities	106,222	81.1%	32.17	12.31	10	60
Student with Disability (SWD)	24,702	18.9%	22.29	13.06	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	134,023	100.0%	747.32	38.98	650	850
Female	65,419	48.8%	753.69	38.28	650	850
Male	68,554	51.2%	741.22	38.67	650	850
American Indian/Alaska Native	286	0.2%	738.13	40.53	650	837
Asian	7,761	5.8%	771.91	38.63	650	850
Black or African American	21,296	15.9%	731.80	35.80	650	850
Hispanic/Latino	38,833	29.0%	735.51	38.87	650	850
Middle Eastern or North African	301	0.2%	751.72	44.31	650	850
Native Hawaiian or Pacific Islander	127	0.1%	753.88	38.80	650	836
Two or More Races	5,548	4.1%	750.63	39.30	650	850
White	59,871	44.7%	757.00	35.48	650	850
Not Economically Disadvantaged	67,325	50.2%	760.63	35.65	650	850
Economically Disadvantaged	66,698	49.8%	733.88	37.56	650	850
Non-English Learner (EL)	112,658	84.1%	753.42	36.76	650	850
English Learner (EL)	21,365	15.9%	715.12	34.30	650	836
Students without Disabilities	109,205	81.5%	753.20	36.64	650	850
Student with Disability (SWD)	24,818	18.5%	721.41	38.42	650	850

Appendix B: Scale Score Performance by Demographic Subgroup

Subgroup	N	%	Mean	SD	Min.	Max.
Reading Claim Score	134,023	100.0%	48.31	15.50	10	90
Female	65,419	48.8%	49.88	15.16	10	90
Male	68,554	51.2%	46.81	15.67	10	90
American Indian/Alaska Native	286	0.2%	45.18	16.44	10	90
Asian	7,761	5.8%	57.69	15.91	10	90
Black or African American	21,296	15.9%	43.06	14.50	10	90
Hispanic/Latino	38,833	29.0%	43.66	15.33	10	90
Middle Eastern or North African	301	0.2%	47.74	16.77	10	90
Native Hawaiian or Pacific Islander	127	0.1%	50.66	15.79	10	85
Two or More Races	5,548	4.1%	49.99	15.61	10	90
White	59,871	44.7%	51.84	14.28	10	90
Not Economically Disadvantaged	67,325	50.2%	53.35	14.44	10	90
Economically Disadvantaged	66,698	49.8%	43.23	14.86	10	90
Non-English Learner	112,658	84.1%	50.75	14.69	10	90
English Learner	21,365	15.9%	35.46	13.14	10	81
Students without Disabilities	109,205	81.5%	50.49	14.65	10	90
Student with Disability (SWD)	24,818	18.5%	38.72	15.51	10	90
Writing Claim Score	134,023	100.0%	32.19	12.84	10	60
Female	65,419	48.8%	34.84	11.95	10	60
Male	68,554	51.2%	29.67	13.14	10	60
American Indian/Alaska Native	286	0.2%	29.49	13.24	10	53
Asian	7,761	5.8%	39.10	10.88	10	60
Black or African American	21,296	15.9%	26.92	12.92	10	60
Hispanic/Latino	38,833	29.0%	29.03	13.13	10	60
Middle Eastern or North African	301	0.2%	34.90	13.45	10	60
Native Hawaiian or Pacific Islander	127	0.1%	34.48	11.96	10	54
Two or More Races	5,548	4.1%	32.60	13.04	10	60
White	59,871	44.7%	35.18	11.53	10	60
Not Economically Disadvantaged	67,325	50.2%	36.16	11.19	10	60
Economically Disadvantaged	66,698	49.8%	28.19	13.14	10	60
Non-English Learner	112,658	84.1%	33.85	12.18	10	60
English Learner	21,365	15.9%	23.48	12.67	10	60
Students without Disabilities	109,205	81.5%	34.02	11.99	10	60
Student with Disability (SWD)	24,818	18.5%	24.15	13.31	10	60

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	131,915	100.0%	731.74	36.35	650	850
Female	65,022	49.3%	729.76	34.61	650	850
Male	66,878	50.7%	733.66	37.87	650	850
American Indian/Alaska Native	291	0.2%	728.35	38.50	650	850
Asian	7,622	5.8%	755.83	39.22	650	850
Black or African American	21,253	16.1%	712.10	31.15	650	850
Hispanic/Latino	36,782	27.9%	720.32	31.61	650	850
Middle Eastern or North African	318	0.2%	732.13	36.42	650	850
Native Hawaiian or Pacific Islander	89	0.1%	733.55	31.64	657	820
Two or More Races	6,432	4.9%	734.32	37.67	650	850
White	59,128	44.8%	742.53	34.61	650	850
Not Economically Disadvantaged	62,934	47.7%	746.34	35.55	650	850
Economically Disadvantaged	68,981	52.3%	718.42	31.66	650	850
Non-English Learner (EL)	103,698	78.6%	735.57	36.76	650	850
English Learner (EL)	28,217	21.4%	717.67	30.98	650	850
Students without Disabilities	107,146	81.2%	735.95	35.06	650	850
Student with Disability (SWD)	24,769	18.8%	713.53	36.22	650	850
Spanish	6,162	4.7%	708.73	27.96	650	814

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,600	100.0%	731.68	33.25	650	850
Female	64,127	49.1%	729.41	32.23	650	850
Male	66,454	50.9%	733.87	34.06	650	850
American Indian/Alaska Native	334	0.3%	728.22	31.81	650	822
Asian	7,836	6.0%	754.80	33.22	650	850
Black or African American	21,026	16.1%	711.28	28.79	650	850
Hispanic/Latino	37,038	28.4%	721.13	29.68	650	850
Middle Eastern or North African	304	0.2%	734.47	32.84	650	822
Native Hawaiian or Pacific Islander	102	0.1%	737.95	29.92	677	822
Two or More Races	6,171	4.7%	734.69	34.24	650	850
White	57,789	44.2%	742.41	30.54	650	850
Not Economically Disadvantaged	63,168	48.4%	745.41	31.35	650	850
Economically Disadvantaged	67,432	51.6%	718.82	29.64	650	850
Non-English Learner (EL)	102,568	78.5%	735.71	33.39	650	850
English Learner (EL)	28,032	21.5%	716.96	28.20	650	850
Students without Disabilities	105,484	80.8%	735.92	31.93	650	850
Student with Disability (SWD)	25,116	19.2%	713.90	32.80	650	850
Spanish	5,241	4.0%	708.76	27.04	650	801

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,283	100.0%	728.07	33.46	650	850
Female	63,725	48.9%	726.44	31.99	650	850
Male	66,536	51.1%	729.64	34.75	650	850
American Indian/Alaska Native	316	0.2%	719.75	32.74	650	835
Asian	7,683	5.9%	753.71	34.60	650	850
Black or African American	20,908	16.0%	708.05	28.29	650	849
Hispanic/Latino	36,670	28.1%	717.10	29.46	650	850
Middle Eastern or North African	278	0.2%	731.45	33.10	650	819
Native Hawaiian or Pacific Islander	96	0.1%	732.06	33.84	650	805
Two or More Races	5,959	4.6%	731.70	34.75	650	850
White	58,373	44.8%	738.42	30.97	650	850
Not Economically Disadvantaged	63,430	48.7%	742.01	32.00	650	850
Economically Disadvantaged	66,853	51.3%	714.85	29.18	650	850
Non-English Learner (EL)	108,463	83.3%	732.17	33.31	650	850
English Learner (EL)	21,820	16.7%	707.71	25.90	650	835
Students without Disabilities	104,700	80.4%	732.35	32.43	650	850
Student with Disability (SWD)	25,583	19.6%	710.57	31.88	650	850
Spanish	4,257	3.3%	701.42	25.71	650	819

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	129,232	100.0%	726.67	32.57	650	850
Female	63,292	49.0%	725.49	31.66	650	850
Male	65,906	51.0%	727.80	33.39	650	850
American Indian/Alaska Native	302	0.2%	723.43	35.87	650	847
Asian	7,733	6.0%	753.95	34.52	650	850
Black or African American	20,511	15.9%	707.64	26.94	650	847
Hispanic/Latino	36,369	28.1%	715.68	28.74	650	850
Middle Eastern or North African	288	0.2%	729.37	33.54	650	847
Native Hawaiian or Pacific Islander	96	0.1%	734.90	34.37	650	806
Two or More Races	5,832	4.5%	730.50	33.54	650	850
White	58,101	45.0%	736.24	30.09	650	850
Not Economically Disadvantaged	63,408	49.1%	740.08	31.26	650	850
Economically Disadvantaged	65,824	50.9%	713.75	28.32	650	850
Non-English Learner (EL)	110,231	85.3%	730.85	32.10	650	850
English Learner (EL)	19,001	14.7%	702.42	23.45	650	828
Students without Disabilities	104,204	80.6%	731.10	31.54	650	850
Student with Disability (SWD)	25,028	19.4%	708.23	30.24	650	850
Spanish	3,700	2.9%	698.63	23.50	650	785

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	130,683	100.0%	735.38	29.24	650	850
Female	63,927	48.9%	734.35	28.52	650	850
Male	66,720	51.1%	736.36	29.88	650	850
American Indian/Alaska Native	271	0.2%	729.94	27.29	675	817
Asian	7,772	5.9%	759.94	31.57	650	850
Black or African American	20,582	15.7%	718.33	23.63	650	850
Hispanic/Latino	37,349	28.6%	726.33	25.49	650	850
Middle Eastern or North African	290	0.2%	736.77	30.01	652	850
Native Hawaiian or Pacific Islander	124	0.1%	743.06	27.54	675	814
Two or More Races	5,640	4.3%	738.01	31.11	650	850
White	58,655	44.9%	743.61	27.52	650	850
Not Economically Disadvantaged	65,214	49.9%	746.83	28.53	650	850
Economically Disadvantaged	65,469	50.1%	723.97	25.20	650	850
Non-English Learner (EL)	110,300	84.4%	739.00	29.08	650	850
English Learner (EL)	20,383	15.6%	715.80	21.23	650	814
Students without Disabilities	106,048	81.1%	739.44	28.09	650	850
Student with Disability (SWD)	24,635	18.9%	717.91	27.61	650	850
Spanish	2,916	2.2%	709.21	19.64	650	779

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

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

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Score	133,755	100.0%	728.24	41.18	650	850
Female	65,288	48.8%	728.77	39.90	650	850
Male	68,417	51.2%	727.72	42.36	650	850
American Indian/Alaska Native	285	0.2%	719.15	41.51	650	850
Asian	7,744	5.8%	765.91	45.57	650	850
Black or African American	21,237	15.9%	705.77	32.84	650	850
Hispanic/Latino	38,752	29.0%	715.63	35.41	650	850
Middle Eastern or North African	301	0.2%	736.14	44.25	650	850
Native Hawaiian or Pacific Islander	128	0.1%	732.92	40.31	650	832
Two or More Races	5,542	4.1%	731.07	43.48	650	850
White	59,766	44.7%	739.24	39.30	650	850
Not Economically Disadvantaged	67,231	50.3%	743.84	40.84	650	850
Economically Disadvantaged	66,524	49.7%	712.46	35.06	650	850
Non-English Learner (EL)	112,427	84.1%	733.30	41.18	650	850
English Learner (EL)	21,328	15.9%	701.52	29.06	650	850
Students without Disabilities	109,022	81.5%	733.91	39.86	650	850
Student with Disability (SWD)	24,733	18.5%	703.21	37.42	650	850
Spanish	2,725	2.0%	690.47	25.46	650	789

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

Appendix C: Differential Item Functioning (DIF) Results

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

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

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	212	2	1	6	3	204	96				
White vs. Black/African American	212	1	0	20	9	191	90				
White vs. Hispanic/Latino	212	4	2	10	5	198	93				
White vs. Asian	212					209	99	2	1	1	0
White vs. AI/AN	212					212	100				
White vs. NH/PI	212					212	100				
White vs. Multiracial	212			1	0	210	99			1	0
NED vs. ED	212					212	100				
ELN vs. ELY	212	2	1	13	6	197	93				
SWDN vs. SWDY	212					211	100	1	0		

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	203	1	0	6	3	193	95	3	1		
White vs. Black/African American	203	1	0	15	7	187	92				
White vs. Hispanic/Latino	203	1	0	5	2	197	97				
White vs. Asian	203					200	99	2	1	1	0
White vs. AI/AN	203					203	100				
White vs. NH/PI	203					203	100				
White vs. Multiracial	203					203	100				
NED vs. ED	203					203	100				
ELN vs. ELY	203	1	0	18	9	184	91				
SWDN vs. SWDY	203	1	0	2	1	200	99				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	211			10	5	200	95	1	0		
White vs. Black/African American	211	3	1	15	7	193	91				
White vs. Hispanic/Latino	211	2	1	21	10	188	89				
White vs. Asian	211					210	100	1	0		
White vs. AI/AN	211					211	100				
White vs. NH/PI	211					211	100				
White vs. Multiracial	211					211	100				
NED vs. ED	211			1	0	210	100				
ELN vs. ELY	211	10	5	32	15	169	80				
SWDN vs. SWDY	211	1	0	11	5	199	94				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	205			11	5	189	92	5	2		
White vs. Black/African American	205	3	1	11	5	190	93	1	0		
White vs. Hispanic/Latino	205	1	0	12	6	192	94				
White vs. Asian	205					205	100				
White vs. AI/AN	205					205	100				
White vs. NH/PI	205					205	100				
White vs. Multiracial	205			1	0	204	100				
NED vs. ED	205					205	100				
ELN vs. ELY	205	22	11	25	12	157	77	1	0		
SWDN vs. SWDY	205			4	2	201	98				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	216	1	0	12	6	196	91	6	3	1	0
White vs. Black/African American	216	4	2	12	6	198	92	2	1		
White vs. Hispanic/Latino	216	1	0	16	7	199	92				
White vs. Asian	216					215	100			1	0
White vs. AI/AN	216					216	100				
White vs. NH/PI	216					216	100				
White vs. Multiracial	216					216	100				
NED vs. ED	216					216	100				
ELN vs. ELY	216	9	4	36	17	171	79				
SWDN vs. SWDY	216	1	0	10	5	205	95				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	215	1	0	12	6	191	89	11	5		
White vs. Black/African American	215	4	2	18	8	190	88	3	1		
White vs. Hispanic/Latino	215	3	1	11	5	201	93				
White vs. Asian	215					214	100	1	0		
White vs. AI/AN	215					215	100				
White vs. NH/PI	215					215	100				
White vs. Multiracial	215					215	100				
NED vs. ED	215					215	100				
ELN vs. ELY	215	14	7	46	21	155	72				
SWDN vs. SWDY	215			14	7	201	93				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	212			10	5	199	94	3	1		
White vs. Black/African American	212			26	12	185	87	1	0		
White vs. Hispanic/Latino	212			10	5	202	95				
White vs. Asian	212			1	0	207	98	4	2		
White vs. AI/AN	212					212	100				
White vs. NH/PI	212					212	100				
White vs. Multiracial	212			1	0	211	100				
NED vs. ED	212					212	100				
ELN vs. ELY	212	1	0	9	4	202	95				
SWDN vs. SWDY	212			11	5	199	94	2	1		

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	217			17	8	197	91	3	1		
White vs. Black/African American	217	1	0	9	4	204	94	3	1		
White vs. Hispanic/Latino	217			5	2	212	98				
White vs. Asian	217					211	97	6	3		
White vs. AI/AN	217					217	100				
White vs. NH/PI	217					217	100				
White vs. Multiracial	217					217	100				
NED vs. ED	217					217	100				
ELN vs. ELY	217	1	0	9	4	202	93	5	2		
SWDN vs. SWDY	217			7	3	209	96	1	0		

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	221	3	1	9	4	208	94	1	0		
White vs. Black/African American	221	3	1	14	6	202	91	2	1		
White vs. Hispanic/Latino	221			3	1	217	98	1	0		
White vs. Asian	221					215	97	4	2	2	1
White vs. AI/AN	221					221	100				
White vs. NH/PI	221					221	100				
White vs. Multiracial	221			1	0	220	100				
NED vs. ED	221			1	0	220	100				
ELN vs. ELY	221	1	0	9	4	208	94	3	1		
SWDN vs. SWDY	221			3	1	215	97	3	1		

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	200			12	6	186	93	2	1		
White vs. Black/African American	200			18	9	182	91				
White vs. Hispanic/Latino	200			9	5	190	95	1	1		
White vs. Asian	200					192	96	7	4	1	1
White vs. AI/AN	200					200	100				
White vs. NH/PI	200					200	100				
White vs. Multiracial	200			1	1	199	100				
NED vs. ED	200			1	1	199	100				
ELN vs. ELY	200			12	6	187	94	1	1		
SWDN vs. SWDY	200			8	4	192	96				

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	199	2	1	10	5	184	92	1	1	2	1
White vs. Black/African American	199	2	1	10	5	186	93	1	1		
White vs. Hispanic/Latino	199			6	3	193	97				
White vs. Asian	199					190	95	5	3	4	2
White vs. AI/AN	199					199	100				
White vs. NH/PI	199					199	100				
White vs. Multiracial	199					199	100				
NED vs. ED	199					199	100				
ELN vs. ELY	199	1	1	10	5	186	93	2	1		
SWDN vs. SWDY	199			2	1	195	98	1	1	1	1

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

DIF Comparison	Total #Unique Items	C- N	C- %	B- N	B- %	A N	A %	B+ N	B+ %	C+ N	C+ %
Female vs. Male	170	1	1	2	1	166	98	1	1		
White vs. Black/African American	170	4	2	8	5	157	92	1	1		
White vs. Hispanic/Latino	170			3	2	167	98				
White vs. Asian	170					158	93	6	4	6	4
White vs. AI/AN	170					170	100				
White vs. NH/PI	170					170	100				
White vs. Multiracial	170					169	99	1	1		
NED vs. ED	170					170	100				
ELN vs. ELY	170			13	8	155	91	2	1		
SWDN vs. SWDY	170			4	2	164	96	2	1		

Appendix D: TCCs, CSEM Curves, and TIF Curves

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

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

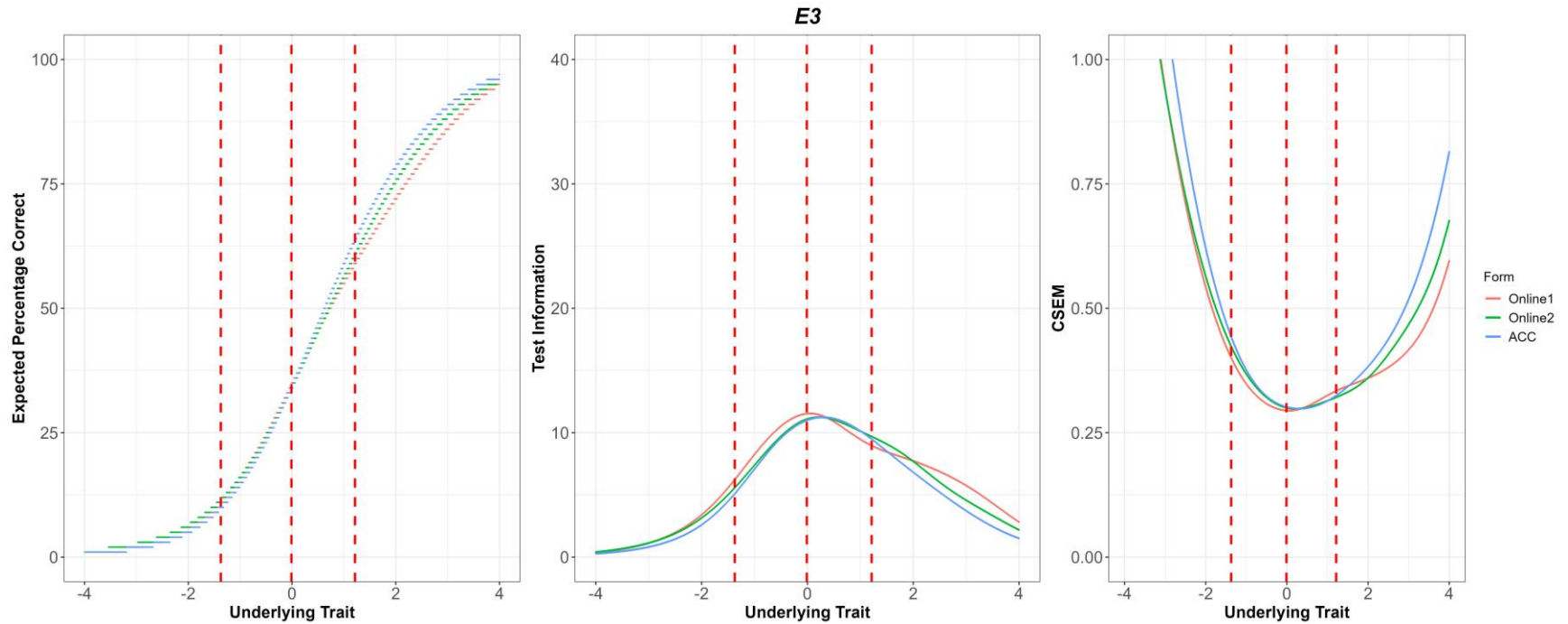


Figure D.2. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 4

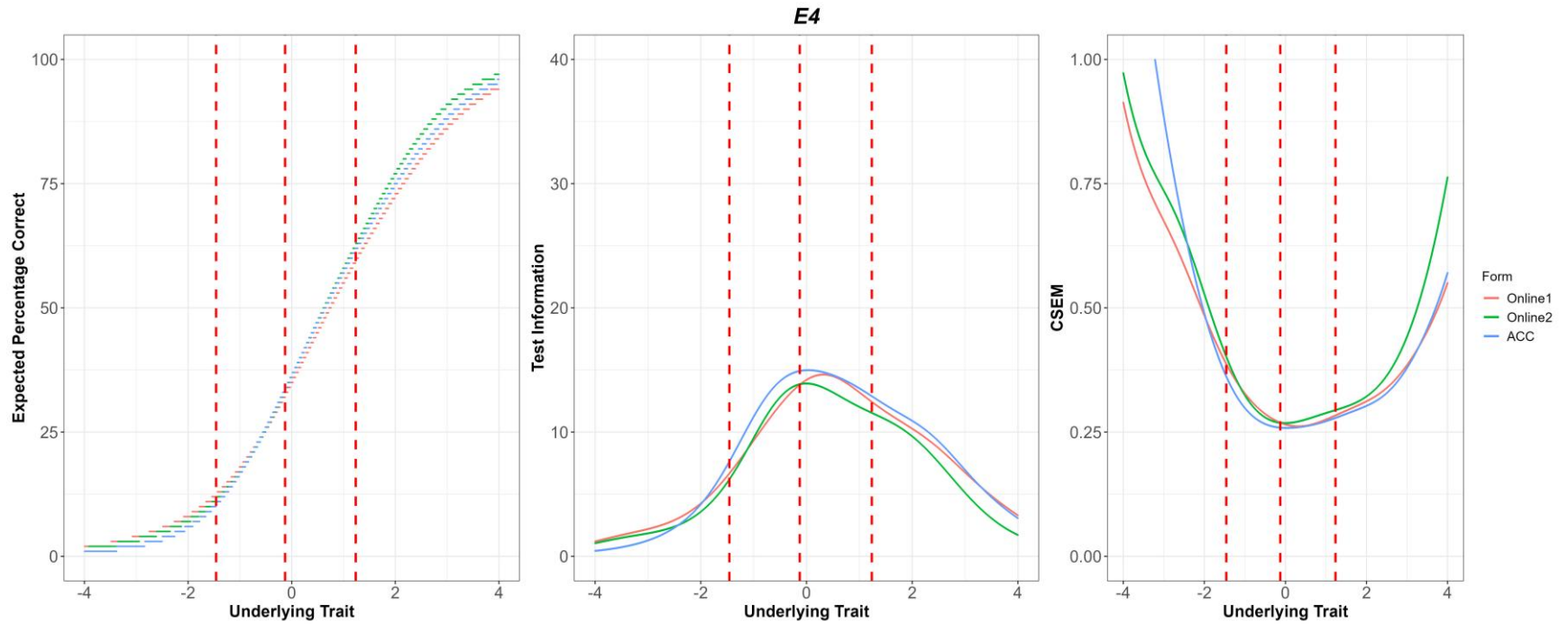


Figure D.3. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 5

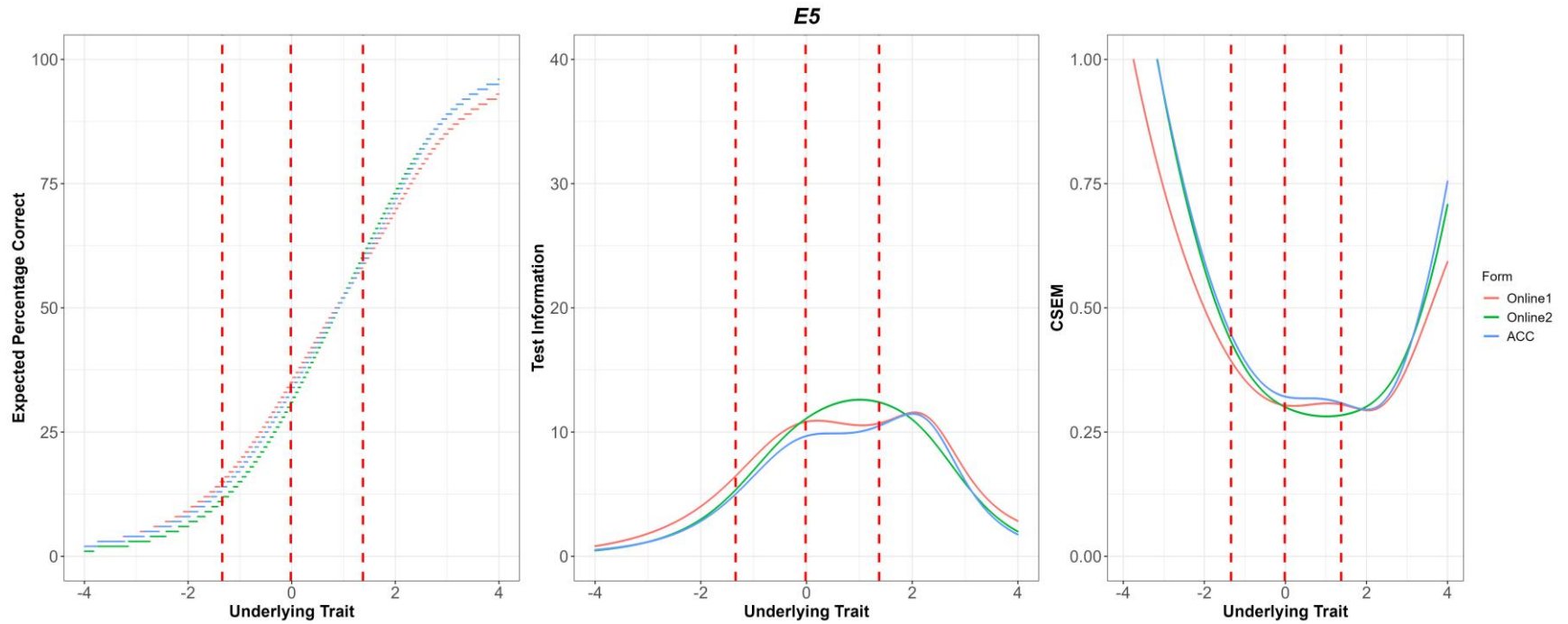


Figure D.4. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 6

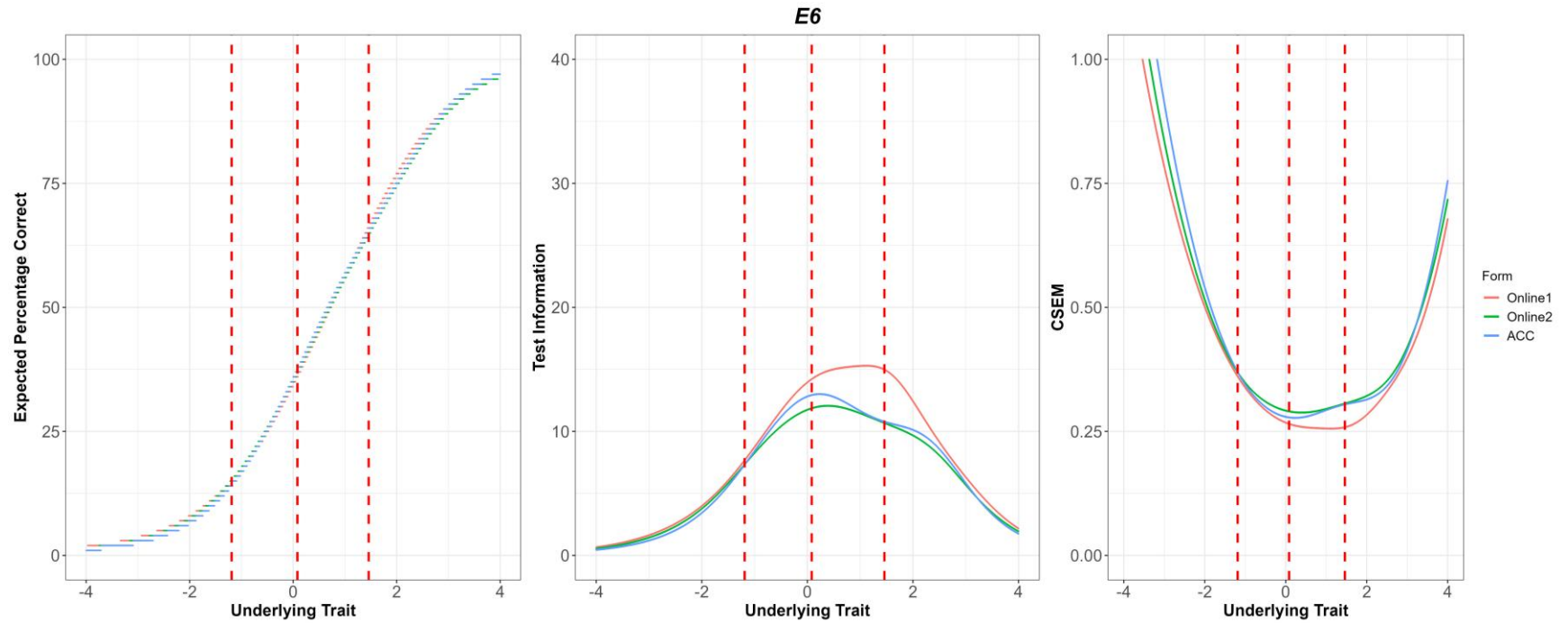


Figure D.5. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 7

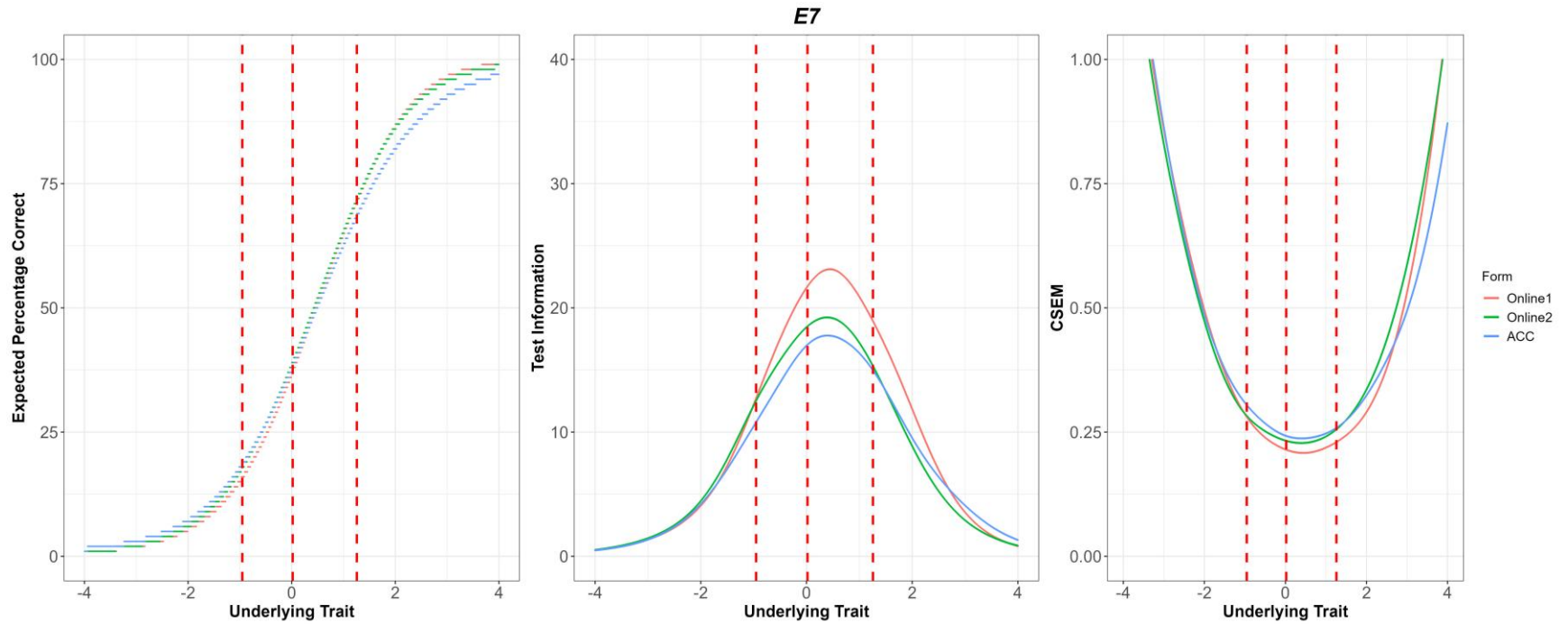


Figure D.6. Pre-Equated TCCs, CSEM Curves, and TIF Curves—ELA/L Grade 8

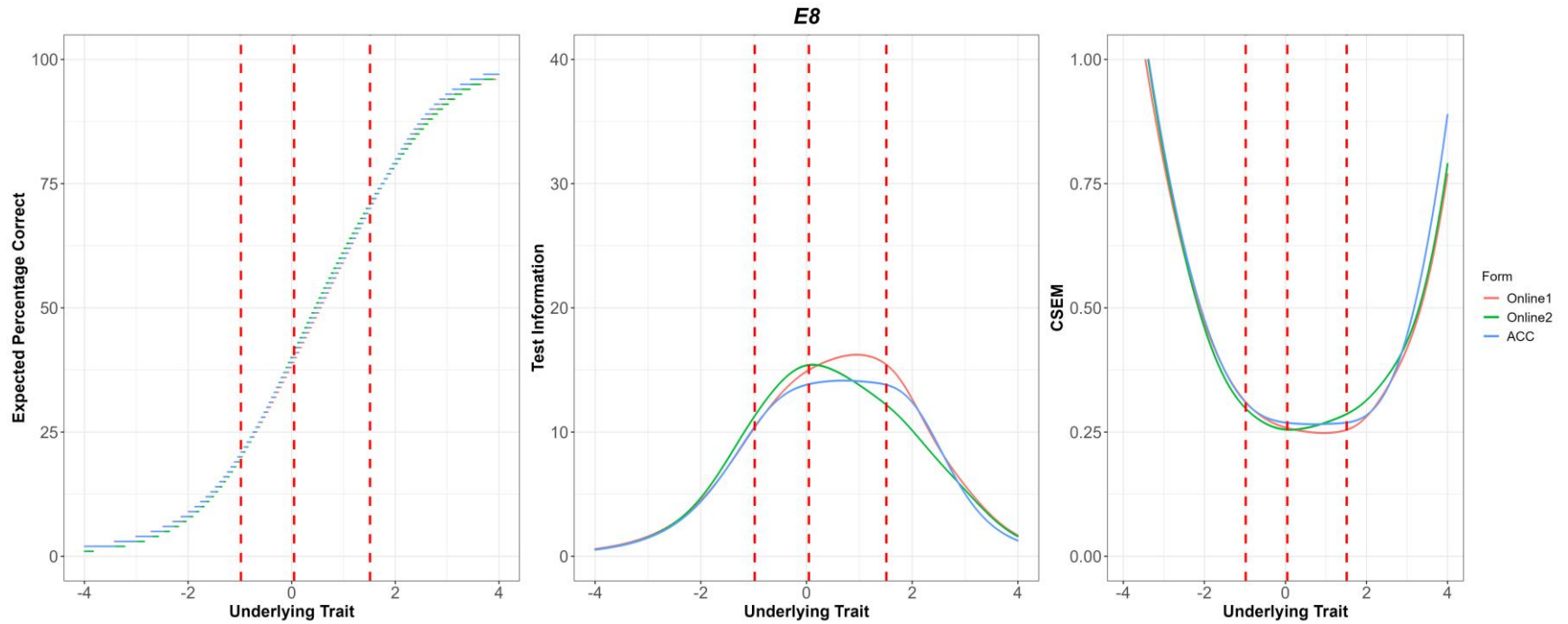


Figure D.7. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 3

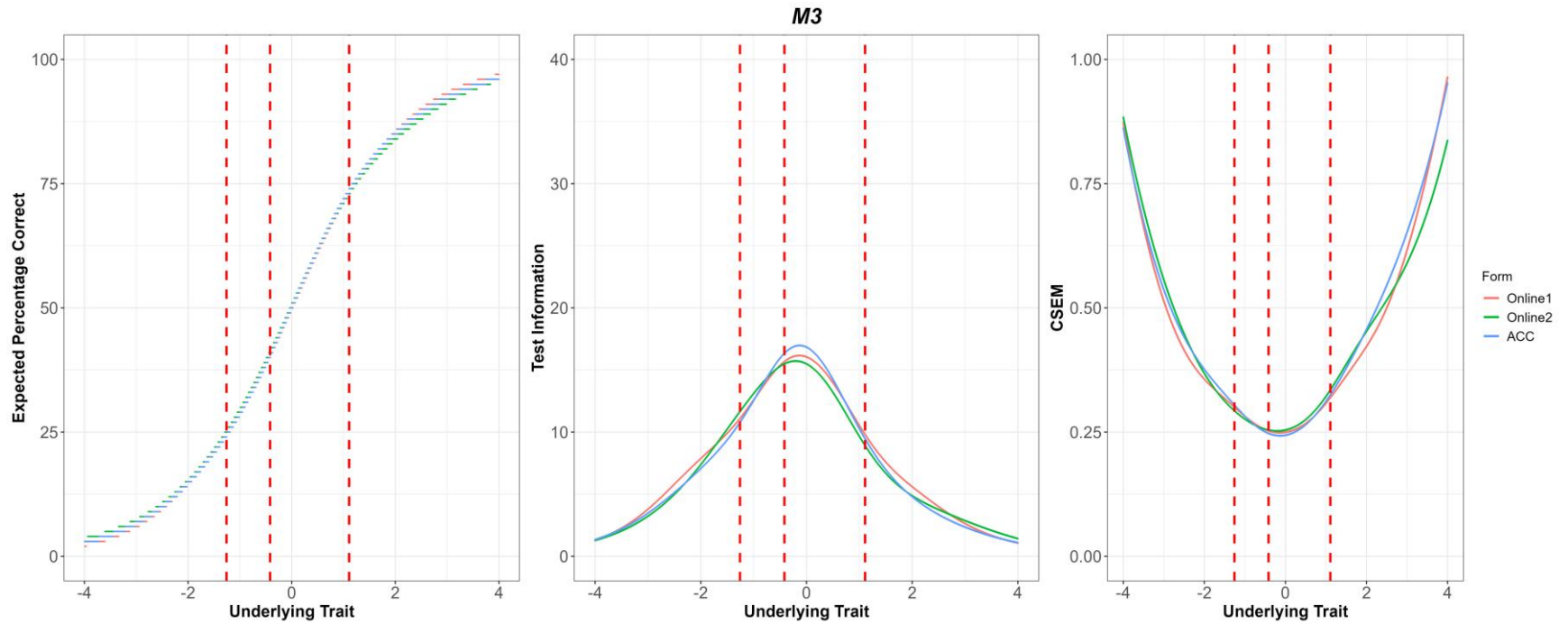


Figure D.8. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 4

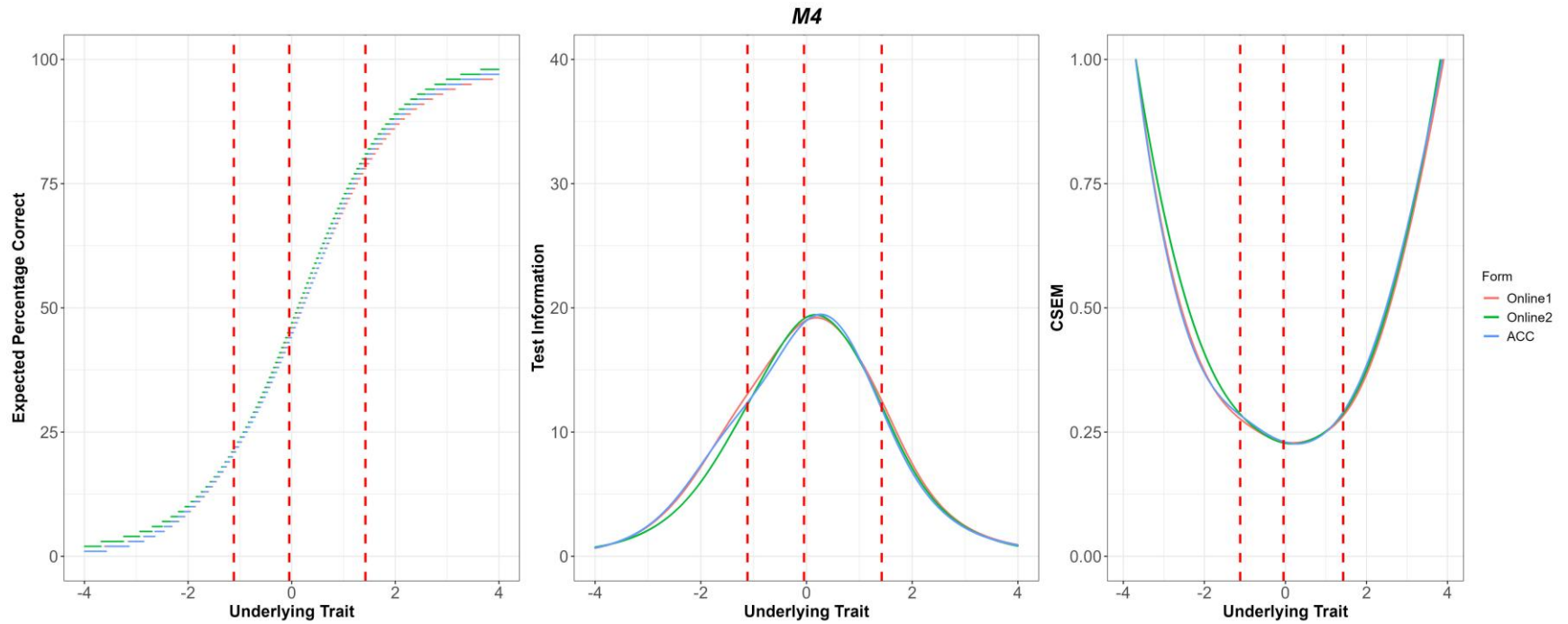


Figure D.9. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 5

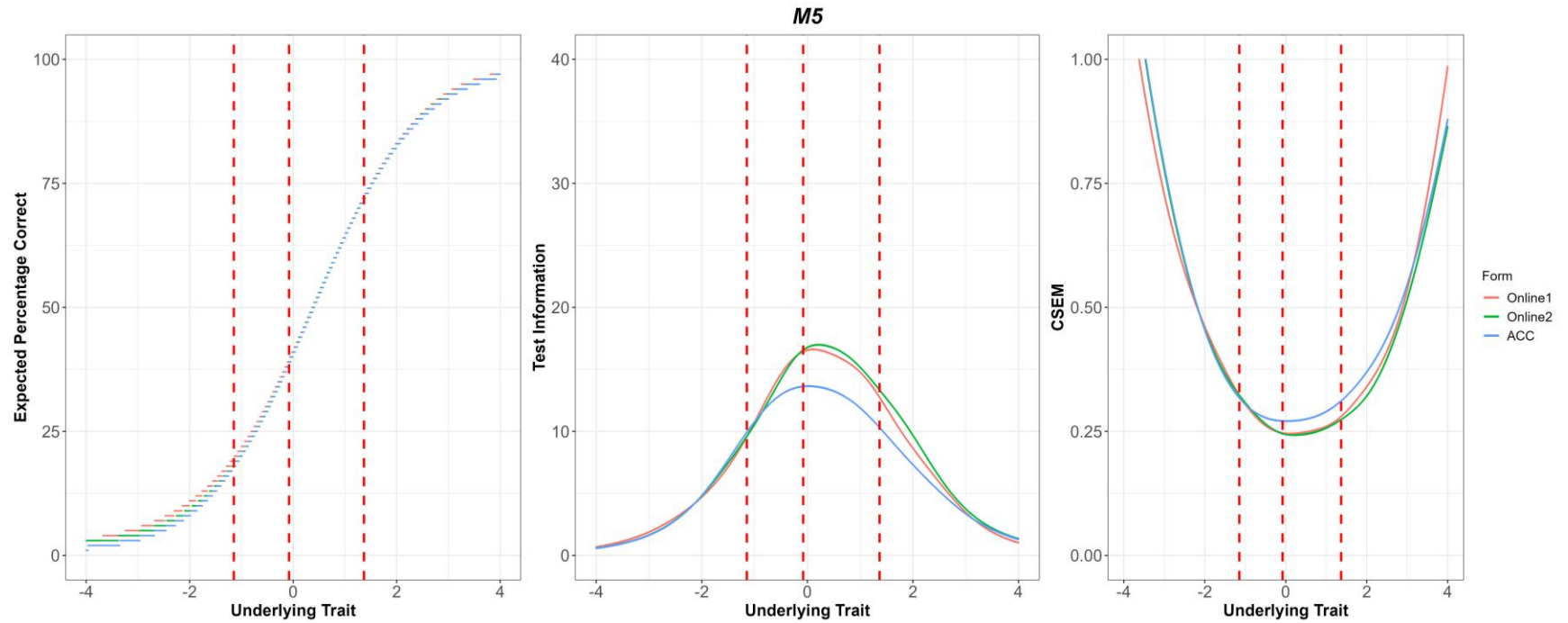


Figure D.10. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 6

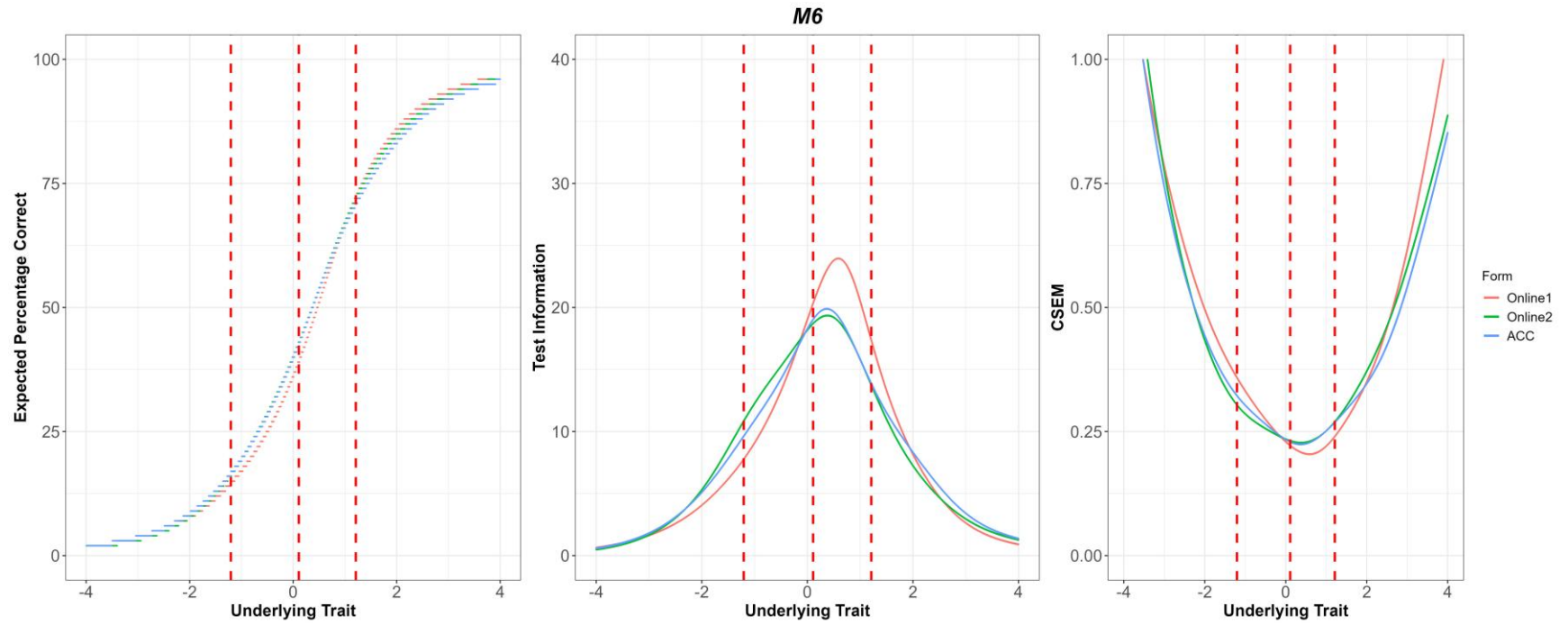


Figure D.11. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 7

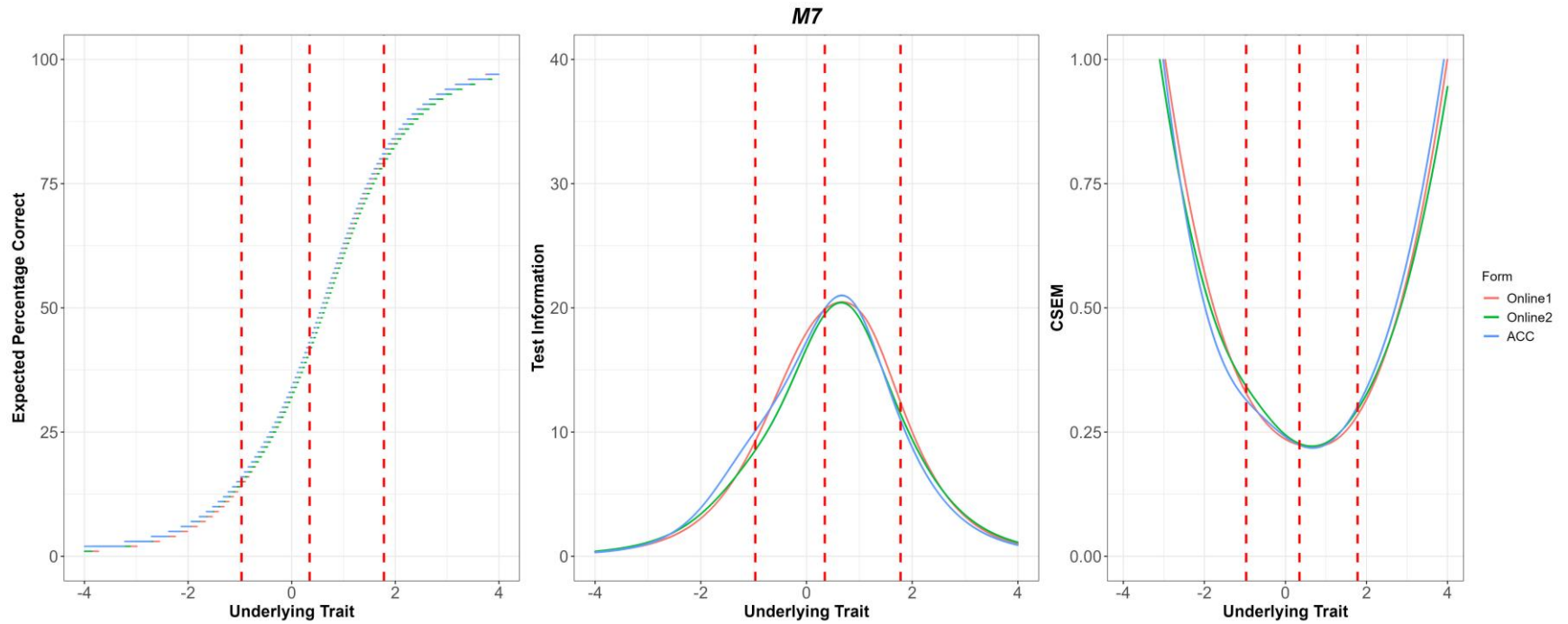
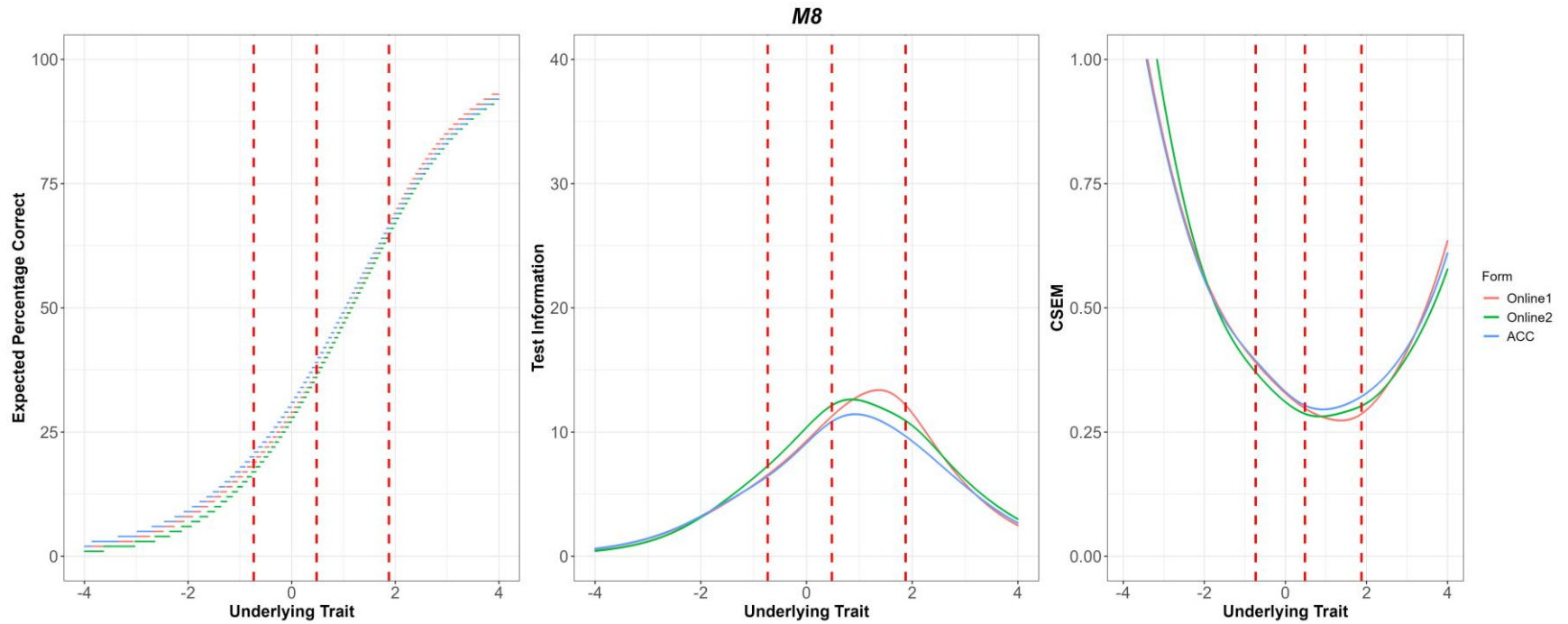


Figure D.12. Pre-Equated TCCs, CSEM Curves, and TIF Curves—Mathematics Grade 8



Appendix E: Reliability by Subgroup

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	55	3.46	0.87	65,523	0.90	65,211	0.88	931	0.82
Male	55	3.37	0.87	33,293	0.90	32,876	0.88	578	0.82
Female	55	3.54	0.87	32,224	0.90	32,326	0.88	353	0.82
American Indian/Alaska Native	55	n/a	n/a	148	0.89	139	0.89	n/a	n/a
Asian	55	n/a	n/a	3,586	0.90	3,986	0.89	n/a	n/a
Black/African American	55	3.26	0.84	10,715	0.88	10,222	0.87	231	0.78
Hispanic/Latino	55	3.30	0.87	18,505	0.89	17,927	0.88	278	0.83
Middle Eastern or North African	55	n/a	n/a	3,180	0.90	3,208	0.88	n/a	n/a
Native Hawaiian/Pacific Islander	55	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	55	n/a	n/a	139	0.90	173	0.90	n/a	n/a
White	55	3.59	0.86	29,208	0.89	29,509	0.86	353	0.83
Economically Disadvantaged	55	3.30	0.85	34,815	0.89	33,358	0.87	621	0.78
Not Economically Disadvantaged	55	3.62	0.87	30,708	0.89	31,853	0.87	310	0.86
English Learner (EL)	55	3.14	0.83	14,101	0.88	13,835	0.86	220	0.76
Non-EL	55	3.53	0.87	51,422	0.90	51,376	0.88	711	0.83
Students with Disabilities (SWD)	55	3.19	0.86	13,709	0.89	9,922	0.89	843	0.82
Students without Disabilities	55	n/a	n/a	51,693	0.89	55,210	0.88	n/a	n/a
American Sign Language (ASL)	55	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	55	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	55	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	55	n/a	n/a	65,523	0.90	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	74	3.82	0.91	64,883	0.93	64,675	0.91	869	0.88
Male	74	3.73	0.91	33,234	0.93	32,577	0.92	555	0.88
Female	74	3.90	0.91	31,640	0.93	32,088	0.91	314	0.89
American Indian/Alaska Native	74	n/a	n/a	169	0.93	161	0.91	n/a	n/a
Asian	74	n/a	n/a	3,751	0.93	4,044	0.91	n/a	n/a
Black/African American	74	3.57	0.89	10,616	0.92	10,166	0.90	190	0.86
Hispanic/Latino	74	3.61	0.88	18,662	0.92	18,056	0.91	270	0.82
Middle Eastern or North African	74	n/a	n/a	2,967	0.93	3,140	0.91	n/a	n/a
Native Hawaiian/Pacific Islander	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	74	n/a	n/a	135	0.92	168	0.92	n/a	n/a
White	74	3.96	0.91	28,526	0.92	28,895	0.90	326	0.90
Economically Disadvantaged	74	3.62	0.89	34,025	0.92	32,745	0.90	564	0.84
Not Economically Disadvantaged	74	3.98	0.91	30,858	0.92	31,930	0.90	305	0.91
English Learner (EL)	74	3.44	0.86	14,023	0.90	13,733	0.88	238	0.82
Non-EL	74	3.90	0.91	50,860	0.93	50,942	0.91	631	0.89
Students with Disabilities (SWD)	74	3.47	0.91	14,065	0.92	9,983	0.92	817	0.88
Students without Disabilities	74	n/a	n/a	50,721	0.92	54,600	0.91	n/a	n/a
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	64,883	0.93	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	74	3.82	0.89	64,800	0.92	64,433	0.90	887	0.84
Male	74	3.73	0.88	33,230	0.92	32,660	0.90	567	0.83
Female	74	3.92	0.89	31,560	0.92	31,761	0.90	320	0.85
American Indian/Alaska Native	74	n/a	n/a	165	0.91	149	0.90	n/a	n/a
Asian	74	n/a	n/a	3,747	0.92	3,896	0.91	n/a	n/a
Black/African American	74	3.62	0.87	10,587	0.91	10,035	0.89	210	0.83
Hispanic/Latino	74	3.66	0.85	18,463	0.91	17,919	0.90	275	0.74
Middle Eastern or North African	74	n/a	n/a	2,885	0.92	3,041	0.91	n/a	n/a
Native Hawaiian/Pacific Islander	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	74	n/a	n/a	126	0.92	148	0.91	n/a	n/a
White	74	3.92	0.89	28,781	0.91	29,196	0.89	332	0.86
Economically Disadvantaged	74	3.69	0.87	33,874	0.91	32,274	0.89	583	0.81
Not Economically Disadvantaged	74	3.92	0.89	30,926	0.91	32,159	0.89	304	0.87
English Learner (EL)	74	3.42	0.82	11,053	0.87	10,544	0.84	227	0.74
Non-EL	74	3.88	0.89	53,747	0.92	53,889	0.90	660	0.85
Students with Disabilities (SWD)	74	3.54	0.89	14,485	0.91	9,979	0.91	829	0.84
Students without Disabilities	74	n/a	n/a	50,210	0.91	54,359	0.90	n/a	n/a
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	64,800	0.92	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	74	4.09	0.90	64,314	0.93	64,019	0.91	662	0.87
Male	74	3.94	0.90	33,068	0.93	32,281	0.91	400	0.86
Female	74	4.22	0.90	31,231	0.92	31,719	0.90	262	0.87
American Indian/Alaska Native	74	n/a	n/a	149	0.93	153	0.91	n/a	n/a
Asian	74	n/a	n/a	3,682	0.92	4,022	0.90	n/a	n/a
Black/African American	74	3.82	0.88	10,481	0.92	9,785	0.89	171	0.83
Hispanic/Latino	74	3.88	0.89	18,335	0.92	17,800	0.91	174	0.85
Middle Eastern or North African	74	n/a	n/a	2,906	0.93	2,863	0.91	n/a	n/a
Native Hawaiian/Pacific Islander	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	74	n/a	n/a	143	0.93	144	0.91	n/a	n/a
White	74	4.21	0.89	28,572	0.92	29,201	0.89	258	0.87
Economically Disadvantaged	74	3.90	0.89	33,346	0.91	31,887	0.90	432	0.85
Not Economically Disadvantaged	74	4.21	0.90	30,968	0.92	32,132	0.89	230	0.89
English Learner (EL)	74	3.48	0.84	9,748	0.87	9,082	0.85	138	0.81
Non-EL	74	4.15	0.90	54,566	0.92	54,937	0.90	524	0.87
Students with Disabilities (SWD)	74	3.75	0.90	14,241	0.92	9,922	0.91	618	0.86
Students without Disabilities	74	n/a	n/a	49,976	0.92	54,005	0.90	n/a	n/a
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	64,314	0.93	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	74	4.37	0.91	65,139	0.93	64,682	0.91	559	0.89
Male	74	4.17	0.91	33,564	0.93	32,661	0.91	341	0.89
Female	74	4.54	0.91	31,563	0.92	31,999	0.91	218	0.89
American Indian/Alaska Native	74	n/a	n/a	136	0.93	130	0.89	n/a	n/a
Asian	74	n/a	n/a	3,819	0.92	3,942	0.90	n/a	n/a
Black/African American	74	4.04	0.90	10,473	0.92	9,862	0.89	121	0.88
Hispanic/Latino	74	4.13	0.90	18,659	0.92	18,451	0.90	165	0.87
Middle Eastern or North African	74	n/a	n/a	2,771	0.93	2,826	0.91	n/a	n/a
Native Hawaiian/Pacific Islander	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	74	n/a	n/a	131	0.92	157	0.91	n/a	n/a
White	74	4.49	0.90	29,089	0.92	29,252	0.90	235	0.89
Economically Disadvantaged	74	4.13	0.90	33,129	0.92	31,747	0.90	369	0.87
Not Economically Disadvantaged	74	4.55	0.91	32,010	0.92	32,935	0.89	190	0.91
English Learner (EL)	74	3.70	0.86	10,334	0.89	9,907	0.86	101	0.82
Non-EL	74	4.45	0.91	54,805	0.92	54,775	0.90	458	0.89
Students with Disabilities (SWD)	74	4.01	0.91	13,916	0.92	9,922	0.91	518	0.89
Students without Disabilities	74	n/a	n/a	51,125	0.92	54,672	0.90	n/a	n/a
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	65,139	0.93	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	74	4.25	0.91	66,565	0.92	66,249	0.91	543	0.90
Male	74	4.11	0.91	34,237	0.92	33,622	0.91	354	0.90
Female	74	4.37	0.91	32,304	0.91	32,602	0.90	189	0.91
American Indian/Alaska Native	74	n/a	n/a	138	0.92	143	0.90	n/a	n/a
Asian	74	n/a	n/a	3,762	0.91	3,971	0.90	n/a	n/a
Black/African American	74	4.09	0.89	10,766	0.91	10,197	0.90	145	0.86
Hispanic/Latino	74	4.11	0.90	19,462	0.91	19,037	0.91	131	0.89
Middle Eastern or North African	74	n/a	n/a	2,686	0.92	2,794	0.91	n/a	n/a
Native Hawaiian/Pacific Islander	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	74	n/a	n/a	146	0.92	153	0.92	n/a	n/a
White	74	4.30	0.90	29,540	0.91	29,893	0.89	229	0.91
Economically Disadvantaged	74	4.10	0.89	33,706	0.91	32,165	0.90	363	0.86
Not Economically Disadvantaged	74	4.33	0.91	32,859	0.91	34,084	0.89	180	0.93
English Learner (EL)	74	3.78	0.87	10,900	0.88	10,248	0.89	107	0.83
Non-EL	74	4.29	0.90	55,665	0.91	56,001	0.90	436	0.91
Students with Disabilities (SWD)	74	4.02	0.91	14,044	0.91	9,927	0.91	497	0.90
Students without Disabilities	74	n/a	n/a	52,432	0.91	56,230	0.90	n/a	n/a
American Sign Language (ASL)	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	74	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	74	n/a	n/a	66,565	0.92	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.96	0.91	87,873	0.92	42,477	0.91	1,118	0.90
Male	52	2.97	0.92	44,914	0.93	21,046	0.92	698	0.91
Female	52	2.95	0.91	42,951	0.92	21,424	0.91	420	0.89
American Indian/Alaska Native	52	n/a	n/a	195	0.93	n/a	n/a	n/a	n/a
Asian	52	n/a	n/a	4,745	0.92	2,813	0.91	n/a	n/a
Black/African American	52	2.73	0.89	15,854	0.90	5,031	0.89	222	0.87
Hispanic/Latino	52	2.87	0.90	29,066	0.90	7,254	0.90	319	0.89
Middle Eastern or North African	52	n/a	n/a	3,932	0.93	2,443	0.92	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	259	0.93	n/a	n/a	n/a	n/a
White	52	3.07	0.91	33,764	0.92	24,760	0.90	484	0.90
Economically Disadvantaged	52	2.85	0.90	50,080	0.90	17,846	0.90	734	0.89
Not Economically Disadvantaged	52	3.07	0.91	37,793	0.92	24,631	0.90	384	0.91
English Learner (EL)	52	2.83	0.89	23,079	0.90	4,729	0.90	292	0.89
Non-EL	52	3.00	0.91	64,794	0.92	37,748	0.91	826	0.90
Students with Disabilities (SWD)	52	2.87	0.91	18,524	0.92	4,939	0.92	964	0.90
Students without Disabilities	52	2.98	0.90	69,211	0.92	37,477	0.91	149	0.87
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	61,690	0.92	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.93	0.93	87,096	0.94	42,087	0.93	992	0.92
Male	52	2.91	0.93	44,539	0.94	21,042	0.93	630	0.92
Female	52	2.95	0.92	42,547	0.93	21,036	0.92	362	0.91
American Indian/Alaska Native	52	n/a	n/a	228	0.93	n/a	n/a	n/a	n/a
Asian	52	n/a	n/a	4,836	0.93	2,951	0.92	n/a	n/a
Black/African American	52	2.67	0.91	15,798	0.92	4,854	0.91	201	0.89
Hispanic/Latino	52	2.82	0.91	29,368	0.92	7,278	0.92	260	0.88
Middle Eastern or North African	52	n/a	n/a	3,712	0.94	2,380	0.93	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	239	0.94	n/a	n/a	n/a	n/a
White	52	3.05	0.92	32,850	0.93	24,427	0.91	423	0.93
Economically Disadvantaged	52	2.82	0.91	49,277	0.92	17,203	0.91	641	0.90
Not Economically Disadvantaged	52	3.04	0.92	37,819	0.93	24,884	0.92	351	0.93
English Learner (EL)	52	2.74	0.89	23,281	0.91	4,422	0.91	242	0.87
Non-EL	52	2.98	0.93	63,815	0.94	37,665	0.93	750	0.92
Students with Disabilities (SWD)	52	2.81	0.93	19,285	0.93	4,659	0.93	876	0.91
Students without Disabilities	52	3.03	0.92	67,680	0.93	37,371	0.92	110	0.92
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	61,224	0.93	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.90	0.89	85,694	0.93	43,199	0.92	999	0.82
Male	52	2.89	0.89	43,984	0.93	21,729	0.93	620	0.81
Female	52	2.92	0.89	41,698	0.92	21,460	0.91	379	0.83
American Indian/Alaska Native	52	n/a	n/a	236	0.92	n/a	n/a	n/a	n/a
Asian	52	n/a	n/a	4,650	0.93	2,983	0.92	n/a	n/a
Black/African American	52	2.68	0.85	15,695	0.88	4,883	0.90	190	0.77
Hispanic/Latino	52	2.73	0.87	28,343	0.90	7,937	0.90	278	0.81
Middle Eastern or North African	52	n/a	n/a	3,529	0.93	2,380	0.93	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	231	0.92	n/a	n/a	n/a	n/a
White	52	3.02	0.88	32,954	0.92	24,858	0.91	450	0.81
Economically Disadvantaged	52	2.76	0.86	48,350	0.90	17,592	0.90	636	0.79
Not Economically Disadvantaged	52	3.02	0.90	37,344	0.92	25,607	0.91	363	0.85
English Learner (EL)	52	2.59	0.83	17,970	0.86	3,545	0.87	248	0.78
Non-EL	52	2.96	0.89	67,724	0.93	39,654	0.92	751	0.82
Students with Disabilities (SWD)	52	2.76	0.88	19,555	0.91	4,832	0.93	875	0.82
Students without Disabilities	52	2.97	0.88	66,005	0.92	38,295	0.92	124	0.81
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	59,088	0.93	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.88	0.92	85,421	0.94	42,417	0.93	712	0.88
Male	52	2.88	0.92	43,614	0.94	21,512	0.93	422	0.90
Female	52	2.88	0.90	41,785	0.93	20,893	0.92	290	0.86
American Indian/Alaska Native	52	n/a	n/a	182	0.95	117	0.94	n/a	n/a
Asian	52	n/a	n/a	4,705	0.94	2,978	0.92	n/a	n/a
Black/African American	52	2.53	0.88	15,339	0.90	4,797	0.90	136	0.82
Hispanic/Latino	52	2.70	0.90	28,061	0.91	7,878	0.91	201	0.87
Middle Eastern or North African	52	n/a	n/a	3,519	0.94	2,246	0.93	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	225	0.94	n/a	n/a	n/a	n/a
White	52	3.04	0.91	33,326	0.93	24,307	0.91	311	0.89
Economically Disadvantaged	52	2.67	0.89	47,330	0.91	17,565	0.91	441	0.84
Not Economically Disadvantaged	52	3.07	0.92	38,091	0.93	24,852	0.92	271	0.91
English Learner (EL)	52	2.41	0.86	15,539	0.83	3,154	0.87	182	0.86
Non-EL	52	2.96	0.91	69,882	0.94	39,263	0.92	530	0.88
Students with Disabilities (SWD)	52	2.65	0.91	19,228	0.92	4,860	0.93	587	0.88
Students without Disabilities	52	2.97	0.91	66,078	0.93	37,486	0.92	121	0.88
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	55,903	0.94	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.97	0.92	85,011	0.94	44,259	0.93	556	0.90
Male	52	2.96	0.93	43,801	0.94	22,158	0.94	348	0.91
Female	52	2.97	0.91	41,194	0.93	22,083	0.93	208	0.88
American Indian/Alaska Native	52	n/a	n/a	181	0.94	n/a	n/a	n/a	n/a
Asian	52	n/a	n/a	4,645	0.94	3,098	0.93	n/a	n/a
Black/African American	52	n/a	n/a	15,301	0.90	4,881	0.91	n/a	n/a
Hispanic/Latino	52	2.81	0.90	28,627	0.92	8,319	0.92	156	0.86
Middle Eastern or North African	52	n/a	n/a	3,330	0.95	2,248	0.94	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	222	0.94	n/a	n/a	n/a	n/a
White	52	3.10	0.92	32,622	0.93	25,524	0.92	275	0.90
Economically Disadvantaged	52	2.81	0.90	46,799	0.92	17,703	0.92	357	0.87
Not Economically Disadvantaged	52	3.10	0.93	38,212	0.94	26,556	0.93	199	0.93
English Learner (EL)	52	2.56	0.84	16,537	0.87	3,593	0.88	112	0.77
Non-EL	52	3.03	0.93	68,474	0.94	40,666	0.93	444	0.91
Students with Disabilities (SWD)	52	2.75	0.92	18,700	0.93	5,053	0.94	484	0.91
Students without Disabilities	52	n/a	n/a	66,187	0.93	39,148	0.93	n/a	n/a
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	55,502	0.94	n/a	n/a	n/a	n/a

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

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

Subgroup	Max. Raw Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	Online2 N	Online2 Alpha	ACC1 N	ACC1 Alpha
Total Group	52	2.79	0.89	94,594	0.92	37,840	0.92	557	0.83
Male	52	2.75	0.89	48,599	0.93	19,078	0.93	346	0.82
Female	52	2.84	0.89	45,970	0.91	18,739	0.92	211	0.84
American Indian/Alaska Native	52	n/a	n/a	205	0.92	n/a	n/a	n/a	n/a
Asian	52	n/a	n/a	5,245	0.93	2,475	0.93	n/a	n/a
Black/African American	52	2.53	0.81	16,699	0.87	4,145	0.88	125	0.68
Hispanic/Latino	52	2.64	0.83	31,088	0.89	7,293	0.90	143	0.70
Middle Eastern or North African	52	n/a	n/a	3,507	0.93	1,972	0.93	n/a	n/a
Native Hawaiian/Pacific Islander	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	52	n/a	n/a	246	0.93	n/a	n/a	n/a	n/a
White	52	2.88	0.90	37,510	0.91	21,791	0.91	251	0.86
Economically Disadvantaged	52	2.59	0.80	50,426	0.89	15,196	0.89	356	0.61
Not Economically Disadvantaged	52	2.96	0.91	44,168	0.92	22,644	0.92	201	0.90
English Learner (EL)	52	2.42	0.80	17,710	0.83	3,347	0.84	129	0.73
Non-EL	52	2.84	0.89	76,884	0.92	34,493	0.92	428	0.84
Students with Disabilities (SWD)	52	2.56	0.88	19,436	0.90	4,438	0.92	462	0.83
Students without Disabilities	52	n/a	n/a	75,028	0.92	33,352	0.92	n/a	n/a
American Sign Language (ASL)	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	52	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Text-to-Speech (TTS)	52	n/a	n/a	56,727	0.92	n/a	n/a	n/a	n/a

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

Appendix F: Decision Accuracy and Consistency by Performance Level

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–684	0.12	0.02	0.00	0.00	0.14
	685–734	0.04	0.31	0.05	0.00	0.41
	735–779	0.00	0.05	0.24	0.04	0.33
	780–850	0.00	0.00	0.03	0.09	0.12
Consistency	650–684	0.11	0.05	0.00	0.00	0.16
	685–734	0.05	0.27	0.07	0.00	0.39
	735–779	0.00	0.07	0.19	0.05	0.31
	780–850	0.00	0.00	0.05	0.08	0.13

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–694	0.11	0.02	0.00	0.00	0.13
	695–736	0.03	0.30	0.05	0.00	0.38
	737–779	0.00	0.05	0.29	0.03	0.36
	780–850	0.00	0.00	0.02	0.10	0.12
Consistency	650–694	0.11	0.04	0.00	0.00	0.15
	695–736	0.04	0.27	0.06	0.00	0.37
	737–779	0.00	0.06	0.25	0.04	0.35
	780–850	0.00	0.00	0.04	0.10	0.13

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–699	0.13	0.02	0.00	0.00	0.15
	700–738	0.03	0.24	0.05	0.00	0.32
	739–779	0.00	0.05	0.30	0.05	0.40
	780–850	0.00	0.00	0.03	0.10	0.13
Consistency	650–699	0.12	0.04	0.00	0.00	0.16
	700–738	0.04	0.21	0.07	0.00	0.32
	739–779	0.00	0.07	0.26	0.05	0.38
	780–850	0.00	0.00	0.05	0.09	0.14

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–704	0.13	0.02	0.00	0.00	0.15
	705–740	0.03	0.25	0.05	0.00	0.33
	741–779	0.00	0.05	0.31	0.04	0.40
	780–850	0.00	0.00	0.03	0.10	0.12
Consistency	650–704	0.12	0.04	0.00	0.00	0.16
	705–740	0.04	0.22	0.07	0.00	0.32
	741–779	0.00	0.07	0.27	0.05	0.38
	780–850	0.00	0.00	0.05	0.09	0.14

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–709	0.14	0.02	0.00	0.00	0.16
	710–742	0.03	0.23	0.04	0.00	0.31
	743–784	0.00	0.05	0.33	0.03	0.41
	785–850	0.00	0.00	0.02	0.09	0.11
Consistency	650–709	0.13	0.04	0.00	0.00	0.17
	710–742	0.04	0.20	0.06	0.00	0.30
	743–784	0.00	0.06	0.29	0.04	0.40
	785–850	0.00	0.00	0.04	0.09	0.13

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–709	0.16	0.02	0.00	0.00	0.18
	710–744	0.03	0.19	0.04	0.00	0.26
	745–794	0.00	0.05	0.37	0.04	0.46
	795–850	0.00	0.00	0.03	0.07	0.10
Consistency	650–709	0.15	0.04	0.00	0.00	0.19
	710–744	0.04	0.16	0.06	0.00	0.26
	745–794	0.00	0.06	0.33	0.05	0.44
	795–850	0.00	0.00	0.05	0.07	0.11

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–704	0.21	0.03	0.00	0.00	0.25
	705–731	0.05	0.19	0.05	0.00	0.28
	732–780	0.00	0.05	0.31	0.03	0.38
	781–850	0.00	0.00	0.02	0.08	0.09
Consistency	650–704	0.20	0.05	0.00	0.00	0.26
	705–731	0.05	0.15	0.06	0.00	0.27
	732–780	0.00	0.06	0.28	0.03	0.37
	781–850	0.00	0.00	0.03	0.08	0.10

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–707	0.22	0.03	0.00	0.00	0.25
	708–739	0.04	0.27	0.04	0.00	0.34
	740–783	0.00	0.04	0.29	0.02	0.35
	784–850	0.00	0.00	0.01	0.04	0.05
Consistency	650–707	0.21	0.05	0.00	0.00	0.26
	708–739	0.05	0.23	0.06	0.00	0.33
	740–783	0.00	0.06	0.27	0.02	0.35
	784–850	0.00	0.00	0.02	0.04	0.06

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–708	0.26	0.04	0.00	0.00	0.30
	709–739	0.05	0.24	0.05	0.00	0.34
	740–781	0.00	0.05	0.24	0.02	0.31
	782–850	0.00	0.00	0.01	0.04	0.05
Consistency	650–708	0.24	0.06	0.00	0.00	0.31
	709–739	0.06	0.20	0.07	0.00	0.33
	740–781	0.00	0.07	0.21	0.02	0.30
	782–850	0.00	0.00	0.02	0.04	0.06

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–704	0.23	0.03	0.00	0.00	0.27
	705–741	0.04	0.33	0.04	0.00	0.41
	742–772	0.00	0.04	0.18	0.02	0.24
	773–850	0.00	0.00	0.01	0.07	0.08
Consistency	650–704	0.22	0.05	0.00	0.00	0.28
	705–741	0.05	0.29	0.06	0.00	0.40
	742–772	0.00	0.06	0.15	0.03	0.23
	773–850	0.00	0.00	0.03	0.06	0.09

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–711	0.19	0.03	0.00	0.00	0.22
	712–744	0.04	0.34	0.04	0.00	0.42
	745–780	0.00	0.04	0.24	0.02	0.30
	781–850	0.00	0.00	0.01	0.05	0.06
Consistency	650–711	0.18	0.05	0.00	0.00	0.23
	712–744	0.05	0.30	0.06	0.00	0.41
	745–780	0.00	0.06	0.22	0.02	0.29
	781–850	0.00	0.00	0.02	0.05	0.07

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

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	650–704	0.27	0.04	0.00	0.00	0.32
	705–744	0.05	0.25	0.05	0.00	0.35
	745–790	0.00	0.04	0.19	0.03	0.26
	791–850	0.00	0.00	0.01	0.06	0.07
Consistency	650–704	0.26	0.07	0.00	0.00	0.33
	705–744	0.07	0.21	0.06	0.00	0.33
	745–790	0.00	0.06	0.16	0.03	0.25
	791–850	0.00	0.00	0.03	0.06	0.08

Appendix G: Student Growth Percentile (SGP) Estimates by Subgroup**Table G.1. SGP Estimates by Subgroup—ELA/L Grade 4**

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	63,245	49.60	13.31	50
Female	61,178	50.10	13.12	50
African American	19,714	43.19	13.80	41
Asian/Pacific Islander	7,289	56.06	12.83	58
American Indian/Alaska Native	311	46.72	13.10	45
Hispanic	35,034	46.56	13.61	45
Middle Eastern	255	51.28	12.47	52
Multiple	5,875	51.55	13.12	53
White	55,961	53.28	12.83	55
Economically Disadvantaged	63,877	45.62	13.60	44
Not Economically Disadvantaged	60,562	54.31	12.82	56
English Learner (EL)	25,615	44.70	13.91	43
Non-EL	98,824	51.18	13.04	52
Students with Disabilities (SWD)	24,160	43.10	14.10	40
Students without Disabilities	100,279	51.47	13.01	52

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	63,460	47.86	14.69	47
Female	60,933	51.73	14.71	52
African American	19,610	46.52	14.65	45
Asian/Pacific Islander	7,208	56.89	14.72	60
American Indian/Alaska Native	296	52.03	14.68	54
Hispanic	34,780	49.37	14.62	49
Middle Eastern	216	56.88	14.11	61
Multiple	5,683	50.51	14.65	51
White	56,620	50.10	14.77	50
Economically Disadvantaged	63,395	47.70	14.59	47
Not Economically Disadvantaged	61,018	51.90	14.82	53
English Learner (EL)	19,593	48.15	14.52	47
Non-EL	104,820	50.06	14.73	50
Students with Disabilities (SWD)	24,631	43.36	14.71	40
Students without Disabilities	99,782	51.34	14.70	52

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	62,912	48.01	14.54	47
Female	60,310	51.58	14.38	52
African American	19,311	48.52	14.76	48
Asian/Pacific Islander	7,265	56.27	14.41	59
American Indian/Alaska Native	282	49.31	14.59	49
Hispanic	34,473	48.31	14.67	48
Middle Eastern	231	54.77	14.20	55
Multiple	5,534	49.56	14.35	49
White	56,158	50.23	14.26	50
Economically Disadvantaged	62,469	47.84	14.61	47
Not Economically Disadvantaged	60,785	51.73	14.31	52
English Learner (EL)	16,879	46.43	15.25	45
Non-EL	106,375	50.29	14.34	50
Students with Disabilities (SWD)	24,085	44.73	15.23	42
Students without Disabilities	99,169	50.98	14.28	51

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	63,720	46.45	14.19	45
Female	61,184	53.22	14.01	55
African American	19,339	48.47	14.23	48
Asian/Pacific Islander	7,420	55.28	14.27	57
American Indian/Alaska Native	251	47.71	14.19	46
Hispanic	35,628	48.27	14.17	48
Middle Eastern	236	54.70	13.90	56
Multiple	5,363	49.19	14.05	49
White	56,698	50.48	14.01	51
Economically Disadvantaged	62,245	47.78	14.12	47
Not Economically Disadvantaged	62,690	51.75	14.09	52
English Learner (EL)	18,398	46.92	14.37	45
Non-EL	106,537	50.26	14.06	50
Students with Disabilities (SWD)	23,668	45.55	14.66	44
Students without Disabilities	101,267	50.76	13.98	51

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	65,498	47.71	14.36	47
Female	62,515	51.82	14.67	53
African American	20,004	47.90	14.15	47
Asian/Pacific Islander	7,400	54.95	15.27	57
American Indian/Alaska Native	260	51.39	14.01	52
Hispanic	36,912	49.21	14.18	49
Middle Eastern	254	56.84	14.27	57.5
Multiple	5,282	49.41	14.47	49
White	57,944	50.00	14.76	50
Economically Disadvantaged	63,164	48.27	14.13	47
Not Economically Disadvantaged	64,892	51.13	14.89	52
English Learner (EL)	19,388	47.27	13.57	46
Non-EL	108,668	50.16	14.68	50
Students with Disabilities (SWD)	23,722	45.27	14.01	43
Students without Disabilities	104,334	50.73	14.63	51

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,104	51.27	12.99	52
Female	58,875	48.54	13.02	48
African American	19,614	42.01	13.72	39
Asian/Pacific Islander	7,278	58.69	12.95	62
American Indian/Alaska Native	289	50.73	13.22	54
Hispanic	30,849	48.23	13.18	48
Middle Eastern	255	54.84	12.56	56
Multiple	5,859	50.88	12.94	51
White	55,851	52.39	12.66	54
Economically Disadvantaged	60,100	46.22	13.26	44
Not Economically Disadvantaged	59,895	53.67	12.74	55
English Learner (EL)	21,425	48.82	13.25	48
Non-EL	98,570	50.18	12.95	50
Students with Disabilities (SWD)	23,648	45.83	13.58	44
Students without Disabilities	96,347	50.94	12.86	51
Spanish Language Form	4,225	46.70	13.55	45

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,704	49.94	14.76	50
Female	59,179	50.16	14.95	50
African American	19,570	45.30	15.97	43
Asian/Pacific Islander	7,205	59.78	14.08	64
American Indian/Alaska Native	259	47.79	15.10	46
Hispanic	31,458	49.88	15.30	50
Middle Eastern	215	59.67	13.48	64
Multiple	5,665	50.25	14.64	50
White	56,529	50.50	14.34	51
Economically Disadvantaged	60,457	47.84	15.41	47
Not Economically Disadvantaged	60,444	52.25	14.29	53
English Learner (EL)	16,348	50.62	15.81	51
Non-EL	104,553	49.96	14.71	50
Students with Disabilities (SWD)	24,285	48.42	15.65	48
Students without Disabilities	96,616	50.46	14.66	51
Spanish Language Form	3,318	48.16	16.24	47

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	61,385	50.06	14.95	50
Female	58,843	49.96	14.89	50
African American	19,243	46.73	16.28	45
Asian/Pacific Islander	7,257	58.08	14.37	62
American Indian/Alaska Native	267	52.48	14.95	56
Hispanic	31,691	49.39	15.35	49
Middle Eastern	231	52.46	13.97	53
Multiple	5,521	49.71	14.92	49
White	56,049	50.45	14.28	51
Economically Disadvantaged	59,981	48.07	15.55	47
Not Economically Disadvantaged	60,278	51.95	14.29	53
English Learner (EL)	14,313	46.80	16.45	45
Non-EL	105,946	50.45	14.71	51
Students with Disabilities (SWD)	23,773	46.18	16.47	44
Students without Disabilities	96,486	50.96	14.54	51
Spanish Language Form	2,779	44.70	16.67	43

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	62,487	50.61	15.44	51
Female	59,961	49.37	15.39	49
African American	19,242	46.93	16.51	46
Asian/Pacific Islander	7,394	55.35	15.11	57
American Indian/Alaska Native	241	50.54	15.31	48
Hispanic	33,443	50.49	15.63	50
Middle Eastern	236	54.85	15.35	55
Multiple	5,342	49.78	15.44	50
White	56,581	50.05	14.95	50
Economically Disadvantaged	60,192	48.69	15.84	48
Not Economically Disadvantaged	62,287	51.27	15.00	52
English Learner (EL)	16,238	49.17	16.43	48
Non-EL	106,241	50.13	15.26	50
Students with Disabilities (SWD)	23,427	45.81	16.84	44
Students without Disabilities	99,052	50.99	15.08	51
Spanish Language Form	2,088	48.36	17.35	48

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

Subgroup	Total Sample Size	Average SGP	Average Standard Error	Median SGP
Male	64,308	48.48	15.90	48
Female	61,466	51.45	15.91	52
African American	19,910	47.48	16.87	46
Asian/Pacific Islander	7,368	56.69	15.01	59
American Indian/Alaska Native	247	49.30	16.97	50
Hispanic	34,972	49.84	16.31	50
Middle Eastern	254	58.27	14.80	61
Multiple	5,269	49.06	15.87	49
White	57,797	50.01	15.45	50
Economically Disadvantaged	61,290	48.39	16.44	48
Not Economically Disadvantaged	64,527	51.39	15.40	52
English Learner (EL)	17,450	48.94	17.04	48
Non-EL	108,367	50.09	15.72	50
Students with Disabilities (SWD)	23,477	44.67	17.09	42
Students without Disabilities	102,340	51.14	15.63	52
Spanish Language Form	1,877	47.03	17.40	44