



**2019 LEAP 2025 Grades 3-8  
Operational Technical Report  
English Language Arts and Mathematics**

Submitted to the  
Louisiana Department of Education

February 2020



This online-only document was published at a cost of \$33,533. This document was published for the Louisiana Department of Education, P.O. Box 94064, Baton Rouge, LA 70804-9064, by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN 55311. This material was printed in accordance with the standards for printing by State Agencies established pursuant to R.S. 43:31.

## Table of Contents

---

Executive Summary.....	7
E.1..... Overview of This Report	7
E.2..... Administration	8
E.3..... Student Performance	9
E.4..... Validity and Test Scores	10
Chapter 1: Introduction .....	11
1.1 Background.....	11
1.2 Purpose of the LEAP 2025 .....	11
1.3 Design of the LEAP 2025.....	12
Chapter 2: The Uses of Test Scores.....	15
2.1 Uses of Test Scores.....	15
2.2 Test-Level Scores .....	15
2.2.1 Scale Scores.....	16
2.2.2 Levels of Achievement .....	16
2.2.3 Use of Test-Level Scores .....	16
2.3 Category- and Subcategory-Level Subscores .....	16
2.3.1 Use of the Reporting Category- and Subcategory- Level Ratings .....	17
Chapter 3: Test Content Development .....	18
3.1 Defining the Specific Test Blueprint .....	20
3.2 English Language Arts Test Blueprints and Test Designs.....	20
3.3 Mathematics Test Blueprints and Test Designs .....	30
3.4 Item Development and Selection .....	43
3.5 Considerations of Test Fairness in Item Development.....	43
3.6 PARCC Item Reviews.....	43
3.7 Louisiana Item Development and Item Review.....	44
3.7.1 ELA Development Process.....	44
3.7.2 Text Complexity Specifications for Field Test Passages .....	44
3.7.3 Passage Review .....	45
3.7.4 ELA Item Writing and Review.....	45

3.8 Mathematics Item Development .....	46
3.9 Guidelines on Bias, Fairness, and Sensitivity .....	47
3.9.1 Louisiana Item Alignment Review .....	48
3.10 Operational Test Selection .....	49
3.10.1 General Item and Passage Set Selection Process and Criteria .....	49
3.10.2 Review of the English Language Arts Items and Forms .....	50
3.10.3 Item-Selection Options for Special Cases.....	51
3.10.4 Psychometric Review .....	51
3.11 Universal Design .....	58
3.12 Accommodations and Designated Supports .....	59
3.13 Item and Task Specifications .....	60
3.14 Summary .....	61
Chapter 4: Test Administration .....	63
4.1 Return Material Forms and Guidelines .....	69
4.2 Security Checklists .....	69
4.3 Interpretive Guides.....	72
4.4 Test Security Measures .....	72
4.4.1 Data Forensic Analyses .....	72
4.4.2 Response Change Analysis .....	72
4.4.3 Score Fluctuation Analysis .....	72
4.4.4 Web Monitoring.....	73
4.4.5 Plagiarism Detection .....	73
4.5 Test Administration .....	73
4.5.1 Time .....	73
4.5.2 Accommodations .....	74
4.6 Summary .....	79
Chapter 5: Scoring of Constructed-Response and Technology-Enhanced Items.....	81
5.1 Constructed-Response Item Scoring Process .....	81
5.1.1 Selection of Scoring Evaluators .....	82
5.1.2 Security .....	82
5.1.3 Handscoring Training Process .....	83
5.1.4 Monitoring the Scoring Process .....	87
5.2 Inter-Rater Reliability .....	89
5.3 Technology-Enhanced Item Scoring Process.....	96

5.4 Multiple-Choice and Multiple-Select Item Scoring Process .....	96
5.5 Summary .....	96
Chapter 6: Operational Data Analyses .....	98
6.1 Test-Level Statistics .....	98
6.2 Item-Level Statistics.....	100
6.3 Item Response Theory.....	125
6.4 Calibration and Linking.....	131
6.4.1 Calibration of the 2019 LEAP 2025 Tests .....	131
6.4.2 Linking 2019 LEAP 2025 Grades 3–8 to PARCC Scale .....	139
6.5 Summary .....	187
Chapter 7: Test Results .....	188
7.1 Current Administration Data .....	195
7.1.1 Description of Each Type of Report .....	198
Chapter 8: Performance-Level Setting .....	200
8.1 PARCC Performance-Level Setting Process for English Language Arts and Mathematics .....	200
8.2 Cut Scores .....	200
8.2.1 Reporting Category Cut Scores .....	201
8.3 Summary .....	202
Chapter 9: Evidence of Validity .....	203
9.1 Construct-Irrelevant Variance and Construct Underrepresentation .....	204
9.2 Reliability .....	204
9.2.1 Test Reliability .....	205
9.2.2 Standard Error of Measurement.....	206
9.2.3 Conditional Standard Error of Measurement .....	207
9.2.4 Classification Accuracy and Consistency.....	211
9.2.1 Convergent Validity.....	214
9.3 Principal Components Analysis .....	214
9.4 Analyses by Reporting Categories and Subcategories .....	217
9.4.1 Correlations among Reporting Categories and Subcategories .....	217
9.4.2 Reliability of Reporting Categories and Subcategories .....	222
9.4.3 Standard Error of Measurement of Reporting Categories and Subcategories .....	223
9.5 Divergent (Discriminant) Validity .....	227
9.6 Regression of LEAP 2025 from 2018 to 2019 .....	227
9.7 Summary .....	231

Chapter 10: Fairness .....	232
10.1 Minimizing Bias through Careful Test Development.....	233
10.2 Evaluating Bias through Differential Item Functioning (DIF) Statistics .....	233
10.2.1 DIF Statistics for Demographic Groups.....	235
10.2.2 DIF Statistics for Test Language.....	241
10.3 Evaluating Bias through Impact Analysis.....	241
10.3.1 Reliability.....	242
10.3.2 Effect Size .....	250
10.4 Mode Effect Study .....	265
10.4.1 Sampling Using Propensity Score Matching.....	265
10.5 Summary .....	267
Appendix A—Text Complexity Placemat Template .....	269
Appendix B—Item Content and Bias Review .....	270
Appendix C—Item Alignment Review Process.....	275
Appendix D—Accommodated Print Form Creation.....	280
Appendix E—Transadaptation Process for Spanish Mathematics Forms.....	282
Appendix F—LEAP 2025 Spring 2018 Handscoring/AI Documentation.....	284
References .....	365

## Executive Summary

---

This report is a technical summary of the 2019 administration of the Louisiana Educational Assessment Program (LEAP 2025) in English Language Arts (ELA) and mathematics for grades 3 through 8. The LEAP 2025 is a summative assessment in ELA and mathematics administered in grades 3 through 8 and high school. These tests are designed to measure students' readiness for the next grade or course of study and proficiency in ELA and mathematics. The ELA and mathematics test forms were developed by Data Recognition Corporation (DRC) test development staff using the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium's item bank as well as items from the Louisiana Department of Education's own item bank. Items taken from these banks were on pre-established item response theory (IRT) scales. This section provides a summary of the 2019 operational technical report.

### E.1 Overview of This Report

This technical report documents the major activities of the testing cycle and provides details that confirm that the processes and procedures applied in the LEAP 2025 assessments adhered to appropriate professional standards and practices of educational assessment. Ultimately, this report serves to document evidence that valid inferences about Louisiana student performance in ELA and mathematics can be derived from the LEAP 2025 assessments. An overview of major activities documented within this report is provided below.

#### *The Uses of Test Scores (Chapter 2)*

Chapter 2 of the technical report discusses the concept of validity evidence. This technical report is composed of evidence that supports the intended uses of the LEAP 2025 test scores, and Chapter 2 discusses some of those uses.

#### *Test Content Development (Chapter 3)*

Chapter 3 of the technical report provides a summary of the test development activities that occurred in order to create the Spring 2019 operational test forms.

#### *Test Administration (Chapter 4)*

Chapter 4 of the technical report describes the processes implemented and the information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students.

#### *Constructed-Response and Technology-Enhanced Scoring (Chapter 5)*

Chapter 5 of the technical report describes the processes used to score constructed-response and technology-enhanced items. This chapter discusses how scorers are trained and the measures used to ensure consistency among scorers. Finally, this chapter presents the results of the inter-rater reliability studies.

#### *Operational Data Analyses (Chapter 6)*

Chapter 6 of the technical report includes a detailed description of the operational data analyses of the 2019 LEAP 2025 assessments, which include the following major parts: the classical item analysis; calibration, scaling, and linking using IRT models; and student scoring. This chapter also describes the demographics of the calibration samples and compares them to state census data. It reports the results of the classical item analysis and the results of the calibration, scaling, and linking processes.

### ***Test Results (Chapter 7)***

Chapter 7 of the technical report contains information on the results of the Spring 2019 LEAP 2025 assessments. Detailed summary statistics based on scale scores and information about achievement-levels are also provided. Finally, this chapter presents information on the score reports sent to school systems.

### ***Performance-Level Setting (Chapter 8)***

Chapter 8 of the technical report briefly discusses performance-level setting. It provides a brief overview of the PARCC procedures for performance-level setting and derivation of the cut scores used to classify students into achievement levels for ELA and mathematics.

### ***Evidence of Construct-Related Reliability (Chapter 9)***

Chapter 9 of the technical report provides evidence of the reliability and validity of the LEAP 2025 test scores. This chapter provides detailed evidence of the reliability of the tests and information on the decision consistency of the cut scores. It also provides evidence of construct validity for the LEAP 2025 test scores.

### ***Fairness (Chapter 10)***

Chapter 10 of the technical report discusses fairness and how the LEAP 2025 assessments are constructed to be fair to all Louisiana students. This chapter summarizes the results of the differential item functioning (DIF) analysis. It also discusses the results of an impact analysis designed to determine whether large differences exist with the test results of different demographic groups in Louisiana. The results of the administration mode study are also summarized.

## **E.2 Administration**

In the spring of 2019, Louisiana administered the LEAP 2025 summative assessments in ELA and mathematics to students in grades 3–8. A paper-based test (PBT) and a computer-based test (CBT) were administered in grades 3–4, and a CBT was administered in grades 5–8. The CBTs were administered from April 1 to May 3, 2019. The PBTs were administered from April 29 to May 3, 2019. Test administration is discussed in Chapter 4 of this report.

A total of 104 school systems and 36 charter schools administered the ELA and mathematics LEAP 2025 tests in grades 3–8. Table E.1 shows participation rates based on census data. For the purposes of this report, participation rate is defined as the percentage of students who earned a valid scale score given the total number of students who were expected to take the test. The “Accountable” column shows the total number of students who were expected to take the test by grade and content area. The “Percentage Reportable” column shows the percentage of students who received a scale score on the LEAP 2025 by grade and content area. Further analysis of participation rates is provided in Chapter 7 of this report. The results presented in Table E.1 and Chapter 7 are presented as evidence of reliability and validity of the scores from the LEAP 2025 assessments and should not be used for state accountability purposes.



**Table E.1 Participation Rates: All Students Participating in LEAP 2025 Grades 3-8**

Grade	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics*
3	≥52,910	99.86%	≥53,440	99.86%
4	≥54,700	99.84%	≥55,170	99.85%
5	≥54,780	99.92%	≥54,790	99.92%
6	≥54,800	99.84%	≥54,800	99.85%
7	≥52,290	99.80%	≥52,300	99.81%
8	≥50,780	99.66%	≥50,800	99.70%

\*Students in grade 8 who were enrolled in Algebra I had the option of taking the LEAP 2025 Algebra I assessment instead of the LEAP 2025 Grade 8 Mathematics test.

### E.3 Student Performance

Tables E.2 and E.3 present the percentage of students who were classified in each of the 2019 achievement levels for ELA and mathematics.

**Table E.2 Percentage of Students Classified in 2019 Achievement Levels Using 2019 Census Data: English Language Arts**

Grade	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
3	13.2	17.2	23.7	39.5	6.4
4	10.3	18.1	26.6	36.1	8.9
5	8.4	21.1	30.0	36.0	4.4
6	9.2	23.5	29.8	32.2	5.3
7	11.6	16.7	25.1	33.0	13.7
8	11.7	16.2	25.4	37.6	9.2

**Table E.3 Percentage of Students Classified in 2019 Achievement Levels Using 2019 Census Data: Mathematics**

Grade	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
3	9.7	20.6	26.4	36.5	6.7
4	11.1	20.5	27.1	38.0	3.3
5	10.3	26.8	28.3	30.5	4.1
6	11.4	26.7	31.7	26.6	3.6
7	9.1	29.5	34.7	24.5	2.3
8	20.9	25.7	25.4	25.7	2.3

More information on student performance may be found in Chapter 7 of this report.

#### E.4 Validity and Test Scores

Most sections of this technical report are designed to provide validity evidence to support the intended uses of the LEAP 2025 test scores. Chapter 2 discusses the intended uses of the LEAP 2025 test scores. Chapter 3 discusses the test development process used to create the LEAP 2025 tests, which is important to the content-related validity of the LEAP 2025 test scores. Chapter 4 presents information on test administration. Chapter 5 discusses the scoring process and the results of the inter-rater reliability studies. Chapter 6 presents the test scaling and linking procedures, student scoring methodology, and the results of other operational data analyses. Chapter 7 reviews the results of the 2019 administration and gives an overview of the score reports that were electronically delivered to the school systems for distribution to schools and parents. Chapter 8 highlights the procedures for performance-level setting implemented by PARCC, which were used because PARCC's standards and achievement levels were used for the LEAP 2025. Chapter 9 discusses reliability and construct-related validity. Chapter 10 gives an overview of the statistical processes used to evaluate bias to ensure fairness of the LEAP 2025 for all examinees.

## Chapter 1: Introduction

---

The LEAP 2025 assessment system is designed to measure students' knowledge of ELA, mathematics, science, and social studies. This report provides a technical overview of the LEAP 2025 ELA and mathematics assessments administered in grades 3 through 8 in the spring of 2019 and presents evidence for the validity of the 2019 LEAP 2025 ELA and mathematics assessment scores.

This chapter describes the background, purpose, and design of the LEAP 2025.

### 1.1 Background

In 2010, the Board of Elementary and Secondary Education (BESE) approved the Common Core State Standards (CCSS) in ELA and mathematics. After adopting the CCSS, Louisiana became a governing member of PARCC, a group of states working to develop high-quality assessments that measure the full range of the CCSS.

To prepare for the PARCC assessments and help ease the transition to the new standards, the Louisiana Department of Education (LDOE) incrementally revised the LEAP and *i*LEAP ELA and mathematics assessments in grades 3 through 8 and administered transitional tests during the 2012–2013 and 2013–2014 school years.

In the 2014–2015 school year, students in grades 3–8, except those qualifying for the LEAP Alternate Assessment, Level 1 (LAA 1), took the PARCC assessments for ELA and mathematics, which included two components: the performance-based assessment (PBA), which was administered in March, and the end-of-year assessment (EOY), which was administered in May.

As a result of a legislative agreement reached during the summer of 2015, and to maintain comparability to the 2015 assessments, the LEAP ELA and mathematics assessments in grades 3–8 for the 2015–2016 school year consisted of items taken from both the PARCC assessments (no more than 49.9%) and DRC's College and Career Readiness item bank.

In March 2016, BESE approved the Louisiana Student Standards in ELA and mathematics. In the 2016–2017, 2017–2018, and 2018–2019 school years, students in grades 3–8, except those qualifying for an alternate assessment for students with the most significant cognitive disabilities (the LAA 1 in 2016–2017 or LEAP Connect in 2017–2018 and 2018–2019), were administered forms for ELA and mathematics that consisted of PARCC assessment items while developing some Louisiana-owned items to enhance the PARCC item bank. This allowed for the continued comparability to forms administered in the 2014–2015 and 2015–2016 school years.

The information that follows describes the technical aspects of the 2019 LEAP 2025 ELA and mathematics assessments and provides information about how to read and interpret the data.

### 1.2 Purpose of the LEAP 2025

The BESE and the LDOE are committed to ensuring that every student is on track to be successful in either postsecondary education or the workforce through their comprehensive plan Louisiana Believes. The LEAP 2025 supports this vision by measuring the full range of student performance, including the performance of high- and low-performing students and providing information for educators and parents about student readiness for college and careers.

### 1.3 Design of the LEAP 2025

Students in grades 3–8 were administered computer-based tests (CBTs) in both ELA and mathematics; some school systems opted to administer paper-based tests (PBTs) to students in grades 3 and 4. All mathematics assessments were translated into Spanish forms. Additionally, a braille form was available for each grade and content area. The braille form was based on the PBT in grades 3 and 4 and was based on the CBT in grades 5–8. Online tools allowed students to magnify assessment items, as needed, and students with visual impairments could also take large-print versions of the PBTs. See Chapter 3, Section 3.4 for more information about the accommodations and designated supports available for students taking the LEAP 2025.

The operational blueprints for the PARCC flagship form are the basis of the design of the 2019 LEAP 2025 test blueprints and test design. The 2019 LEAP 2025 test blueprints and test design for ELA and mathematics differ from the PARCC blueprints and design in order to reduce testing time while maintaining full coverage and including a variety of standards.

The 2019 LEAP 2025 ELA blueprints kept a similar design as the design of PARCC’s flagship form, which includes both performance-based tasks and stand-alone passage sets, and a higher percentage of reading points to writing points. However, only two of the three types of performance tasks—Research Simulation Task (RST) and Literary Analysis Task (LAT) or Narrative Writing Task (NWT)—are included on each of the grade-level tests. All three task types are represented across grades 3–8, which allows Louisiana flexibility in the choice of the tasks administered for each grade from year to year and encourages teachers to focus equally on all three writing types. Besides having two (instead of three) performance tasks, the 2019 LEAP 2025 Spring ELA blueprints are also different from the PARCC blueprints with respect to testing time and percentage of reading and writing points. Since the choice of Literary Analysis Task or Narrative Writing Task is determined during the forms construction process, alternative blueprints—one with a Literary Analysis Task and a Research Simulation Task and the other with a Research Simulation Task and a Narrative Writing Task—were created for each grade’s assessment.

The passages chosen for the 2019 LEAP 2025 ELA assessments contain a variety of text types, including texts that diverse populations will find engaging and that have a balance of gender and ethnicity among authors. Chosen passages are authentic, contain a variety of different genres and varying degrees of text complexity, and are content-rich, engaging, high-quality, and challenging. Additionally, paired passages are selected with careful consideration of the purpose of the standards that require the use of more than one text to be assessed. This combination of criteria during passage selection allows students to demonstrate their ability to read and comprehend a range of complex texts. With respect to an overall passage set and form, the goal is to ensure as much coverage of standards as possible.

The LEAP 2025 ELA assessments focus on an integrated approach to reading and writing that reflects instruction in an effective ELA classroom and measures students’ ability to understand what they read and express that understanding in writing. This means careful, close reading of complex grade-level literary and informational texts; a full range of texts from across the disciplines, including science, social studies, and the arts; tasks that integrate key ELA skills by asking students to read texts, answer reading and vocabulary questions about the texts, and then write using evidence from what they have read; questions worth answering, ordered in a way that builds meaning; a focus on students citing evidence from texts when answering questions about a specific passage or when writing about a set of related passages; and a focus on words that matter most in texts, are essential to understanding a particular text, and include context that allows students to determine literal and figurative meanings.

In mathematics, the test blueprints are similar to those of the flagship PARCC test design with a few notable exceptions:

- In grades 3-5, the LEAP 2025 blueprints make use of three sessions with a total testing time of 235 minutes, instead of four sessions with a total testing time of 240 minutes.
  - In grade 3, the difference in items is a reduction of 1 Type II item worth 4 points and an increase of 2 Type I items worth 1 point with a corresponding decrease of 1 Type I item worth 2 points. Therefore, the total number of items is the same across both designs, but LEAP 2025 has 4 fewer points.
  - In grades 4 and 5, there is a bigger difference, as LEAP 2025 uses the same test design for grades 3-5, so the increase in type I 1-point items is 8 with a decrease in 4 2-point items in addition to the reduction of 1 Type II item worth 4 points.
- In grades 6-8, both LEAP 2025 and PARCC have three sessions and a total testing time of 240 minutes. However, PARCC uses three sessions of equal testing time with 80 minutes each, while LEAP 2025 has a shorter non-calculator session 1 (60 minutes) followed by two 90-minute calculator sections. PARCC has a split session in grade 7 mathematics for session 1 in which the non-calculator and calculator sections are split within the same session/unit. In grades 6 and 8, the entire first session/unit is designated as non-calculator. The LEAP 2025 test design has consistency across grades 6-8 in testing time per session and has either non-calculator or calculator as the designation for the entire session for ease of administration.
  - In grades 6 and 7, the LEAP 2025 design uses 8 more type I items worth 1 point, 2 fewer type I items worth 2 points, and 1 fewer type I item worth 4 points. (LEAP 2025 does not use any type I items worth 4 points.) Grades 6-8 use the same number of type II and III items in both PARCC and LEAP 2025 test designs.
  - LEAP 2025 uses the same test design for grade 8, so there are 8 more type I items worth 1 point and 2 fewer type I items worth 4 points (but the same number of type I items worth 2 points).

The LEAP 2025 mathematics assessments focus on testing the Louisiana Student Standards for Mathematics (LSSM) according to the components of rigor reflected in high-quality mathematics instructional tasks that

- require students to demonstrate understanding of mathematical reasoning in mathematical and applied contexts;
- assess accurate, efficient, and flexible application of procedures and algorithms;
- rely on application of procedural skill and fluency to solve complex problems; and
- require students to demonstrate mathematical reasoning and modeling in real-world contexts.

The LSSM support students to become mathematically proficient by focusing on three components of rigor: conceptual understanding, procedural skill and fluency, and application.

- Conceptual understanding refers to understanding mathematical concepts, operations, and relations. It is more than knowing isolated facts and methods. Students should be able to make sense of why a mathematical idea is important and the kinds of contexts in which it is useful. It also allows students to connect prior knowledge to new ideas and concepts.
- Procedural skill and fluency is the ability to apply procedures accurately, efficiently, and flexibly. It requires speed and accuracy in calculation while giving students opportunities to practice basic skills. Students' ability to solve more complex application tasks is dependent on procedural skill and fluency.
- Application provides a valuable context for learning and the opportunity to solve problems in a relevant and a meaningful way. It is through real-world application that students learn to select an efficient method to find a solution, determine whether the solution(s) makes sense by reasoning, and develop critical thinking skills.

Each item on the LEAP 2025 mathematics assessment is referred to as a task and is identified by one of three types: Type I, Type II, or Type III. The tasks on the LEAP 2025 mathematics test are aligned directly to the LSSM for all reporting categories.

- **Type I** tasks, designed to assess conceptual understanding, fluency, and application, are aligned to the major, additional, and supporting content for each grade. Some Type I tasks may be further aligned to LEAP 2025 evidence statements for the Major Content and Additional & Supporting reporting categories and allow for the testing of more than one of the student standards on a single task.
- **Type II** tasks are designed to assess student reasoning ability of selected major content for the grade or the previous grade in applied contexts.
- **Type III** tasks are designed to assess student modeling ability of selected content for the grade or the previous grade in applied contexts. Type II and III tasks are further aligned to LEAP 2025 evidence statements for the Expressing Mathematical Reasoning and Modeling & Application reporting categories.

Each of the three task types is aligned to one of four reporting categories: Major Content, Additional & Supporting Content, Expressing Mathematical Reasoning, or Modeling & Application. Each task type is designed to align with at least one of the Louisiana Student Standards for Mathematical Practice (MP).

Additional details about the design of the ELA and mathematics assessments can be found in Chapter 3.

## Chapter 2: The Uses of Test Scores

---

Validity is the central component of any analysis of the LEAP 2025 assessments. The following excerpt is from the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014):

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (22)

As stated by the *Standards*, the validity of a testing program hinges on the use of the test scores. Validity evidence that supports the uses of the LEAP 2025 test scores is provided in this technical report. This chapter examines some possible uses of the LEAP 2025 test scores. However, this technical report cannot anticipate all possible interpretations and uses of the LEAP 2025 test scores.

### 2.1 Uses of Test Scores

To understand whether a test score is being used properly, one must understand the purpose of the test. The intended uses of the LEAP 2025 test scores include the following:

- evaluating students' overall proficiency of the Louisiana Student Standards
- identifying students' strengths and weaknesses
- evaluating programs at the school, school system, and/or state level
- informing stakeholders, including students, teachers, school administrators, school system administrators, LDOE staff members, parents, and the public, of the status of students' progress toward meeting college and career readiness standards

This technical report refers to the uses of the test-level scores (i.e., scale scores and achievement levels), category-level scores and achievement-level classifications, and subcategory-level scores and achievement-level classifications.

### 2.2 Test-Level Scores

At the test level, an overall scale score that is based on student performance on the entire test is reported. In addition, an associated level of achievement is reported. These scores and achievement levels indicate, in varying ways, a student's achievement in ELA or mathematics. Test-level scores are reported at four reporting levels: the state, the school system, the school, and the student.

The LEAP 2025 high school ELA and mathematics test forms were developed by DRC's test development staff using the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium's item bank as well as items from the Louisiana Department of Education's own item bank. Items taken from these banks were on pre-established item response theory (IRT) scales for ELA and mathematics and were reviewed and approved for use by LDOE content experts and committees of Louisiana educators. Braille forms and Spanish translations of mathematics forms were also developed. See Chapter 3, "Test Content Development," for additional details about the processes used to develop these test forms.

The following sections discuss two types of test-level scores that are reported that indicate a student's achievement on the LEAP 2025 assessments: the scale score and its associated level of achievement.

### 2.2.1 Scale Scores

A scale score indicates a student’s total performance for each content area on the LEAP 2025 assessments. The overall scale score for a content area quantifies the achievement being measured by the ELA or mathematics assessments. In other words, the scale score represents the student’s level of achievement, where higher scale scores indicate higher levels of achievement on the test and lower scale scores indicate lower levels of achievement. For all LEAP 2025 test forms, the lowest obtainable scale score (LOSS) is 650 and the highest obtainable scale score (HOSS) is 850.

Scale scores are derived from raw scores (i.e., the number of items answered correctly). Raw scores depend on the items in a particular form of a test and can only be interpreted in terms of that particular set of test questions. This does not allow year-to-year or form-to-form comparison. Scale scores are more meaningful than raw scores because they maintain their meaning year-to-year, thus allowing comparisons of different test forms across the entire range of the ability scale.

### 2.2.2 Levels of Achievement

A student’s performance on the ELA or mathematics LEAP 2025 assessments is reported in one of five levels of achievement: *Advanced*, *Mastery*, *Basic*, *Approaching Basic*, or *Unsatisfactory*. The cut scores for the ELA and mathematics achievement levels were established by PARCC using the Evidence-Based Standard Setting (EBSS) method (Beimers, Way, McClarty, & Miles, 2012) for the PARCC Performance-Level Setting (PLS) process. Details regarding the PLS process can be found in the [Performance Level Setting Technical Report](#) (Pearson, 2015).

Descriptions of each level of achievement in terms of what a student should know and be able to do are provided with the LEAP *Interpretive Guide* (see Chapter 7).

### 2.2.3 Use of Test-Level Scores

The LEAP 2025 scale scores and achievement levels provide summary evidence of student performance in ELA or mathematics relative to the Louisiana Student Standards. Classroom teachers may use these scores as evidence of student achievement in these content areas. At the aggregate level, school system and school administrators may use this information for activities such as curriculum planning. The results presented in this technical report provide evidence that the scale scores and achievement levels are valid and reliable indicators of what students know, understand, and are able to do relative to the Louisiana Student Standards in ELA and mathematics.

## 2.3 Category- and Subcategory-Level Subscores

A student’s performance on the ELA categories (i.e., reading and writing) is reported by one of three ratings: *Strong*, *Moderate*, or *Weak*. Additionally, performance on the subcategories is reported at the student level for ELA and mathematics. ELA has three subcategories for reading and two subcategories for writing, as described in Table 3.1, *ELA Categories and Subcategories*. Mathematics has four subcategories, as described in Table 3.8, *Overview of LEAP 2025 Mathematics Task Types and Reporting Categories*. Subcategory performance is reported in one of three ratings: *Strong*, *Moderate*, or *Weak*.

Although the performance ratings are determined only by the items included within a category or subcategory, the level of knowledge and ability needed to demonstrate a performance rating is connected to the level of knowledge and ability required by the content-level assessments; a *Strong* rating requires similar knowledge and ability as the Mastery or Advanced achievement levels, a *Moderate* rating requires similar knowledge and ability as the Basic achievement level, and a *Weak* rating requires similar knowledge and ability as the Unsatisfactory and Approaching Basic achievement levels.



### 2.3.1 Use of the Reporting Category- and Subcategory-Level Ratings

The purpose of reporting category- or subcategory-level performance ratings on LEAP 2025 assessments is to show, for each student, the relationship between the overall achievement being measured and the skills in each of the areas defined by the categories and subcategories. Teachers may use these ratings for individual students as indicators of strengths and weaknesses, but they are best corroborated by other evidence, such as grades, teacher feedback, and scores on other tests. Chapter 3 of this technical report provides evidence of content validity that supports the use of the category- or subcategory-level performance ratings. Chapter 9 of this technical report provides evidence of construct-related validity that further supports the use of these performance ratings.

## Chapter 3: Test Content Development

---

Content-related validity in achievement tests is evidenced by a correspondence between test content and the range of knowledge and skills that compose the construct the assessment is designed to measure, i.e., the ELA or mathematics Louisiana Student Standards. Content-related validity can be demonstrated through consistent adherence to test blueprints, through a high-quality test development process that includes review of items for accessibility to English learners and students with disabilities, and through alignment studies performed by independent groups. This chapter provides a detailed discussion of the test development process. In particular, it shows how rigorous procedures were followed to construct tests that reflect the full range of content that the 2019 LEAP 2025 assessments were expected to cover.

This chapter is particularly relevant to the following sections of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014): Standards 4.0, 4.1, and 4.7. It also addresses Standards 3.1, 3.2, 3.9, and 4.12, which are discussed in pertinent sections of this chapter.

Standard 4.0 states the following:

Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (85)

Standard 4.1 states the following:

Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (85)

The 2019 LEAP 2025 test specifications consisted of a test blueprint and a test design for each grade and content area. The 2019 blueprints and test designs were closely aligned to the PARCC flagship blueprint that was used for their spring 2019 test administrations. The specific content area and grade-level test blueprints for the 2019 LEAP 2025 ELA assessments for grades 3–8 were designed with the goal for all students to read, understand, and express understanding of complex, grade-level texts. The specific content area and grade-level test blueprints for the 2019 LEAP 2025 mathematics assessments for grades 3–8 were designed with the goal of supporting students to become mathematically proficient by focusing on three components of rigor: conceptual understanding, procedural skill and fluency, and application. The 2019 LEAP 2025 ELA and mathematics assessments for grades 3–8 provide questions that have been reviewed by Louisiana educators to ensure their alignment to the Louisiana Student Standards and appropriateness for Louisiana students; measure the full range of student performance, including the performance of high- and low-performing students; and inform educators and parents about student readiness in ELA and mathematics and whether students are “on track” for college and careers. For ELA and mathematics, the 2019 LEAP 2025 assessments for grades 3–8 use the same reporting categories that were used in spring 2018. Subcategories in mathematics were introduced for spring 2018 in response to requests from school systems. In ELA, the type and/or number of reading literary and informational passage sets changed from the 2017 LEAP 2025 assessments to the 2018 LEAP 2025 ELA assessments to reflect a similar change made in the PARCC blueprints. This change was continued for the 2019 LEAP 2025 ELA assessments.

To construct the assessments after the test blueprints and test designs were approved, the LDOE and DRC collaborated to use items, aligned to the Louisiana Student Standards, from the PARCC and the Louisiana-owned item banks. DRC contracted with PARCC and was provided access to the entire bank of items and passage sets that could potentially be used on operational forms. These and Louisiana-owned items and passage sets make up the available item pool for the 2019 LEAP 2025 forms construction. Please refer to the [PARCC Model Content Frameworks for ELA/Literacy](#) (Grades 3-11) and [PARCC Model Content Frameworks for Mathematics](#) (Grades 3-11) for additional information about the development of item specifications and blueprints for the PARCC flagship assessments. These resources can be accessed via the [Research](#) page of [New Meridian's website](#). The LDOE and DRC confirmed that all items selected for use on the LEAP 2025 forms were appropriate for use on Louisiana assessments by convening committees of Louisiana educators who reviewed and approved items from the item banks prior to form selection.

The ELA and mathematics LEAP 2025 assessments for grades 3–8 were developed based on the requirements of “RFP #678PUR-LEAP 2025 English Language Arts and Mathematics Assessment System” as follows:

The assessments shall be

- aligned to the ELA and mathematics Louisiana Student Standards;
- designed to be accessible for use by the widest possible range of students, including, but not limited to, students with disabilities and students with limited English proficiency;
- constructed to yield valid and reliable test results;
- constructed to report student performance using achievement level policy definitions and reporting categories that are comparable to a significant number of other states and, for grades 3 through 8 assessments, to Louisiana’s 2015–2018 assessments;
- constructed to use Louisiana’s grades 3 through 8 ELA and mathematics assessments as the baseline scale<sup>1</sup> to report test results for grades 3 through 8 students;
- developed to limit the amount of testing time required and to be in compliance with state law regarding testing time;
- developed and reviewed with Louisiana educators;
- non-computer adaptive;
- used in assessing students’ readiness to successfully transition to postsecondary education and the workplace; and
- administered, scored, and reported through a separate administration contract in both paper- and computer-based formats.

The products of the above requirements are dual-mode assessments—paper-based tests (PBTs) and computer-based tests (CBTs)—comprised of PARCC and Louisiana-owned items aligned to the Louisiana Student Standards. Louisiana had access to the complete PARCC item bank when selecting items to build the forms needed for the 2019 LEAP 2025 ELA and mathematics assessments. For grades 3 and 4, the contract with New Meridian provided for the use of enough PARCC items and passage sets, which had been approved during Item Alignment Reviews, combined with additional items and passage sets developed specifically for Louisiana, to create one complete operational test form for each content area and grade that can be administered in a dual-mode testing environment (i.e., PBT and CBT). For grades 5–8, Louisiana selected one CBT form per grade from the content that was reviewed during Item Alignment Reviews in addition to items

---

<sup>1</sup> In the spring of 2016 and 2017, PARCC item parameters were used to place the LEAP 2025 assessments on the PARCC scale. In the spring of 2018, PARCC items that had been previously administered in Louisiana were available, so the item parameters generated from Louisiana students were used to create the LEAP 2025 scale. The LEAP 2025 scale is comparable to the PARCC scale. Future LEAP 2025 assessments will be linked to the Spring 2018 LEAP 2025 scale, which is considered the baseline.

and passage sets developed specifically for Louisiana. These items and passage sets became the available item pool used to construct the 2019 forms. DRC and LDOE content experts scrutinized each final blueprint to ensure optimal content coverage and prudent use of time and resources. In general, the blueprints represent content sampling proportions that reflect intended emphasis in instruction and mastery at each grade level and are comparable to PARCC 2019 flagship test blueprints. The test specifications provide the numbers of items by reporting category, assessment focus, or item type, and they demonstrate the desired proportions within test delivery and available item pool constraints. These specifications can be found in the *2018–2019 LEAP 2025 Grades 3-8 English Language Arts and Mathematics Assessment Frameworks*. All assessments were fixed forms, which means that all students who received the same form were administered the same set of items, as the forms were not adaptive.

### 3.1 Defining the Specific Test Blueprint

The specific content area and grade-level test blueprints were designed based on two primary factors: (1) the content requirements of the Louisiana Student Standards and (2) the reporting needs of the assessments.

### 3.2 English Language Arts Test Blueprints and Test Designs

The ELA test was administered during a CBT testing window (April 1-May 3, 2019) and during a PBT testing window (April 29-May 3, 2019). Only two of the three types of performance tasks—Research Simulation Task, Literary Analysis Task, and Narrative Writing Task—were included on each of the Louisiana grade-level tests; however, all three types were represented across grades 3 through 8. This allows Louisiana to rotate the tasks given for each grade from administration to administration and encourages educators to focus on all three performance task types. As the choice of Literary Analysis Task or Narrative Writing Task would be made during the forms construction process, alternative blueprints—one with a Literary Analysis Task and a Research Simulation Task and the other with a Research Simulation Task and a Narrative Writing Task—were created for each grade. During forms construction, the Narrative Writing Task was selected for grades 3, 6, and 7 and the Literary Analysis Task was selected for grades 4, 5, and 8, based on item performance and the quality of the available passage sets for each performance task.

Student performance on the LEAP 2025 ELA assessments is reported by category and subcategory as outlined in the following table.

**Table 3.1 ELA Categories and Subcategories**

Category	Subcategory	Subcategory Description
<b>Reading</b>	Reading Literary Text	Students read and demonstrate comprehension of grade-level fiction, drama, and poetry.
	Reading Informational Text	Students read and demonstrate comprehension of grade-level nonfiction, including texts about history, science, art, and music.
	Reading Vocabulary	Students use context to determine the meaning of words and phrases in grade-level texts.
<b>Writing</b>	Written Expression	Students use details from provided texts to compose well-developed, organized, clear writing.
	Knowledge and Use of Language Conventions	Students use the rules of standard English (grammar, mechanics, and usage) to compose writing.

These reporting categories are the same as the reporting categories on the spring 2015-2018 ELA student reports and provide parents and educators with valuable information about

- overall student performance, including readiness to continue further studies in English language arts;
- student performance broken down by subcategory which may help identify when students need additional support or more challenging work in reading and writing; and
- how well schools and school systems help students achieve expectations.

The session testing times shown in the ELA test blueprints (see Tables 3.2 through 3.6) are based on PARCC testing times proportioned to be comparable based on the passage type being tested. The passage set that comes after the Narrative Writing Task is designed to balance the reading load between the Literary Analysis Task and the Narrative Writing Task. It is also designed to provide consistent timing in sessions 1 and 2.

Table 3.2 Grade 3 English Language Arts Test Blueprint and Test Design

Session	Content	Number of Passages	Categories/ Subcategories	Number of Two-Point SR Items	Number of Points from Two-Point SR Items	Number of PCR Items	Number of Points from PCR Items	Total Items	Total Points	Assessable ELA Student Standards (by subcategory)	Testing Time (minutes)
1	Research Simulation Task	2	Reading: Reading Informational Text/Reading Vocabulary*	6	12	1	3	6	15	RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	75
			Writing: Written Expression	0	0		9	9	Writing standards W.1-2, 7-8, 10		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Totals	2		6	12	1	15	7	27		
2	Narrative Writing Task	1	Reading: Reading Literary Text/Reading Vocabulary*	4	8	1	0	4	8	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	75
			Writing: Written Expression	0	0		9	9	Writing standards W.3, 10		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Reading Literary/ Informational Texts	1		4	8	0	0	4	8	RL Standards 1-3, 5-10; RI standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	
	Totals	2		8	16	1	12	9	28		
3	Reading Literary Texts	2	Reading: Reading Literary Text/Reading Vocabulary*	8	16	0	0	8	16	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	60**
	Reading Informational Texts		Reading: Reading Informational Text/Reading Vocabulary*							RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	
	Totals	2		8	16	0	0	8	16		
Grade 3 Totals		6	Reading: Reading Literary Text/Reading Vocab*	22	44	2	0	22	47	47	210
			Reading: Reading Informational Text/Reading Vocab*				3				
			Writing: Written Expression	0	0		18	18			
			Writing: Knowledge and Use of Language Conventions	0	0		6		2	6	
			Total	22	44		2	27	24	71	

\*Reading vocabulary items must constitute at least eight points on the test.

\*\*The time in session 3 allows for an additional passage set that is being field tested.

**Table 3.3 Grade 4 English Language Arts Test Blueprint and Test Design**

Session	Content	Number of Passages	Categories/ Subcategories	Number of Two-Point SR Items	Number of Points from Two-Point SR Items	Number of PCR Items	Number of Points from PCR Items	Total Items	Total Points	Assessable ELA Student Standards (by subcategory)	Testing Time (minutes)
1	Literary Analysis Task	2	Reading: Reading Literary Text/Reading Vocabulary*	6	12	1	4	6	16	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 9, 10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Totals	3		10	20	1	19	11	39		
	Reading Literary/ Informational Texts	1	Reading (Reading Informational Text/Reading Literature Text/Reading Vocabulary)	4	8	0	0	4	8	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5 RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	
2	Research Simulation Task	3	Reading: Reading Informational Text/ Reading Vocabulary*	8	16	1	4	8	20	RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 7-10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Totals	3		8	16	1	19	9	35		
3	Reading Literary Texts	1-2	Reading: Reading Literary Text/Reading Vocabulary*	6	12	0	0	6	12	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	60**
	Reading Informational Texts		Reading: Reading Informational Text/Reading Vocab*							RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	
	Totals	1-2		6	12					0	
Grade 4 Totals		7-8	Reading: Reading Literary Text/Reading Vocab*	24	48	2	4	24	56	56	240
			Reading: Reading Informational Text/Reading Vocab*				4				
			Writing: Written Expression	0	0		24	24	30		
			Writing: Knowledge and Use of Language Conventions	0	0		6			6	
			Total	24	48		2	38	26	86	

\*Reading vocabulary items must constitute at least eight points on the test.

\*\*The time in session 3 allows for an additional passage set that is being field tested.

**Table 3.4 Grade 5 English Language Arts Test Blueprint and Test Design**

Session	Content	Number of Passages	Categories/ Subcategories	Number of Two-Point SR Items	Number of Points from Two-Point SR Items	Number of PCR Items	Number of Points from PCR Items	Total Items	Total Points	Assessable ELA Student Standards (by subcategory)	Testing Time (minutes)
1	Literary Analysis Task	2	Reading: Reading Literary Text/Reading Vocabulary*	6	12	1	4	6	16	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 9, 10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Reading (Reading Literary Text/Reading Informational Text/Reading Vocabulary)	4	8	0	0	4	8	RL Standards 1-3, 5-10; RI standards 1-3, 5-10; vocabulary standards RL.4, RI.4, L.4, L.5			
Totals	3		10	20	1	19	11	39			
2	Research Simulation Task	3	Reading: Reading Informational Text/ Reading Vocabulary*	8	16	1	4	8	20	RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 7- 10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Totals	3		8	16	1	19	9	35		
3	Reading Informational Texts	1-2	Reading: Reading Informational Text/Reading Vocab*	6	12	0	0	6	12	RI standards 1-3, 5, 7-10; vocabulary standards RI.4, L.4, L.5	60**
	Totals	1-2		6	12	0	0	6	12		
Grade 5 Totals		8	Reading: Reading Literary Text/Reading Vocab*	10	20	2	4	10	24	56	240
			Reading: Reading Informational Text/Reading Vocab*	14	28		4	14	32		
			Writing: Written Expression	0	0		24	24	30		
			Writing: Knowledge and Use of Language Conventions	0	0		6	6			
			Total	24	48		2	38	26	86	

\*Reading vocabulary items must constitute at least eight points on the test.

\*\*The time in session 3 allows for an additional passage set that is being field tested.



Table 3.5 Grades 6 and 7 English Language Arts Test Blueprint and Test Design

Session	Content	Number of Passages	Categories/ Subcategories	Number of Two-Point SR Items	Number of Points from Two-Point SR Items	Number of PCR Items	Number of Points from PCR Items	Total Items	Total Points	Assessable ELA Student Standards (by subcategory)	Testing Time (minutes)
1	Research Simulation Task	3	Reading: Reading Informational Text/Reading Vocabulary*	8	16	1	4	8	20	RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 7-10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Totals	3		8	16	1	19	9	35		
2	Narrative Writing Task	1	Reading: Reading Literary Text/Reading Vocabulary*	4	8	1	0	4	8	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.3, 4, 10		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Reading Literary / Informational Texts	1-2		6	12	0	0	6	12	RL Standards 1-3, 5-10; RI standards 1-3, 5-10; vocabulary standards RL.4, RI.4, L.4, L.5	
Totals	2-3		10	20	1	15	11	35			
3	Reading Literary Texts	2	Reading: Reading Literary Text/Reading Vocabulary*	10	20	0	0	10	20	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	80**
	Reading Informational Texts		0			0	RI.1-3, 5, 7-10; vocabulary standards RI.4, L.4, L.5				
	Totals	2		10	20	0	0	10	20		
Grade 6 and 7 Totals		7-8	Reading: Reading Literary Text/Reading Vocab*	28	56	2	0	28	60	60	260
			Reading: Reading Informational Text/Reading Vocab*				4				
			Writing: Written Expression	0	0		24	24			
			Writing: Knowledge and Use of Language Conventions	0	0		6	6	30		
			Total	28	56		2	34	30	90	

\*Reading vocabulary items must constitute at least eight points on the test.

\*\*The time in session 3 allows for an additional passage set that is being field tested.

Table 3.6 Grade 8 English Language Arts Test Blueprint and Test Design

Session	Content	Number of Passages	Categories/ Subcategories	Number of Two-Point SR Items	Number of Points from Two-Point SR Items	Number of PCR Items	Number of Points from PCR Items	Total Items	Total Points	Assessable ELA Student Standards (by subcategory)	Testing Time (minutes)
1	Literary Analysis Task	2	Reading: Reading Literary Text/Reading Vocabulary*	6	12	1	4	6	16	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 9, 10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	Reading (Reading Literary Text/Reading Informational Text/Reading Vocabulary)	4	8	0	0	4	8	RL Standards 1-3, 5-10; RI standards 1-3, 5-10; vocabulary standards RL.4, RI.4, L.4, L.5			
	<b>Totals</b>	<b>3</b>		<b>10</b>	<b>20</b>	<b>1</b>	<b>19</b>	<b>11</b>	<b>39</b>		
2	Research Simulation Task	3	Reading: Reading Informational Text/ Reading Vocabulary*	8	16	1	4	8	20	RI standards 1-3, 5-10; vocabulary standards RI.4, L.4, L.5	90
			Writing: Written Expression	0	0		12	12	Writing standards W.1-2, 4, 7- 10,		
			Writing: Knowledge and Use of Language Conventions	0	0		3	3	Convention standards L.1, 2, plus language skills from previous grades		
	<b>Totals</b>	<b>3</b>		<b>8</b>	<b>16</b>	<b>1</b>	<b>19</b>	<b>9</b>	<b>35</b>		
3	Reading Literary Texts	2	Reading: Reading Literary Text/Reading Vocabulary*	10	20	0	0	10	20	RL Standards 1-3, 5-10; vocabulary standards RL.4, L.4, L.5	80**
	Reading Informational Texts		0			0	RI standards 1-3, 5, 7-10; vocabulary standards RI.4, L.4, L.5				
	<b>Totals</b>	<b>2</b>		<b>10</b>	<b>20</b>	<b>0</b>	<b>0</b>	<b>10</b>	<b>20</b>		
Grade 8 Totals		8	Reading: Reading Literary Text/Reading Vocab*	28	56	2	4	28	64	64	260
			Reading: Reading Informational Text/Reading Vocab*				4				
			Writing: Written Expression	0	0		24	2	24	30	
			Writing: Knowledge and Use of Language and Conventions	0	0		6				
			<b>Total</b>	<b>28</b>	<b>56</b>		<b>2</b>	<b>38</b>	<b>30</b>	<b>94</b>	

\*Reading vocabulary items must constitute at least eight points on the test.

\*\*The time in session 3 allows for an additional passage set that is being field tested.

The LEAP 2025 ELA assessments consist of tasks and reading passage sets. The tasks are described below.

- **Narrative Writing Task**
  - This task asks students to read a literary text, answer a set of selected-response questions about the text, and create a narrative related to the text (e.g., finish the story or retell the story in another narrative form, such as a journal entry).
  - This task focuses on students' ability to use narrative elements (e.g., dialogue, description) when writing.
- **Literary Analysis Task**
  - This task provides students with an opportunity to show their understanding of literature. It asks students to read two literary texts, answer a set of selected-response questions about the texts, and write an extended response that compares and/or explains key ideas or elements in the texts (e.g., central idea/message, contribution of illustrations, characterization).
  - This task focuses on students' ability to read complex text closely and asks them to carefully consider literature worthy of close study.
- **Research Simulation Task**
  - This task mirrors the research process by presenting three texts on a given topic. Students answer a set of selected-response questions about the texts and then write an extended response about some aspect of the related texts (e.g., relationship between a series of events, ideas, or concepts; comparison/contrast of key details; presentation of information).
  - This task requires students to synthesize information from related informational resources.

The following item types were included in the 2019 LEAP 2025 ELA assessments:

- **Selected-Response Items:**
  - **Evidence-based selected response (EBSR):** This item type consists of two parts. One part asks students to show their understanding of a text, and the other part asks students to identify evidence to support that understanding. The evidence supports a generalization, conclusion, or inference. This type of item is designed to provide students with opportunities to make explicit the evidence that supports their close analysis of a specific text.
  - **Multiple select (MS):** This item type requires students to select more than one correct answer and may appear as a one-part question or as part of an EBSR item. This type of item allows for the assessment of students' ability to identify multiple pieces of evidence to support a claim.
  - **Technology enhanced (TE):** This item type allows measurement of learning that may not be sufficiently measured by traditional multiple-choice items. TE items can measure the ordering of ideas within a summary; ordering of steps in a process; sorting, classifying, and categorizing ideas; matching of two themes/ideas to their unique evidence, etc. The technology used in TE items offers students additional ways to show understanding that parallels the classroom instructional techniques that teachers use to determine whether students are able to comprehend complex, grade-level text. TE Items may involve any of the following:

- Highlighting text: requires students to select text-based answer(s) from within a larger text
- Drag and drop: requires students to move draggable elements (e.g., words, phrases, or sentences) into one or more drop boxes (e.g., cells within a table or part[s] of a diagram)
- Drop-down menu: requires students to select from one or more drop-down menus to complete a phrase or sentence
- Match interaction table: requires students to select a checkbox in each row from two or more columns to classify statements presented in each row
- Prose constructed response (PCR): This item type appears at the end of each task and asks students to create an extended, complete written response. It elicits evidence that students have understood a text or texts they have read and can communicate that understanding well, both in terms of written expression and in terms of knowledge and use of language conventions.

A variety of item types allows for the measurement of the full range of student performance, including the performance of high- and low-performing students. Items and tasks should be clearly aligned to specific standards. Some items and tasks may ask students to draw evidence from one specific standard, while others may ask students to draw evidence from several standards.

The following table details the number of items and points by session and item type for each of the PBT (grades 3 and 4) and CBT (grades 3–8) forms.

**Table 3.7 Distribution of ELA Items and Points by Session and Item Type**

	Sub	Gr	Session	EBSR		MS		TE		PCR		Total No. of Pts.
				No. of Items	No. of Pts.	No. of Items	No. of Pts.	No. of Items	No. of Pts.	No. of Items	No. of Pts.	
Paper-Based Test (PBT)	ELA	3	1. Research Simulation Task	6	12					1	15	71
			2. Narrative Writing Task/Reading Passage	6	12	2	4			1	12	
			3. Reading Literary/Informational Texts	7	14	1	2					
	ELA	4	1. Literary Analysis Task/Reading Passage	9	18	1	2			1	19	86
			2. Research Simulation Task	7	14	1	2			1	19	
			3. Reading Literary/Informational Texts	6	12							
Computer-Based Tests (CBT)	ELA	3	1. Research Simulation Task	4	8			2	4	1	15	71
			2. Narrative Writing Task/Reading Passage	5	10	1	2	2	4	1	12	
			3. Reading Literary/Informational Texts	5	10	1	2	2	4			
	ELA	4	1. Literary Analysis Task/Reading Passage	6	12	1	2	3	6	1	19	86
			2. Research Simulation Task	5	10	1	2	2	4	1	19	
			3. Reading Literary/Informational Texts	5	10			1	2			
	ELA	5	1. Literary Analysis Task/Reading Passage	6	12	2	4	2	4	1	19	86
			2. Research Simulation Task	5	10	1	2	2	4	1	19	
			3. Reading Literary/Informational Texts	3	6	1	2	2	4			
	ELA	6	1. Research Simulation Task	5	10	1	2	2	4	1	19	90
			2. Narrative Writing Task/Reading Passage	3	6	3	6	4	8	1	15	
			3. Reading Literary/Informational Texts	4	8	3	6	3	6			
	ELA	7	1. Research Simulation Task	5	10	1	2	2	4	1	19	90
			2. Narrative Writing Task/Reading Passage	5	10	1	2	4	8	1	15	
			3. Reading Literary/Informational Texts	6	10	4	8	1	2			
ELA	8	1. Literary Analysis Task/Reading Passage	5	10	2	4	3	6	1	19	94	
		2. Research Simulation Task	5	10	1	2	2	4	1	19		
		3. Reading Literary/Informational Texts	5	10	3	6	2	4				

### 3.3 Mathematics Test Blueprints and Test Designs

The mathematics assessments were administered during a CBT testing window (April 1-May 3, 2019) or during a PBT testing window (April 29-May 3, 2019). The 2019 mathematics assessment had a similar structure to the 2018 assessment: each test session included the four mathematics subcategories, using the three mathematics task types (see Table 3.8).

Each item on the LEAP 2025 mathematics assessment is referred to as a task and is identified by one of three types: Type I, Type II, and Type III. As shown in the following table, each task type is aligned to one or two of four reporting categories: Major Content, Additional & Supporting Content, Expressing Mathematical Reasoning, or Modeling & Application. Each task type is designed to align with at least one of the [Standards for Mathematical Practice](#) (MP).

**Table 3.8 Overview of LEAP 2025 Mathematics Task Types and Reporting Categories**

Task Type	Description	Reporting Categories	Mathematical Practice(s)
Type I	Conceptual understanding, fluency, and application	<i>Major Content:</i> solve problems involving the <u>major content</u> for the grade level.  <i>Additional &amp; Supporting Content:</i> solve problems involving the <u>additional and supporting content</u> for the grade level.	Can involve any or all practices
Type II	Written arguments/justifications, critique of reasoning, or precision in mathematical statements	<i>Expressing Mathematical Reasoning:</i> express mathematical <u>reasoning</u> by constructing mathematical arguments and critiques.	Primarily MP.3 and MP.6 but may also involve any of the other practices
Type III	Modeling/application in a real-world context or scenario	<i>Modeling &amp; Application:</i> solve real-world problems engaging particularly in the <u>modeling</u> practice.	Primarily MP.4 but may also involve any of the other practices

These reporting categories provide parents and educators with valuable information about

- overall student performance, including readiness to continue further studies in mathematics;
- student performance broken down by mathematics subcategory, which may help identify when students need additional support or more challenging work; and
- how well schools and school systems help students achieve higher expectations.

Table 3.9 provides the distribution of operational points by subcategory, or reporting category, by grade.

**Table 3.9 Distribution of Points by Reporting Category—Mathematics**

Reporting Category	Grade					
	3	4	5	6	7	8
Major Content	30	30	30	30	30	30
Additional & Supporting Content	10	10	10	10	10	10
Expressing Mathematical Reasoning	10	10	10	14	14	14
Modeling & Application	12	12	12	12	12	12
<b>Total</b>	<b>62</b>	<b>62</b>	<b>62</b>	<b>66</b>	<b>66</b>	<b>66</b>

The Major Content areas for mathematics are broken into subcategories by grade as follows:

**Table 3.10 Major Content Subcategories by Grade**

Grade	Major Content Subcategory
3	<ul style="list-style-type: none"> <li>• Products and Quotients/Solve Multiplication and Division Problems</li> <li>• Solve Problems with Any Operation</li> <li>• Fractions as Numbers and Equivalence</li> <li>• Solve Time, Area, Measurement, and Estimation Problems</li> </ul>
4	<ul style="list-style-type: none"> <li>• Compare and Solve Problems with Fractions</li> <li>• Solve Multi-step Problems</li> <li>• Multiplicative Comparison and Place Value</li> </ul>
5	<ul style="list-style-type: none"> <li>• Operations with Decimals/Read, Write, and Compare Decimals</li> <li>• Solve Fraction Problems</li> <li>• Interpret Fractions, Place Value, and Scaling</li> <li>• Recognize, Represent, and Determine Volume/Multiply and Divide Whole Numbers</li> </ul>
6	<ul style="list-style-type: none"> <li>• Rational Numbers/Multiply and Divide Fractions</li> <li>• Ratio and Rate</li> <li>• Expressions, Inequalities, and Equations</li> </ul>
7	<ul style="list-style-type: none"> <li>• Analyze Proportional Relationships and Solve Problems</li> <li>• Operations with Rational Numbers</li> <li>• Expressions, Inequalities, and Equations</li> </ul>
8	<ul style="list-style-type: none"> <li>• Radicals, Integer Exponents, and Scientific Notation</li> <li>• Proportional Relationships, Linear Equations, and Functions</li> <li>• Solving Linear Equations/Systems of Linear Equations</li> <li>• Congruence and Similarity/Pythagorean Theorem</li> </ul>

The resulting 2019 LEAP 2025 mathematics test blueprints are shown in Tables 3.11–3.16.

**Table 3.11 Grade 3 Mathematics Test Blueprint**

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	27–30	30					Louisiana Student Standards for Mathematics (LSSM):  3.OA.A.1-4, 3.OA.B.6,  3.OA.C.7, 3.OA.D.8,  3.NF.A.1-3, 3.MD.A.1-2,  3.MD.C.5-7  LEAP 2025 Evidence Statements: LEAP.I.3.1-4
Additional & Supporting Content	7–10	10					LSSM:  3.NBT.A.1-3, 3.MD.B.3-4,  3.MD.D.8, 3.MD.E.9, 3.G.A.1-2  LEAP 2025 Evidence Statements: LEAP.I.3.5-6
Expressing Mathematical Reasoning			3	10			LEAP 2025 Evidence Statements: LEAP.II.3.1-8
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.3.1-2
<b>TOTAL</b>	<b>37</b>	<b>40</b>	<b>3</b>	<b>10</b>	<b>3</b>	<b>12</b>	
	<b>TOTAL TASKS</b>		<b>43</b>	<b>TOTAL POINTS</b>		<b>62</b>	



Table 3.12 Grade 4 Mathematics Test Blueprint

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	27–30	30					LSSM: 4.OA.A.1-3, 4.NBT.A.1-3 4.NBT.B.4-6, 4.NF.A.1-2, 4.NF.B.3-4, 4.NF.C.5-7  LEAP 2025 Evidence Statements:  LEAP.I.4.1-8
Additional & Supporting Content	7–10	10					LSSM: 4.OA.B.4, 4.OA.C.5, 4.MD.A.1-3, 4.MD.B.4, 4.MD.C.5-7, 4.MD.D.8, 4.G.A.1-3
Expressing Mathematical Reasoning			3	10			LEAP 2025 Evidence Statements: LEAP.II.4.1-7
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.4.1-2
<b>TOTAL</b>	<b>37</b>	<b>40</b>	<b>3</b>	<b>10</b>	<b>3</b>	<b>12</b>	
	<b>TOTAL TASKS</b>		<b>43</b>	<b>TOTAL POINTS</b>		<b>62</b>	

Table 3.13 Grade 5 Mathematics Test Blueprint

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	27–30	30					LSSM: 5.NBT.A.1-4, 5.NBT.B.5-7 5.NF.A.1-2, 5.NF.B.3-7 5.MD.C.3-5 LEAP 2025 Evidence Statements: LEAP.I.5.1-2
Additional & Supporting Content	7–10	10					LSSM: 5.OA.A.1-2, 5.OA.B.3 5.MD.A.1, 5.MD.B.2 5.G.A.1-2, 5.G.B.3-4
Expressing Mathematical Reasoning			3	10			LEAP 2025 Evidence Statements: LEAP.II.5.1-9
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.5.1-2
<b>TOTAL</b>	<b>37</b>	<b>40</b>	<b>3</b>	<b>10</b>	<b>3</b>	<b>12</b>	
	<b>TOTAL TASKS</b>		<b>43</b>	<b>TOTAL POINTS</b>		<b>62</b>	

Table 3.14 Grade 6 Mathematics Test Blueprint

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	26–30	30					LSSM: 6.RP.A.1-3, 6.NS.A.1, 6.NS.C.5-8, 6.EE.A.1-2,4, 6.EE.B.5-8, 6.EE.C.9
Additional & Supporting Content	6–10	10					LSSM: 6.NS.B.2-4, 6.G.A.1-4, 6.SP.A.1-3, 6.SP.B.4-5
Expressing Mathematical Reasoning			4	14			LEAP 2025 Evidence Statements: LEAP.II.6.1-9
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.6.1-3
<b>TOTAL</b>	<b>36</b>	<b>40</b>	<b>4</b>	<b>14</b>	<b>3</b>	<b>12</b>	
	<b>TOTAL TASKS</b>		<b>43</b>	<b>TOTAL POINTS</b>		<b>66</b>	

Table 3.15 Grade 7 Mathematics Test Blueprint

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	26–30	30					LSSM: 7.RP.A.1-3, 7.NS.A.1-3, 7.EE.A.1-2, 7.EE.B.3-4
Additional & Supporting Content	6–10	10					LSSM: 7.G.A.1-3, 7.G.B.4-6, 7.SP.A.1-2, 7.SP.B.3-4, 7.SP.C.5-8
Expressing Mathematical Reasoning			4	14			LEAP 2025 Evidence Statements: LEAP.II.7.1-7
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.7.1-4
<b>TOTAL</b>	<b>36</b>	<b>40</b>	<b>4</b>	<b>14</b>	<b>3</b>	<b>12</b>	
	<b>TOTAL TASKS</b>		<b>43</b>	<b>TOTAL POINTS</b>		<b>66</b>	

**Table 3.16 Grade 8 Mathematics Test Blueprint**

Reporting Category	Task Types						Assessable Content
	Type I		Type II		Type III		
	Tasks	Points	Tasks	Points	Tasks	Points	
Major Content	25-30	30					LSSM: 8.EE.A.1-4, 8.EE.B.5-6 8.EE.C.7-8, 8.F.A.1-3 8.G.A.1-4, 8.G.B.7-8
Additional & Supporting Content	5-10	10					LSSM: 8.F.B.4-5, 8.G.C.9 8.SP.A.1-4, 8.NS.A.1-2
Expressing Mathematical Reasoning			4	14			LEAP 2025 Evidence Statements: LEAP.II.8.1-5
Modeling & Application					3	12	LEAP 2025 Evidence Statements: LEAP.III.8.1-4
TOTAL	35	40	4	14	3	12	
	<b>TOTAL TASKS</b>		<b>42</b>	<b>TOTAL POINTS</b>		<b>66</b>	

Unlike the ELA test blueprints, which were organized by test sessions one through three, the mathematics test blueprints were organized by reporting categories, so it was necessary to define the general structure of the test forms by test session. The design goal was to have balanced test sessions with a variety of task types and equivalent testing times. For all forms in grades 3–5, students were prohibited from using calculators, except for those students with a calculator accommodation. For session one of the mathematics test in grades 6–8, students are prohibited from using calculators, except those students with a calculator accommodation. Calculators were allowed to be used by all students in grades 6–8 in sessions two and three. The general test structures (see Tables 3.17–3.22) guided test form sequencing and design. The LEAP 2025 [Calculator Policy](#) provided the basis for calculator designation of tasks and items.

Table 3.17 General Mathematics Test Structure—Grade 3

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 No Calculator		Session 3 No Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	9–10	10	8–10	10	10	10	27–30	30
Additional & Supporting Content	3–4	4	2–4	4	2	2	7–10	10
Expressing Mathematical Reasoning	1	4	1	3	1	3	3	10
Modeling & Application	1	3	1	3	1	6	3	12
<b>TOTAL (Operational Only)</b>	<b>15</b>	<b>21</b>	<b>14</b>	<b>20</b>	<b>14</b>	<b>21</b>	<b>43</b>	<b>62</b>
<b>Test Duration (minutes)*</b>	<b>75</b>		<b>85</b>		<b>75</b>		<b>235</b>	

\*The testing time includes items that are being field tested.

Table 3.18 General Mathematics Test Structure—Grade 4

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 No Calculator		Session 3 No Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	9–10	10	8–10	10	10	10	27–30	30
Additional & Supporting Content	3–4	4	2–4	4	2	2	7–10	10
Expressing Mathematical Reasoning	1	4	1	3	1	3	3	10
Modeling & Application	1	3	1	3	1	6	3	12
<b>TOTAL (Operational Only)</b>	<b>15</b>	<b>21</b>	<b>14</b>	<b>20</b>	<b>14</b>	<b>21</b>	<b>43</b>	<b>62</b>
<b>Test Duration (minutes)*</b>	<b>75</b>		<b>85</b>		<b>75</b>		<b>235</b>	

\*The testing time includes items that are being field tested.

Table 3.19 General Mathematics Test Structure—Grade 5

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 No Calculator		Session 3 No Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	9–10	10	8–10	10	10	10	27–30	30
Additional & Supporting Content	3–4	4	2–4	4	2	2	7–10	10
Expressing Mathematical Reasoning	1	4	1	3	1	3	3	10
Modeling & Application	1	3	1	3	1	6	3	12
<b>TOTAL (Operational Only)</b>	<b>15</b>	<b>21</b>	<b>14</b>	<b>20</b>	<b>14</b>	<b>21</b>	<b>43</b>	<b>62</b>
<b>Test Duration (minutes)*</b>	<b>75</b>		<b>85</b>		<b>75</b>		<b>235</b>	

\*The testing time includes items that are being field tested.

Table 3.20 General Mathematics Test Structure—Grade 6

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 Calculator		Session 3 Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	10–12	12	6–8	8	8–10	10	26–30	30
Additional & Supporting Content	6–8	8	1–2	2	0	0	6–10	10
Expressing Mathematical Reasoning	0	0	2	7	2	7	4	14
Modeling & Application	0	0	2	9	1	3	3	12
<b>TOTAL (Operational Only)</b>	<b>16–20</b>	<b>20</b>	<b>12–13</b>	<b>26</b>	<b>11–13</b>	<b>20</b>	<b>43</b>	<b>66</b>
<b>Test Duration (minutes)*</b>	<b>60</b>		<b>90</b>		<b>90</b>		<b>240</b>	

\*The testing time includes items that are being field tested.

Table 3.21 General Mathematics Test Structure—Grade 7

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 Calculator		Session 3 Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	16–20	20	3–5	5	3–5	5	26–30	30
Additional & Supporting Content	0	0	3–5	5	3–5	5	6–10	10
Expressing Mathematical Reasoning	0	0	2	7	2	7	4	14
Modeling & Application	0	0	2	9	1	3	3	12
<b>TOTAL (Operational Only)</b>	<b>16–20</b>	<b>20</b>	<b>12-13</b>	<b>26</b>	<b>11–13</b>	<b>20</b>	<b>43</b>	<b>66</b>
<b>Test Duration (minutes)*</b>	<b>60</b>		<b>90</b>		<b>90</b>		<b>240</b>	

\*The testing time includes items that are being field tested.

Table 3.22 General Mathematics Test Structure—Grade 8

Reporting Category	Test Session						TOTAL (Operational Only)	
	Session 1 No Calculator		Session 2 Calculator		Session 3 Calculator			
	Tasks	Points	Tasks	Points	Tasks	Points	Tasks	Points
Major Content	13–18	18	3–6	6	4–6	6	25–30	30
Additional & Supporting Content	2–4	4	2–3	3	2–3	3	5–10	10
Expressing Mathematical Reasoning	0	0	2	7	2	7	4	14
Modeling & Application	0	0	2	9	1	3	3	12
<b>TOTAL (Operational Only)</b>	<b>15–20</b>	<b>22</b>	<b>10–13</b>	<b>25</b>	<b>10–12</b>	<b>19</b>	<b>42</b>	<b>66</b>
<b>Test Duration (minutes)*</b>	<b>60</b>		<b>90</b>		<b>90</b>		<b>240</b>	

\*The testing time includes items that are being field tested.



The following item types were used in the 2019 LEAP 2025 mathematics assessments:

- Multiple choice: This item type requires students to select one correct answer from four answer choices. It may appear as a one-part question, as part of a two-part question, or as a part of a constructed-response item. The multiple choice items are worth one point.
- Multiple select: This item type requires students to select more than one correct answer from more than four answer choices. It may appear as a one-part question, as part of a two-part question, or as a part of a constructed-response item. The multiple select items are worth one point. Students must choose all correct answers and no incorrect answer to receive credit.
- Short answer: This item type requires students to enter a numeric response by typing from the keyboard; it allows a decimal and numbers for grades 3–8 and a negative sign for grades 6–8. It may appear as a one-part question, as part of a two-part question, or as a part of a constructed-response item. The short answer items are worth one point. Unless specified in the question, a student will earn credit for an answer that is equivalent to the correct numerical answer and proper rounding may be required.
- Keypad input: This item type requires students to enter a mathematical response using a customized pallet of numbers, operations, variables, and/or mathematical symbols; allows all rational and irrational numbers as well as expressions and equations; and scores all equivalent responses as correct unless noted otherwise. This item type may appear as a one-part question, as part of a two-part question, or as a part of a constructed-response item.
- Constructed response: This item type requires students to respond to an open-ended question which must be typed into a response box; students may use the equation builder tool (specific to the grade or grade span) to insert mathematical characters. This item type can be a single- or multi-part item. Constructed-response items ask students to write explanations or justifications, model a process, and/or solve real-world, multi-step contextual problems. A student may receive partial or full credit on constructed-response items, and maximum point values will vary by constructed-response task. Maximum values for constructed-response items are 3, 4, or 6 points.
- Technology enhanced: This item type uses technology to capture student responses. Technology-enhanced items may appear as a one-part question, as part of a two-part question, or as a part of a constructed-response item. The technology-enhanced items are worth one point. Technology-enhanced items may involve any of the following:
  - Bar graph: requires students to complete a bar graph or histogram by raising/lowering each bar to a value
  - Drag and drop: requires students to move draggable elements into one or more drop boxes
  - Dropdown menu: requires students to select from one or more dropdown menus to complete a sentence, phrase, or expression/equation/inequality
  - Hot spot: requires students to select one or more responses by choosing selectable areas on the screen

- Match interaction table: requires students to select a checkbox in each row from two or more columns
- Graph input: requires students to enter a response on a coordinate grid
- Number line input: requires a student to enter a response on a number line
- Line plot: requires students to complete a line plot with “X” as the input

A variety of item types allows for the measurement of the full range of student performance, including that of high- and low-performing students.

The following table details the number of items by point value and task type as well as the number of points per task type for each of the PBT (grades 3 and 4) and CBT (grades 3–8) forms.

**Table 3.23 Distribution of Mathematics Tasks and Points by Task Type**

	Content Area	Grade	Type I			Type II			Type III			Total Points
			1 pt Tasks	2 pt Tasks	Points	3 pt Tasks	4 pt Tasks	Points	3 pt Tasks	6 pt Tasks	Points	
Paper-Pencil (PBT)	Math	3	34	3	40	2	1	10	2	1	12	62
	Math	4	34	3	40	1	1	10	2	1	12	62
Online (CBT)	Math	3	34	3	40	2	1	10	2	1	12	62
	Math	4	34	3	40	2	1	10	2	1	12	62
	Math	5	34	3	40	2	1	10	2	1	12	62
	Math	6	32	4	40	2	2	14	2	1	12	66
	Math	7	32	4	40	2	2	14	2	1	12	66
	Math	8	30	5	40	2	2	14	2	1	12	66

### 3.4 Item Development and Selection

The processes of item development and selection are discussed in this section in compliance with the *Standards*.

Standard 4.7 states the following:

The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (87)

The items used in the 2019 LEAP 2025 ELA and mathematics assessments came from the PARCC consortium's and Louisiana-owned item banks.

The items selected for use on the 2019 LEAP forms were used to equate to the LEAP 2025 scale. Operational forms were selected based on LEAP 2025 test blueprint specifications, which were supported by statistical data from PARCC operational testing.

### 3.5 Considerations of Test Fairness in Item Development

Standard 3.2 is particularly relevant to fairness in item development:

Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

Bias and sensitivity guidelines used to develop the PARCC and Louisiana-owned items help ensure the assessments are fair for all groups of test takers, despite differences in characteristics that include, but are not limited to, disability status, ethnic group, race, gender, regional background, native language, religion, sexual orientation, and socioeconomic status. DRC relied strongly on the bias and sensitivity guidelines in the development of the assessments, particularly in item selection and review. To be included in the assessments, items had to comply with the bias and sensitivity guidelines and be approved by Louisiana educators involved in the Louisiana alignment and item review meetings.

### 3.6 PARCC Item Reviews

As part of PARCC's ongoing item development practices, several educator committees had already been convened to conduct rigorous reviews of every passage and item developed for the PARCC assessment system prior to the items becoming a part of the item bank that included items and passages available for selection on Louisiana forms. These reviews include

- text reviews of all passages (during which participants review and edit passages independently and then discuss content and bias concerns as a grade-level group),
- item reviews (during which committees review and edit items for adherence to PARCC foundational documents, basic principles of universal design, PARCC accessibility guidelines, selected metadata fields, and the PARCC style guide),
- bias and sensitivity reviews (during which educators and community members review items and tasks to confirm the absence of issues relating to bias, fairness, and sensitivity to ensure that items and tasks do not unfairly advantage or disadvantage any student subgroup over another subgroup),
- editorial reviews (during which the review committee completes a copy edit review and records member comments), and
- data reviews (during which educators evaluate item-level statistics to determine eligibility of items and tasks to move forward to the operational assessments).

Additional information on PARCC’s item review processes and procedures can be found in their [technical reports](#). Only items that have been approved by expert reviewers during text reviews (ELA only), item reviews, bias and sensitivity reviews, and editorial reviews are moved forward for field testing by PARCC affiliate states. Of the field tested items, only those determined to have acceptable statistics, either by having acceptable item parameters according to the data review flagging criteria or by being approved by expert reviewers during data review, are eligible for review by Louisiana educators for potential use on an operational assessment. These processes follow the criteria set forth by the *Standards*.

Standard 3.1 states the following:

Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

Standard 3.2 states the following:

Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

Independent studies of PARCC passages and items have found that the content being licensed assesses the skills that matter most and is rigorous, aligned to standards, and accessible to students with disabilities and English learners. For more information on the studies performed, refer to New Meridian’s website: <https://resources.newmeridiancorp.org/research/>.

## 3.7 Louisiana Item Development and Item Review

### 3.7.1 ELA Development Process

Item development for ELA began with a detailed analysis of the acquired item bank to determine the needs of the pool. This analysis resulted in development targets for each grade beginning with the selection of passages. Development targets indicated whether passage sets should be short, long, or paired. They also determined whether passages should be literary or informational. Additional traits for each target, such as text complexity, standards that should be assessed, genre coverage, gender representation, topic variety, and ways to add diversity to the pool of passages available, were also provided. Once the targets were identified and approved by LDOE ELA content experts, DRC’s ELA test development team worked to provide options for each target for LDOE review. Experienced passage finders recommended authentic texts, including permissioned and/or public domain passages. These initial selections were reviewed by DRC’s ELA test development team members, who then analyzed the text complexity of each passage. The passages, any associated graphics, and the results of text complexity analyses were provided to LDOE. LDOE’s ELA team reviewed the options and provided feedback to ensure that two options for each target were ready for review by Louisiana educators.

### 3.7.2 Text Complexity Specifications for Field Test Passages

As part of the passage development process, a passage’s text complexity is analyzed so that an appropriate grade-level placement for each passage can be made. The analysis of the passage’s text complexity was captured on a placemat. (Please see Appendix A for a sample placemat.) DRC used a process that includes (1) a quantitative evaluation of the text and (2) a qualitative evaluation of the text. Passages and their respective placemats are submitted to LDOE during initial passage reviews. In addition, a third component, matching reader to text and task, is also taken into consideration during passage evaluation.

### 3.7.3 Passage Review

In June 2018, in conjunction with the alignment reviews of items from the acquired item bank, passage reviews were conducted by Louisiana educator committees. During the review process, the committees, which represented a variety of perspectives, reviewed the proposed literary and informational passages to ensure the texts used to develop passage sets on the LEAP 2025 ELA tests were fair and appropriate for all students and would allow an opportunity for students to demonstrate their knowledge and skills in ELA. Educators reviewed the passages and provided feedback and a consensus decision about the status of each passage. The status identified whether the passage was acceptable to move forward with development or not acceptable to move forward with development. Educators also provided individual rankings of the preference for passages of each target type. Upon conclusion of the passage reviews, members of the LDOE and DRC's ELA test development teams met to discuss the results. Decisions regarding which passages would move forward to item development were made at that time.

Table 3.24 provides the count of passages brought to passage review and the status of those passages after passage review.

**Table 3.24 LEAP 2025 June 2018 Passage Review**

Content Area	Grade	Number of Passage Sets Reviewed	Accepted	Rejected
ELA	3	6	6	0
ELA	4	7	5	2
ELA	5	7	6	1
ELA	6	6	5	1
ELA	7	7	6	1
ELA	8	6	5	1

### 3.7.4 ELA Item Writing and Review

Once passage sets were approved for development by Louisiana educators and content experts from the LDOE, the passage sets were provided to experienced item writers for development. Item writers participated in item writing training with DRC prior to developing items. The training involved discussion of the following:

- Passage set quality:
  - Passage sets have value.
  - Passages sets are cohesive.
  - Items are text dependent.
  - Items are aligned to and reflect the rigor of the Louisiana Student Standards.
- Item type descriptions and examples of each:
  - Evidence-Based Selected Response

- Multiple Select
- Technology Enhanced
- Resources to support item writing:
  - PARCC practice tests
  - PARCC released items
- Training on universal design and bias, fairness, and sensitivity
- Training on security and confidentiality

Once the items were written, they were revised as necessary by DRC prior to delivery to the LDOE. The LDOE reviewed each batch of items and provided feedback that was implemented prior to the passage sets being prepared in mock test forms for review by committees of Louisiana educators at item content and bias review meetings.

At the item content and bias review meetings, grade-level committees of ELA educators met to provide feedback on the alignment and appropriateness of items for use on Louisiana assessments. Louisiana educators reviewed items for alignment to content standards; grade appropriateness; issues of bias, fairness, and sensitivity; difficulty and cognitive complexity; and clarity of language. The discussion about difficulty and cognitive complexity included not only approving the cognitive complexity levels assigned to each item but also ensuring that the difficulty and cognitive complexity were appropriate for the grade level. Louisiana educators also reviewed the items to ensure they represented a range of difficulty and cognitive complexity. Louisiana educators edited items as needed to ensure they were appropriate for use on Louisiana assessments, which allowed the items to move forward for possible field-testing. Any items deemed inappropriate were rejected if educators were not able to revise or recommend appropriate revisions for those items. Items that successfully passed through the item content and bias reviews were then embedded within operational test forms for field-testing, and data was collected on each field test item. For a detailed description of the process followed during the content and bias review meetings, see Appendix B.

### 3.8 Mathematics Item Development

To determine the mathematics item development needs for field-testing in the Spring 2019 administration, the LDOE determined the count of items needed per grade and then DRC content experts analyzed the item pool to determine the number of type II or type III items and the evidence statements/standards based on that analysis. DRC content experts reviewed standards coverage on the previous year's test by looking at the number and types of items used to cover each content standard, the difficulty range, the level of cognitive complexity covered by each content standard, and the topic/material presented in items (to ensure a variety of engaging topics are included). DRC determined gaps or holes in coverage, based on these criteria, to create an item development plan for the number and types of items to be newly developed for possible field-testing in spring 2019. DRC presented the item development plan to LDOE content experts, who then provided feedback to DRC. DRC and the LDOE collaborated to finalize the item development plan. DRC contracted with content experts to have items written. Item writers participated in item writing training with DRC and the LDOE prior to developing items. The training included:

- an overview of the assessable content and task types,
- a description of the type II and type III items,
- an explanation of how to use the standards and evidence statements when writing items,
- examples of type II and type III items,
- a discussion that covered item writing guidelines
- examples of items with issues,

- training on security and confidentiality, and
- training on universal design and bias, fairness, and sensitivity

These items were reviewed by the LDOE and revised by DRC. Once items were approved by the LDOE, they became part of the set of items that were taken to item content and bias reviews with Louisiana educators in summer 2018. Refer to Appendix B “Item Content and Bias Review,” for counts of the items developed for content and bias reviews and field-testing.

At the mathematics item content and bias reviews, committees met to provide feedback on the alignment and appropriateness of items. Louisiana educators reviewed items for alignment to content standards; grade appropriateness; issues of bias, fairness, and sensitivity; and difficulty and cognitive complexity, which included determining whether the difficulty and cognitive complexity were appropriate for each item and whether the items available represented a range of difficulty and cognitive complexity. For a detailed description of the process followed during the item content and bias reviews, see Appendix B. Louisiana educators edited items as needed to ensure they were appropriate for use on Louisiana assessments, which allowed the items to move forward for possible field-testing. Any items deemed inappropriate were rejected if educators were not able to revise those items. Items that successfully passed through the content and bias reviews were then placed on a test form in a field test position, and data was collected on each field test item. Once field-testing was complete, the items were taken to range-finding, where committees of Louisiana educators reviewed Louisiana student responses to assign true scores to responses that would be used in training materials for the scoring of items. The field-tested constructed response items were then scored, and the data were analyzed by DRC psychometricians.

For a detailed description of the process followed during the Item Content and Bias Reviews, see Appendix B “Item Content and Bias Review.”

### 3.9 Guidelines on Bias, Fairness, and Sensitivity

ELA and mathematics item writers and content and bias committee members were provided with guidelines on bias, fairness, and sensitivity issues as they pertain to testing. The information included definitions of bias and sensitivity, examples of different types of bias, and topics of concern, which were specific to given content areas. Writers were also provided with sample items that contained bias, fairness, and sensitivity issues and examples of how to revise items and graphics to ensure universal design is applied. The writers were also given information on accessibility and accommodations, including information on how to address language, visual elements, and design issues when considering students in special populations (e.g., students with disabilities and English learners).

#### **Types of Bias:**

- Stereotyping
  - may result when an image is formed by relating certain characteristics to ALL members of a group and may include physical characteristics, intellectual characteristics, emotions, careers, activities, and domestic or social roles
- Gender Bias
  - may result when members of either sex are unnecessarily presented in stereotypical activities, occupations, and/or situations or are unnecessarily presented as having stereotypical emotions or characteristics
- Regionalism

- may result from the inclusion of terms that are not commonly used nationwide or within a particular region of the state in which the test will be given
- Ethnic or Cultural Bias
  - may result from the inclusion of terms, concepts, or situations that are demeaning and/or offensive to a particular ethnic group or culture
- Socioeconomic or Class Bias
  - may result from the inclusion of activities, possessions, or ideas that may not be common to all students
- Religious Bias
  - may result from the inclusion of terms, concepts, or situations that are demeaning and/or offensive to a particular religious group
- Ageism
  - may result from the inclusion of terms, concepts, or situations that are demeaning and/or offensive to elders or to older persons (defined as people older than the reference group) and may also involve issues of bias with other age groups, including teenagers and young children, or even with the age of the reference group itself, where the grade (age) of a student is depicted negatively
- Bias against Persons with Disabilities
  - may result from the inclusion of terms, concepts, or situations that are demeaning and/or offensive to persons with disabilities

### 3.9.1 Louisiana Item Alignment Review

Independent of PARCC reviews, DRC conducts the Louisiana Item Alignment Reviews, during which Louisiana educators review items and passage sets for alignment to the Louisiana Student Standards and for appropriateness of the items and tasks for students in Louisiana, including being free of issues of bias, fairness, and sensitivity.

DRC, with guidance from the LDOE, conducted the Louisiana Item Alignment Review in June 2018 with committees of Louisiana educators. Grade-level committees met for two and a half days (mathematics) or two to three days (ELA) to provide feedback on the alignment and appropriateness of items that made up the PARCC item bank. To the extent possible, each committee included educators from different parts of Louisiana, who represent all Louisiana students (e.g., special education, English learners, students with disabilities, etc.). Committee members are also representative of the diverse demographics of the state.

As described in the preceding sections, items presented at these reviews went through a rigorous review process before and after the items were field-tested by PARCC to ensure quality and appropriateness. Items were selected for inclusion in the form selection pool, imported into IDEAS (DRC's item banking system), and formatted for use on Louisiana test forms. They were placed on mock test forms to allow them to be reviewed as students would see them. Louisiana educators reviewed these items to confirm they were acceptable for use on a Louisiana assessment. Educators reviewed items individually to verify that each item aligned to the Louisiana Student Standard(s) for that item prior to discussing the items as a group. In addition, educators reviewed item keys and discussed the difficulty and cognitive complexity of each item and task. The groups came to a consensus regarding the status of each item: Accepted with Current Alignment, Accepted with Realignment, or Rejected. Items that were accepted were determined to appropriately measure the intended standard(s) and be free of issues of bias, fairness, or sensitivity that could impact student responses to the item. For a detailed description of the process followed during the



item alignment reviews, including results and descriptions of the demographics of each committee, see Appendix C, “Item Alignment Review Process.”

### 3.10 Operational Test Selection

Operational item selection for 2019 took place from September 2018 through November 2018 by the LDOE and DRC. The PARCC and Louisiana-owned item banks were used to select fixed LEAP 2025 ELA and mathematics forms.

The LEAP 2025 assessments were given in two modalities: computer-based test (CBT) or paper-based test (PBT). For both ELA and mathematics, students in grades 3 through 8 took the CBTs; some school systems elected to administer the PBTs to students in grades 3 and 4. For ELA, the dual-mode forms were identical except for a small quantity (four to five items) of technology-enhanced items (TE) in each CBT. Items used on PBTs as replacements for the TE items were evidence-based selected-response items that addressed the same content standards and were of similar rigor as the TE items, when possible. For mathematics, short-answer (SA) items were reformatted as gridded-response (GR) items for use on PBTs.

#### 3.10.1 General Item and Passage Set Selection Process and Criteria

The item and passage selection process used for forms construction was a content-focused, collaborative process between the LDOE and DRC ELA and mathematics content specialists, and it was followed by a psychometric evaluation of each selection. The critical psychometric consideration, other than individual item performance, was the degree to which the selected items reflected the 2018 target’s test characteristic curve (TCC), standard error of measurement (SEM) and test information function (TIF). Although the item pool was limited, items that were determined to be very difficult (i.e., IRT difficulty parameter  $b > 2.0$ ) and/or not discriminating (i.e., IRT discrimination parameter  $a < 0.3$ ) were avoided when possible.

##### *Item Selection Guidelines*

- Using the acquired pool of items, content-area assessment specialists select ELA passage sets and tasks that consist of quality texts displaying diversity in topics and authors and mathematics tasks that match the blueprint. The sets and/or tasks include items that cover a range of Louisiana Student Standards and/or Evidence Statements (mathematics only) and address the appropriate reporting categories.
- Content-area assessment specialists and research analysts verify that each item meets psychometric guidelines for excellence as available item-performance data allows.
- Forms include adequate content coverage, as required by the detailed test blueprint.
- Each form contains an anchor set that includes passage sets/items from a previously administered form. The anchor set, which is a mini-blueprint of the form, ensures comparability between the previous form and the 2019 form. The remaining items selected for a form complete blueprint requirements.
- No item in a form should “clue” (or provide the answer to) another item on that same form.
- Clang association should be avoided. Clang is when a distractor can be associated with a stem word by sound rather than meaning (e.g., rhyming, alliteration).
- Passage sets in ELA forms should be diverse.
- Forms should be diverse, including a variety of text types, including texts that appeal to a diverse student population (see the [PARCC Passage Selection Guidelines](#)).
- Forms should include a wide range of topics and a variety of questions.
- Correct answer distributions should follow best practice (no more than 3 keys of the same answer option in a row).
- Forms **must not** contain any items that have been released to the public.

### 3.10.2 Review of the English Language Arts Items and Forms

DRC and LDOE ELA content specialists and members of educator committees verified that the items were in compliance with the guidelines provided by LDOE, including alignment to the content standards and appropriateness for Louisiana students. Because establishing content validity is one of the most important aspects in the legal defensibility of a test, the alignment of the items to the content standards must be reviewed and verified at every stage of the test development process. As a result, it is essential that an item selected for a form link directly to the content standard that it purports to measure. The ELA content specialists also verified all items against their designated content codes and metadata, both to evaluate the correctness of the coding and to ensure that the given item measures what it purports to measure.

In addition, the ELA content specialists reviewed each item for item quality, ensuring that the items were in compliance with industry guidelines for clarity, style, accuracy, and appropriateness for Louisiana students. While there are many published guidelines for reviewing assessment items, the following list serves to summarize the major considerations content specialists followed when reviewing items to ensure the items conformed to item quality standards for good, reliable, and fair test questions.

#### ***Guidelines for Reviewing Items Selected for Forms***

##### *A good item should*

- have only one clear, correct answer and contain answer choices that are reasonably parallel in length and structure (multiple choice);
- have only the indicated number of correct answers and contain answer choices that are reasonably parallel in length and structure (multiple select);
- have a correctly assigned content code (item map);
- measure one main idea or standard, unless it is a complex item, such as a prose constructed-response item (PCR);
- measure the objective or content standard(s) it is designed to measure;
- be at the appropriate level of rigor;
- be simple, direct, and free of ambiguity;
- make use of vocabulary and sentence structure that is appropriate to the grade level of the student being tested;
- be based on content that is accurate and current;
- when appropriate, contain stimulus material that is clear and concise and provides all of the information needed;
- when appropriate, contain graphics that are clearly labeled;
- contain answer choices that are plausible and reasonable in terms of the requirements of the question, as well as a student's level of knowledge;
- contain distractors that relate to the question in the same way and can be supported by a rationale;
- reflect current teaching and learning practices in the subject area; and
- be free of gender, ethnic, cultural, socioeconomic, and regional bias.

#### 3.10.2.1. Review of the Mathematics Items and Forms

DRC and LDOE mathematics content specialists also ensured the items were in compliance with the guidelines provided by LDOE, including alignment to the content standards and appropriateness for Louisiana students. Since establishing content validity is one of the most important aspects in the legal defensibility of a test, the alignment of the items to the content standards must be reviewed and verified at every stage of the test development process. As a result, it is essential that an item selected for a form link directly to the content standard that it purports to measure. The mathematics content specialists also verified all items against their designated content codes and metadata, both to evaluate the accuracy of the coding and to ensure that the given item measures what it purports to measure.

In addition, the mathematics content specialists reviewed each item for item quality, ensuring that the test items are in compliance with industry guidelines for clarity, style, accuracy, and appropriateness for Louisiana students. While there were many published guidelines for reviewing assessment items, the list below serves to summarize the major considerations mathematics content specialists followed when reviewing items to ensure they conformed to item quality standards for good, reliable, and fair test questions.

### ***Guidelines for Reviewing Items Selected for Forms***

A good item should

- contain answer choices that are reasonably parallel in length and structure;
- have the appropriate number of correct answer(s) based on item type:
  - only one clear, correct answer for a multiple-choice (MC) item
  - only the indicated number of correct answers for a multiple select (MS) item;
- have a correctly assigned content code (i.e., item map);
- measure one content standard or evidence statement;
- measure the content standard or evidence statement it is designed to measure;
- be at the appropriate level of rigor;
- be simple, direct, and free of ambiguity;
- make use of vocabulary and sentence structure that is appropriate for the grade level assessed;
- be based on content that is accurate and current;
- when appropriate, contain stimulus material that is clear and concise and provides all the necessary information;
- when appropriate, contain graphics that are clearly labeled;
- contain answer choices that are plausible and reasonable in terms of the requirements of the question as well as the student’s level of knowledge;
- contain distractors that relate to the question in the same way and can be supported by a rationale;
- reflect current teaching and learning practices in the content area; and
- be free of gender, ethnic, racial, cultural, socioeconomic, regional, and other forms of bias.

### 3.10.3 Item-Selection Options for Special Cases

While every effort is made to select a test form that meets all psychometric guidelines for excellence, it may not be possible to comply with all the psychometric criteria for item/form difficulty due to item pool limitations. In these cases, critical psychometric guidelines are followed while allowing some tolerance on less critical item-selection guidelines. The tolerance of meeting target characteristics, the relative exposure of previously used operational items, and other considerations (e.g., content coverage) may possibly be affected in such cases.

### 3.10.4 Psychometric Review

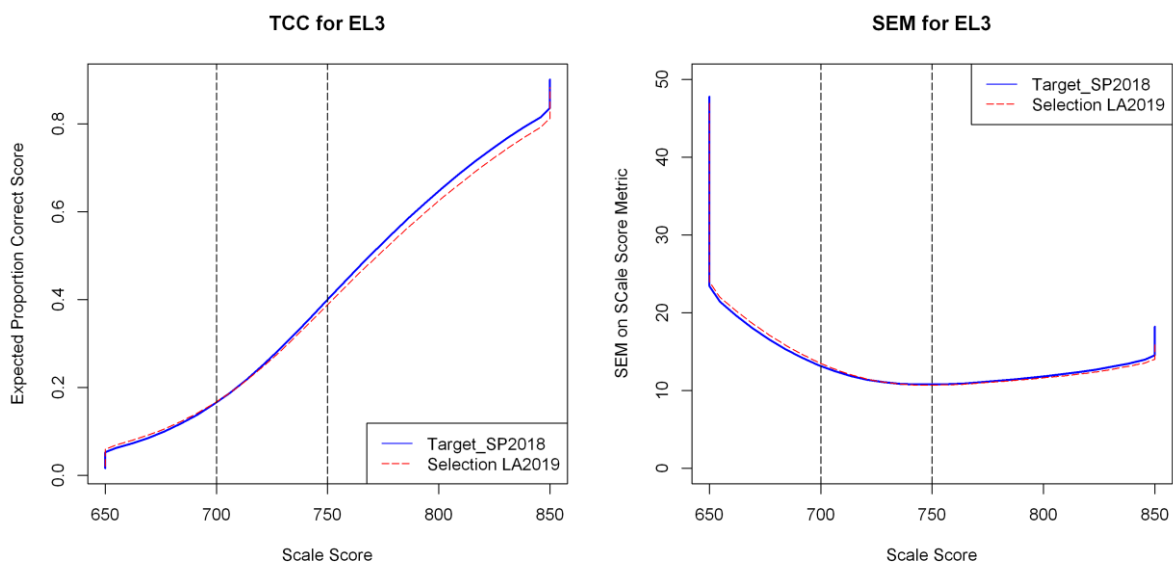
The psychometric evaluation of each selection was centered on reviewing the items with operational item parameters.

#### 3.10.4.1. Selecting Targets

The 2018 LEAP 2025 operational form was used as the target to select the items in the 2019 operational form. The rationale for the choice of the targets was that each 2018 LEAP 2025 form should be on the PARCC scale and be closely comparable to PARCC assessments, and using the 2018 forms as the target accomplishes that. Figures 3.1 through 3.6 for ELA and Figures 3.7 through 3.12 for mathematics show the test characteristic curves (TCCs) and standard errors of measurement (SEMs) of the final forms compared to those of the target forms. The left line graph displays the TCC of the target form and the selected 2019 form, summarizing the expected proportion of the maximum raw score needed to achieve the scale score. The

right line graph displays the SEM of the scale score of the target form and the selected 2019 form. This summarizes the amount of measurement error surrounding a scale score.

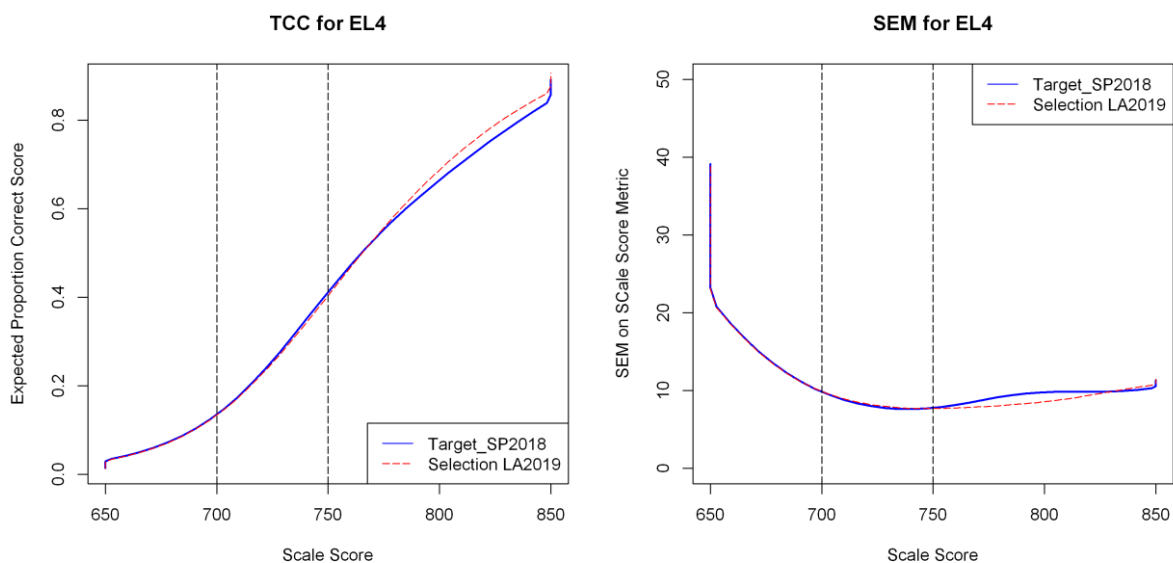
**Figure 3.1 2019 ELA Form Evaluation—Grade 3**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

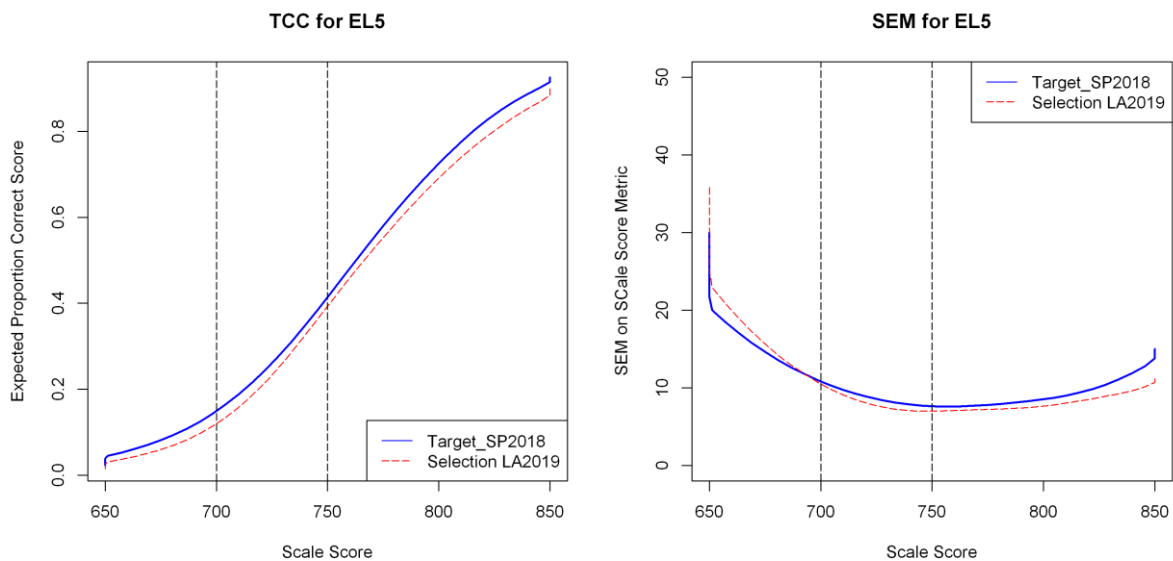
**Figure 3.2 2019 ELA Form Evaluation—Grade 4**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

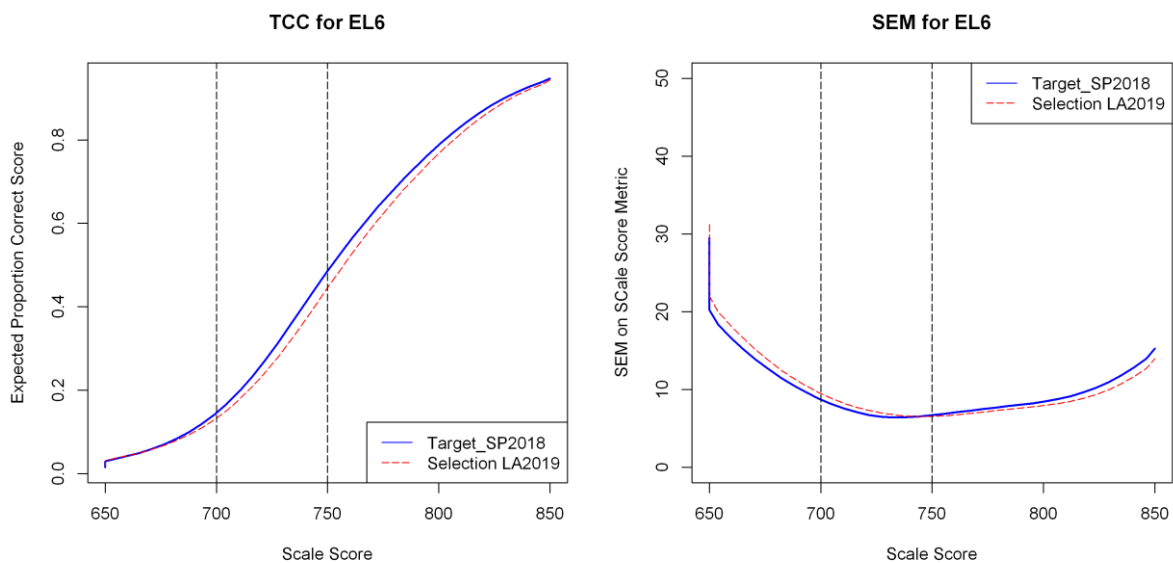
**Figure 3.3 2019 ELA Form Evaluation—Grade 5**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

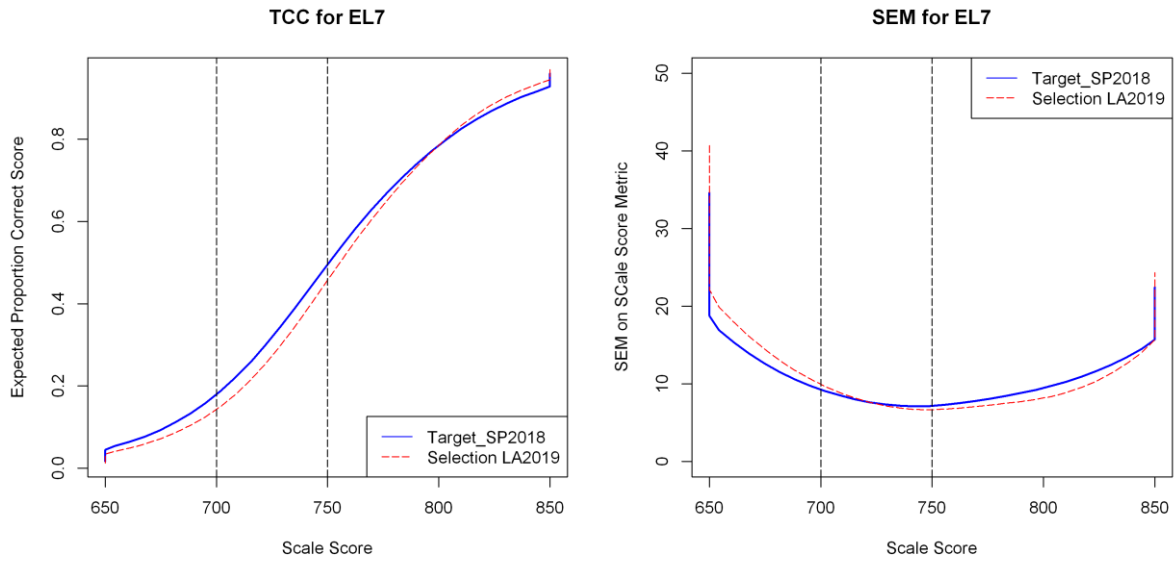
**Figure 3.4 2019 ELA Form Evaluation—Grade 6**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

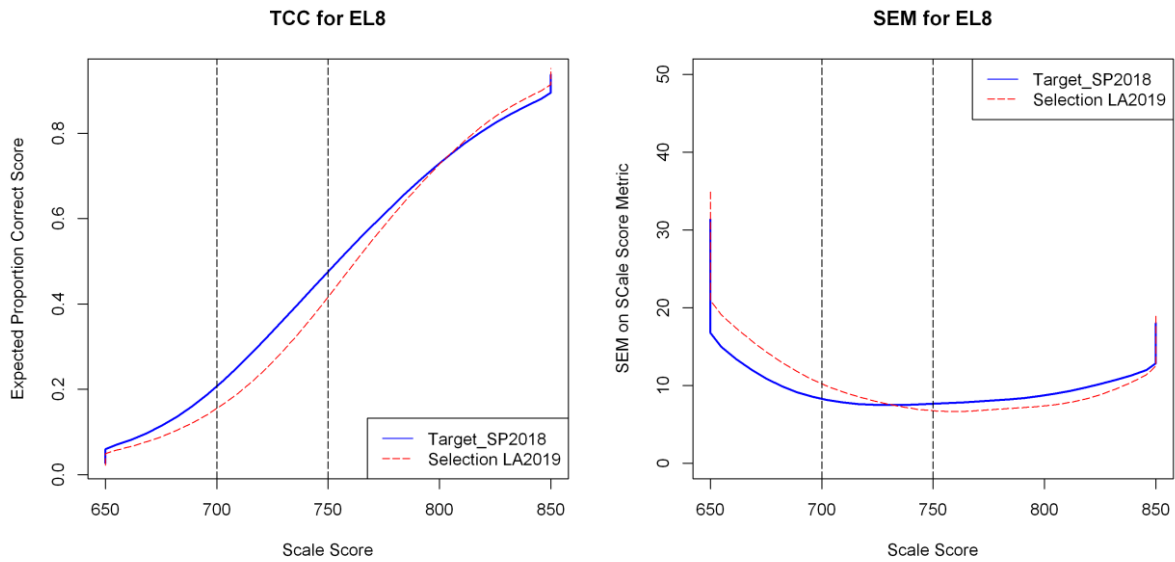
**Figure 3.5 2019 ELA Form Evaluation—Grade 7**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

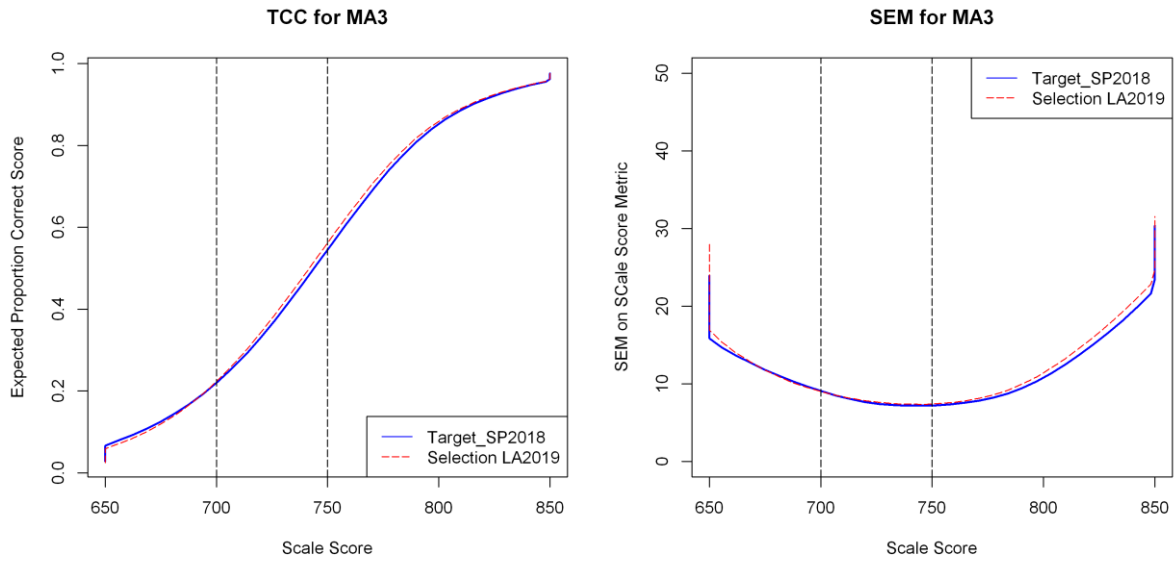
**Figure 3.6 2019 ELA Form Evaluation—Grade 8**



**NOTE:**

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

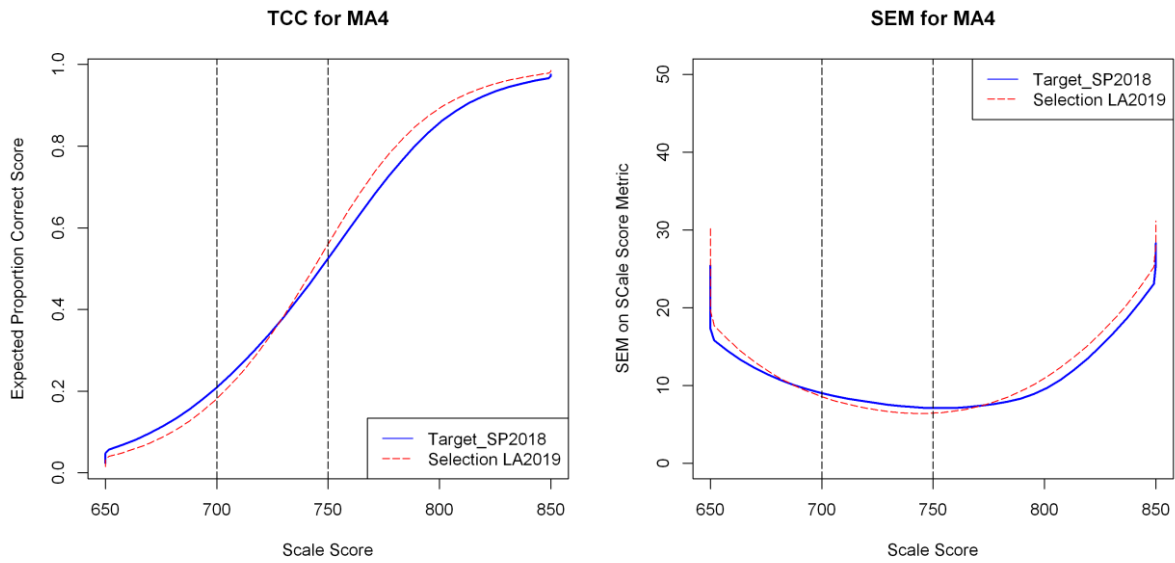
**Figure 3.7 2019 Mathematics Form Evaluation—Grade 3**



**NOTE:**

- *LEAP2018\_Target is the 2018 LEAP 2025 intact test form.*
- *Selection LA2019 is the selected 2019 LEAP 2025 test form.*

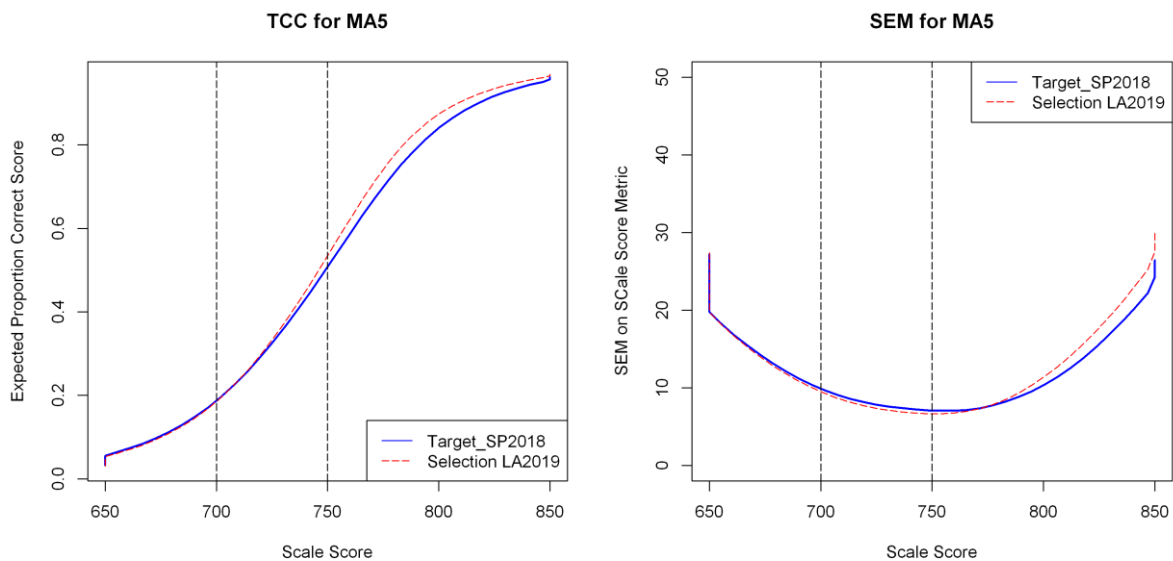
**Figure 3.8 2019 Mathematics Form Evaluation—Grade 4**



**NOTE:**

- *LEAP2018\_Target is the 2018 LEAP 2025 intact test form.*
- *Selection LA2019 is the selected 2019 LEAP 2025 test form.*

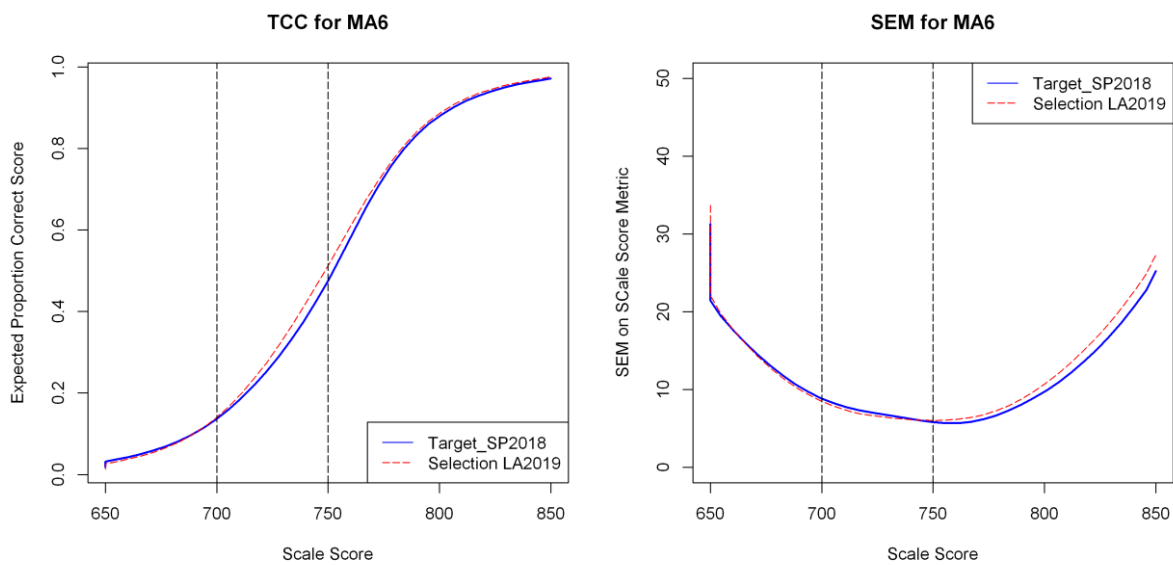
Figure 3.9 2019 Mathematics Form Evaluation—Grade 5



## NOTE:

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.

Figure 3.10 2019 Mathematics Form Evaluation—Grade 6

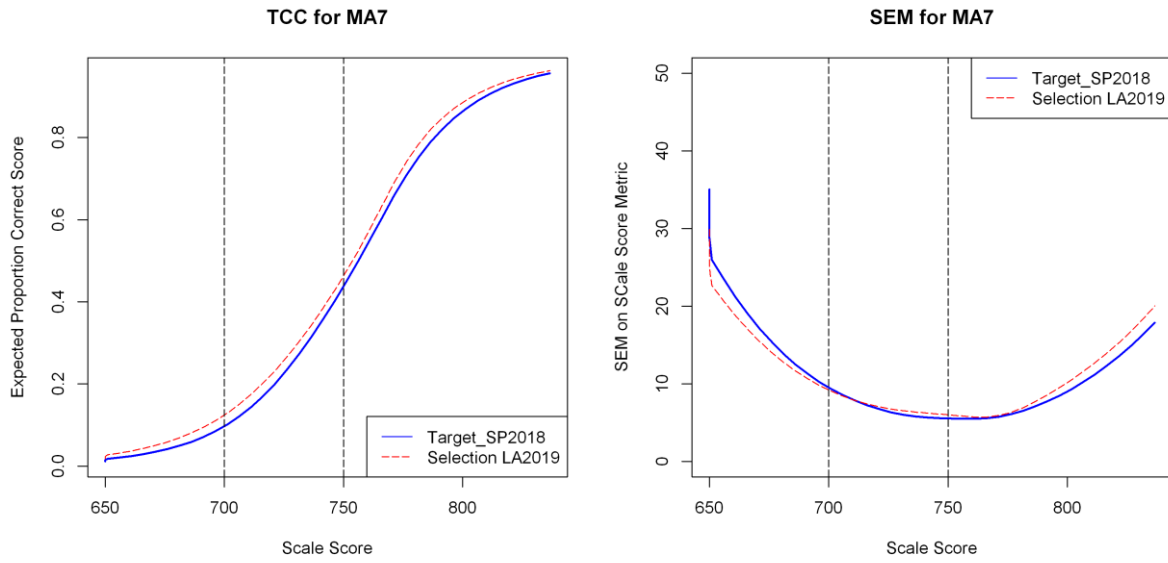


## NOTE:

- *LEAP2018\_Target* is the 2018 LEAP 2025 intact test form.
- *Selection LA2019* is the selected 2019 LEAP 2025 test form.



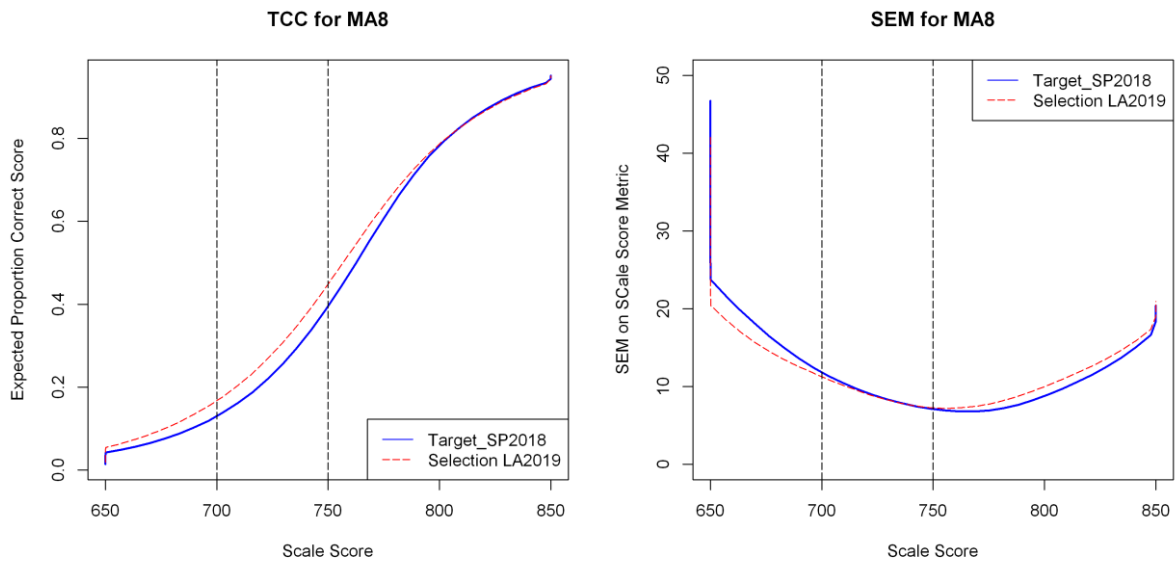
**Figure 3.11 2019 Mathematics Form Evaluation—Grade 7**



**NOTE:**

- LEAP2018\_Target is the 2018 LEAP 2025 intact test form.
- Selection LA2019 is the selected 2019 LEAP 2025 test form.

**Figure 3.12 2019 Mathematics Form Evaluation—Grade 8**



**NOTE:**

- LEAP2018\_Target is the 2018 LEAP 2025 intact test form.
- Selection LA2019 is the selected 2019 LEAP 2025 test form.

### 3.10.4.2. Selecting Anchors

Anchor sets used in the common item nonequivalent group design underwent considerable scrutiny due to the generally accepted guideline that the anchor set should mirror the total (or reference) test in terms of content and item characteristics. One of the critical psychometric considerations for an anchor set, other than individual item performance, is the extent to which the TCC and SEM of the anchor set aligns to that of the total test. This alignment is carefully considered during the item selection process. Three anchor sets were selected for post-equating purposes, all of which are representative of the full-form blueprint:

- Anchor 1: PARCC item parameters
  - Intact PARCC item parameters
  - Most items on test are part of anchor selection
- Anchor 2: LEAP 2025 item parameters
  - Estimated item parameters based on Louisiana student responses from LEAP 2025 administrations
  - Parameters on PARCC scale from previous years' post-equating
  - Approximately 12–14 items for ELA and 15 items for Math
- Anchor 3: Mixed item parameters
  - All Anchor 2 items and Anchor 1 items for the remaining items
  - Item parameters used for item selection

## 3.11 Universal Design

Grade-level assessments that follow universal design guidelines allow participation of the widest possible range of students, resulting in more valid inferences about students' performances. Such assessments may reduce the need for accommodations by reducing or eliminating access barriers associated with the tests themselves. Table 3.25 presents the elements of universal design (Thompson & Thurlow, 2002). The elements of universal design are relevant to both item development and form construction. This section describes how the elements of universal design were addressed in the construction of the Spring 2018 test forms in compliance with AERA, APA, & NCME (2014) Standard 3.1, which states the following:

Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

Universal design requires that grade-level assessments measure the performance of students with a wide range of abilities and skills, ensuring that students with diverse learning needs receive opportunities to demonstrate competence on the same content. To ensure that students can access the tests, the LEAP 2025 assessments include simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. All these design components were addressed primarily through the CBTs. The online test specifications define how directions and test items are formatted online, including the spacing between an item stem and answer choices, and other page elements (such as online tools and Help files) to ensure consistent, clean visual appearance of CBTs. Test directions at the beginning of each test session were clearly and simply stated, and the wording of such instructions is standardized as much as possible across content areas and grade levels to ensure clarity and consistency while being comparable to PARCC.

**Table 3.25 Elements of Universal Design**

Element	Explanation
Inclusive Assessment Population	Tests designed for state, school system, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Non-Biased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations (e.g., all items can be in braille form).
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	A variety of readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, tables, figures, illustrations, and response formats.

### 3.12 Accommodations and Designated Supports

AERA, APA, & NCME (2014) Standard 3.9 states the following:

Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs. (67)

Students with disabilities, students with 504 plans, and English learners (ELs) may be provided test administration accommodations as documented on their accommodation plan. More information on accommodations can be found in Section 4.3.2 of Chapter 4. Accommodation code definitions can be found in the *Paper-Based Test Administration Manual*.

Accommodated print forms were developed in grades 5–8 of ELA and mathematics for those students who were unable to participate in an online administration. For a detailed description of the process used to develop the accommodated print forms and how to modify technology-enhanced items for use in an accommodated print form, see Appendix D, *Accommodated Print Form Creation*.

Braille and large-print test forms were constructed for each grade and content area to enable students with visual impairments to participate in the LEAP 2025 assessments. Braille and large-print forms for grades 3 and 4 of ELA and mathematics were based on the standard-print forms. Braille forms for grades 5–8 of ELA and mathematics were based on the accommodated print forms. There are no large-print versions of the grades 5–8 accommodated print forms. Instead, students needing a large-print version in grades 5–8 use larger-sized monitors and/or the magnification features of the online testing system. All online test content has been developed to scale in relation to the available area on larger monitors while maintaining the correct aspect ratio. Specific recommendations on how to transcribe items into braille were provided by the braille

publisher to produce the braille version of the LEAP 2025 assessments and the test administrator’s notes that accompany the braille forms. The goal was to maximize the number of items on the braille forms that could be transcribed into braille.

The following assessment features were available to all students and do not require any documentation either prior to or during the assessment:

- blank scratch paper and graph paper
- calculators (to be used in the calculator section only)
- color overlay
- contrasting colors/reverse colors
- directions in native language
- equation builder
- bookmark
- general administration directions clarified
- general administration directions read aloud and repeated as necessary
- general masking
- headphones
- highlighters
- line guides
- magnifiers/variable zoom
- measurement tools
- redirection of student to the test
- specialized furniture or equipment
- sticky note/notepad
- strikethrough
- and writing/formatting tools (for ELA constructed response items only).

Accessibility features were available for all students with the particular need documented in their Individualized Education Programs (IEPs), Individual Accommodation Plans (IAPs), English Learner (EL) plans, or Personal Needs Profiles (PNPs). The following accessibility features were available: individual testing, small group testing, student reads assessment aloud to himself or herself, adaptive and specialized equipment or furniture, and math read aloud (text-to-speech or human reader).

Accommodations were available for students who have an IEP, IAP, or EL plan, including: braille test materials, calculation device and math tools for non-calculator sections of mathematics assessments, transferred answers, recorded answers, large print test materials (mathematics Spanish), mathematics Spanish read aloud, translated mathematics test, test read aloud (text-to-speech, Kurzweil, recorded audio file). For details on how these assessment and accessibility features and accommodations should be used with PBTs and CBTs, see the [LEAP 2025 Accommodations and Accessibility Features User Guide](#).

For a detailed description of the process used to develop the Spanish translation forms of the mathematics tests, see Appendix E, “Forms Development Process for Spanish Translations Forms.”

### 3.13 Item and Task Specifications

AERA, APA, & NCME (2014) Standard 4.12 states the following:

Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (89)

The item and task specifications are designed to ensure that the assessment items measure the assessment’s claims. The purpose of the item and task specifications is to define the characteristics of the items and tasks that will provide the evidence to support one or more claims. To do this, the item and task specifications delineate the types of evidence, or targets, that should be elicited for each reporting category within a grade level. Then, the specifications provide explicit guidance on how to write items to elicit the desired evidence. To address LEAP 2025 assessment comparability goals with PARCC 2019, PARCC claims, subclaims, and evidence statements, along with guidance provided by the *Louisiana Student Standards for ELA and Mathematics*, were used as item and task specifications.

The item and task specifications provide guidance on how to measure the targets (i.e., standards) first found in the content specifications and guidelines on how to create the items that are specific to each assessment target and reporting category. In ELA and mathematics, item specifications describe the knowledge, skills, and processes being measured by each item type aligned to particular standards.

These item specifications were developed for each grade and standard to delineate the expectations of knowledge and skill to be included on test questions. In addition, the ELA and mathematics item and stimulus specifications provide guidance on determining the appropriateness of task and stimulus materials (i.e., the materials that a student must refer to when working on a test question). The stimulus specifications also provide information on the characteristics of stimuli or activities that should be avoided because they are not important to the knowledge, skill, or process being measured. This underscores DRC’s efforts to select items that are accessible to the widest range of students possible; in other words, 2019 LEAP 2025 items were selected according to the elements of universal design.

### 3.14 Summary

In summary, the overall purpose of this chapter is to explicate the procedures used in the development of the 2019 LEAP 2025 grade-level assessments. The efforts by the LDOE and DRC in developing the LEAP 2025 assessments are in alignment with multiple best practices of the test industry but, in particular, support the following AERA, APA, & NCME (2014) standards:

**Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

**Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

**Standard 3.9** Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees’ ability to demonstrate their standing on the target constructs. (67)

**Standard 4.0** Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (85)

**Standard 4.1** Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended

uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (85)

**Standard 4.7** The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (87)

**Standard 4.12** Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (89)

## Chapter 4: Test Administration

---

Chapter 4 of the technical report describes the processes implemented and the information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), “The usefulness and interpretability of test scores require that a test be administered and scored according to the test developer’s instructions” (111). This chapter examines how test administration procedures implemented for the 2019 Louisiana Education Assessment Program (LEAP 2025) strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations.

Chapter 4 demonstrates how the LEAP 2025 assessments adhere to AERA, APA, & NCME (2014) Standards 4.15, 6.1, 6.2, 6.3, 6.4, 6.6, and 6.7. Each standard will be explicated within the relevant section of this chapter.

To ensure that the LEAP 2025 assessments are administered in accordance with the department’s mandates, the LDOE takes a primary role in communicating with and training school system personnel. The development of the assessments is a collaborative effort between the LDOE and DRC. The LDOE conveys to school systems the purpose of the assessments and the importance of test administration being consistent with test industry standards. The tests and administration standards must also meet the State Board of Elementary and Secondary Education policies and the mandates of both state and federal legislation.

To accomplish these goals, the LDOE provides train-the-trainer opportunities for school system test coordinators, who, in turn, administer test-administration training to schools within their school systems. The LDOE conducts quality assurance visits during testing to ensure that school systems adhere to the standardized administration of the tests.

The district test coordinators are responsible for the schools within their school systems. They disseminate information to each school, offer assistance with test administration, and serve as liaisons between the LDOE and their school systems. The LDOE also provides assistance with and interpretation of assessment data and test results.

Ancillary materials for the LEAP 2025 test administration contribute to the body of evidence of the validity of score interpretation. This section examines how the test materials address the standards related to test administration procedures.

For the Spring 2019 administration of the LEAP 2025 assessments, DRC produced the following administration manuals: *LEAP 2025 Grades 3 – 4 Paper-Based Test Administration Manual* and *LEAP 2025 Grades 3 – 8 Computer-Based Test Administration Manual* (TAMs). DRC also produced the following district Test Coordinators Manuals: *LEAP 2025 Computer-Based Test Coordinators Manual* and *LEAP 2025 Paper-Based Test Coordinators Manual* (TCMs). LDOE assessment administration and development staff review these manuals, provide feedback, and give final approval. The TCMs include ELA, mathematics, social studies, and science in grades 3 through 8. They provide detailed instructions for district and school test coordinators’ on distributing and collecting test materials and for returning them to DRC.

### Paper-Based Administration *Test Coordinators Manual* Table of Contents

1. Key Dates
2. Alerts
3. Oath of Security and Confidentiality Statements

4. General Information
5. LEAP 2025
6. Test Security
  - 6.1. Key Definitions
  - 6.2. Violations of Test Security
  - 6.3. Answer Change Analysis
  - 6.4. Voiding Student Tests
7. Testing Guidelines
  - 7.1. Testing Eligibility
  - 7.2. Testing Conditions
  - 7.3. Testing in Class-sized Groups
  - 7.4. Test Schedule
  - 7.5. Extended Time for Testing
  - 7.6. Extended Breaks
  - 7.7. Makeup Testing
  - 7.8. Test Administration Resources
  - 7.9. Testing Times
8. District Test Coordinator
  - 8.1. Conduct Training Session
  - 8.2. Receive Test Materials
  - 8.3. Large-print, Braille, and CAS Test Materials
  - 8.4. Accommodated Materials
  - 8.5. Verify and Distribute Test Materials to School Test Coordinators
  - 8.6. Request Additional Test Materials and Bar-code Labels
  - 8.7. Collect Materials from Schools After Testing
  - 8.8. Used and Unused Consumable Test Booklets (Defined)
  - 8.9. Unscorable Documents and Unscorable Document Labels
9. Directions for Returning Test Materials to DRC in May
  - 9.1. Pickup 1
  - 9.2. Pickup 2
  - 9.3. Pickup 3
  - 9.4. Final Checklist for Returning Test Materials to DRC
10. School Test Coordinator
  - 10.1. Receive and Verify Test Materials
  - 10.2. Conduct Test Administration and Security Training Session
  - 10.3. Supervise Application of Bar-code Labels and Coding of Consumable Test Booklets
  - 10.4. Soiled, Damaged, and Other Unscorable Consumable Test Booklets
  - 10.5. Verify and Distribute Materials to Test Administrators
  - 10.6. Supervise Test Administration
  - 10.7. Collect Test Materials
  - 10.8. Used and Unused Consumable Test Booklets (Defined)
  - 10.9. Coding Responsibilities of Principals—Before Testing
  - 10.10. Coding Responsibilities of Principals—Before and After Testing
  - 10.11. Coding Responsibilities of Principals—After Testing
11. Directions for Returning Test Materials to the DTC
  - 11.1. Pickup 1
  - 11.2. Pickup 2
  - 11.3. Pickup 3
  - 11.4. Final Checklist for Returning Materials to the DTC
12. Void Notification



## 13. Index

### Computer-Based Administration *Test Coordinators Manual* Table of Contents

1. Key Dates
2. Resources Available in eDIRECT Spring 2019
3. Alerts
4. Oath of Security and Confidentiality Statements
5. General Information
  - a. eDIRECT and INSIGHT
6. LEAP 2025
7. Test Security
  - a. Key Definitions
  - b. Violations of Test Security
8. Testing Guidelines
  - a. Testing Eligibility
  - b. Testing Conditions
  - c. Testing in Class-sized Groups
  - d. Test Schedule
  - e. Extended Time for Testing
  - f. Extended Breaks
  - g. Makeup Testing
  - h. Test Administration Resources
9. Testing Times
10. Roles and Responsibilities
  - a. District Test Coordinator
  - b. School Test Coordinator
  - c. Technology Coordinator
11. Managing Test Tickets
  - a. Student Transfers
  - b. Locked Test Tickets
  - c. Technical Issues
  - d. Invalidating Test Tickets
12. Resources for Online Testing
  - a. Test Administration Manuals
  - b. eDIRECT User Guides
  - c. LEAP 2025 Accommodations and Accessibility Features User Guide
  - d. INSIGHT Technology User Guide
  - e. Online Tools Training (OTT)
  - f. Student Tutorials

The TAMs are specific to grades, content areas, and modes of administration (i.e., online or paper). They provide detailed instructions for administering the LEAP 2025 assessments. The manuals include instructions for test security, test administrator responsibilities, test preparation, administration of tests (i.e., online or paper), and post-test procedures. Information included in the TAMs is listed below.

## Paper Administration Table of Contents

1. Spring Notes and Reminders
2. Test Administrator Oath of Security and Confidentiality Statements
3. Overview
4. Test Security
  - 4.1. Secure Test Materials
  - 4.2. Testing Irregularities and Security Breaches
  - 4.3. Testing Environment
  - 4.4. Violations of Test Security
  - 4.5. Answer Change Analysis
  - 4.6. Voiding Student Tests
5. Test Administrator Responsibilities
6. Test Administration Checklists
  - 6.1. Before Testing
  - 6.2. During Testing
  - 6.3. After Testing (Daily)
  - 6.4. After Testing (Last Day)
7. Test Administrators' Frequently Asked Questions
8. Test Materials
  - 8.1. Receipt of Test Materials
9. Testing Guidelines
  - 9.1. Testing Eligibility
  - 9.2. Test Schedule
  - 9.3. Extended Time for Testing
10. Testing Times for Grades 3–4
  - 10.1. Makeup Testing
  - 10.2. Testing Conditions
11. Special Populations and Accommodations
  - 11.1. IDEA Special Education Students
  - 11.2. Students with One or More Disabilities According to Section 504
  - 11.3. Gifted and Talented Special Education Students
  - 11.4. Test Accommodations for Special Education and Section 504 Students
  - 11.5. Special Considerations for Deaf and Hard of Hearing Students
  - 11.6. English Learner (EL) Students
12. Hand-coded Consumable Test Booklets
13. Students Absent from Testing
14. Consumable Test Booklet Coding
  - 14.1. Coding the Demographic Section
15. Sample Grade 3 English Language Arts Consumable Test Booklet
16. General Instructions for LEAP 2025
  - 16.1. Student Marking/Erasing on Consumable Test Booklet
  - 16.2. Reading Directions to Students
  - 16.3. Special Instructions
17. Directions for Administering LEAP 2025 Tests

18. Post-test Procedures
  - 18.1. Test Administrator Oath of Security and Confidentiality Statement
  - 18.2. Used and Unused Consumable Test Booklets (Defined)
  - 18.3. Transferring Student Responses
  - 18.4. Returning Test Materials to the School Test Coordinator
19. Index

#### Online Administration Table of Contents

1. Spring Notes and Reminders
2. Test Administrator Oath of Security and Confidentiality Statements
3. Overview
4. Test Security
  - 4.1. Secure Test Materials
  - 4.2. Testing Irregularities and Security Breaches
  - 4.3. Testing Environment
  - 4.4. Violations of Test Security
  - 4.5. Voiding Student Tests
5. Test Administrator Responsibilities
  - 5.1. Software Tools and Features for Test Administrators
6. Test Administration Checklists
  - 6.1. Before Testing
  - 6.2. During Testing
  - 6.3. After Testing (Daily)
  - 6.4. After Testing (Last Day)
7. Test Administrators' Frequently Asked Questions
8. Testing Guidelines
  - 8.1. Testing Eligibility
  - 8.2. Test Schedule
  - 8.3. Extended Time for Testing
9. Testing Times for Grades 3–8
  - 9.1. Makeup Testing
  - 9.2. Testing Conditions
10. Online Tools Training
11. Student Tutorials
12. Directions for Administering the Grades 3–8 LEAP 2025 Tests
13. Special Populations and Accommodations
  - 13.1. IDEA Special Education Students
  - 13.2. Students with One or More Disabilities According to Section 504
  - 13.3. Gifted and Talented Special Education Students
  - 13.4. Test Accommodations for Special Education and Section 504 Students
  - 13.5. Special Considerations for Deaf and Hard of Hearing Students
  - 13.6. English Learner (EL)
14. Students Absent from Testing
15. Test Materials
  - 15.1. Receipt Directions to Students
16. Post-test Procedures
  - 16.1. Test Administrator Oath of Security and Confidentiality Statement
  - 16.2. Returning Test Materials to the School Test Coordinator
17. Index

The *Standards* contain multiple references that are relevant to test administration. Information in the TAMs addresses these standards.

The directions for test administration found in the manual address Standard 4.15, which states:

The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented. (90)

The LEAP 2025 Test Administration Manuals provide instructions for activities conducted before, during, and after testing with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the manuals describe the following: general rules of paper and online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

Furthermore, the standardized procedures addressed in the test administration manual need to be followed, as the *Standards* state in Standard 6.1:

Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user (114).

It was essential that the LEAP 2025 was administered according to the prescribed test administration manual to ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance. It should be noted that adhering to the test schedule is also a critical component. The test administration manuals include instructions for scheduling the test within the state testing window. The test administration manual also contains the schedule for timing each test session. The test timing schedule is presented in Table 4.1.

**Standard 6.3** Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (115)

The LDOE test administration staff reports on testing concerns that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; using a calculator on parts of the test where it is not allowed; allowing students to correct or complete answers after tests have been submitted; splitting sessions into two parts; ignoring the standardized directions in the online assessment; reading the ELA assessment to students with the exception of those students with the read-aloud accommodation; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the accommodation plan; allowing accommodations for students who do not have an accommodation plan; or defining terms on the test.

**Standard 6.4** The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (116)

Test administration manuals outline the steps that teachers should take to prepare classroom environment testing for administering the LEAP 2025 assessments. These steps include the following:

- Determine the layout of the classroom environment.

- Plan seating arrangements. Allow enough space between students to prevent the sharing of answers.
- Eliminate distractions such as bells or telephones.
- Use a Do Not Disturb sign on the door of the testing room.
- Make sure classroom maps, charts, and any other materials that relate to the content and processes of the test are covered, removed, or out of the students' view.

**Standard 6.6** Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)

The test administration manuals present instructions for post-test activities to ensure that online tests are submitted and printed test materials are handled properly to maintain the integrity of student information and test scores. Detailed instructions guide test examiners in submitting all online test records. For students who were administered a large-print or braille test form, examiners are instructed to transcribe students' responses from the large-print test or braille test form into a consumable test booklet for grades 3 and 4, and the online testing system (INSIGHT) for grades 5 through 8, exactly as the responses appear in the original form.

**Standard 6.7** Test users have the responsibility of protecting the security of test materials at all times. (117)

Throughout the manuals, test coordinators and examiners are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted. Detailed information about test security procedures are presented under "Test Security" in the test administration manuals.

## 4.1 Return Material Forms and Guidelines

The *Test Coordinators Manual* instructs test coordinators on how to organize, pack, and return testing materials to DRC for secure inventory purposes. The LDOE assessment administration and development staff have opportunities to review these materials, provide feedback, and give final approval. The purpose of the instructions is to ensure the secure test materials are properly accounted for and organized appropriately for return shipment.

## 4.2 Security Checklists

As soon as printed test materials are received by a school system, the district test coordinator confirms the receipt and count of the school system materials and completes the Receipt Notice in eDIRECT to confirm all school system materials have been received. The district test coordinator then packages the tests to be sent to schools. Upon returning secure test materials to DRC, district test coordinators are required to complete and submit a materials accountability form that details the number of consumable test booklets or secure accommodated test materials returned. This materials accountability form also requires that school systems document nonstandard situations, including lost, damaged, destroyed, extra, or missing test books. This form ensures all materials are accounted for. Any material not accounted for on this form is placed on a missing materials list which is used by DRC and the LDOE to follow up with all districts to ensure security of all materials. A sample accountability form is shown in Figure 4.1.

**Figure 4.1 Sample Accountability Form**

Administration  District  School

Enter Counts | Summary | Status Report

Accountability Form Data for District 999 has been completed. You may continue making changes through the end of the accountability form window.

Reference the *Instructional Text* below for the reasons for any return material discrepancies.

[Instructions](#)

This form may be updated throughout the testing window, but it MUST be completed by the end of the testing window when all materials have been returned to Data Recognition Corporation.  
 All secure materials received from Data Recognition Corporation should be included in the box counts provided in the "Returned to DRC" column.  
 Any secure documents (test booklets, answer documents, or consumable test booklets) soiled with bodily fluids must be listed in the "Record reasons for discrepancies here:" field to ensure they are not reported as missing materials.  
 Always provide both the security barcode number AND the date the document was destroyed.

Accountability Form for <input type="text"/>		Exact Number of Boxes Shipped to DRC
Science and ELA/Math Test Materials		
Pickup 1: UPS Ground Service (automatic pickup date)	<b>SCORABLE MATERIALS:</b>	5
	Used Science answer documents	
	Used ELA and Math consumable test booklets	
Pickup 2: UPS Ground Service (automatic pickup date)	<b>SCORABLE MATERIALS:</b>	
	Used Science makeup answer documents	
	Used ELA/Math makeup consumable test booklets	
	Used Science answer documents and ELA/Math consumable test booklets for home study program students	
	Used ELA/Math consumable test booklets for nonpublic school students	
	Accountability-coded answer documents and consumable test booklets	
	<b>NONSCORABLE MATERIALS:</b>	
Pickup 3: Assessment Distribution Services (ADS)	<b>NONSCORABLE MATERIALS:</b>	
	All unused bar-code labels for Science and ELA/Math	
	All used and unused Science test booklets, including large print and braille	
	All ELA and Math large print and braille test booklets	

Accountability Form for <input type="text"/>		Exact Number of Boxes Shipped to DRC
Social Studies Test Materials		
Pickup 1: UPS Ground Service (automatic pickup date)	<b>SCORABLE AND NONSCORABLE MATERIALS:</b>	
	All used consumable test booklets	
	All used consumable test booklets for homestudy students	
	All unused consumable test booklets	
	All used and unused large-print and braille test booklets	

Record reasons for discrepancies here:

Enter Counts | Summary | Status Report

[Instructions](#)

Previously entered accountability form data will display. The accountability form summary information can be printed by clicking the **Print** button.  
 Note: The accountability form summary information is view only and cannot be edited.

Summary for District [REDACTED]		
Science and ELA/Math Test Materials		Exact Number of Boxes Shipped to DRC
Pickup 1: UPS Ground Service (automatic pickup date)	<b>SCORABLE MATERIALS:</b>	5
	Used Science answer documents	
	Used ELA and Math consumable test booklets	
Pickup 2: UPS Ground Service (automatic pickup date)	<b>SCORABLE MATERIALS:</b>	
	Used Science makeup answer documents	
	Used ELA/Math makeup consumable test booklets	
	Used Science answer documents and ELA/Math consumable test booklets for home study program students	
	Used ELA/Math consumable test booklets for nonpublic school students	
	Accountability-coded answer documents and consumable test booklets	
	<b>NONSCORABLE MATERIALS:</b>	
	All unused Science answer documents	
Pickup 3: Assessment Distribution Services (ADS)	<b>NONSCORABLE MATERIALS:</b>	
	All unused bar-code labels for Science and ELA/Math	
	All used and unused Science test booklets, including large print and braille	
	All ELA and Math large print and braille test booklets	

Summary for District [REDACTED]		
Social Studies Test Materials		Exact Number of Boxes Shipped to DRC
Pickup 1: UPS Ground Service (automatic pickup date)	<b>SCORABLE AND NONSCORABLE MATERIALS:</b>	
	All used consumable test booklets	
	All used consumable test booklets for homestudy students	
	All unused consumable test booklets	
	All used and unused large-print and braille test booklets	

Record reasons for discrepancies here:

[Print](#)

Enter Counts | **Summary** | Status Report

[Instructions](#)

The progress status of the accountability form is displayed at the district level. Use this key to evaluate the status for your site:

- Not Started – District has not completed data entry
- Completed – District has completed data entry

The accountability form status can be exported to Excel by clicking the **Export to Excel** button.

[Click here](#) to access a report of Users that clicked the Complete button and their information.

Overall Status for District [REDACTED]	
District	Status
[REDACTED]	Completed

[Export to Excel](#)

### 4.3 Interpretive Guides

An understanding of what test scores mean and how to interpret score reports is essential to making valid interpretations of the test scores. The *Interpretive Guide* is written for Louisiana teachers and administrators who receive the LEAP 2025 score reports. More details about the guide can be found in Chapter 7.

### 4.4 Test Security Measures

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors, such as unauthorized exposure of test items that would affect the valid interpretation of test scores. Several test security measures are implemented for the LEAP 2025 assessments. Test security procedures are discussed throughout the Test Coordinators Manuals and Test Administration Manuals.

Test coordinators and administrators are instructed to keep all test materials in locked storage, except during actual test administration, and access to secure materials must be restricted to authorized individuals only (e.g., test administrators and the school test coordinator). During testing sessions, the test administrators are directly responsible for the security of the LEAP 2025 assessments and must account for all test materials and supervise the test administration at all times.

#### 4.4.1 Data Forensic Analyses

Due to the importance of the LEAP 2025 assessment, it is prudent to ensure that the results from the assessments are based on effective instruction and true student achievement. While there are many ways to achieve meaningful understanding of student knowledge via test scores, there are also ways to obtain higher test scores that are not related to actual learning. To assist ensuring that assessment results are valid, data forensic analyses are conducted to help separate meaningful gains from spurious gains. It is important to note that although the results may be used to identify potential problems within a school, the identification of a problem is not an accusation of misconduct.

Multiple methods were incorporated into the forensic analysis. The following methods were applied:

- Response-Change Analysis
- Score Change Analysis
- Web Monitoring
- Plagiarism Detection

#### 4.4.2 Response Change Analysis

Students make changes to answer choices when taking the LEAP 2025, and this is expected behavior. Unfortunately, changing student answers is also an opportunity for school personnel to improve classroom performance and, therefore, the response change analysis focuses on identifying school- and test-administrator level response-change patterns that are statistically improbable when compared to the expected pattern at the state level.

#### 4.4.3 Score Fluctuation Analysis

It is anticipated that performance on the LEAP 2025 will improve over time from legitimate sources such as changes in the curriculum and improvement in instruction. However, large and unexpected score changes may be a sign of testing impropriety. The LDOE applied an approach where the state's level of change in performance from one year to the next is compared to a schools' and test administrators' change in performance during the same time frame. Schools and test administrators were identified when the level of change was statistically unexpected.



#### 4.4.4 Web Monitoring

LEAP 2025 operational test content should not appear outside the boundaries of the forms administered. To protect Louisiana test content, the internet is monitored for postings which contain, or appear to contain, potentially exposed and/or copied LDOE test content. When test content is verified, steps are taken so that the infringing content is removed quickly.

#### 4.4.5 Plagiarism Detection

The LDOE monitors for two different plagiarism situations: copying from student to student and copying from an outside source, such as Wikipedia or another internet sources. Instances of plagiarism are identified regardless if an item is scored by human scorers or artificial intelligence. Alerts are set to identify responses that may indicate the possibility of teacher interference, plagiarism, or disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). Alerted responses are given additional review so the appropriate response can be taken.

### 4.5 Test Administration

The 2019 assessments were administered to students within the state testing window of April 1 through May 3, 2019. The paper testing window was April 29 through May 3, 2019. Each session of the assessment within each content area of the LEAP 2025 assessments was required to be administered in one block of time.

#### 4.5.1 Time

All sessions of the ELA and mathematics LEAP 2025 assessments were timed. Only students with an extended time accommodation were permitted to exceed the established time limits of any given session. The timing schedule of the LEAP 2025 assessments is presented in Table 4.1.

**Table 4.1 LEAP 2025 Administration Schedule Timing Guidelines by Session (Time in Minutes)**

Grade	Session	English Language Arts	Mathematics
3	1	75	75
	2	75	85
	3	60	75
4	1	90	75
	2	90	85
	3	60	75
5	1	90	75
	2	90	85
	3	60	75
6	1	90	60
	2	90	90
	3	80	90
7	1	90	60
	2	90	90
	3	80	90
8	1	90	60
	2	90	90
	3	80	90

### 4.5.2 Accommodations

Accommodations are allowed on the LEAP 2025 assessments. Accommodations may be used by a student who qualifies under the Individual with Disabilities Act (IDEA), has an IEP or a Section 504 plan of the Americans with Disabilities Act, or identifies as an English learner (EL). Accommodations must be specified in the qualifying student's individual plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student information sheet at the time of test administration. AERA, APA, & NCME Standard 6.2 states:

When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (115)

In compliance with this standard, the LEAP 2025 *Test Administration Manual* contains the list of universal tools, designated supports, and accommodations permissible for the LEAP 2025 assessments. Further guidance can be found in the [LEAP 2025 Accommodations and Accessibility Features User Guide](#).

Visually impaired students may be provided braille forms for any assessment and large print forms for the PBT.

Tables 4.2 through 4.5 summarize the numbers of reportable students receiving accommodations by accommodation type for the 2019 LEAP 2025. Accommodation assignment guidance is provided in the LEAP 2025 Accommodations and Accessibility User Guide. Accommodations are grouped into four sections: special education accommodation, English learner status accommodation, Section 504 status accommodation, and online accommodation. The analyses are based on census data and the number includes only those students who received accommodations and received a scale score on the ELA or mathematics LEAP 2025 assessments. The percentage represents the percentage of the census population receiving that accommodation. The students who are included in the “No Accommodation” category are students who are eligible for an accommodation but have indicated that none was used.

**Table 4.2 Number and Percentage of Students Receiving Special Education Accommodations by Accommodation Type, as Bubbled on the Test Booklet**

Special Education Accommodation Type					
		English Language Arts		Mathematics	
Grade	Accommodation	Number	Percentage	Number	Percentage
3	No Accommodation	≥2,130	4.16%	≥2,100	4.09%
3	Braille	<50	NR	<50	NR
3	Large Print	<50	NR	<50	NR
3	Answers Recorded	≥640	1.25%	≥630	1.24%
3	Extended Time	≥4,630	9.01%	≥4,660	9.08%
3	Transferred Answers	≥200	0.39%	≥200	0.40%
3	Individual/Small Group Administration	≥4,500	8.76%	≥4,520	8.81%
3	Tests Read Aloud	≥3,270	6.37%	≥3,770	7.36%
4	No Accommodation	≥1,910	4.05%	≥1,910	4.06%
4	Braille	<50	NR	<50	NR
4	Large Print	<50	NR	<50	NR
4	Answers Recorded	≥550	1.18%	≥540	1.16%
4	Extended Time	≥4,610	9.78%	≥4,610	9.79%
4	Transferred Answers	≥250	0.53%	≥250	0.53%
4	Individual/Small Group Administration	≥4,430	9.39%	≥4,430	9.40%
4	Tests Read Aloud	≥3,260	6.92%	≥3,720	7.90%

**Table 4.3 Number and Percentage of Students Receiving English Learner Accommodations by Accommodation Type, as Bubbled on the Test Booklet**

EL Accommodation Type					
Grade	Accommodation	English Language Arts		Mathematics	
		Number	Percentage	Number	Percentage
3	No Accommodation	≥280	0.55%	≥260	0.51%
3	Extended Time	≥2,050	3.99%	≥1,990	3.89%
3	Individual/Small Group Administration	≥1,530	2.99%	≥1,500	2.92%
3	English/Native Language Word-to-Word Dictionary	≥290	0.58%	≥250	0.5%
3	Test Administered by ESL Teacher	≥170	0.34%	≥150	0.3%
3	Directions Read Aloud/Clarified in Native Language	≥110	0.23%	≥80	0.16%
4	No Accommodation	≥210	0.46%	≥190	0.42%
4	Extended Time	≥1,520	3.24%	≥1,490	3.16%
4	Individual/Small Group Administration	≥1,090	2.33%	≥1,060	2.26%
4	English/Native Language Word-to-Word Dictionary	≥340	0.72%	≥270	0.57%
4	Test Administered by ESL Teacher	≥130	0.28%	≥80	0.19%
4	Directions Read Aloud/Clarified in Native Language	≥110	0.24%	≥60	0.13%

**Table 4.4 Number and Percentage of Students Receiving Section 504 Status by Accommodation Type, as Bubbled on the Test Booklet**

Section 504 Status Accommodation Type					
		English Language Arts		Mathematics	
Grade	Accommodation	Number	Percentage	Number	Percentage
3	No Accommodation	≥330	0.65%	≥4,160	8.12%
3	Large Print	<50	NR	<50	NR
3	Answers Recorded	≥120	0.23%	≥120	0.24%
3	Extended Time	≥3,730	7.27%	≥3,740	7.30%
3	Transferred Answers	<50	NR	<50	NR
3	Individual/Small Group Administration	≥3,080	6.00%	≥3,100	6.06%
3	Tests Read Aloud	≥1,310	2.55%	≥1,580	3.08%
4	No Accommodation	≥340	0.74%	≥4,780	10.14%
4	Large Print	<50	NR	<50	NR
4	Answers Recorded	≥100	0.22%	≥90	0.21%
4	Extended Time	≥4,370	9.25%	≥4,350	9.23%
4	Transferred Answers	≥50	0.12%	≥60	0.13%
4	Individual/Small Group Administration	≥3,550	7.53%	≥3,560	7.56%
4	Tests Read Aloud	≥1,550	3.28%	≥1,130	2.40%

**Table 4.5 Number and Percentage of Students Receiving Online Accommodations by Accommodation Type, as valued in eDIRECT**

Online Accommodation Type					
Grade	Accommodation	English Language Arts		Mathematics	
		Number	Percentage	Number	Percentage
3	Text-to-Speech	≥160	10.88%	≥280	18.59%
3	Human Read Aloud	≥60	4.17%	≥80	5.32%
3	Native Language Word-to-Word Dictionary	≥50	3.65%	<50	NR
3	Directions in Native Language	<50	NR	<50	NR
3	Transferred Answers	<50	NR	<50	NR
3	Answers Recorded	<50	NR	<50	NR
3	Extended Time	≥350	23.00%	≥330	22.21%
3	Individual/Small Group Administration	≥310	20.39%	≥310	20.57%
3	Accommodated Paper	<50	NR	<50	NR
3	Braille	<50	NR	<50	NR
3	Communication Assistance Scripts	<50	NR	<50	NR
4	Text-to-Speech	≥620	8.20%	≥1,370	18.27%
4	Human Read Aloud	≥280	3.71%	≥340	4.62%
4	Native Language Word-to-Word Dictionary	≥110	1.45%	≥100	1.32%
4	Directions in Native Language	<50	NR	<50	NR
4	Transferred Answers	<50	NR	<50	NR
4	Answers Recorded	≥110	1.53%	≥110	1.54%
4	Extended Time	≥1,700	22.48%	≥1,690	22.39%
4	Individual/Small Group Administration	≥1,520	20.06%	≥1,530	20.39%
4	Accommodated Paper	<50	NR	<50	NR
4	Braille	<50	NR	<50	NR
4	Communication Assistance Scripts	<50	NR	<50	NR
5	Text-to-Speech	≥5,320	9.69%	≥8,560	15.65%
5	Human Read Aloud	≥1,990	3.63%	≥2,360	4.32%
5	Native Language Word-to-Word Dictionary	≥430	0.80%	≥360	0.66%
5	Directions in Native Language	≥170	0.32%	≥110	0.21%
5	Transferred Answers	≥300	0.55%	≥300	0.56%
5	Answers Recorded	≥690	1.27%	≥690	1.27%
5	Extended Time	≥12,520	22.81%	≥12,450	22.75%
5	Individual/Small Group Administration	≥10,280	18.72%	≥10,260	18.76%
5	Accommodated Paper	<50	NR	<50	NR
5	Braille	<50	NR	<50	NR
5	Communication Assistance Scripts	<50	NR	<50	NR

Online Accommodation Type					
		English Language Arts		Mathematics	
Grade	Accommodation	Number	Percentage	Number	Percentage
6	Text-to-Speech	≥5,260	9.60%	≥7,800	14.27%
6	Human Read Aloud	≥750	1.38%	≥940	1.72%
6	Native Language Word-to-Word Dictionary	≥810	1.48%	≥680	1.26%
6	Directions in Native Language	≥260	0.47%	≥200	0.38%
6	Transferred Answers	≥130	0.24%	≥120	0.23%
6	Answers Recorded	≥270	0.50%	≥270	0.50%
6	Extended Time	≥11,770	21.49%	≥11,710	21.40%
6	Individual/Small Group Administration	≥8,290	15.13%	≥8,320	15.21%
6	Accommodated Paper	<50	NR	<50	NR
6	Braille	<50	NR	<50	NR
6	Communication Assistance Scripts	<50	NR	<50	NR
7	Text-to-Speech	≥5,070	9.70%	≥7,240	13.91%
7	Human Read Aloud	≥570	1.09%	≥730	1.41%
7	Native Language Word-to-Word Dictionary	≥780	1.50%	≥600	1.17%
7	Directions in Native Language	≥250	0.48%	≥180	0.35%
7	Transferred Answers	≥70	0.15%	≥70	0.15%
7	Answers Recorded	≥160	0.31%	≥160	0.32%
7	Extended Time	≥11,120	21.24%	≥10,950	21.03%
7	Individual/Small Group Administration	≥7,320	13.99%	≥7,280	13.99%
7	Accommodated Paper	<50	NR	<50	NR
7	Braille	<50	NR	<50	NR
7	Communication Assistance Scripts	<50	NR	<50	NR
8	Text-to-Speech	≥4,640	9.15%	≥6,700	15.05%
8	Human Read Aloud	≥520	1.03%	≥580	1.32%
8	Native Language Word-to-Word Dictionary	≥870	1.73%	≥730	1.64%
8	Directions in Native Language	≥230	0.47%	≥170	0.39%
8	Transferred Answers	≥70	0.15%	≥70	0.17%
8	Answers Recorded	≥130	0.26%	≥120	0.29%
8	Extended Time	≥10,630	20.96%	≥10,250	23.04%
8	Individual/Small Group Administration	≥6,660	13.14%	≥6,480	14.57%
8	Accommodated Paper	<50	NR	<50	NR
8	Braille	<50	NR	<50	NR
8	Communication Assistance Scripts	<50	NR	<50	NR

## 4.6 Summary

In summary, the overall purpose of each of the test administration trainings and the ancillary materials is to keep school systems informed about policies and procedures related to testing in general and the LEAP 2025 program in particular. The information imparted is clearly related to standardizing the administration of the LEAP 2025, maintaining the security of the assessment, allowing access to the assessments for special

populations by clearly delineating appropriate accommodations, and maintaining integrity of the scores. These communication and training efforts by the LDOE and the ancillary information developed by DRC address multiple best practices of the testing industry but, in particular, are related to the following standards:

**Standard 4.15** The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented. (90)

**Standard 6.1** Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (114)

**Standard 6.3** Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (115)

**Standard 6.4** The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (116)

**Standard 6.6** Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)

**Standard 6.7** Test users have the responsibility of protecting the security of test materials at all times. (117)



## Chapter 5: Scoring of Constructed-Response and Technology-Enhanced Items

In this chapter, the scoring process used for the 2019 LEAP 2025 ELA and mathematics assessment is described, with a particular focus on the handscoring of constructed-response items and the automated scoring of technology-enhanced items. At the end of this section, the results of the inter-rater reliability for the handscoring of the LEAP 2025 constructed-response items are presented.

Chapter 5 adheres to the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014) Standards 4.18, 4.20, 6.8, and 6.9. Each standard is presented in the pertinent section of this chapter. Standard 4.18 provides some general guidance for Chapter 5:

Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays. (91)

Chapter 5 explains the procedures used for scoring the LEAP 2025 ELA and mathematics constructed-response items and technology-enhanced items. The scoring criteria used for each item are not presented in this chapter to preserve the integrity of the items for future use.

### 5.1 Constructed-Response Item Scoring Process

Constructed-response items were scored by human raters who were trained by DRC. Handscoring and Artificial Intelligence (AI) processing rules are detailed in Appendix F. Four ELA items across grades 5-8 ELA (noted in the table below) were scored by an AI engine, Pearson's Intelligent Essay Assessor (IEA), using scoring models previously developed by Pearson. Second reads of 10% of these responses were completed by human scorers; handscoring supervisors also reviewed the responses that IEA was not able to score.

**Table 5.1 Constructed-Response Scoring**

Subject and Grade	Handscoring Only	AI Scoring	AI Vendor
ELA grade 3	Q7, Q12	N/A	
ELA grade 4	Q7, Q20	N/A	
ELA grade 5	Q7, Q20	N/A	
ELA grade 6	N/A	Q9, Q14	Pearson
ELA grade 7	Q9	Q14	Pearson
ELA grade 8	Q20	Q7	Pearson
Math grades 3-8	All CRs	N/A	

### 5.1.1 Selection of Scoring Evaluators

Standard 4.20 states the following:

The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring. (92)

The following sections explain how scorers were selected and trained for the LEAP 2025 ELA and mathematics handscoring process. Section 5.1.3 describes how the scorers were monitored throughout the handscoring process.

#### ***The Recruitment and Interview Process***

DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects.

All readers hired by DRC to score 2019 LEAP 2025 ELA and mathematics test responses had at least a four-year college degree. DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, or recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to demonstrate their proficiency in writing by responding to a DRC writing topic and their proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and demonstrate the ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

### 5.1.2 Security

Each DRC scoring center is a secure facility. All employees are issued photo identification badges and are required to wear them in plain view at all times. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials may leave the scoring center and all readers must sign legally binding confidentiality agreements before work begins. DRC retains these agreements for the duration of the contract. To prevent the unauthorized duplication of secure materials, cell phone and camera use within the scoring rooms is strictly forbidden. Readers only have access to the student responses they are qualified to score. Each scorer is assigned a unique username and password to access the DRC imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

### 5.1.3 Handscoring Training Process

Standard 6.9 specifies:

Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (118)

#### ***Training Material Development***

DRC scoring supervisors trained scorers using training materials from two sources.

1. PARCC-approved training materials provided by PARCC for ELA and math. These materials were developed according to processes described in [PARCC technical reports](#) and include the following:
  - Passages, prompts, and associated stimuli
  - Rubrics
  - Anchor sets
  - Practice sets
  - Qualifying sets (for prototype items only)
  
2. Math training materials developed by DRC in conjunction with and approved by the LDOE. These materials were made for use with DRC-developed math items (which were newly operational in the spring of 2019) according to processes described in DRC's response to the LDOE's "REQUEST FOR PROPOSALS For LEAP 2025 Assessment Administration (RFP #: 815200-20150723001)".
  - Prompts
  - Rubrics
  - Anchor sets
  - Practice sets
  - Qualifying sets (for all DRC-developed items)

#### ***Training and Qualifying Procedures***

Handscoring involves training and qualifying team leaders and evaluators, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities. The LDOE visits the scoring centers to review training materials and oversee the training process. An explanation of the training and qualification procedures follows.

DRC used the PARCC-approved mathematics and ELA training and qualifying materials to score two categories of items: "prototype" items and "abbreviated" items. Note that, like the PARCC "prototype" items for math, full sets of training and qualifying materials were also developed for all DRC-developed math items. The training and qualifying procedures DRC used for these items was the same process outlined below for PARCC-approved "prototype" math items.

#### ***Prototype Items***

Only one item (for grade 7 math) included in the 2019 Louisiana forms was a prototype item, meaning it had a full set of associated training materials, including anchor set, practice sets, and qualifying sets. DRC started the training process with a review of the item, rubric, and anchor set, followed by the scoring and discussion

of practice sets and qualifying sets. Once this process was completed, qualified readers started scoring live student responses for that item.

### **Abbreviated Items**

Abbreviated items required a two-step training and qualifying process. First, scorers trained and qualified as described above using PARCC-approved materials for an associated prototype item that was similar to the abbreviated one they would be scoring on the Louisiana form.<sup>2</sup> Readers who did not qualify on the prototype item training were not allowed to continue the training.

After qualifying on the associated prototype item training, a reader received additional item-specific training on the abbreviated item he or she was going to score. This consisted of an item-specific anchor set and two item-specific practice sets. After completing the abbreviated item training, the reader could begin scoring live student responses for the abbreviated item.

The following tables detail the composition of the training materials provided by Pearson for mathematics and ELA.

**Table 5.2 Mathematics Training Set Composition**

<b>Set Type</b>	<b>Prototype Item Training</b>	<b>Abbreviated Item Training</b>	<b>Annotated</b>
Anchor Set	3 responses per score point (Composite items had 3 responses per composite score.)	3 responses per score point (Composite items had 3 responses per composite score.)	Yes
Practice Set 1	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Practice Set 2	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Qualifying Set 1	10 responses comparable to the anchor set responses		No
Qualifying Set 2	10 responses comparable to the anchor set responses		No
Qualifying Set 3	10 responses comparable to the anchor set responses		No
*For DRC-developed math items, examples of responses at the top score points may not have been present in some anchor, training, and qualifying sets as there were few or no examples found during rangefinding or subsequent field test scoring. In such cases, DRC Scoring Directors identified examples of these scores during live scoring to supplement reader training.			

<sup>2</sup> Item associations were determined by PARCC/Pearson with the understanding that aspects of training are generalizable across similar items. For mathematics, the determination of prototype versus abbreviated items was made by PARCC and Pearson based on similar item types and by evidence statements. For ELA items, this determination by PARCC and Pearson was based on grade and task type.

**Table 5.3 ELA Training Set Composition**

Set Type	Prototype Item Training	Abbreviated Item Training	Annotated
Anchor Set*	3 responses per score point	16 responses per item: Anchor Sets for abbreviated RST and LAT item training included scores for the combined trait Reading Comprehension and Written Expression (RCWE). Anchor Sets for abbreviated NWT item training included scores for Written Expression (WE).	Yes
Practice Set 1	5 responses representing the range of responses for the Reading Comprehension and Written Expression (RCWE) trait (for LAT and RST items) the Written Expression trait (for NWT items)	10 responses representing the range of responses for the trait appropriate to the task type	Yes
Practice Set 2	5 responses representing the range of responses for the Knowledge and Use of Language Conventions trait	10 responses representing the range of responses for the conventions trait	Yes
Practice Set 3	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Practice Set 4	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Qualifying Set 1	10 responses comparable to the anchor set responses (included both traits appropriate to the task type)		No
Qualifying Set 2	10 responses comparable to the anchor set responses (included both traits appropriate to the task type)		No
Qualifying Set 3	10 responses comparable to the anchor set responses (included both traits appropriate to the task type)		No
Direct Copy Set**	3-5 responses composed entirely or partially of text copied from passage or passages (included both traits appropriate to the task type)	3-5 responses composed entirely or partially of text copied from passage or passages (included both traits appropriate to the task type)	Yes

\*For the ELA Knowledge and Use of Language Conventions trait, there were two mixed-prompt anchor sets per grade level (one for the narrative task and the other for the literary analysis and research simulation tasks). In addition to the mixed-prompt anchor set, depending on the task, the practice sets for prototype and abbreviated items required readers to practice scoring the Knowledge and Use of Language Conventions trait along with the Reading Comprehension and Written Expression trait (for LAT and RST items) or with the Written Expression trait (NWT). Readers were also required to qualify on the Knowledge and Use of Language Conventions trait during each prototype item qualifying session.

\*\*These PARCC-approved sets provided additional annotated sample responses explaining the scoring rationale for responses composed entirely or partially of text copied from the source passage(s) associated with an item. DRC scoring supervisors reviewed these item-specific sets with the readers prior to scoring the associated item.

Some items selected for use on the spring 2019 administration were previously only field tested by PARCC. Consequently, the abbreviated training materials available for use with these items were abridged versions of typical abbreviated sets of materials. They consisted of:

- An Anchor Set (for ELA, some have annotations and some lack examples of the top scores)
- One Practice Set of 5 responses (scored but not annotated in the case of ELA)
- Approximately 10 validity responses

Since these materials were somewhat limited compared to typical abbreviated materials (the main difference being a lack of formal written annotations and fewer practice responses), DRC bolstered the training by using the PARCC-approved field test validity responses provided by New Meridian as additional practice responses. DRC Scoring Directors then pulled additional responses from operational Louisiana student responses to use as validity responses during the scoring window. The Scoring Directors also found examples of higher-scoring responses that might be missing from the field test anchors. The validity and additional exemplar responses, along with the DRC Scoring Directors' notes for all papers used during the training of the abbreviated field-test only items, were submitted to the LDOE for approval. It is important to note that readers still had to qualify via standard qualification procedures on the prototype items for all items by first going through full training with the appropriate prototype Anchor Set, Practice Sets 1-4, and Qualifying Sets 1-3 (as well as the Conventions sets).

### **Qualifying Standards**

DRC followed the same qualification standards that Pearson used for PARCC. A description of these PARCC qualifying standards follows.

Scorers demonstrated their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with true scores on qualifying sets). After each qualifying set was scored, the DRC scoring director responsible for training led the scorers in a discussion of the set.

Any scorer who did not qualify by the end of the qualifying process for an item was not allowed to score live student responses.

**Table 5.4 Mathematics Qualifying Standards**

	<b>Perfect Agreement</b>	<b>Perfect Plus Adjacent Agreement</b>
0, 1, 2 Rubric	80% on two of three sets	96% on two of three sets
0, 1, 2, 3 Rubric	70% on two of three sets	96% on two of three sets
0, 1, 2, 3, 4 Rubric	70% on two of three sets	95% on two of three sets

**Table 5.5 Mathematics Qualifying Standards (Composite Items)\***

<b>Composite (multipart) Items</b>	<b>Perfect Agreement</b>	<b>Perfect Plus Adjacent Agreement</b>
0, 1 Rubric	90% on two of three sets	100% on two of three sets
0, 1, 2 Rubric	80% on two of three sets	96% on two of three sets
0, 1, 2, 3 Rubric	70% on two of three sets	96% on two of three sets
0, 1, 2, 3, 4 Rubric	70% on two of three sets	95% on two of three sets

*\*For mathematics composite items, the appropriate qualifying standard had to be achieved on each part of the item. For example, if an item had Part A with a top score of 1, Part B with a top score of 2, and Part C with a top score of 3, a scorer/supervisor would need to achieve 90% perfect agreement on Part A, 80% perfect agreement on Part B, and 70% perfect agreement on Part C, with no more than one nonadjacent score per part across all three qualifying sets.*

**Table 5.6 ELA Qualifying Standards**

Perfect Agreement	Perfect Plus Adjacent Agreement
70% average for both traits on two of three qualifying sets	96% across the three qualifying sets combined on both traits
70% on each trait at least once across three qualifying sets	

ELA readers were required to meet all three of the qualifications listed in Table 5.6. Perfect plus adjacent agreement of 96% means that out of the entire pool of scores that a reader gave across the three qualifying sets for an item, no more than 4% of those scores could be nonadjacent. In other words, no more than 2 of the 60 applied scores could be nonadjacent (3 sets x 10 responses/set x 2 traits = 60 applied scores).

### 5.1.4 Monitoring the Scoring Process

Standard 6.8 states:

Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (118)

Section 5.1.4 explains the monitoring procedures that DRC uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored. Detailed scoring rubrics, which specify the criteria for scoring, are available for handscoring evaluators for all constructed-response items.

#### ***Reader Monitoring Procedures***

Throughout the handscoring process, DRC project managers, scoring directors, and team leaders reviewed the statistics that were generated on a daily basis. DRC used one team leader for every 10 to 12 readers, which was the same ratio that Pearson used for PARCC. If scoring concerns were apparent among individual scorers, team leaders dealt with those issues on an individual basis. If a scorer appeared to need clarification of the scoring rules, DRC supervisors typically monitored one out of five of the scorer’s readings, making adjustments to that ratio as needed. If a supervisor disagreed with a reader’s scores during monitoring, he or she provided retraining in the form of direct feedback to the reader, using rubric language and applicable training responses.

#### ***Validity Sets and Inter-Rater Reliability***

In addition to the feedback that supervisors provided to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducted validity scoring using PARCC-approved validity responses supplied by PARCC and LDOE-approved validity responses identified by DRC scoring supervisors during live scoring for newly operational DRC-developed math items and PARCC field test-only items. Validity responses were inserted among the live student responses.

The validity responses were added to DRC’s image handscoring system prior to the beginning of scoring. Validity reports compared readers’ scores to pre-determined scores and were used to help detect potential room drift and individual scorer drift. This data was used to make decisions regarding the retraining and/or release of scorers, as well as the rescoring of responses.

Approximately 10% of all live student responses were scored by a second reader to establish inter-rater reliability statistics for all constructed-response items. This procedure is called a “double-blind read” because

the second reader does not know the first reader’s score. DRC monitored inter-rater reliability based on the responses that were scored by two readers. If a scorer fell below the expected rate of agreement, the team leader or scoring director retrained the scorer. If a scorer failed to improve after retraining and feedback, DRC removed the scorer from the project. In this situation, DRC removed all scores assigned by the scorer in question. The responses were then reassigned and rescored.

To monitor inter-rater reliability, DRC produced scoring summary reports on a daily basis. DRC’s scoring summary reports display exact, adjacent, and nonadjacent agreement rates for each reader. These rates are calculated based on responses that are scored by two readers, and their definitions are included below.

- **Percentage Exact (%EX)**—total number of responses by reader where scores are the same, divided by the number of responses that were scored twice
- **Percentage Adjacent (%AD)**—total number of responses by reader where scores are one point apart, divided by the number of responses that were scored twice
- **Percentage Nonadjacent (%NA)**—total number of responses by reader where scores are more than one score point apart, divided by the number of responses that were scored twice

The following table provided by Pearson shows the expectations for validity and inter-rater reliability:

**Table 5.7 Expectations for Validity and Inter-Rater Reliability**

Agreement Rate Requirements for Validity and Inter-Rater Reliability			
Content Area	Score Point Range	Perfect Agreement	Perfect Agreement + Adjacent
Mathematics	0–1	90%	100%
Mathematics	0–2	80%	95%
Mathematics	0–3	70%	95%
Mathematics	0–4	65%	95%
ELA	Multi-trait 0–3 or 0–4 (varies by grade and trait)	65% (each trait)	96% (each trait)

Each reader was required to maintain a level of exact agreement on validity responses and on inter-rater reliability as shown under “Perfect Agreement” in the table above. Additionally, readers were required to maintain an acceptably low rate of nonadjacent agreement. To monitor this, DRC summed each reader’s exact and adjacent agreement rates and required each reader to maintain the levels shown under “Perfect Agreement + Adjacent” in the table above.

### **Calibration Sets**

PARCC provided DRC with PARCC-approved calibration responses for all operational items that came from the PARCC item pool. DRC pulled calibration responses for DRC-developed math items as well as additional responses for field-test only items from PARCC. DRC used these sets to perform calibration across the entire scorer population for an item if trends were detected (e.g., low agreement between certain score points if a certain type of response was missing from initial training). These calibrations were designed to help refocus scorers on how to properly use the scoring guidelines. They were selected to help illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After readers



scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and scoring concepts exemplified by the anchor responses to explain the reasoning behind each response's score.

### ***Reports and Reader Feedback***

Reader performance and intervention information were recorded in reader feedback logs. These logs tracked information about actions taken with individual readers to ensure scoring consistency in regard to reliability, score point distribution, and validity performance. In addition to the reader feedback logs, DRC provided the LDOE with handscoring quality control reports for review throughout the scoring window. Further detail about these reports can be found in Appendix F.

## 5.2 Inter-Rater Reliability

A minimum of 10% of the constructed responses in ELA and mathematics were scored independently by a second reader. This was the case regardless of whether the first reader was human or AI. The statistics for inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the first and second scores was examined.

A total of 51 operational items were scored by human readers across all grades and both content areas. The inter-rater reliability rates and the total numbers of reads are shown in Table 5.8 for ELA items, Table 5.9 for operational mathematics items, Table 5.10 for Spanish mathematics items, and Table 5.11 for mathematics field test items.

As shown in Table 5.8, raters demonstrated at least 98% perfect and adjacent agreement for all ELA handscored items. As shown in Tables 5.9 and 5.11, raters demonstrated at least 97% perfect and adjacent agreement for mathematics items. As shown in Table 5.10, raters demonstrated 100% perfect and adjacent agreement for Spanish mathematics items.

**Table 5.8 Inter-Rater Agreement, English Language Arts Items**

Grade	Task Type	Question	Trait	Total Reads	Read 2x	Inter-Rater Reliability %		
						EX	AD	EX + AD
3	Research Simulation	7	Reading Comprehension and Written Expression	≥59,490	≥12,410	80	19	99
			Knowledge and Use of Language Conventions	≥59,490	≥12,410	80	20	100
	Narrative Writing	12	Written Expression	≥59,340	≥12,100	86	14	100
			Knowledge and Use of Language Conventions	≥59,340	≥12,100	77	22	99
4	Literary Analysis	7	Reading Comprehension and Written Expression	≥60,400	≥10,670	81	19	100
			Knowledge and Use of Language Conventions	≥60,400	≥10,670	79	21	100
	Research Simulation	20	Reading Comprehension and Written Expression	≥62,100	≥14,080	83	17	100
			Knowledge and Use of Language Conventions	≥62,100	≥14,080	83	17	100
5	Literary Analysis	7	Reading Comprehension and Written Expression	≥61,200	≥12,480	77	23	100
			Knowledge and Use of Language Conventions	≥61,200	≥12,480	75	25	100
	Research Simulation	20	Reading Comprehension and Written Expression	≥62,770	≥15,450	80	20	100
			Knowledge and Use of Language Conventions	≥62,770	≥15,450	80	20	100
6	Research Simulation (AI)	9	Reading Comprehension and Written Expression	≥61,420	≥12,870	71	28	99
			Knowledge and Use of Language Conventions	≥61,420	≥12,870	68	30	98
	Narrative Writing (AI)	14	Written Expression	≥61,220	≥12,420	79	21	100
			Knowledge and Use of Language Conventions	≥61,220	≥12,420	76	24	100
7	Research Simulation	9	Reading Comprehension and Written Expression	≥57,940	≥11,070	76	23	99
			Knowledge and Use of Language Conventions	≥57,940	≥11,070	75	23	98* (na = 1)
	Narrative Writing (AI)	14	Written Expression	≥58,490	≥12,160	76	23	99
			Knowledge and Use of Language Conventions	≥58,490	≥12,160	74	25	99
8	Literary Analysis (AI)	7	Reading Comprehension and Written Expression	≥57,100	≥12,670	75	24	99
			Knowledge and Use of Language Conventions	≥57,100	≥12,670	74	25	99
	Research Simulation	20	Reading Comprehension and Written Expression	≥56,130	≥10,710	77	23	100
			Knowledge and Use of Language Conventions	≥56,130	≥10,710	74	25	99

\*Total Exact (EX) + Adjacent (AD) + Non-adjacent (na) does not add up to 100% due to rounding

Table 5.9 Inter-Rater Agreement, Mathematics Items

Grade	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
					EX	AD	EX + AD
3	17	Part A	≥58,720	≥11,030	86	3	99
		Part B	≥58,720	≥11,030	94	6	100
	18	N/A	≥58,640	≥10,940	92	8	100
	32	Part A	≥58,720	≥11,070	96	4	100
		Part B	≥58,720	≥11,070	99	1	100
	33	Part B (CBT)	≥1,760	≥340	98	2	100
		Part B (PBT)	≥56,870	≥10,540	96	3	99
	48	N/A	≥58,670	≥10,950	94	6	100
	49	Part B (CBT)	≥1,780	≥340	97	3	100
		Part C (CBT)	≥1,780	≥340	97	3	100
Part B (PBT)		≥56,800	≥10,410	95	4	99	
Part C (PBT)		≥56,800	≥10,410	96	4	100	
4	17	Part C (CBT)	≥8,370	≥1,560	94	6	100
		Part C (PBT)	≥52,260	≥9,870	92	8	100
	18	N/A	≥60,520	≥11,430	96	3	99
	32	N/A	≥60,530	≥11,300	89	10	99* (na = 0)
	33	N/A	≥60,600	≥11,410	89	11	100
	48	Part A	≥60,610	≥11,640	94	6	100
		Part B	≥60,610	≥11,640	97	3	100
	49	Part A	≥60,410	≥11,300	93	7	100
Part B		≥60,410	≥11,300	93	7	100	
Part C		≥60,410	≥11,300	95	5	100	
5	17	N/A	≥60,820	≥11,120	82	17	99
	18	N/A	≥60,400	≥11,580	90	9	99
	32	Part B	≥60,430	≥11,000	92	8	100
	33	N/A	≥60,210	≥11,750	92	8	100
	48	Part B	≥60,440	≥11,010	94	6	100
	49	Part B	≥60,390	≥11,010	94	6	100
		Part C	≥60,390	≥11,010	89	10	99

\*Total Exact (EX) + Adjacent (AD) + Non-adjacent (na) does not add up to 100% due to rounding

\*\*N/A if an item does not have multiple parts

Table 5.10 Inter-Rater Agreement, Mathematics Items, continued

Grade	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
					EX	AD	EX + AD
6	30	N/A	≥60,800	≥12,440	75	23	98
	34	Part A	≥60,290	≥11,790	92	7	99
		Part B	≥60,290	≥11,790	96	4	100
	35	Part A	≥59,910	≥11,600	93	6	99
		Part B	≥59,910	≥11,600	87	11	99* (na = 1)
	36	Part B	≥59,730	≥10,890	91	8	99
	47	N/A	≥60,540	≥12,100	77	20	97
	48	Part B	≥60,450	≥10,990	91	9	100
49	N/A	≥60,150	≥11,660	92	7	99	
7	31	Part A	≥57,520	≥10,990	95	4	99
		Part B	≥57,520	≥10,990	96	4	100
	34	N/A	≥56,960	≥11,710	94	5	99* (na = 2)
	36	N/A	≥56,940	≥12,200	97	3	100
	37	Part A	≥56,650	≥10,260	88	11	99
		Part B	≥56,650	≥10,260	96	4	100
	47	N/A	≥56,780	≥11,970	96	3	99
	48	Part A	≥57,320	≥11,320	98	2	100
Part B		≥57,320	≥11,320	98	2	100	
49	N/A	≥56,930	≥11,590	92	8	100	
8	31	Part A	≥49,260	≥9,480	92	8	100
		Part B	≥49,260	≥9,480	80	18	98* (na = 1)
	34	Part A	≥48,840	≥9,980	90	9	99
		Part B	≥48,840	≥9,980	89	9	98
	35	N/A	≥48,410	≥10,110	92	7	99
	36	Part A	≥47,700	≥9,770	95	4	99
		Part B	≥47,700	≥9,770	96	4	100
	42	Part B	≥49,180	≥9,000	91	9	100
46	N/A	≥49,070	≥10,610	93	7	100	
48	Part B	≥49,150	≥8,960	94	6	100	
	Part C	≥49,150	≥8,960	95	5	100	

\*Total Exact (EX) + Adjacent (AD) + Non-adjacent (na) does not add up to 100% due to rounding

\*\*N/A if an item does not have multiple parts

Table 5.11 Inter-Rater Agreement, Spanish Mathematics Items

Grade	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
					EX	AD	EX + AD
3	17	Part A	≥100	≥10	100	0	100
		Part B	≥100	≥10	100	0	100
	18	N/A	≥110	≥30	94	6	100
	32	Part A	≥110	≥30	100	0	100
		Part B	≥110	≥30	100	0	100
	33	Part B (CBT)	≥10	<10	NR	NR	NR
		Part B (PBT)	≥90	≥10	100	0	100
	48	N/A	≥110	≥30	100	0	100
	49	Part B (CBT)	≥10	<10	NR	NR	NR
		Part C (CBT)	≥10	<10	NR	NR	NR
49	Part B (PBT)	≥80	≥10	100	0	100	
	Part C (PBT)	≥80	≥10	100	0	100	
4	17	Part C (CBT)	≥30	<10	NR	NR	NR
		Part C (PBT)	≥130	≥20	100	0	100
	18	N/A	≥160	≥30	100	0	100
	32	N/A	≥160	≥50	92	8	100
	33	N/A	≥160	≥40	100	0	100
	48	Part A	≥160	≥40	100	0	100
		Part B	≥160	≥40	100	0	100
	49	Part A	≥160	≥40	100	0	100
		Part B	≥160	≥40	100	0	100
		Part C	≥160	≥40	100	0	100
5	17	N/A	≥140	≥30	100	0	100
	18	N/A	≥150	≥40	95	5	100
	32	Part B	≥140	≥20	100	0	100
	33	N/A	≥140	≥30	100	0	100
	48	Part B	≥140	≥20	100	0	100
	49	Part B	≥140	≥20	100	0	100
		Part C	≥140	≥20	93	7	100

\*Total Exact (EX) + Adjacent (AD) does not add up to 100% due to rounding

\*\*N/A if an item does not have multiple parts

Table 5.12 Inter-Rater Agreement, Spanish Mathematics Items, continued

Grade	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
					EX	AD	EX + AD
6	30	N/A	≥160	≥30	94	6	100
	34	Part A	≥160	≥30	100	0	100
		Part B	≥160	≥30	94	6	100
	35	Part A	≥160	≥30	100	0	100
		Part B	≥160	≥30	94	6	100
	36	Part B	≥160	≥30	100	0	100
	47	N/A	≥170	≥40	100	0	100
	48	Part B	≥160	≥30	100	0	100
49	N/A	≥160	≥30	100	0	100	
7	31	Part A	≥200	≥30	100	0	100
		Part B	≥200	≥30	100	0	100
	34	N/A	≥210	≥50	100	0	100
	36	N/A	≥220	≥60	100	0	100
	37	Part A	≥210	≥40	95	5	100
		Part B	≥210	≥40	100	0	100
	47	N/A	≥220	≥60	100	0	100
	48	Part A	≥210	≥50	100	0	100
Part B		≥210	≥50	100	0	100	
49	N/A	≥210	≥40	100	0	100	
8	31	Part A	≥180	≥30	100	0	100
		Part B	≥180	≥30	95	5	100
	34	Part A	≥170	≥30	100	0	100
		Part B	≥170	≥30	100	0	100
	35	N/A	≥170	≥40	100	0	100
	36	Part A	≥170	≥40	100	0	100
		Part B	≥170	≥40	100	0	100
	42	Part B	≥180	≥30	89	11	100
46	N/A	≥170	≥30	100	0	100	
48	Part B	≥180	≥30	100	0	100	
	Part C	≥180	≥30	100	0	100	

\*Total Exact (EX) + Adjacent (AD) does not add up to 100% due to rounding

\*\*N/A if an item does not have multiple parts

Table 5.13 Inter-Rater Agreement, Mathematics Field Test Items

Grade	Question	Part(s)*	Total Reads	Read 2x	Inter-Rater Reliability %		
					EX	AD	EX + AD
3	979827	Part A	≥1,760	≥320	88	12	100
		Part B	≥1,760	≥320	86	14	100
	979828	Part B	≥1,770	≥340	98	2	100
	979829	Part B	≥1,760	≥330	93	7	100
		Part C	≥1,760	≥330	93	7	100
		Part D	≥1,760	≥330	93	7	100
	979830	Part A	≥1,760	≥320	92	8	100
Part B		≥1,760	≥320	98	2	100	
4	981277	Part B	≥1,760	≥330	97	3	100
	981279	Part B	≥1,750	≥310	89	11	100
		Part C	≥1,750	≥310	94	6	100
	981574	Part A	≥1,760	≥330	98	2	100
		Part B	≥1,760	≥330	94	6	100
	981575	Part A	≥1,750	≥310	95	5	100
Part B		≥1,750	≥310	99	1	100	
5	982589	Part A	≥1,670	≥170	98	2	100
		Part B	≥1,670	≥170	95	5	100
	982590	Part B	≥1,760	≥320	96	4	100
	982591	Part B	≥1,750	≥320	92	8	100
983509	Part B	≥1,760	≥330	100	0	100	
6	982443	Part B	≥1,760	≥320	90	10	100
	982561	Part B	≥1,750	≥320	98	2	100
	982562	Part C	≥1,760	≥330	95	4	99
		Part D	≥1,760	≥330	98	2	100
	982573	Part B	≥1,750	≥310	99	1	100
		Part C	≥1,750	≥310	98	1	99
Part D		≥1,750	≥310	97	3	100	
7	983507	Part A	≥1,750	≥340	96	4	100
		Part B	≥1,750	≥340	100	0	100
	983508	Part A	≥1,760	≥360	97	3	100
		Part B	≥1,760	≥360	95	4	99
	983510	N/A	≥1,730	≥360	96	4	100
	984138	Part A	≥1,770	≥360	99	1	100
Part B		≥1,770	≥360	97	2	99	
8	983809	Part A	≥1,740	≥350	99	1	100
		Part B	≥1,740	≥350	98	2	100
	984008	Part A	≥1,750	≥350	100	0	100
		Part B	≥1,750	≥350	97	3	100
		Part C	≥1,750	≥350	100	0	100
	984137	N/A	≥1,730	≥350	99	1	100
984139	Part B	≥1,740	≥340	98	2	100	
	Part C	≥1,740	≥340	98	1	99	

### 5.3 Technology-Enhanced Item Scoring Process

All technology-enhanced items, as well as EBSR, MPSR, and SA items, were processed through DRC’s autoscoring engine and scored according to the assigned scoring rules as established during content creation by PARCC or DRC as applicable in conjunction with the LDOE. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring any technology-enhanced items. DRC established an adjudication process for technology-enhanced items and short-answer responses to verify that correct answers were identified. DRC’s technology-enhanced scoring process included the following procedures:

- A scoring rubric was created for each technology-enhanced item. The rubric described the one and only correct answer for dichotomously scored items (i.e., items scored as either right or wrong). If partial credit was possible, the rubric described in detail the type of response that could receive credit for each score point.
- The information from the scoring rubric was entered into the scoring system within the item banking system so that the truth resided in one place along with the item image and other metadata. This scoring information included details that varied by item type. For example, for a drag-and-drop item, the information included which objects are to be placed in each drop region to receive credit.
- The information was then verified by another autoscoring expert.
- After testing started, reports were generated that showed every response, how many students gave that response, and the score the scoring system provided for that response.
- The scoring was then checked against the scoring rubric using two levels of verification.
- If any discrepancies were found, the scoring information was modified and verified again. The scoring process was then rerun. This checking and modification process continued until no other issues were found.
- As a final check, a final report was generated that showed all student responses, their frequencies, and their received scores.

In the case of braille and large-print test forms, student responses to items were transcribed into the online system by a test administrator.

### 5.4 Multiple-Choice and Multiple-Select Item Scoring Process

Responses to multiple-choice and multiple-select items were captured during the CBT administration and during scanning of the PBT answer documents. In the case of braille and large-print test forms, student responses to these items were transcribed into the online system by a test administrator.

### 5.5 Summary

The information presented in this chapter summarizes the scoring procedures for different types of items and the steps taken by DRC to ensure accuracy in the autoscoring and handscoring processes. The inter-rater reliability statistics presented in Section 5.4 demonstrate that the items were scored reliably. These efforts by DRC address multiple best practices of the testing industry but are particularly related to AERA, APA, & NCME (2014) Standards 4.18, 4.20, 6.8, and 6.9:

**Standard 4.18** Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for



using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays. (91)

**Standard 4.20** The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring. (92)

**Standard 6.8** Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (118)

**Standard 6.9** Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (118)

## Chapter 6: Operational Data Analyses

---

This chapter of the LEAP 2025 technical report describes the analyses that were conducted on the operational data. These include a classical item analysis and examination of the raw scores and an item response theory (IRT) analysis involving calibrating, scaling, and linking.

This section presents the classical item statistics, including aggregate raw score statistics and individual item-level statistics. Next, this section discusses the IRT models used for calibrating the data and addresses the purpose of data calibration and scaling for each content area is addressed. The calibration samples are presented next, followed by the data calibration results, including the model-data fit for the Louisiana data. If the IRT models fit the empirical item response distributions for the population about which generalizations are to be made (i.e., Louisiana students), then the claim that the scores are valid indicators of an underlying ability is strengthened. The lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for the LEAP 2025 tests are also presented.

Chapter 6 demonstrates how LEAP 2025 assessments adhere to American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014) Standards 1.8, 4.14, 5.2, 5.13, 5.15, and 7.2. Each standard is explicated within the appropriate section of this chapter. Standard 7.2 provides general guidance that is relevant to this chapter. It states the following:

The population for whom a test is intended and specifications for the test should be documented.  
(126)

For all 2019 LEAP 2025 analyses, the Louisiana student population was used. In Section 6.3, the characteristics of calibration samples, such as subgroups, are discussed. Chapter 3 presents the test specifications. Information regarding reported data is discussed in detail in Chapter 7.

In this section, summary test statistics for each form, grade, and content area of LEAP 2025 are presented. These statistics are followed by item-level statistics for each grade and content area of LEAP 2025. These statistics were produced using census data.

### 6.1 Test-Level Statistics

Table 6.1 presents the number of items, score points, mean and standard deviation of the raw scores, and average form difficulty for each test form at each grade level of the ELA and mathematics assessments, respectively. Form difficulty for an examinee was calculated by dividing the raw score of the student by total score points of the test.

As can be seen in the table, average form difficulty for ELA ranged from 0.28 to 0.45. Average form difficulty for mathematics ranged from 0.33 to 0.53. In general, the 2019 LEAP 2025 tests were relatively difficult tests across all subjects and grades. For ELA, the grade 3 computer-based test (CBT) was the most difficult, with 0.28 average form difficulty, and the grade 7 was the easiest, with 0.45 average form difficulty. For mathematics, the grade 8 test was the most difficult, with 0.33 average form difficulty, and the grade 3 paper-based test (PBT) test was the easiest, with 0.53 average form difficulty.

**Table 6.1 LEAP 2025 Means and Standard Deviations for Raw Scores and Form Difficulty**

Content	Grade	Mode	Total Items	Total Points	Mean Raw Score (Std. Dev.)	Average Form Difficulty (Std. Dev.)
ELA	3	CBT	27	71	19.45 (10.83)	0.28 (0.13)
	3	PBT	27	71	26.73 (12.52)	0.38 (0.13)
	4	CBT	30	86	28.77 (15.46)	0.34 (0.14)
	4	PBT	30	86	32.31 (15.90)	0.38 (0.12)
	5	CBT	30	86	29.41 (15.42)	0.34 (0.16)
	6	CBT	33	90	35.60 (17.26)	0.40 (0.12)
	7	CBT	33	90	40.00 (19.10)	0.45 (0.11)
	8	CBT	34	94	38.32 (17.88)	0.41 (0.11)
Mathematics	3	CBT	43	62	23.55 (12.39)	0.38 (0.18)
	3	PBT	43	62	32.31 (13.82)	0.53 (0.16)
	4	CBT	43	62	27.03 (13.79)	0.44 (0.20)
	4	PBT	43	62	29.39 (13.76)	0.48 (0.19)
	5	CBT	41	60	26.36 (13.05)	0.44 (0.16)
	6	CBT	42	65	25.18 (14.01)	0.39 (0.18)
	7	CBT	43	66	23.09 (13.45)	0.35 (0.18)
	8	CBT	41	65	21.32 (12.35)	0.33 (0.18)

Table 6.2 presents the number of items, mean and standard deviation of the item  $p$ -values, and item-total correlations (i.e., item discrimination values) for each test form at each grade level of the ELA and mathematics assessments, respectively.

The mean  $p$ -value is the average of all item  $p$ -values of a specific grade and content area. The mean item-total correlation ( $R_{it}$ ) is the average of all item point-biserial correlations of a specific grade and content area. The  $p$ -value and item-total correlation are explained in the next section.

**Table 6.2 LEAP 2025 Means, Standard Deviations for Raw Scores,  $p$ -Values, Item-Total Correlation ( $R_{it}$ )**

Content	Grade	Mode	N of Items	Item $p$ -Value				Item-Total Correlation			
				Mean	Std. Dev.	Min.	Max	Mean	Std. Dev.	Min.	Max
ELA	3	CBT	27	0.32	0.12	0.11	0.57	0.42	0.12	0.19	0.61
	3	PBT	27	0.42	0.13	0.26	0.67	0.43	0.12	0.23	0.65
	4	CBT	30	0.38	0.14	0.20	0.71	0.49	0.15	0.27	0.79
	4	PBT	30	0.42	0.12	0.23	0.67	0.48	0.16	0.28	0.77
	5	CBT	30	0.40	0.16	0.13	0.76	0.48	0.16	0.22	0.78
	6	CBT	33	0.43	0.12	0.24	0.74	0.47	0.15	0.21	0.77
	7	CBT	33	0.47	0.12	0.27	0.70	0.50	0.14	0.27	0.80
	8	CBT	34	0.44	0.11	0.26	0.67	0.47	0.16	0.12	0.81
Mathematics	3	CBT	43	0.44	0.18	0.16	0.85	0.45	0.13	0.18	0.72
	3	PBT	43	0.58	0.16	0.27	0.91	0.48	0.11	0.27	0.76
	4	CBT	43	0.50	0.19	0.18	0.85	0.50	0.10	0.32	0.70
	4	PBT	43	0.53	0.19	0.22	0.88	0.49	0.10	0.32	0.68
	5	CBT	41	0.50	0.16	0.15	0.77	0.47	0.12	0.26	0.71
	6	CBT	42	0.44	0.17	0.09	0.76	0.49	0.11	0.27	0.67
	7	CBT	43	0.42	0.18	0.08	0.84	0.44	0.14	0.09	0.68
	8	CBT	41	0.37	0.17	0.09	0.81	0.46	0.12	0.10	0.67

## 6.2 Item-Level Statistics

Tables 6.3–6.10 present the item statistics for each operational item included in regular test forms organized by grade for ELA. Tables 6.11–6.18 show the item statistics for each item included in regular test forms organized by grade for mathematics. The tables include administration mode, item number,  $p$ -value, item-total correlation ( $R_{it}$ ), omit rates, total N, adjusted N (adjusted N excludes items with multiple responses [PBT only], omitted responses, responses that were not scored, or responses that received a non-score code), and the percentage at each score point, if applicable, for each item by grade and content area.

### ***p*-Value**

The  $p$ -value is a measure of item difficulty. For a multiple-choice (MC) item, the  $p$ -value is calculated by dividing the number of students who correctly responded to an item by the total number of students who attempted the item. The value is reported as a proportion. For a non-MC item, the  $p$ -value is calculated by dividing the average score for the item by the maximum points possible. This value is also reported as a proportion.

In terms of  $p$ -values, test scores tend to be more precise when their average  $p$ -values are between the mid-0.50s and the low 0.70s. However, it is important to select items on the basis of content rather than on purely statistical criteria when building a criterion-referenced test. As shown in Table 6.2, the average  $p$ -values associated with the ELA forms range from 0.32 in the grade 3 CBT form to 0.47 in grade 7. The average  $p$ -values associated with the mathematics forms range from 0.37 in grade 8 CBT to 0.58 in grade 3 PBT.

It is important that one examines the range of  $p$ -values, not just the average  $p$ -value, to determine whether a test measures well. It is desirable for a test to measure well throughout the range of skills present at a given grade. That is, it is important that the items measure the performance of both low-scoring and high-scoring

students, not just students in the center of the distribution. Having a range of  $p$ -values also helps to prevent floor and/or ceiling effects so that the test does not have large numbers of students at the minimum or maximum possible scores. The ELA forms have items with  $p$ -values ranging from 0.11 to 0.76 (see Tables 6.3–6.10) across all grade levels. The  $p$ -values on the mathematics forms range from 0.08 to 0.91 (see Tables 6.11–6.18). Such a broad range of  $p$ -values, which indicates the items measure well throughout the range of skill levels at a given grade, supports the accuracy of the LEAP 2025 test scores.

### ***Item-Total Correlations***

An item-total correlation is the correlation between an item score and the total test score, where the item score is not included in the total score. It indicates how well an item differentiates between low-scoring and high-scoring students. In general, items with correlations below 0.20 are said to be poorly discriminating. The majority of the items in the LEAP 2025 had item-total correlations above this threshold. Any item with an item-total correlation below the 0.20 threshold was further analyzed to ensure that the item was correctly keyed.

### ***Omit Rates***

The omit rate for each item indicates the percentage of students who did not answer the item. Omit rates can be used to examine possible speededness issues on tests. A test may be speeded if students do not have adequate time to answer all questions on the test. In general, an item is said to have a high omit rate if more than 5% of students failed to respond to the item. Evidence of speededness is considered a threat to validity because student test scores may not reflect their ability. Additionally, content validity may be threatened because the items that were not completed are needed to fulfill content blueprint specifications (Lu & Sireci, 2007).

This examination of omit rates complies with Standard 4.14 of the *Standards*. This standard is concerned with the speededness of a test and states the following:

For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (90)

The results in this section will show that, overall, student test scores are not adversely affected by the rate at which the students complete the test. In general, students have ample time to complete all sections of the test and there is not a threat to construct or content validity.

The results presented in Tables 6.3–6.18 show that the omit rates for most of the items on the LEAP 2025 regular forms are less than 5%, suggesting that the majority of students were able to complete the test in the prescribed amount of time. There is not an omit rate higher than 9%, and the omit rates for the last items in the tests do not exceed 3%. These omit rates indicate that 97% of the students completed the test. Lu & Sireci (2007) report that the Education Testing Service has used an approach where a test was considered unspeeded if at least 80% of the examinees reach the last item and all testers reach at least 75% of the items. The reported omit rates fall within these ranges.

Table 6.3 Operational Item Statistics—English Language Arts Grade 3 CBT Administration

ELA Grade 3 Computer-Based Test Administration										
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3
915222	ESR	≥1,620	≥1,620	0.39	0.44	0.12	46.62	28.29	24.97	
915224	ESR	≥1,620	≥1,610	0.37	0.25	0.55	53.75	17.40	28.29	
915228	TE	≥1,620	≥1,620	0.31	0.35	0.37	51.35	34.44	13.84	
915230	ESR	≥1,620	≥1,610	0.41	0.33	0.43	44.71	27.43	27.43	
915220	TE	≥1,620	≥1,580	0.34	0.41	2.40	39.24	49.45	8.92	
915219	ESR	≥1,620	≥1,610	0.31	0.19	0.49	57.01	22.45	20.05	
91522702	CR	≥1,620	≥1,530	0.11	0.55	1.48	63.41	29.64	1.05	0.06
91522703	CR	≥1,620	≥1,530	0.12	0.53	1.48	64.51	26.57	2.83	0.25
936916	MS	≥1,620	≥1,620	0.22	0.39	0.06	66.85	21.77	11.32	
913494	ESR	≥1,620	≥1,620	0.36	0.46	0.12	56.95	13.10	29.83	
913495	TE	≥1,620	≥1,600	0.57	0.44	1.60	19.19	46.31	32.90	
913493	ESR	≥1,620	≥1,620	0.33	0.40	0.18	58.49	15.93	25.40	
91349702	CR	≥1,620	≥1,530	0.19	0.61	1.60	46.43	42.25	5.04	0.74
91349703	CR	≥1,620	≥1,530	0.20	0.59	1.60	45.69	41.45	6.64	0.68
913318	TE	≥1,620	≥1,610	0.37	0.37	0.49	30.44	64.76	4.31	
913308	ESR	≥1,620	≥1,610	0.37	0.53	0.49	53.94	17.10	28.47	
913314	ESR	≥1,620	≥1,610	0.38	0.53	0.68	51.78	18.88	28.66	
913310	ESR	≥1,620	≥1,610	0.23	0.19	0.55	63.16	26.38	9.90	
934821	ESR	≥1,620	≥1,620	0.30	0.23	0.06	58.98	21.71	19.25	
934823	ESR	≥1,620	≥1,620	0.47	0.43	0.00	29.95	45.69	24.35	
934822	TE	≥1,620	≥1,600	0.49	0.55	1.23	38.25	24.35	36.16	
934802	ESR	≥1,620	≥1,620	0.45	0.48	0.12	44.46	21.83	33.58	
915910	ESR	≥1,620	≥1,610	0.26	0.38	0.55	65.74	16.24	17.47	
915902	TE	≥1,620	≥1,610	0.47	0.38	0.86	43.48	19.07	36.59	
915908	MS	≥1,620	≥1,610	0.21	0.39	0.92	61.44	32.90	4.74	
915905	ESR	≥1,620	≥1,600	0.27	0.34	1.05	59.90	23.74	15.31	

Table 6.4 Operational Item Statistics—English Language Arts Grade 3 PBT Administration

ELA Grade 3 Paper-Based Test Administration										
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3
915222	ESR	≥51,490	≥51,350	0.54	0.47	0.27	31.98	28.58	39.17	
915224	ESR	≥51,490	≥51,200	0.42	0.26	0.55	49.00	18.25	32.19	
915225	ESR	≥51,490	≥51,230	0.27	0.29	0.49	65.53	14.07	19.92	
915230	ESR	≥51,490	≥51,220	0.51	0.36	0.52	34.42	28.02	37.04	
915229	ESR	≥51,490	≥51,110	0.27	0.35	0.73	67.83	9.77	21.67	
915219	ESR	≥51,490	≥51,020	0.36	0.27	0.90	54.05	18.02	27.03	
915227P2	CR	≥51,490	≥50,250	0.26	0.65	1.44	35.13	48.67	13.39	0.40
915227P3	CR	≥51,490	≥50,250	0.26	0.59	1.44	36.34	47.40	12.86	1.00
936916	MS	≥51,490	≥51,320	0.31	0.38	0.31	58.12	20.86	20.71	
913494	ESR	≥51,490	≥51,210	0.47	0.40	0.53	47.46	9.65	42.36	
913496	ESR	≥51,490	≥51,250	0.67	0.54	0.45	27.20	10.47	61.88	
913493	ESR	≥51,490	≥51,200	0.42	0.39	0.55	52.25	10.21	36.99	
913497P2	CR	≥51,490	≥50,620	0.28	0.61	0.72	30.32	54.46	11.74	1.80
913497P3	CR	≥51,490	≥50,620	0.32	0.57	0.72	24.29	54.30	18.07	1.66
913315	MS	≥51,490	≥50,040	0.46	0.41	2.80	18.90	67.08	11.21	
913308	ESR	≥51,490	≥49,740	0.55	0.54	3.39	35.45	15.36	45.80	
913314	ESR	≥51,490	≥49,860	0.54	0.53	3.15	33.20	22.67	40.98	
913310	ESR	≥51,490	≥49,620	0.27	0.23	3.63	59.70	22.19	14.48	
934821	ESR	≥51,490	≥51,190	0.36	0.27	0.57	55.36	17.54	26.53	
934823	ESR	≥51,490	≥51,120	0.60	0.38	0.70	16.30	46.74	36.25	
934806	ESR	≥51,490	≥51,190	0.47	0.47	0.57	50.64	4.62	44.16	
934802	ESR	≥51,490	≥51,060	0.61	0.52	0.83	28.41	21.30	49.47	
915910	ESR	≥51,490	≥50,560	0.37	0.42	1.80	54.15	15.15	28.90	
915909	ESR	≥51,490	≥50,390	0.62	0.26	2.13	32.18	9.36	56.34	
915908	MS	≥51,490	≥50,440	0.34	0.47	2.04	48.69	31.76	17.52	
915905	ESR	≥51,490	≥50,020	0.42	0.45	2.85	47.62	16.87	32.66	

Table 6.5 Operational Item Statistics—English Language Arts Grade 4 CBT Administration

ELA Grade 4 Computer-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
913561	ESR	≥7,610	≥7,610	0.56	0.47	0.03	35.49	16.85	47.63		
913562	TE	≥7,610	≥7,600	0.71	0.55	0.25	15.32	27.29	57.15		
913563	ESR	≥7,610	≥7,610	0.46	0.44	0.08	43.40	21.00	35.52		
946024	TE	≥7,610	≥7,600	0.25	0.43	0.24	62.27	24.75	12.74		
913564	ESR	≥7,610	≥7,610	0.54	0.49	0.04	40.82	9.91	49.23		
913566	MS	≥7,610	≥7,610	0.46	0.45	0.09	42.34	22.82	34.74		
91356702	CR	≥7,610	≥7,530	0.21	0.73	0.38	35.06	45.40	16.96	1.51	0.03
91356703	CR	≥7,610	≥7,530	0.29	0.73	0.38	34.06	45.28	17.78	1.82	
913592	ESR	≥7,610	≥7,590	0.41	0.37	0.30	48.25	20.17	31.28		
913594	TE	≥7,610	≥7,550	0.36	0.34	0.84	40.36	45.22	13.58		
998347	ESR	≥7,610	≥7,580	0.21	0.29	0.51	70.98	15.72	12.78		
913595	ESR	≥7,610	≥7,580	0.34	0.28	0.51	58.37	14.03	27.09		
982220	ESR	≥7,610	≥7,610	0.55	0.43	0.01	20.23	50.02	29.74		
982222	ESR	≥7,610	≥7,610	0.38	0.34	0.07	55.32	13.23	31.38		
982223	TE	≥7,610	≥7,600	0.41	0.51	0.14	37.97	42.12	19.77		
982225	ESR	≥7,610	≥7,610	0.44	0.50	0.09	45.32	20.49	34.10		
982227	TE	≥7,610	≥7,610	0.20	0.36	0.11	73.58	12.53	13.78		
982230	MS	≥7,610	≥7,610	0.32	0.49	0.07	51.74	31.76	16.43		
982228	ESR	≥7,610	≥7,610	0.43	0.48	0.12	53.88	6.71	39.30		
982229	ESR	≥7,610	≥7,610	0.64	0.44	0.07	30.78	9.98	59.18		
98223302	CR	≥7,610	≥7,530	0.22	0.79	0.38	35.69	41.90	18.26	2.63	0.38
98223303	CR	≥7,610	≥7,530	0.29	0.79	0.38	37.25	39.87	18.72	3.01	
915315	ESR	≥7,610	≥7,600	0.64	0.51	0.18	24.11	24.52	51.19		
915319	ESR	≥7,610	≥7,600	0.22	0.27	0.13	62.41	30.06	7.40		
915322	ESR	≥7,610	≥7,600	0.37	0.37	0.16	57.72	9.95	32.17		
915325	TE	≥7,610	≥7,590	0.36	0.48	0.29	41.21	44.81	13.69		
915316	ESR	≥7,610	≥7,590	0.39	0.46	0.28	53.04	15.83	30.86		
915317	ESR	≥7,610	≥7,590	0.42	0.49	0.35	50.10	16.28	33.27		



Table 6.6 Operational Item Statistics—English Language Arts Grade 4 PBT Administration

ELA Grade 4 Paper-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
913561	ESR	≥47,250	≥47,120	0.64	0.43	0.26	28.45	15.08	56.21		
946021	ESR	≥47,250	≥47,030	0.46	0.30	0.46	47.67	11.36	40.51		
913563	ESR	≥47,250	≥47,070	0.49	0.41	0.37	41.61	19.06	38.96		
946023	ESR	≥47,250	≥47,050	0.28	0.38	0.41	61.45	20.06	18.07		
913564	ESR	≥47,250	≥47,060	0.52	0.49	0.40	43.83	8.92	46.86		
913566	MS	≥47,250	≥47,070	0.47	0.44	0.38	41.73	22.46	35.42		
913567P2	CR	≥47,250	≥46,780	0.28	0.74	0.76	25.24	39.57	30.25	3.88	0.07
913567P3	CR	≥47,250	≥46,780	0.40	0.73	0.76	21.54	40.84	32.37	4.26	
913592	ESR	≥47,250	≥46,100	0.45	0.33	2.43	44.72	18.64	34.20		
913593	ESR	≥47,250	≥45,960	0.38	0.52	2.72	51.53	17.70	28.06		
998347	ESR	≥47,250	≥45,780	0.24	0.31	3.10	65.79	16.27	14.83		
913595	ESR	≥47,250	≥45,300	0.39	0.31	4.13	52.64	12.15	31.08		
982220	ESR	≥47,250	≥47,110	0.62	0.44	0.29	14.85	45.90	38.96		
982222	ESR	≥47,250	≥47,030	0.45	0.38	0.47	49.10	11.26	39.17		
982221	ESR	≥47,250	≥47,040	0.45	0.35	0.44	31.14	47.70	20.72		
982225	ESR	≥47,250	≥46,980	0.52	0.52	0.56	38.30	19.16	41.98		
982226	ESR	≥47,250	≥46,970	0.43	0.39	0.59	51.58	9.28	38.54		
982230	MS	≥47,250	≥46,960	0.34	0.40	0.60	50.14	31.29	17.97		
982228	ESR	≥47,250	≥47,000	0.49	0.46	0.52	47.90	5.68	45.90		
982229	ESR	≥47,250	≥46,930	0.67	0.39	0.68	28.98	8.53	61.81		
982233P2	CR	≥47,250	≥46,820	0.27	0.77	0.62	26.55	42.48	26.58	3.24	0.23
982233P3	CR	≥47,250	≥46,820	0.36	0.77	0.62	26.59	42.24	26.82	3.44	
915315	ESR	≥47,250	≥46,960	0.66	0.48	0.60	22.46	22.50	54.43		
915319	ESR	≥47,250	≥46,740	0.23	0.28	1.07	62.22	28.07	8.64		
915322	ESR	≥47,250	≥46,780	0.41	0.40	0.99	53.63	9.02	36.36		
915321	ESR	≥47,250	≥46,700	0.37	0.39	1.16	58.58	7.44	32.83		
915316	ESR	≥47,250	≥46,720	0.41	0.45	1.11	50.94	15.56	32.38		
915317	ESR	≥47,250	≥46,480	0.42	0.46	1.62	50.29	13.64	34.45		

Table 6.7 Operational Item Statistics—English Language Arts Grade 5 CBT Administration

ELA Grade 5 Computer-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
799888	ESR	≥55,090	≥55,050	0.42	0.31	0.07	51.94	12.40	35.60		
799889	MS	≥55,090	≥54,920	0.26	0.29	0.30	61.00	26.12	12.57		
799890	ESR	≥55,090	≥55,000	0.58	0.43	0.17	36.27	11.42	52.14		
799891	ESR	≥55,090	≥55,010	0.32	0.27	0.14	55.91	24.74	19.21		
799892	ESR	≥55,090	≥54,970	0.30	0.43	0.22	63.07	12.72	23.99		
995980	TE	≥55,090	≥54,980	0.64	0.39	0.20	17.21	37.90	44.70		
80131002	CR	≥55,090	≥54,690	0.13	0.69	0.40	54.80	36.55	7.29	0.62	0.01
80131003	CR	≥55,090	≥54,690	0.22	0.70	0.40	46.43	42.46	9.45	0.94	
932836	ESR	≥55,090	≥54,140	0.50	0.36	1.73	29.42	38.91	29.95		
932839	ESR	≥55,090	≥54,010	0.56	0.56	1.96	36.75	12.21	49.09		
932840	MS	≥55,090	≥53,800	0.45	0.57	2.33	42.56	22.79	32.32		
932837	TE	≥55,090	≥53,540	0.44	0.58	2.82	43.11	23.46	30.61		
915501	ESR	≥55,090	≥55,070	0.50	0.37	0.03	36.47	26.66	36.84		
915500	ESR	≥55,090	≥55,040	0.61	0.44	0.08	35.94	6.48	57.49		
915507	ESR	≥55,090	≥55,040	0.45	0.45	0.09	49.28	11.74	38.89		
915497	ESR	≥55,090	≥55,050	0.76	0.46	0.08	18.57	10.09	71.26		
915499	ESR	≥55,090	≥55,050	0.46	0.47	0.07	47.15	13.78	39.00		
915511	TE	≥55,090	≥55,060	0.28	0.22	0.06	71.90	0.01	28.03		
915512	TE	≥55,090	≥55,030	0.47	0.55	0.11	33.65	37.87	28.37		
915508	MS	≥55,090	≥55,030	0.27	0.35	0.11	60.93	24.06	14.90		
91551002	CR	≥55,090	≥54,710	0.26	0.78	0.22	32.09	36.14	25.07	5.32	0.70
91551003	CR	≥55,090	≥54,710	0.35	0.77	0.22	32.01	36.09	25.08	6.14	
913665	ESR	≥55,090	≥54,990	0.38	0.46	0.19	50.69	22.35	26.77		
913664	ESR	≥55,090	≥55,010	0.37	0.36	0.14	52.50	21.03	26.33		
913666	TE	≥55,090	≥54,960	0.69	0.42	0.23	9.83	42.84	47.10		
913668	ESR	≥55,090	≥54,950	0.50	0.50	0.26	46.55	5.91	47.28		
913667	MS	≥55,090	≥54,940	0.27	0.37	0.27	59.38	26.38	13.96		
913669	TE	≥55,090	≥54,860	0.27	0.50	0.42	59.24	26.65	13.69		

Table 6.8 Operational Item Statistics—English Language Arts Grade 6 CBT Administration

ELA Grade 6 Computer-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
913709	ESR	≥54,990	≥54,950	0.49	0.40	0.07	39.59	23.39	36.95		
913708	ESR	≥54,990	≥54,910	0.59	0.45	0.14	34.93	12.97	51.95		
913710	ESR	≥54,990	≥54,930	0.43	0.38	0.12	47.08	20.47	32.33		
913711	TE	≥54,990	≥54,870	0.43	0.31	0.22	30.01	53.80	15.98		
980309	ESR	≥54,990	≥54,950	0.51	0.43	0.08	43.50	10.91	45.50		
913712	TE	≥54,990	≥54,860	0.60	0.63	0.24	31.99	15.50	52.26		
913713	MS	≥54,990	≥54,840	0.34	0.42	0.28	53.24	24.25	22.24		
913714	ESR	≥54,990	≥54,810	0.45	0.50	0.33	44.65	19.46	35.56		
91371502	CR	≥54,990	≥54,530	0.34	0.77	0.42	19.92	33.34	38.44	6.80	0.67
91371503	CR	≥54,990	≥54,530	0.47	0.75	0.42	20.15	30.14	37.61	11.26	
913690	MS	≥54,990	≥54,980	0.42	0.42	0.03	37.56	40.83	21.59		
913691	TE	≥54,990	≥54,930	0.39	0.44	0.11	47.39	26.53	25.97		
913692	MS	≥54,990	≥54,920	0.29	0.35	0.13	56.19	30.41	13.27		
913693	TE	≥54,990	≥54,910	0.26	0.44	0.15	64.47	19.29	16.08		
91369402	CR	≥54,990	≥54,550	0.26	0.75	0.36	39.99	23.76	26.86	6.99	1.61
91369403	CR	≥54,990	≥54,550	0.36	0.76	0.36	29.73	38.99	24.33	6.14	
917785	ESR	≥54,990	≥54,890	0.44	0.50	0.18	40.67	30.71	28.44		
917781	ESR	≥54,990	≥54,920	0.37	0.34	0.13	50.51	24.48	24.88		
917755	MS	≥54,990	≥54,910	0.36	0.37	0.15	53.79	20.94	25.12		
917763	TE	≥54,990	≥54,860	0.34	0.37	0.23	47.16	36.80	15.80		
917778	TE	≥54,990	≥54,840	0.59	0.53	0.28	14.00	53.72	32.01		
917721	ESR	≥54,990	≥54,820	0.42	0.21	0.32	46.41	22.86	30.41		
913752	ESR	≥54,990	≥54,980	0.40	0.44	0.02	51.82	17.19	30.97		
913753	TE	≥54,990	≥54,950	0.72	0.52	0.08	18.13	19.54	62.24		
913754	TE	≥54,990	≥54,920	0.74	0.40	0.13	23.92	4.32	71.63		
913755	ESR	≥54,990	≥54,880	0.39	0.40	0.21	54.76	13.08	31.96		
913757	MS	≥54,990	≥54,880	0.34	0.52	0.20	56.34	18.72	24.74		
913756	MS	≥54,990	≥54,900	0.24	0.34	0.16	68.49	15.77	15.58		
980274	TE	≥54,990	≥54,760	0.30	0.24	0.42	52.87	34.30	12.42		
980271	ESR	≥54,990	≥54,710	0.45	0.46	0.51	38.77	31.95	28.77		
980273	MS	≥54,990	≥54,700	0.49	0.39	0.53	18.05	64.74	16.67		
980276	ESR	≥54,990	≥54,660	0.59	0.57	0.60	34.97	12.41	52.02		

Table 6.9 Operational Item Statistics—English Language Arts Grade 7 CBT Administration

ELA Grade 7 Computer-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
915570	ESR	≥52,510	≥52,490	0.70	0.36	0.04	17.70	24.72	57.54		
915572	ESR	≥52,510	≥52,430	0.56	0.43	0.15	33.29	22.05	44.52		
915573	ESR	≥52,510	≥52,410	0.45	0.32	0.19	47.34	15.01	37.46		
915574	TE	≥52,510	≥52,450	0.43	0.39	0.12	50.42	12.28	37.18		
915578	ESR	≥52,510	≥52,460	0.61	0.46	0.11	25.32	27.65	46.91		
915576	TE	≥52,510	≥52,440	0.51	0.52	0.14	34.69	28.04	37.13		
915579	ESR	≥52,510	≥52,390	0.65	0.46	0.23	19.38	31.72	48.67		
915583	MS	≥52,510	≥52,420	0.57	0.54	0.17	23.26	39.60	36.97		
91558202	CR	≥52,510	≥52,030	0.35	0.78	0.45	18.01	33.03	38.39	9.03	0.63
91558203	CR	≥52,510	≥52,030	0.45	0.77	0.45	20.33	32.09	37.34	9.32	
913840	TE	≥52,510	≥52,490	0.33	0.49	0.05	52.90	29.01	18.04		
913839	ESR	≥52,510	≥52,450	0.61	0.44	0.12	28.89	20.03	50.96		
913841	MS	≥52,510	≥52,480	0.29	0.42	0.07	50.09	41.84	8.00		
913838	TE	≥52,510	≥52,500	0.67	0.58	0.03	19.15	27.32	53.50		
91384202	CR	≥52,510	≥52,010	0.39	0.80	0.43	34.47	11.72	24.71	19.00	9.12
91384203	CR	≥52,510	≥52,010	0.49	0.80	0.43	26.60	22.00	29.02	21.41	
913807	ESR	≥52,510	≥52,460	0.64	0.41	0.11	32.84	5.54	61.51		
913808	ESR	≥52,510	≥52,480	0.57	0.50	0.07	33.41	18.85	47.67		
913811	ESR	≥52,510	≥52,480	0.34	0.35	0.07	55.17	20.70	24.06		
913810	ESR	≥52,510	≥52,450	0.33	0.44	0.12	56.83	20.41	22.65		
913812	TE	≥52,510	≥52,470	0.41	0.49	0.08	41.16	35.93	22.82		
913809	TE	≥52,510	≥52,450	0.27	0.47	0.12	56.24	32.56	11.09		
932822	ESR	≥52,510	≥52,490	0.48	0.39	0.05	39.72	25.31	34.92		
932782	ESR	≥52,510	≥52,460	0.61	0.36	0.10	35.60	6.89	57.41		
932785	ESR	≥52,510	≥52,430	0.39	0.44	0.16	54.77	12.42	32.65		
932810	MS	≥52,510	≥52,450	0.49	0.51	0.13	35.94	30.01	33.92		
932791	ESR	≥52,510	≥52,430	0.42	0.27	0.17	51.12	13.29	35.42		
932789	ESR	≥52,510	≥52,430	0.33	0.39	0.17	52.96	28.32	18.55		
932827	ESR	≥52,510	≥52,360	0.50	0.46	0.30	43.10	13.50	43.09		
953139	TE	≥52,510	≥52,370	0.62	0.56	0.27	30.63	15.24	53.86		
932821	MS	≥52,510	≥52,350	0.35	0.55	0.31	46.60	36.72	16.37		
933576	MS	≥52,510	≥52,320	0.39	0.40	0.37	50.03	21.10	28.50		

Table 6.10 Operational Item Statistics—English Language Arts Grade 8 CBT Administration

ELA Grade 8 Computer-Based Test Administration											
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4
913952	ESR	≥51,080	≥51,040	0.36	0.32	0.07	55.38	17.92	26.63		
913953	ESR	≥51,080	≥51,000	0.39	0.38	0.14	48.04	25.26	26.56		
913954	MS	≥51,080	≥51,020	0.47	0.45	0.12	46.34	12.66	40.89		
913955	TE	≥51,080	≥50,960	0.62	0.37	0.23	11.31	53.23	35.22		
913956	ESR	≥51,080	≥51,020	0.47	0.42	0.11	48.57	9.02	42.29		
913957	TE	≥51,080	≥50,990	0.39	0.38	0.17	33.47	54.62	11.74		
91395802	CR	≥51,080	≥50,290	0.38	0.81	0.85	17.54	29.42	36.25	13.69	1.55
91395803	CR	≥51,080	≥50,290	0.53	0.79	0.85	13.92	28.22	40.40	15.92	
982279	ESR	≥51,080	≥50,730	0.58	0.42	0.68	28.70	25.93	44.69		
982281	MS	≥51,080	≥50,680	0.35	0.42	0.78	46.52	35.45	17.25		
982276	ESR	≥51,080	≥50,610	0.58	0.39	0.91	33.83	15.68	49.57		
982278	TE	≥51,080	≥50,300	0.41	0.50	1.52	38.20	39.00	21.28		
982294	MS	≥51,080	≥51,040	0.33	0.39	0.06	43.67	46.75	9.51		
982297	TE	≥51,080	≥51,030	0.45	0.44	0.08	51.64	6.22	42.05		
982299	ESR	≥51,080	≥51,000	0.48	0.38	0.15	43.14	18.24	38.47		
982301	ESR	≥51,080	≥51,010	0.60	0.36	0.12	36.33	7.52	56.03		
982300	ESR	≥51,080	≥51,040	0.51	0.47	0.08	46.89	4.35	48.67		
982302	ESR	≥51,080	≥51,030	0.54	0.44	0.09	41.47	8.70	49.74		
982303	TE	≥51,080	≥50,990	0.60	0.42	0.16	26.84	26.46	46.54		
982304	ESR	≥51,080	≥51,000	0.41	0.43	0.15	55.38	6.14	38.33		
98232702	CR	≥51,080	≥50,330	0.27	0.76	0.90	27.68	41.65	23.36	5.08	0.77
98232703	CR	≥51,080	≥50,330	0.37	0.75	0.90	31.67	31.84	26.69	8.33	
982331	TE	≥51,080	≥50,980	0.67	0.46	0.19	12.92	40.19	46.70		
982330	ESR	≥51,080	≥51,000	0.47	0.52	0.15	50.30	5.00	44.54		
982333	ESR	≥51,080	≥51,010	0.38	0.28	0.13	54.48	15.16	30.23		
982332	TE	≥51,080	≥51,020	0.31	0.44	0.11	56.13	24.65	19.10		
913974	ESR	≥51,080	≥51,000	0.40	0.42	0.14	52.90	13.75	33.20		
913970	MS	≥51,080	≥51,010	0.50	0.29	0.13	22.29	55.67	21.91		
913971	ESR	≥51,080	≥50,970	0.26	0.12	0.22	67.28	13.81	18.69		
913972	MS	≥51,080	≥50,970	0.36	0.45	0.21	39.19	48.83	11.77		
913973	ESR	≥51,080	≥50,970	0.34	0.39	0.21	52.35	27.43	20.01		
913975	MS	≥51,080	≥50,970	0.49	0.43	0.21	31.85	37.48	30.46		

Table 6.11 Operational Item Statistics—Mathematics Grade 3 CBT Administration

Mathematics Grade 3 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
896892	MC	≥1,610	≥1,610	0.68	0.45	0.06							
913997	SA	≥1,610	≥1,600	0.41	0.59	0.87	58.80	40.33					
896772	MC	≥1,610	≥1,610	0.30	0.40	0.19							
914024	SA	≥1,610	≥1,600	0.42	0.29	0.56	58.12	41.33					
904404	MC	≥1,610	≥1,610	0.52	0.49	0.19							
914038	SA	≥1,610	≥1,600	0.38	0.40	0.37	61.52	38.10					
981774	MC	≥1,610	≥1,610	0.37	0.35	0.12							
981799	MC	≥1,610	≥1,600	0.54	0.31	0.50							
896859	SA	≥1,610	≥1,600	0.33	0.55	0.62	66.73	32.65					
981778	MC	≥1,610	≥1,610	0.58	0.32	0.06							
906209	MPSR	≥1,610	≥1,610	0.42	0.30	0.19	35.25	44.92	19.64				
981751	MC	≥1,610	≥1,610	0.65	0.36	0.06							
913987	MC	≥1,610	≥1,600	0.59	0.39	0.37							
981736	CR	≥1,610	≥1,560	0.21	0.56	0.87	50.43	22.61	14.99	6.69	2.11		
914048	CR	≥1,610	≥1,520	0.23	0.62	2.42	58.92	16.42	9.42	9.48			
981762	SA	≥1,610	≥1,610	0.71	0.19	0.12	29.37	70.51					
906210	MC	≥1,610	≥1,600	0.78	0.39	0.31							
896684	SA	≥1,610	≥1,610	0.23	0.36	0.25	76.58	23.17					
916044	SA	≥1,610	≥1,600	0.32	0.48	0.31	48.02	39.53	12.14				
935017	MS	≥1,610	≥1,610	0.16	0.32	0.19	84.20	15.61					
896862	MC	≥1,610	≥1,610	0.61	0.24	0.25							
981795	MC	≥1,610	≥1,610	0.58	0.41	0.06							
981767	MC	≥1,610	≥1,600	0.57	0.42	0.37							
914023	SA	≥1,610	≥1,610	0.55	0.56	0.19	45.42	54.40					
896902	SA	≥1,610	≥1,600	0.33	0.61	0.31	46.10	41.64	11.96				
914007	SA	≥1,610	≥1,600	0.19	0.53	0.74	80.30	18.96					
896860	SA	≥1,610	≥1,600	0.26	0.51	0.81	73.42	25.77					
898001	CR	≥1,610	≥1,540	0.17	0.56	1.86	63.44	15.18	15.55	1.24			
981742	CR	≥1,610	≥1,590	0.18	0.67	1.43	61.96	25.90	5.82	4.89			
981784	MC	≥1,610	≥1,610	0.56	0.53	0.06							
896770	SA	≥1,610	≥1,610	0.33	0.42	0.19	66.67	33.15					
981791	MC	≥1,610	≥1,610	0.53	0.47	0.12							
896868	MC	≥1,610	≥1,610	0.32	0.18	0.19							
896867	SA	≥1,610	≥1,600	0.60	0.46	0.37	39.59	60.04					
896863	MC	≥1,610	≥1,610	0.85	0.33	0.12							
896679	MC	≥1,610	≥1,600	0.58	0.46	0.37							
913991	MC	≥1,610	≥1,610	0.49	0.43	0.19							
914001	MS	≥1,610	≥1,610	0.26	0.40	0.25	73.85	25.90					
878608	MC	≥1,610	≥1,610	0.69	0.40	0.12							

Mathematics Grade 3 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
896760	SA	≥1,610	≥1,600	0.43	0.63	0.56	56.69	42.75					
914036	MS	≥1,610	≥1,610	0.35	0.52	0.12	64.99	34.88					
914039	CR	≥1,610	≥1,560	0.33	0.61	0.87	35.19	32.47	25.59	3.66			
981747	CR	≥1,610	≥1,610	0.27	0.72	0.25	24.23	35.63	16.36	11.65	5.51	3.66	2.73

Table 6.12 Operational Item Statistics—Mathematics Grade 3 PBT Administration

Mathematics Grade 3 Paper-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
896892	MC	≥51,380	≥51,180	0.78	0.46	0.22							
913997	SA	≥51,380	≥50,000	0.52	0.59	2.67	46.76	50.57					
896772	MC	≥51,380	≥50,850	0.47	0.49	0.69							
914024	SA	≥51,380	≥50,390	0.48	0.29	1.92	50.96	47.11					
904404	MC	≥51,380	≥49,800	0.66	0.49	1.67							
914038	SA	≥51,380	≥50,420	0.52	0.43	1.86	46.80	51.34					
981774	MC	≥51,380	≥50,460	0.49	0.45	0.75							
981799	MC	≥51,380	≥48,840	0.68	0.42	0.76							
896859	SA	≥51,380	≥50,340	0.49	0.55	2.01	50.40	47.59					
981778	MC	≥51,380	≥50,860	0.68	0.30	0.63							
906209	MPSR	≥51,380	≥51,000	0.54	0.40	0.75	25.98	39.43	33.85				
981751	MC	≥51,380	≥50,520	0.76	0.36	1.51							
913987	MC	≥51,380	≥50,060	0.72	0.43	1.63							
981736	CR	≥51,380	≥50,730	0.38	0.54	1.07	29.58	19.71	25.68	16.76	7.01		
914048	CR	≥51,380	≥46,900	0.45	0.63	8.53	33.60	17.58	15.83	24.29			
981762	SA	≥51,380	≥50,920	0.73	0.28	0.90	27.23	71.87					
906210	MC	≥51,380	≥50,970	0.89	0.36	0.51							
896684	SA	≥51,380	≥50,610	0.38	0.45	1.49	61.42	37.09					
916044	SA	≥51,380	≥50,970	0.46	0.57	0.80	33.23	39.83	26.13				
935017	MS	≥51,380	≥50,970	0.33	0.44	0.80	66.82	32.38					
896862	MC	≥51,380	≥50,680	0.67	0.29	0.98							
981795	MC	≥51,380	≥49,650	0.71	0.48	1.51							
981767	MC	≥51,380	≥50,970	0.75	0.45	0.66							
914023	SA	≥51,380	≥50,780	0.71	0.53	1.17	28.73	70.10					
896902	SA	≥51,380	≥51,020	0.49	0.67	0.69	29.27	43.03	27.02				
914007	SA	≥51,380	≥50,450	0.37	0.61	1.81	61.89	36.30					
896860	SA	≥51,380	≥50,430	0.40	0.55	1.84	59.12	39.04					
898001	CR	≥51,380	≥50,400	0.27	0.56	1.70	49.41	18.94	27.25	2.49			
981742	CR	≥51,380	≥50,930	0.33	0.68	0.87	41.41	32.32	9.18	16.22			
981784	MC	≥51,380	≥50,960	0.74	0.52	0.62							
896770	SA	≥51,380	≥50,740	0.55	0.39	1.25	44.17	54.58					
981791	MC	≥51,380	≥50,740	0.67	0.52	0.95							
896868	MC	≥51,380	≥50,800	0.43	0.27	0.83							
896867	SA	≥51,380	≥50,510	0.73	0.40	1.68	26.35	71.97					
896863	MC	≥51,380	≥51,040	0.91	0.31	0.56							
896679	MC	≥51,380	≥50,240	0.71	0.49	1.80							
913991	MC	≥51,380	≥50,720	0.66	0.49	1.18							
914001	MS	≥51,380	≥50,830	0.45	0.47	1.07	54.32	44.61					
878608	MC	≥51,380	≥50,920	0.82	0.46	0.67							



Mathematics Grade 3 Paper-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
896760	SA	≥51,380	≥50,540	0.60	0.61	1.64	38.93	59.43					
914036	MS	≥51,380	≥50,340	0.49	0.53	2.02	49.52	48.46					
914039	CR	≥51,380	≥50,680	0.47	0.56	1.24	17.61	27.99	49.08	3.95			
981747	CR	≥51,380	≥51,280	0.47	0.76	0.18	12.65	17.86	15.10	18.74	12.82	8.50	14.14

Table 6.13 Operational Item Statistics—Mathematics Grade 4 CBT Administration

Mathematics Grade 4 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
870707	MC	≥7,590	≥7,580	0.69	0.49	0.14							
870319	SA	≥7,590	≥7,570	0.47	0.48	0.20	52.82	46.98					
981843	MS	≥7,590	≥7,580	0.40	0.55	0.12	59.48	40.40					
981835	SA	≥7,590	≥7,560	0.20	0.54	0.32	80.23	19.45					
897478	MC	≥7,590	≥7,580	0.57	0.32	0.07							
981874	MPSR	≥7,590	≥7,590	0.52	0.63	0.01	29.70	36.83	33.46				
981867	SA	≥7,590	≥7,570	0.41	0.58	0.20	59.11	40.69					
897446	SA	≥7,590	≥7,570	0.64	0.54	0.29	35.51	64.20					
914137	MC	≥7,590	≥7,570	0.61	0.44	0.26							
944080	MC	≥7,590	≥7,580	0.64	0.43	0.08							
981844	SA	≥7,590	≥7,550	0.26	0.56	0.55	74.06	25.38					
914080	MS	≥7,590	≥7,580	0.73	0.48	0.08	26.83	73.09					
914070	SA	≥7,590	≥7,560	0.40	0.63	0.40	59.81	39.79					
914084	CR	≥7,590	≥7,580	0.34	0.70	0.05	24.25	33.88	26.15	13.94	1.74		
914086	CR	≥7,590	≥7,370	0.19	0.59	1.78	64.75	19.30	3.71	9.33			
914101	MC	≥7,590	≥7,580	0.79	0.39	0.09							
897470	MC	≥7,590	≥7,580	0.50	0.64	0.13							
897468	MC	≥7,590	≥7,580	0.38	0.43	0.04							
914082	SA	≥7,590	≥7,580	0.29	0.43	0.16	70.40	29.44					
897302	MC	≥7,590	≥7,580	0.52	0.47	0.12							
914121	SA	≥7,590	≥7,580	0.71	0.39	0.11	29.11	70.79					
914088	MC	≥7,590	≥7,580	0.52	0.51	0.07							
897444	SA	≥7,590	≥7,590	0.71	0.53	0.03	13.76	29.66	56.55				
878669	SA	≥7,590	≥7,580	0.57	0.47	0.09	18.28	49.45	32.18				
897475	SA	≥7,590	≥7,570	0.68	0.45	0.18	32.43	67.39					
897291	MS	≥7,590	≥7,580	0.67	0.55	0.05	32.61	67.33					
981838	MC	≥7,590	≥7,580	0.30	0.47	0.09							
981831	CR	≥7,590	≥7,460	0.25	0.69	0.72	55.90	18.35	16.32	7.76			
899959	CR	≥7,590	≥7,460	0.37	0.66	0.51	39.99	26.80	12.38	19.20			
897434	MC	≥7,590	≥7,580	0.83	0.37	0.05							
981850	MC	≥7,590	≥7,570	0.48	0.42	0.24							
898008	SA	≥7,590	≥7,570	0.66	0.43	0.24	34.30	65.46					
981890	MS	≥7,590	≥7,580	0.77	0.44	0.13	23.35	76.51					
914135	MC	≥7,590	≥7,570	0.85	0.36	0.24							
897305	MC	≥7,590	≥7,570	0.33	0.33	0.17							
897438	MC	≥7,590	≥7,580	0.75	0.42	0.12							
914099	SA	≥7,590	≥7,550	0.27	0.54	0.51	72.87	26.62					
897471	SA	≥7,590	≥7,510	0.49	0.49	1.00	50.96	48.04					
981866	SA	≥7,590	≥7,550	0.45	0.55	0.49	54.86	44.65					

Mathematics Grade 4 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
981853	SA	≥7,590	≥7,550	0.46	0.55	0.55	53.35	46.10					
914108	MS	≥7,590	≥7,560	0.39	0.44	0.37	61.12	38.51					
899955	CR	≥7,590	≥7,350	0.18	0.63	1.83	58.05	26.03	10.71	2.03			
981827	CR	≥7,590	≥7,350	0.18	0.66	2.27	57.19	8.61	14.81	3.71	7.60	2.40	2.58

Table 6.14 Operational Item Statistics—Mathematics Grade 4 PBT Administration

Mathematics Grade 4 Paper-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
870707	MC	≥47,150	≥46,630	0.73	0.50	1.01							
870319	SA	≥47,150	≥46,320	0.44	0.47	1.76	54.99	43.25					
981843	MS	≥47,150	≥46,810	0.43	0.55	0.73	56.44	42.83					
981835	SA	≥47,150	≥46,110	0.22	0.52	2.22	76.12	21.65					
897478	MC	≥47,150	≥46,770	0.62	0.32	0.45							
981874	MPSR	≥47,150	≥46,910	0.56	0.63	0.52	25.65	36.63	37.20				
981867	SA	≥47,150	≥45,940	0.42	0.57	2.58	56.08	41.34					
897446	SA	≥47,150	≥46,230	0.68	0.48	1.96	31.77	66.28					
914137	MC	≥47,150	≥46,440	0.66	0.43	1.41							
944080	MC	≥47,150	≥46,830	0.68	0.43	0.57							
981844	SA	≥47,150	≥45,040	0.29	0.53	4.48	68.05	27.46					
914080	MS	≥47,150	≥46,560	0.80	0.46	1.25	20.19	78.56					
914070	SA	≥47,150	≥45,700	0.43	0.60	3.08	55.68	41.23					
914084	CR	≥47,150	≥47,020	0.38	0.68	0.29	20.67	29.87	28.07	17.35	3.75		
914086	CR	≥47,150	≥44,980	0.26	0.59	4.47	52.23	25.11	5.68	12.36			
914101	MC	≥47,150	≥46,940	0.83	0.37	0.29							
897470	MC	≥47,150	≥46,890	0.55	0.62	0.45							
897468	MC	≥47,150	≥46,840	0.40	0.45	0.56							
914082	SA	≥47,150	≥46,470	0.27	0.41	1.45	71.56	26.99					
897302	MC	≥47,150	≥46,760	0.57	0.49	0.54							
914121	SA	≥47,150	≥46,310	0.73	0.38	1.78	26.22	72.00					
914088	MC	≥47,150	≥46,280	0.59	0.50	0.77							
897444	SA	≥47,150	≥46,960	0.74	0.54	0.42	11.98	27.08	60.52				
878669	SA	≥47,150	≥46,920	0.61	0.43	0.50	14.98	48.42	36.09				
897475	SA	≥47,150	≥46,640	0.64	0.40	1.10	35.23	63.67					
897291	MS	≥47,150	≥46,730	0.69	0.52	0.91	30.65	68.45					
981838	MC	≥47,150	≥46,550	0.33	0.48	1.02							
981831	CR	≥47,150	≥46,690	0.27	0.68	0.88	50.42	21.56	21.64	5.39			
899959	CR	≥47,150	≥46,420	0.40	0.61	1.46	29.81	33.32	20.18	15.14			
897434	MC	≥47,150	≥46,940	0.86	0.35	0.37							
981850	MC	≥47,150	≥46,820	0.50	0.41	0.49							
898008	SA	≥47,150	≥46,260	0.69	0.41	1.90	30.18	67.92					
981890	MS	≥47,150	≥46,750	0.78	0.43	0.87	21.36	77.78					
914135	MC	≥47,150	≥46,660	0.88	0.34	0.95							
897305	MC	≥47,150	≥46,570	0.35	0.34	0.84							
897438	MC	≥47,150	≥46,770	0.77	0.43	0.72							
914099	SA	≥47,150	≥45,900	0.30	0.53	2.66	68.21	29.13					
897471	SA	≥47,150	≥45,430	0.54	0.48	3.65	44.58	51.77					
981866	SA	≥47,150	≥45,610	0.52	0.51	3.28	46.47	50.26					

Mathematics Grade 4 Paper-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
981853	SA	≥47,150	≥46,230	0.53	0.50	1.96	45.79	52.26					
914108	MS	≥47,150	≥46,520	0.42	0.44	1.35	57.09	41.56					
899955	CR	≥47,150	≥46,400	0.35	0.66	1.46	44.41	9.08	39.50	5.41			
981827	CR	≥47,150	≥46,550	0.24	0.65	1.22	48.98	8.75	18.50	4.52	11.05	2.99	3.94

Table 6.15 Operational Item Statistics—Mathematics Grade 5 CBT Administration

Mathematics Grade 5 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
898155	MC	≥54,900	≥54,880	0.63	0.43	0.05							
903245	MC	≥54,900	≥54,860	0.62	0.29	0.08							
914214	TE	≥54,900	≥54,890	0.76	0.35	0.03	24.25	75.72					
898173	SA	≥54,900	≥54,810	0.43	0.61	0.17	56.52	43.31					
800136	MC	≥54,900	≥54,850	0.64	0.37	0.10							
898145	MS	≥54,900	≥54,880	0.67	0.30	0.05	32.96	66.99					
898144	MC	≥54,900	≥54,830	0.60	0.55	0.14							
982506	SA	≥54,900	≥54,800	0.42	0.51	0.20	58.09	41.72					
898141	SA	≥54,900	≥54,890	0.48	0.62	0.02	35.37	32.62	31.99				
914209	MC	≥54,900	≥54,860	0.47	0.46	0.07							
898159	SA	≥54,900	≥54,830	0.47	0.58	0.14	52.56	47.30					
898151	MC	≥54,900	≥54,840	0.64	0.32	0.11							
914152	CR	≥54,900	≥54,370	0.35	0.68	0.44	28.70	28.89	20.55	14.91	5.99		
914148	CR	≥54,900	≥54,190	0.27	0.71	0.72	49.26	26.68	16.14	6.63			
914150	MC	≥54,900	≥54,860	0.77	0.26	0.08							
870762	SA	≥54,900	≥54,770	0.25	0.57	0.23	60.69	27.48	11.60				
982499	SA	≥54,900	≥54,860	0.47	0.48	0.08	52.62	47.30					
914190	SA	≥54,900	≥54,660	0.54	0.45	0.44	45.34	54.22					
898152	MS	≥54,900	≥54,870	0.34	0.49	0.06	66.12	33.82					
898011	MC	≥54,900	≥54,820	0.51	0.42	0.15							
914187	MPSR	≥54,900	≥54,870	0.51	0.54	0.06	33.33	32.11	34.50				
914215	MC	≥54,900	≥54,850	0.63	0.50	0.11							
897984	MC	≥54,900	≥54,800	0.45	0.53	0.19							
903244	MC	≥54,900	≥54,830	0.39	0.36	0.14							
982518	MS	≥54,900	≥54,840	0.77	0.44	0.12	23.40	76.48					
902410	CR	≥54,900	≥54,840	0.39	0.54	0.12	30.15	36.01	21.32	12.40			
902414	CR	≥54,900	≥53,810	0.15	0.53	1.23	72.81	9.61	11.73	3.86			
914140	SA	≥54,900	≥54,850	0.42	0.39	0.11	58.22	41.67					
914171	SA	≥54,900	≥54,840	0.66	0.49	0.11	34.14	65.74					
982538	MC	≥54,900	≥54,860	0.62	0.37	0.09							
914580	TE	≥54,900	≥54,860	0.64	0.41	0.09	35.75	64.17					
898162	MC	≥54,900	≥54,870	0.47	0.45	0.06							
914164	SA	≥54,900	≥54,750	0.29	0.30	0.27	70.99	28.74					
914155	TE	≥54,900	≥54,880	0.49	0.34	0.05	51.36	48.59					
914184	SA	≥54,900	≥54,800	0.69	0.37	0.19	31.11	68.70					
914198	SA	≥54,900	≥54,820	0.42	0.64	0.14	58.33	41.53					
982534	MS	≥54,900	≥54,840	0.26	0.51	0.11	73.77	26.13					
914203	MC	≥54,900	≥54,870	0.69	0.31	0.06							
870786	SA	≥54,900	≥54,720	0.56	0.48	0.33	43.38	56.29					

Mathematics Grade 5 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
914195	CR	≥54,900	≥54,850	0.31	0.69	0.10	47.23	25.32	14.80	12.56			
934015	CR	≥54,900	≥54,810	0.29	0.65	0.18	16.82	41.27	17.82	10.01	5.29	5.64	2.96

Table 6.16 Item Statistics—Mathematics Grade 6 Computer-Based Test Administration

Mathematics Grade 6 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
903096	MS	≥54,910	≥54,900	0.70	0.52	0.02	29.74	70.24					
901541	SA	≥54,910	≥54,640	0.75	0.37	0.48	24.93	74.58					
914223	TE	≥54,910	≥54,850	0.60	0.47	0.10	39.67	60.23					
914260	SA	≥54,910	≥54,590	0.41	0.41	0.58	58.76	40.67					
981981	TE	≥54,910	≥54,840	0.48	0.53	0.13	51.89	47.99					
916476	SA	≥54,910	≥54,760	0.34	0.32	0.27	65.99	33.75					
900521	SA	≥54,910	≥54,690	0.32	0.51	0.39	68.02	31.59					
901543	MS	≥54,910	≥54,860	0.48	0.43	0.08	51.51	48.41					
901534	MPSR	≥54,910	≥54,850	0.71	0.47	0.11	13.06	32.65	54.18				
878302	MC	≥54,910	≥54,830	0.65	0.47	0.14							
914259	TE	≥54,910	≥54,840	0.61	0.44	0.13	38.93	60.94					
800198	MC	≥54,910	≥54,780	0.18	0.28	0.22							
914237	TE	≥54,910	≥54,810	0.65	0.42	0.17	34.59	65.23					
914268	TE	≥54,910	≥54,780	0.40	0.45	0.23	59.47	40.30					
914230	SA	≥54,910	≥54,630	0.59	0.52	0.50	40.97	58.54					
878299	MC	≥54,910	≥54,690	0.50	0.34	0.39							
903077	SA	≥54,910	≥54,560	0.43	0.33	0.64	56.65	42.71					
914257	TE	≥54,910	≥54,570	0.72	0.54	0.61	28.09	71.30					
903099	MS	≥54,910	≥54,890	0.62	0.46	0.03	37.54	62.43					
982013	MC	≥54,910	≥54,870	0.38	0.44	0.07							
982025	TE	≥54,910	≥54,600	0.36	0.57	0.56	64.10	35.34					
901547	SA	≥54,910	≥54,780	0.52	0.62	0.23	37.59	20.29	41.89				
982019	SA	≥54,910	≥54,760	0.35	0.61	0.26	64.38	35.36					
903092	MC	≥54,910	≥54,850	0.44	0.27	0.11							
981963	CR	≥54,910	≥54,020	0.28	0.63	0.84	39.68	25.93	18.59	10.70	3.49		
982011	SA	≥54,910	≥54,800	0.31	0.50	0.19	68.38	31.43					
945486	SA	≥54,910	≥54,750	0.23	0.59	0.28	68.94	15.23	15.55				
981961	CR	≥54,910	≥53,840	0.26	0.61	1.15	47.60	30.31	14.76	5.39			
981954	CR	≥54,910	≥53,540	0.09	0.50	1.72	70.69	14.52	5.22	2.12	2.55	1.01	1.39
981956	CR	≥54,910	≥54,180	0.43	0.66	1.32	33.60	17.99	31.55	15.54			
914249	MC	≥54,910	≥54,840	0.55	0.35	0.11							
914271	SA	≥54,910	≥54,700	0.33	0.60	0.37	67.18	32.45					
901536	SA	≥54,910	≥54,850	0.54	0.56	0.10	46.16	53.74					
914273	SA	≥54,910	≥54,790	0.52	0.62	0.21	25.03	46.31	28.44				
914233	MS	≥54,910	≥54,870	0.22	0.52	0.06	77.47	22.47					
902741	TE	≥54,910	≥54,820	0.76	0.39	0.16	24.18	75.66					
903102	SA	≥54,910	≥54,740	0.29	0.59	0.30	70.61	29.09					
902748	MC	≥54,910	≥54,850	0.48	0.39	0.11							
914280	SA	≥54,910	≥54,780	0.26	0.58	0.23	73.82	25.95					



Mathematics Grade 6 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	<i>p</i> -Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
914231	CR	≥54,910	≥53,700	0.36	0.67	0.98	43.55	18.85	18.49	16.92			
903511	CR	≥54,910	≥54,850	0.25	0.55	0.10	31.76	49.21	8.72	7.47	2.75		
914281	CR	≥54,910	≥53,740	0.27	0.66	1.29	59.54	13.31	8.87	16.14			

Table 6.17 Item Statistics—Mathematics Grade 7 Computer-Based Test Administration

Mathematics Grade 7 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
915100	MC	≥52,260	≥52,220	0.76	0.43	0.07							
914294	MS	≥52,260	≥52,130	0.23	0.52	0.25	76.63	23.12					
870847	SA	≥52,260	≥52,180	0.51	0.35	0.15	48.46	51.39					
982970	TE	≥52,260	≥52,170	0.20	0.49	0.18	79.50	20.33					
914299	MC	≥52,260	≥52,170	0.38	0.40	0.16							
914324	SA	≥52,260	≥51,990	0.56	0.42	0.51	43.74	55.76					
899318	MS	≥52,260	≥52,200	0.32	0.18	0.11	67.87	32.02					
983000	TE	≥52,260	≥52,200	0.48	0.36	0.11	52.32	47.57					
914359	SA	≥52,260	≥52,110	0.53	0.53	0.29	47.09	52.62					
914293	MS	≥52,260	≥52,180	0.33	0.55	0.16	66.42	33.42					
899322	SA	≥52,260	≥52,200	0.45	0.67	0.12	36.44	36.69	26.76				
983019	MC	≥52,260	≥52,160	0.69	0.40	0.20							
983004	SA	≥52,260	≥51,930	0.56	0.36	0.63	43.51	55.86					
914340	MC	≥52,260	≥52,000	0.45	0.45	0.50							
982988	MS	≥52,260	≥52,050	0.25	0.54	0.40	75.14	24.46					
983009	MC	≥52,260	≥51,950	0.29	0.29	0.59							
897990	SA	≥52,260	≥51,590	0.40	0.56	1.28	58.84	39.87					
983024	MC	≥52,260	≥51,880	0.44	0.33	0.72							
798344	MC	≥52,260	≥51,830	0.50	0.51	0.81							
899859	MC	≥52,260	≥52,230	0.33	0.32	0.06							
914330	MS	≥52,260	≥52,180	0.73	0.32	0.15	27.03	72.83					
982964	TE	≥52,260	≥52,220	0.46	0.38	0.07	54.00	45.93					
914633	MS	≥52,260	≥52,230	0.71	0.40	0.06	29.32	70.63					
899323	SA	≥52,260	≥52,120	0.57	0.49	0.26	42.89	56.86					
982941	MC	≥52,260	≥52,150	0.39	0.09	0.22							
982954	TE	≥52,260	≥52,180	0.33	0.59	0.15	58.14	17.41	24.30				
914362	CR	≥52,260	≥51,580	0.14	0.64	0.67	77.49	2.72	2.15	4.06	2.57	3.73	5.99
914316	TE	≥52,260	≥52,110	0.36	0.47	0.29	63.74	35.97					
902446	MC	≥52,260	≥52,210	0.43	0.28	0.10							
982922	CR	≥52,260	≥50,190	0.25	0.68	2.50	58.58	9.48	21.30	6.67			
868848	CR	≥52,260	≥49,580	0.08	0.55	2.95	80.29	6.53	6.81	1.26			
900539	CR	≥52,260	≥51,420	0.33	0.68	1.61	38.15	21.61	14.98	14.37	9.28		
914342	MC	≥52,260	≥52,210	0.48	0.37	0.10							
914319	SA	≥52,260	≥52,230	0.35	0.54	0.06	44.22	41.79	13.93				
982947	MC	≥52,260	≥52,200	0.39	0.34	0.10							
898444	SA	≥52,260	≥52,160	0.67	0.54	0.19	33.30	66.51					
900174	MC	≥52,260	≥52,220	0.84	0.37	0.07							
982935	MC	≥52,260	≥52,220	0.47	0.21	0.07							
914335	MPSR	≥52,260	≥52,220	0.36	0.45	0.07	43.58	41.57	14.78				

Mathematics Grade 7 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
870880	MC	≥52,260	≥52,200	0.55	0.27	0.11							
900520	CR	≥52,260	≥49,710	0.12	0.60	3.02	80.03	4.28	3.36	7.46			
914339	CR	≥52,260	≥51,150	0.18	0.57	1.13	62.44	9.78	19.76	2.47	3.43		
982929	CR	≥52,260	≥50,260	0.25	0.62	2.37	55.26	16.15	18.08	6.69			

Table 6.18 Item Statistics—Mathematics Grade 8 Computer-Based Test Administration

Mathematics Grade 8 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
983101	MC	≥44,850	≥44,710	0.62	0.37	0.30							
983049	MS	≥44,850	≥44,680	0.37	0.52	0.39	62.78	36.84					
914366	SA	≥44,850	≥43,820	0.29	0.46	2.30	69.73	27.98					
983076	MC	≥44,850	≥44,650	0.45	0.26	0.44							
897458	SA	≥44,850	≥43,380	0.22	0.48	3.28	75.02	21.70					
903089	MS	≥44,850	≥44,630	0.47	0.49	0.48	52.44	47.08					
914367	MC	≥44,850	≥44,710	0.40	0.29	0.32							
983117	MC	≥44,850	≥44,750	0.71	0.27	0.22							
983063	TE	≥44,850	≥44,720	0.47	0.53	0.30	33.38	38.90	27.42				
897074	MS	≥44,850	≥44,790	0.28	0.42	0.14	72.08	27.78					
914438	MC	≥44,850	≥44,630	0.46	0.23	0.49							
914427	MS	≥44,850	≥44,500	0.41	0.56	0.77	58.83	40.40					
868884	MS	≥44,850	≥44,630	0.32	0.55	0.49	68.11	31.40					
896995	MS	≥44,850	≥44,620	0.32	0.42	0.51	68.15	31.34					
983032	SA	≥44,850	≥43,570	0.18	0.56	2.84	79.31	17.84					
891485	MS	≥44,850	≥44,450	0.20	0.40	0.89	78.98	20.14					
914431	SA	≥44,850	≥43,550	0.26	0.58	2.90	71.55	25.55					
896996	MC	≥44,850	≥44,260	0.53	0.30	1.30							
944912	MPSR	≥44,850	≥44,220	0.38	0.49	1.41	40.82	40.42	17.35				
914370	MS	≥44,850	≥43,890	0.29	0.56	2.15	69.08	28.77					
983074	MC	≥44,850	≥44,740	0.44	0.39	0.24							
903088	MPSR	≥44,850	≥44,840	0.81	0.45	0.02	9.20	18.75	72.03				
914433	MC	≥44,850	≥44,720	0.42	0.10	0.28							
914420	MPSR	≥44,850	≥44,810	0.54	0.43	0.09	23.45	45.18	31.28				
983034	TE	≥44,850	≥44,720	0.26	0.62	0.28	73.72	26.00					
983010	CR	≥44,850	≥44,130	0.21	0.59	0.91	38.91	25.29	16.05	11.40	4.62	1.62	0.50
897072	SA	≥44,850	≥44,480	0.24	0.58	0.82	75.38	23.80					
982987	CR	≥44,850	≥43,140	0.16	0.54	2.40	64.85	13.20	9.98	2.89	5.27		
982999	CR	≥44,850	≥42,480	0.14	0.52	3.52	67.48	17.37	5.78	4.08			
870899	CR	≥44,850	≥42,080	0.10	0.52	4.70	74.80	11.54	5.38	2.11			
914429	MC	≥44,850	≥44,800	0.57	0.32	0.10							
983109	TE	≥44,850	≥44,790	0.70	0.42	0.14	29.82	70.04					
914436	SA	≥44,850	≥44,550	0.30	0.67	0.68	56.87	25.78	16.68				
914396	MC	≥44,850	≥44,780	0.48	0.52	0.16							
914397	MC	≥44,850	≥44,790	0.30	0.49	0.14							
899312	CR	≥44,850	≥44,630	0.38	0.64	0.48	35.76	27.07	23.13	13.56			
914430	MS	≥44,850	≥44,760	0.25	0.55	0.19	74.66	25.16					
914426	MC	≥44,850	≥44,790	0.62	0.41	0.13							
914381	CR	≥44,850	≥42,690	0.17	0.63	2.57	56.91	14.82	21.10	1.67	0.68		

Mathematics Grade 8 Computer-Based Test Administration													
Item ID	Item Type	Total N	Adj. N	p-Value	Pbis	Omit Rate	% at 0	% at 1	% at 2	% at 3	% at 4	% at 5	% at 6
982967	TE	≥44,850	≥44,710	0.09	0.32	0.31	90.53	9.15					
899329	CR	≥44,850	≥44,620	0.30	0.50	0.51	42.51	33.05	15.05	8.88			

These item level statistics are reviewed at the beginning of the operational analyses process to ensure that items are unflawed and a careful review is given to determine that the answer key is correct.

A multiple-choice (MC) item is reviewed during the key check process if

- it has a  $p$ -value less than 0.25 or more than .95,
- greater number of high-performing students (top 20%) choosing a distractor than are choosing the key,
- the item-total correlation of the keyed response is less than 0.20,
- any of the incorrect answer options yields a positive distractor-total correlation, or
- the percentage of students omitting or not reaching each item is 5 or greater.

Other types of autoscored items are also flagged during the key check for review if

- they have a  $p$ -value less than 0.30 or more than .80,
- the percentage of students who reached any possible score point is less than 3,
- the item-total correlation is less than 0.20, or
- the flagging criteria for omit item is 15%.

### 6.3 Item Response Theory

Item parameters for items included in the ELA and mathematics tests were estimated using a marginal maximum-likelihood (MML) procedure and the 2-parameter logistic (2PL) model for MC items and the generalized partial credit (GPC) model (Muraki, 1992) for non-MC items. Under the 2PL model, the probability that a student with a trait or scale score of  $\theta$  will respond correctly to MC item  $j$  is

$$P_j(\theta) = 1/[1 + \exp(-1.7a_j(\theta - b_j))].$$

In the equation,  $a_j$  is the item discrimination and  $b_j$  is the item difficulty. Under the GPC model, the probability that a student with a trait or scale score of  $\theta$  will respond in category  $x$  to partial-credit item  $j$  is

$$P_{jx}(\theta) = \exp\left[\sum_{k=0}^x (Z_{jk}(\theta))\right] / \sum_{h=0}^{m_j} \exp\left[\sum_{k=0}^h (Z_{jk}(\theta))\right],$$

$$\text{where } z_{jk}(\theta) = Da_j(\theta - b_j + d_{jx}),$$

where  $d_{jx}$  is the relative difficulty of score category  $x$  of item  $j$ .

The software IRTPRO (Cai, Thissen, & du Toit, 2011) was used for the IRT calibrations. IRTPRO is a multipurpose program that implements a variety of IRT models associated with mixed-item formats and

associated statistics. IRTPRO has been used to calibrate large data sets, such as those of PARCC assessments. The program implements MML estimation techniques for items and MLE estimation of theta.

This section describes the calibration sample in adherence to Standard 1.8 of the AERA, APA, & NCME (2014) *Standards for Educational and Psychological Testing*. Standard 1.8 states the following:

The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics. (25)

Sample data was used for grades 3 and 4 PBT calibration. All student data available at the time of calibration was used for grades 3 to 8 CBT calibration, resulting in a near-census data file. Tables 6.19 and 6.20 show the representativeness of the calibration samples compared to the census data. These tables demonstrate that the calibration sample was representative of the state. Grade 3 and 4 counts include both CBT and PBT students.

Table 6.19 Summary of Calibration and Census Data: English Language Arts

Calibration and Census Data: English Language Arts						
Grade		Calibration Sample		Census Data		Census % - Calib %
		N	%	N	%	
3	All Students	≥18,570	100.00%	≥53,210	100.00%	0.00%
	Gender					
	Male	≥9,500	51.14%	≥27,370	51.44%	0.30%
	Female	≥9,070	48.82%	≥25,820	48.52%	-0.30%
	Race Ethnicity					
	Hispanic/Latino	≥1,270	6.86%	≥4,640	8.73%	1.87%
	American Indian or Alaska Native	≥90	0.50%	≥320	0.61%	0.10%
	Asian	≥260	1.45%	≥840	1.59%	0.14%
	Black or African American	≥8,290	44.67%	≥22,840	42.92%	-1.75%
	Native Hawaiian or Other Pacific	≥10	0.09%	≥40	0.09%	-0.01%
	White	≥7,970	42.93%	≥22,780	42.82%	-0.11%
	Two or More Races	≥620	3.37%	≥1,670	3.14%	-0.23%
4	All Students	≥25,350	100.00%	≥54,950	100.00%	0.00%
	Gender					
	Male	≥12,900	50.89%	≥27,970	50.91%	0.02%
	Female	≥12,440	49.08%	≥26,960	49.07%	-0.01%
	Race Ethnicity					
	Hispanic/Latino	≥2,100	8.28%	≥4,550	8.29%	0.01%
	American Indian or Alaska Native	≥120	0.50%	≥320	0.60%	0.10%
	Asian	≥360	1.45%	≥770	1.40%	-0.05%
	Black or African American	≥10,840	42.76%	≥23,980	43.65%	0.89%
	Native Hawaiian or Other Pacific	≥20	0.09%	≥40	0.08%	-0.01%
	White	≥11,070	43.67%	≥23,460	42.70%	-0.97%
	Two or More Races	≥790	3.12%	≥1,760	3.21%	0.09%
5	All Students	≥55,110	100.00%	≥55,130	100.00%	0.00%
	Gender					
	Male	≥28,090	50.97%	≥28,100	50.98%	0.01%
	Female	≥27,020	49.03%	≥27,020	49.02%	-0.01%
	Race Ethnicity					
	Hispanic/Latino	≥4,500	8.18%	≥4,500	8.18%	-0.00%
	American Indian or Alaska Native	≥340	0.63%	≥340	0.63%	-0.00%
	Asian	≥850	1.55%	≥850	1.55%	-0.00%
	Black or African American	≥23,890	43.36%	≥23,910	43.38%	0.01%
	Native Hawaiian or Other Pacific	≥50	0.09%	≥50	0.09%	-0.00%
	White	≥23,720	43.04%	≥23,720	43.03%	-0.00%
	Two or More Races	≥1,720	3.13%	≥1,720	3.13%	-0.00%

Calibration and Census Data: English Language Arts						
Grade		Calibration Sample		Census Data		Census % - Calib %
		N	%	N	%	
6	All Students	≥54,990	100.00%	≥55,160	100.00%	0.00%
	Gender					
	Male	≥27,940	50.80%	≥28,030	50.83%	0.02%
	Female	≥27,050	49.20%	≥27,120	49.17%	-0.02%
	Race Ethnicity					
	Hispanic/Latino	≥4,180	7.62%	≥4,190	7.61%	-0.01%
	American Indian or Alaska Native	≥350	0.65%	≥350	0.65%	0.00%
	Asian	≥810	1.48%	≥810	1.48%	-0.00%
	Black or African American	≥23,840	43.35%	≥23,900	43.33%	-0.02%
	Native Hawaiian or Other Pacific	≥40	0.09%	≥40	0.09%	-0.00%
	White	≥24,140	43.90%	≥24,230	43.94%	0.03%
	Two or More Races	≥1,590	2.91%	≥1,600	2.90%	-0.00%
7	All Students	≥52,520	100.00%	≥52,640	100.00%	0.00%
	Gender					
	Male	≥26,890	51.22%	≥26,970	51.24%	0.02%
	Female	≥25,620	48.78%	≥25,670	48.76%	-0.02%
	Race Ethnicity					
	Hispanic/Latino	≥3,800	7.25%	≥3,810	7.25%	-0.01%
	American Indian or Alaska Native	≥340	0.66%	≥340	0.66%	0.00%
	Asian	≥810	1.55%	≥810	1.55%	-0.00%
	Black or African American	≥22,980	43.76%	≥23,060	43.81%	0.05%
	Native Hawaiian or Other Pacific	≥40	0.09%	≥40	0.09%	-0.00%
	White	≥23,130	44.06%	≥23,170	44.02%	-0.04%
	Two or More Races	≥1,370	2.62%	≥1,370	2.62%	-0.00%
8	All Students	≥51,080	100.00%	≥51,180	100.00%	0.00%
	Gender					
	Male	≥26,140	51.18%	≥26,200	51.21%	0.02%
	Female	≥24,930	48.82%	≥24,970	48.79%	-0.02%
	Race Ethnicity					
	Hispanic/Latino	≥3,670	7.19%	≥3,680	7.19%	0.00%
	American Indian or Alaska Native	≥350	0.69%	≥350	0.69%	-0.00%
	Asian	≥810	1.59%	≥810	1.58%	-0.00%
	Black or African American	≥22,110	43.30%	≥22,170	43.33%	0.04%
	Native Hawaiian or Other Pacific	≥20	0.05%	≥20	0.05%	-0.00%
	White	≥22,960	44.96%	≥22,990	44.93%	-0.03%
	Two or More Races	≥1,120	2.21%	≥1,120	2.21%	-0.00%



Table 6.20 Summary of Calibration and Census Data: Mathematics

Calibration and Census Data: Mathematics						
Grade		Calibration Sample		Census Data		Census % - Calib %
		N	%	N	%	
3	All Students	≥12,110	100.00%	≥53,180	100.00%	0.00%
	Gender					
	Male	≥6,180	51.09%	≥27,360	51.44%	0.36%
	Female	≥5,920	48.88%	≥25,810	48.53%	-0.35%
	Race Ethnicity					
	Hispanic/Latino	≥880	7.30%	≥4,640	8.74%	1.44%
	American Indian or Alaska Native	≥70	0.62%	≥320	0.60%	-0.02%
	Asian	≥180	1.49%	≥840	1.59%	0.10%
	Black or African American	≥5,510	45.55%	≥22,820	42.91%	-2.63%
	Native Hawaiian or Other Pacific	≥10	0.09%	≥40	0.08%	-0.01%
	White	≥5,010	41.42%	≥22,770	42.82%	1.40%
	Two or More Races	≥410	3.42%	≥1,670	3.14%	-0.28%
4	All Students	≥9,460	100.00%	≥54,980	100.00%	0.00%
	Gender					
	Male	≥4,730	50.04%	≥27,980	50.90%	0.86%
	Female	≥4,720	49.94%	≥26,970	49.07%	-0.87%
	Race Ethnicity					
	Hispanic/Latino	≥600	6.40%	≥4,570	8.32%	1.92%
	American Indian or Alaska Native	≥50	0.56%	≥320	0.60%	0.04%
	Asian	≥120	1.33%	≥770	1.40%	0.07%
	Black or African American	≥3,870	40.96%	≥23,980	43.62%	2.65%
	Native Hawaiian or Other Pacific	≥10	0.19%	≥40	0.08%	-0.11%
	White	≥4,470	47.32%	≥23,470	42.69%	-4.63%
	Two or More Races	≥300	3.19%	≥1,760	3.21%	0.02%
5	All Students	≥54,950	100.00%	≥55,070	100.00%	0.00%
	Gender					
	Male	≥28,010	50.98%	≥28,070	50.98%	-0.00%
	Female	≥26,940	49.02%	≥27,000	49.02%	0.00%
	Race Ethnicity					
	Hispanic/Latino	≥4,380	7.97%	≥4,500	8.18%	0.21%
	American Indian or Alaska Native	≥340	0.63%	≥340	0.63%	-0.00%
	Asian	≥850	1.56%	≥850	1.55%	-0.00%
	Black or African American	≥23,880	43.46%	≥23,870	43.35%	-0.11%
	Native Hawaiian or Other Pacific	≥50	0.09%	≥50	0.09%	-0.00%
	White	≥23,700	43.13%	≥23,700	43.04%	-0.09%
	Two or More Races	≥1,720	3.14%	≥1,720	3.14%	-0.01%

Calibration and Census Data: Mathematics						
Grade		Calibration Sample		Census Data		Census % - Calib %
		N	%	N	%	
6	All Students	≥54,930	100.00%	≥55,130	100.00%	0.00%
	Gender					
	Male	≥27,910	50.80%	≥28,020	50.83%	0.03%
	Female	≥27,020	49.20%	≥27,110	49.17%	-0.03%
	Race Ethnicity					
	Hispanic/Latino	≥4,040	7.36%	≥4,190	7.62%	0.25%
	American Indian or Alaska Native	≥350	0.65%	≥350	0.65%	0.00%
	Asian	≥810	1.49%	≥810	1.48%	-0.01%
	Black or African American	≥23,850	43.42%	≥23,880	43.31%	-0.11%
	Native Hawaiian or Other Pacific	≥40	0.09%	≥40	0.09%	0.00%
	White	≥24,210	44.07%	≥24,230	43.95%	-0.13%
Two or More Races	≥1,600	2.91%	≥1,600	2.90%	-0.01%	
7	All Students	≥52,270	100.00%	≥52,580	100.00%	0.00%
	Gender					
	Male	≥26,770	51.22%	≥26,940	51.24%	0.02%
	Female	≥25,490	48.78%	≥25,630	48.76%	-0.02%
	Race Ethnicity					
	Hispanic/Latino	≥3,600	6.89%	≥3,810	7.25%	0.35%
	American Indian or Alaska Native	≥340	0.66%	≥340	0.66%	-0.00%
	Asian	≥810	1.56%	≥810	1.55%	-0.01%
	Black or African American	≥22,960	43.93%	≥23,040	43.82%	-0.11%
	Native Hawaiian or Other Pacific	≥40	0.09%	≥40	0.09%	-0.00%
	White	≥23,110	44.22%	≥23,140	44.01%	-0.22%
Two or More Races	≥1,370	2.63%	≥1,370	2.62%	-0.01%	
8	All Students	≥44,850	100.00%	≥45,110	100.00%	0.00%
	Gender					
	Male	≥23,250	51.85%	≥23,410	51.89%	0.04%
	Female	≥21,590	48.15%	≥21,700	48.11%	-0.04%
	Race Ethnicity					
	Hispanic/Latino	≥3,150	7.03%	≥3,330	7.38%	0.35%
	American Indian or Alaska Native	≥330	0.74%	≥330	0.73%	-0.00%
	Asian	≥570	1.29%	≥570	1.28%	-0.01%
	Black or African American	≥20,750	46.27%	≥20,810	46.14%	-0.13%
	Native Hawaiian or Other Pacific	≥20	0.05%	≥20	0.05%	-0.00%
	White	≥19,020	42.42%	≥19,040	42.22%	-0.20%
Two or More Races	≥980	2.20%	≥980	2.19%	-0.01%	

All 2019 LEAP 2025 item calibration and linking were performed based on IRT. The calibration and linking methodology used for the Spring 2019 LEAP 2025 administration closely followed most of the PARCC methods referenced in the PARCC document *Final Technical Report for 2015 Administration*. To maintain comparability to PARCC, the 2PL/GPC IRT model was applied to item calibration using the software IRTPRO (Cai et al., 2011). To avoid local independence between traits, the writing traits written expression (WE) and written knowledge and use of language (WKL) were separately calibrated using the sparse matrix method.

The Stocking & Lord (1983) procedure was applied using the transformation and scaling software STUIRT (Kim & Kolen, 2004), which can be downloaded at <https://www.education.uiowa.edu/centers/casma/computer-programs#c0748e48-f88c-6551-b2b8-ff00000648cd>. PARCC scale score transformation constants for the PARCC 2016 baseline scale were used to generate final scoring tables. All IRTPRO and STUIRT command files were prepared following PARCC examples.

Descriptions of the PARCC calibration and equating approach can be found in the PARCC documents *Final Technical Report for 2015 Administration* and *Final Technical Report for 2016 Administration*.

There were two test forms, CBT and PBT, for the 2019 LEAP 2025 grades 3 and 4 ELA and mathematics assessments. Only CBT forms were administered for the grades 5 through 8 ELA and mathematics assessments. In general, a school administered the same test mode for ELA and mathematics. Table 6.21 summarizes the student count and item count by test mode for each grade and content area.

The following two steps were taken to place the 2019 LEAP 2025 tests on the 2018 LEAP 2025 scale, which are on the 2016 PARCC baseline scale:

1. Calibrate the 2019 LEAP 2025 tests.
2. Link 2019 LEAP 2025 tests, to the LEAP 2025 scale under the non-equivalent common item design.

PARCC established a new baseline scale using 2016 PARCC spring tests. The 2016 and 2017 LEAP 2025 tests were directly linked to this new PARCC 2016 baseline scale using PARCC item parameters as anchor item parameters. Therefore, LEAP 2016 and 2017 were placed on the PARCC scale. Since the 2016 and 2017 LEAP 2025 tests were calibrated with Louisiana students, the scale for these tests will be referred to as the LEAP 2025 scale, although its scale was placed on PARCC scales built with PARCC associated states' data. The 2019 LEAP 2025 forms were linked to the LEAP 2025 scale using LEAP items, which were administered in LEAP 2025 forms in 2016-2019 as anchors by the Stocking & Lord procedure. Since the 2019 anchor items are on the PARCC scale, 2019 LEAP 2025 forms continue to be considered on the PARCC scale.

#### 6.4.1 Calibration of the 2019 LEAP 2025 Tests

For 2019 LEAP 2025 item calibration, the 2PL/GPC IRT model was applied to the Louisiana students' calibration samples using the software IRTPRO (Cai et al., 2011). Table 6.21 shows the number of students in the calibration samples and number of calibration items by mode. About 90% of grade 3 students took the PBT, and about 10% of grade 3 students took the CBT. About 80% of grade 4 students took PBT, and about 20% of grade 4 students took CBT. More students in grade 8 took the ELA assessment than the mathematics assessment because high-performing students were allowed to take the LEAP 2025 HS Algebra I test instead of the mathematics grade 8 test. For ELA, reading items (RL/RI) in writing prompts are not counted in the N-Items columns because calibration does not include reading item scores; it only includes WE item scores. A reading item score and a WE item score for the same writing prompt are the same. There were between 26 and 32 ELA items and between 41 and 43 mathematics items across grades.

**Table 6.21 Summary of Student Count in Calibration Sample and Item Count by Test Mode**

Content	Grade	N			Percentage		N-Items	
		All	CBT	PBT	CBT	PBT	CBT	PBT
ELA	3	≥18,570	≥1,620	≥16,950	8.76	91.24	26	26
	4	≥25,350	≥7,620	≥17,730	30.05	69.95	28	28
	5	≥55,110	≥55,110	*	100.00	*	28	*
	6	≥54,990	≥54,990	*	100.00	*	32	*
	7	≥52,520	≥52,520	*	100.00	*	32	*
	8	≥51,080	≥51,080	*	100.00	*	21	*
Mathematics	3	≥12,110	≥1,600	≥10,500	13.28	86.72	43	43
	4	≥9,460	≥2,110	≥7,340	22.38	77.62	43	43
	5	≥55,080	≥55,080	*	100.00	*	41	*
	6	≥55,080	≥55,080	*	100.00	*	42	*
	7	≥52,470	≥52,470	*	100.00	*	43	*
	8	≥45,020	≥45,020	*	100.00	*	41	*

\* Grades 5–8 did not have a PBT form.

#### 6.4.1.1. Concurrent Calibration for PBT and CBT

For 2019 LEAP 2025 calibration, CBT and PBT were combined and calibrated together for grades 3 and 4 based on mode effect study (section 10.4).

#### 6.4.1.2. Separate Calibration for ELA Prose Constructed-Response Tasks

To address the issue of local independence for ELA prose-constructed response (PCR) tasks, the sparse matrix method was applied for grades 5 to 8. Each ELA test consisted of two PCR tasks; each task had a written expression (WE) and a written knowledge and use of the language (WKL) trait. As can be seen in Table 6.22, a single calibration was performed for grades 5 to 8 by randomly splitting the students into two groups. Almost half of the data set included responses to other items and responses to two WE traits, and the other calibration data set included the same responses to other items and responses to two WKL traits. Therefore, WE item parameters were estimated using the responses from the first group and WKL item parameters were estimated using the responses from the second group. Because these two sets of item responses were calibrated together, there is only one unique set of item parameters for each item. PARCC took this sparse matrix approach for all grades.

**Table 6.22 Calibration Data Structure for ELA WE and WKL Traits with Sparse Matrix**

Group	Other Items	WE	WKL
I	XXXXXXXX	XX	
II	XXXXXXXX		XX

For ELA grades 3 and 4, the same sample repeated design was applied to the WE and WKL calibration due to the small number of students in maximum score points of WE items. For each grade, two datasets of responses were generated. One calibration dataset included two WE responses and responses of the other items and excluded the WKL responses. The other calibration dataset included two WKL responses and the

same responses of the other items and excluded the WE responses. In these datasets, the responses of the other items, except for WE and WKL, were the same for both forms, and each dataset included either WE or WKL responses. After each dataset was separately calibrated, the item parameters with WKL responses were equated to those with WE responses using all common items as anchor items. Table 6.23 illustrates this design.

**Table 6.23 Calibration Data Structure for ELA WE and WKL Traits with the Same Sample Repeated Design**

Dataset	Group	Other Items	WE	WKL
1	All Students	XXXXXXXX	XX	
2	All Students	XXXXXXXX		XX

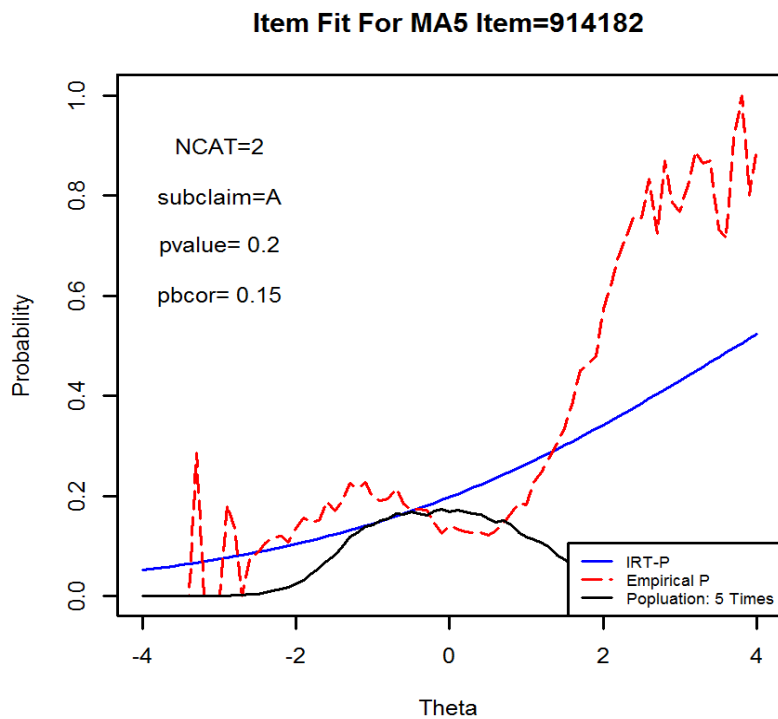
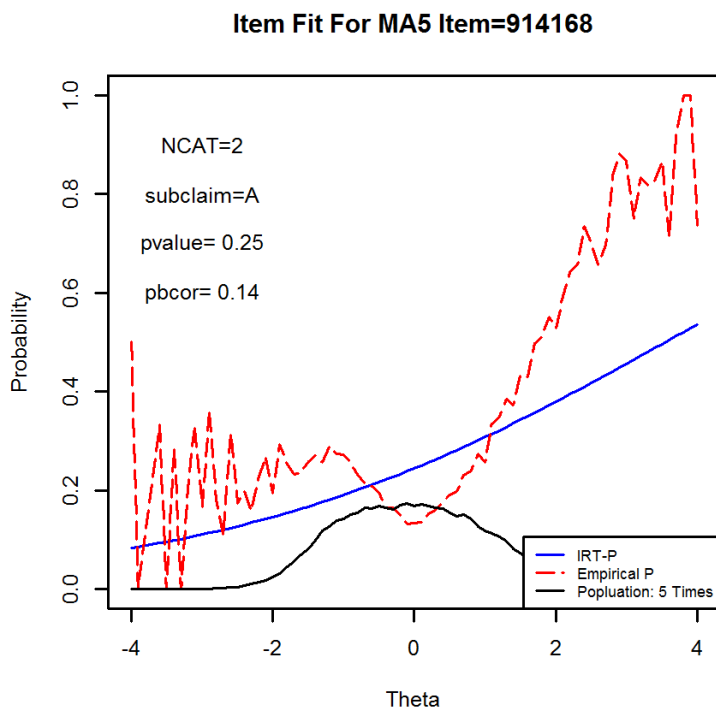
#### 6.4.1.3. IRT Item Fit

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

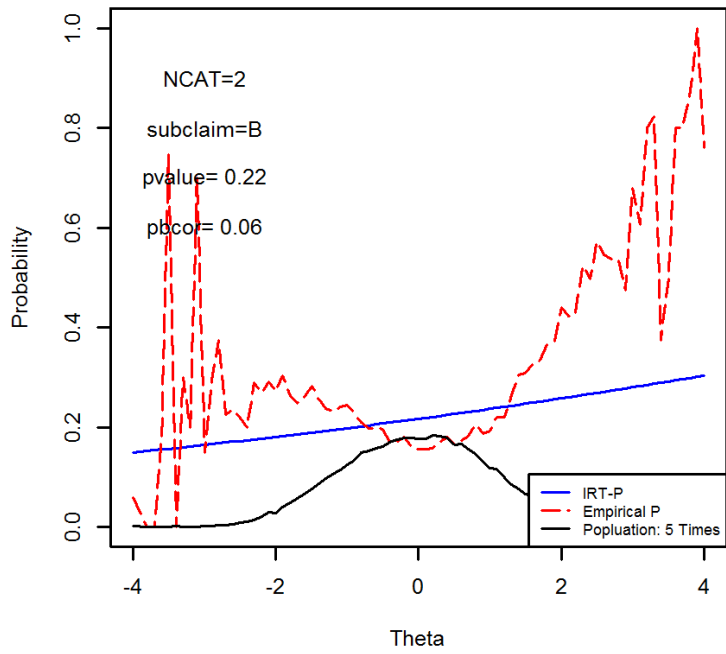
It is important to note that while items may be flagged for misfit, these flags may not be of practical importance. Misfitting items that have content validity are often retained for use in one assessment and monitored over a period of usage. A large number of misfitting items in an assessment would indicate that caution should be exercised in the interpretation of the overall score.

After convergence was achieved for each IRT data set, an item characteristic curve (ICC) for each item was plotted with empirical students’ performances from theta ability -4 to 4. Four items were suppressed from calibration and scoring due to poor fit: two items in grade 5 mathematics, one item in grade 6 mathematics, and one item in grade 8 mathematics. Six additional items were removed from the anchor sets due to poor fit. The fit plots for the items removed from calibration are seen in Figure 6.1. Figure 6.2 displays the fit plots for the items removed from the anchor set.

Figure 6.1 Item Fit Plots of Items Removed from Calibration and Scoring



Item Fit For MA6 Item=981978



Item Fit For MA8 Item=983048

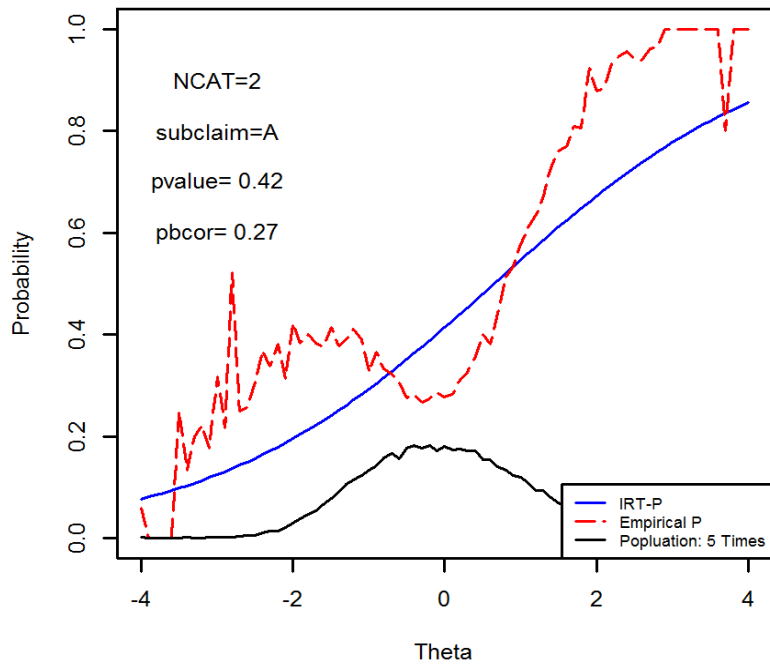
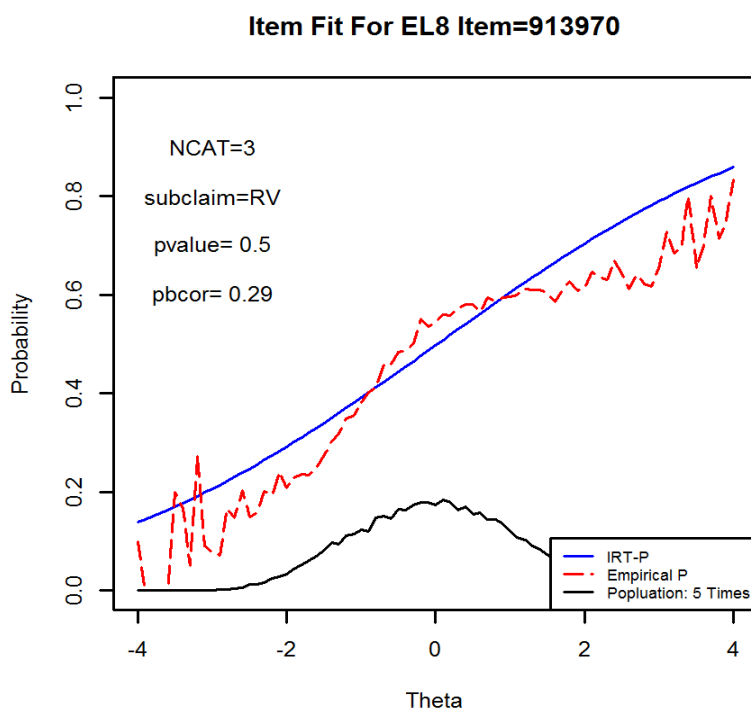
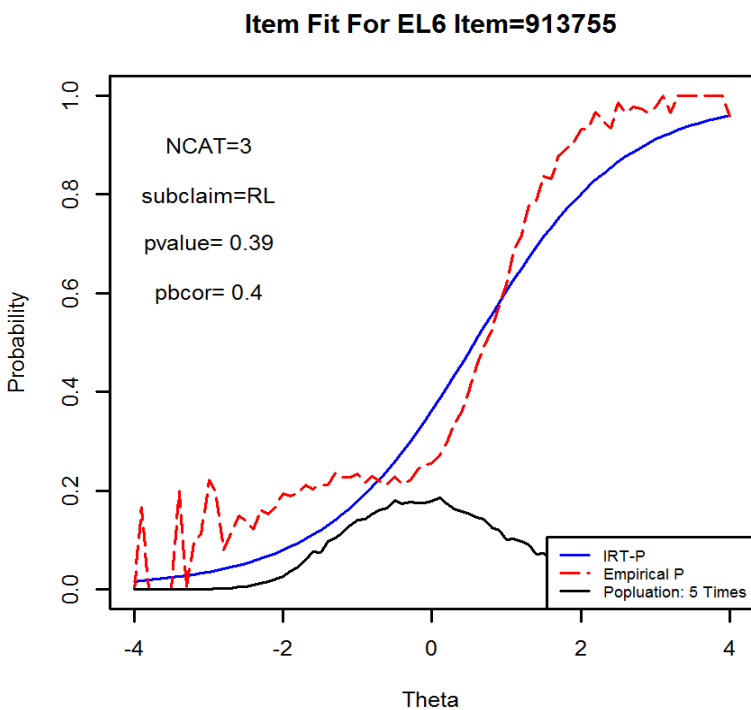
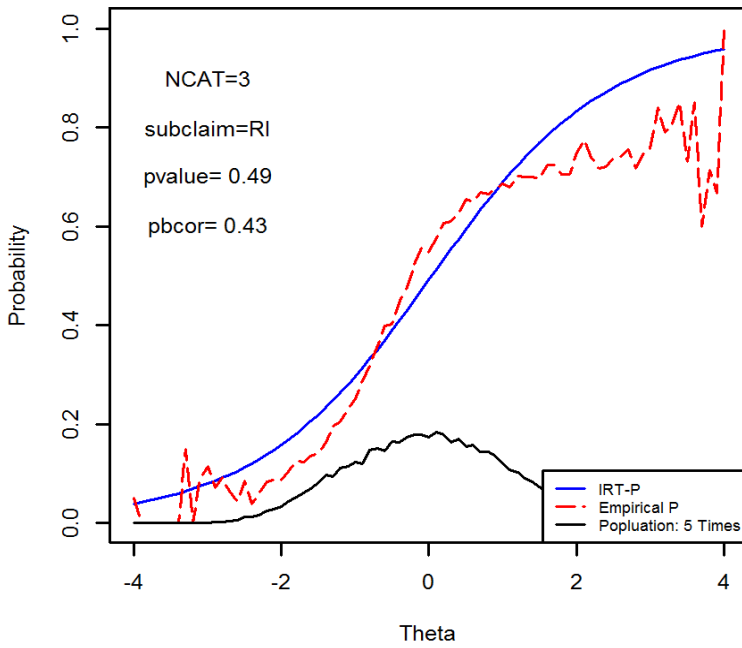


Figure 6.2 Item Fit Plots of Items Removed from Anchor Sets

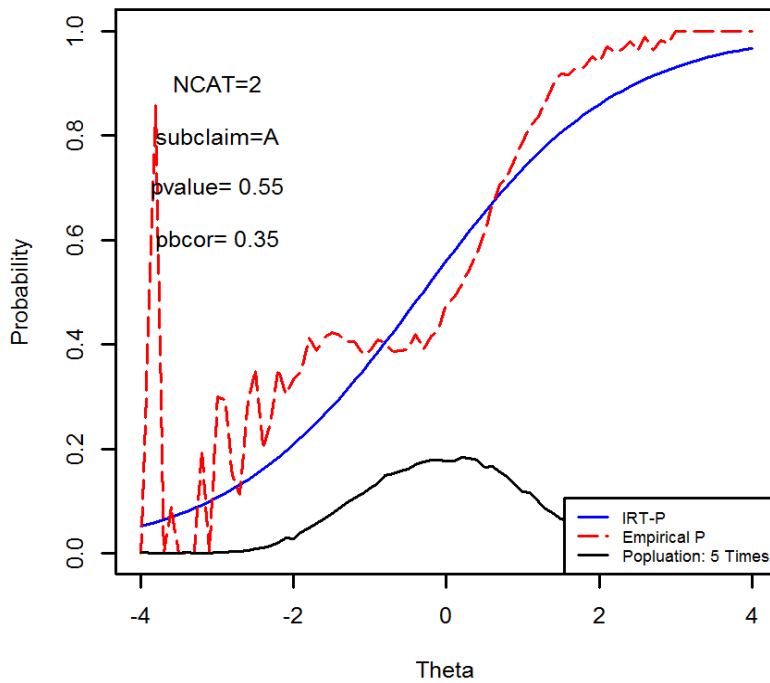




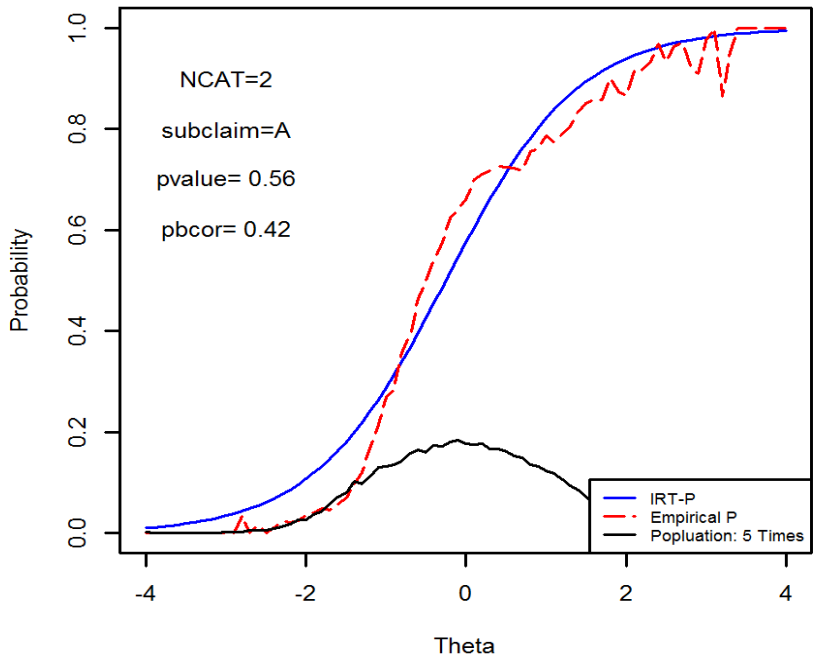
**Item Fit For EL8 Item=913975**



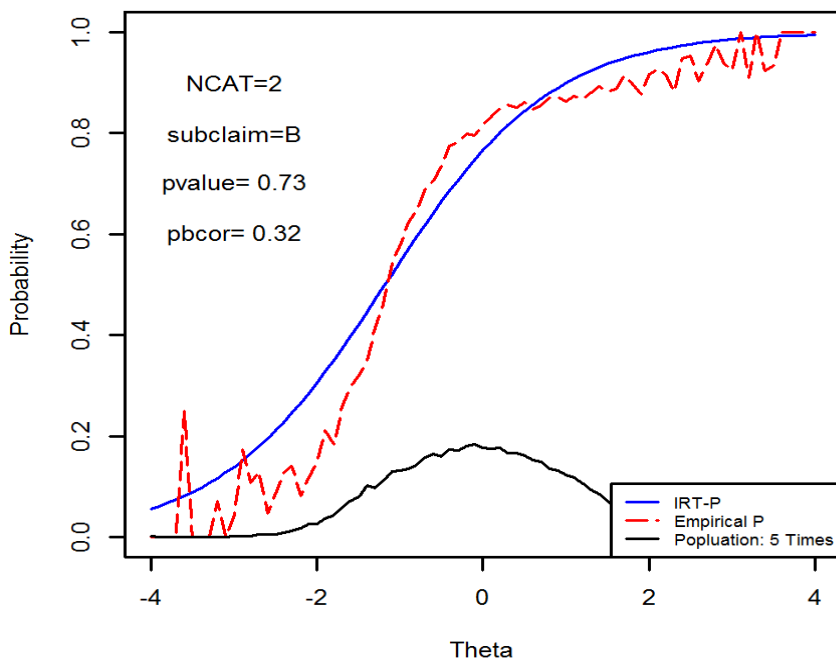
**Item Fit For MA6 Item=914249**



Item Fit For MA7 Item=914324



Item Fit For MA7 Item=914330



After calibration, the IRT model fit was evaluated by reviewing item chi-squared statistic that were calculated using IRTPRO item parameters and item responses from students in the calibration sample. Adjusted fit values were calculated and flagged if they exceeded 0.35 (Pearson, 2018).

Since chi-square values are sensitive to sample size, these statistics are not easily compared when the number of students varies across items. As a result, adjusted fit values were calculated by dividing the chi-square fit statistic by the sample size using the following formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Tables 6.24 and 6.25 show the adjusted item fit C values using the chi-square statistics and calibration sample sizes for ELA and mathematics, respectively. The average adjusted fit ranged from 0.10 to 0.13 for ELA and 0.06 to 0.08 for mathematics. No items were excluded based on model fit statistics because the adjusted item fits for all items were lower than the criterion value of 0.35, as can be seen in the maximum values for both ELA and mathematics. The largest adjusted fit value was 0.32 for ELA grade 6.

**Table 6.24 Summary of Adjusted Fit for ELA**

Grade	Mode	No. Items	Mean	Std. Dev.	Min.	Max.	No. Flagged Items
3	CBT/PBT	30	0.12	0.05	0.04	0.22	0
4	CBT/PBT	32	0.13	0.06	0.05	0.27	0
5	CBT	28	0.12	0.07	0.02	0.27	0
6	CBT	32	0.1	0.06	0.04	0.32	0
7	CBT	32	0.11	0.06	0.02	0.23	0
8	CBT	32	0.11	0.06	0.02	0.23	0

**Table 6.25 Summary of Adjusted Fit for Mathematics**

Grade	Mode	No. Items	Mean	Std. Dev.	Min.	Max.	No. Flagged Items
3	CBT/PBT	43	0.06	0.03	0.02	0.15	0
4	CBT/PBT	43	0.06	0.04	0.02	0.16	0
5	CBT	41	0.07	0.03	0.01	0.15	0
6	CBT	42	0.07	0.04	0.01	0.19	0
7	CBT	43	0.08	0.04	0.01	0.14	0
8	CBT	41	0.06	0.04	0.02	0.16	0

### 6.4.2 Linking 2019 LEAP 2025 Grades 3–8 to PARCC Scale

The 2016 and 2017 LEAP 2025 forms were linked to the PARCC scale using intact PARCC items embedded into the LEAP 2025 forms by using the Stocking & Lord procedure (1983). Therefore, these item parameters were placed on the PARCC scale. However, these equated Louisiana item parameters are based on only Louisiana students' responses while intact PARCC item parameters were estimated based on PARCC associated states' responses. To distinguish these two sets of item parameters, item parameters based on only Louisiana student responses will be called LEAP 2025 item parameters and its scale is referred to as the LEAP 2025 scale.

Three anchor sets were created for the 2019 Spring LEAP 2025 ELA and mathematics assessments equating process. Anchor 1 items were intact PARCC items embedded in the 2019 LEAP 2025 form. Anchor 2 items were items common to the 2019 LEAP 2025 spring forms and previous years' forms, and their item parameters were from previously operational LEAP 2025 item parameters. Anchor 3 item parameters consisted of all Anchor 2 item parameters and Anchor 1 item parameters were used for the remaining items. Anchor 2 was used in the operational analyses to link to the LEAP 2025 scale, which is the same as the PARCC scale, and Anchor 1 and Anchor 3 were used to help evaluate drift from the PARCC scale. Table 6.26 provides the Stocking & Lord transformation constants that were used to link to scale. Table 6.27 summarizes the number and score points of the initial anchor item selection before equating. Table 6.27 also summarizes the number and score points of the final anchor item selections. The difference between the initial number of anchor items and the final number of anchor items is the number of anchor items that were dropped.

**Table 6.26 Stocking & Lord Transformation Constants**

Content	Grade	Slope	Intercept
ELA	3	1.037978	0.299272
	4	0.981004	0.129841
	5	0.92898	0.05861
	6	0.97163	0.04688
	7	0.97168	0.07153
	8	0.97092	0.07147
Mathematics	3	0.962352	-0.12307
	4	1.010616	-0.02758
	5	0.96965	-0.17845
	6	0.99545	-0.1714
	7	0.97284	-0.10087
	8	0.97994	0.02407

**Table 6.27 Number and Score Points of Initial and Final Anchor Item Sets**

Content	Grade	Anchor Set	Anchor 1		Anchor 2		Anchor 3		
			Number of Items	Score Points	Number of Items	Score Points	Number of Items	Score Points	
ELA	3	Initial	18	52	10	28	26	68	
		Final	15	46	8	16	24	64	
	4	Initial	20	62	14	39	28	78	
		Final	18	58	12	25	17	45	
	5	Initial	23	68	13	37	28	78	
		Final	19	49	8	16	19	39	
	6	Initial	25	72	11	33	31	84	
		Final	19	49	10	30	23	57	
	7	Initial	21	64	12	35	31	84	
		Final	16	54	11	33	24	70	
	8	Initial	29	80	12	35	30	82	
		Final	20	50	10	31	19	49	
	Mathematics	3	Initial	25	40	15	20	43	62
			Final	20	35	11	14	36	48
4		Initial	28	42	15	21	41	56	
		Final	22	36	14	20	32	47	
5		Initial	22	31	12	17	39	58	
		Final	15	23	12	17	33	47	
6		Initial	24	42	14	20	38	61	
		Final	17	27	13	19	27	40	
7		Initial	27	44	13	23	37	60	
		Final	19	34	9	13	25	43	
8		Initial	28	45	15	22	37	59	
		Final	20	34	11	16	29	47	

*\*Following OP2 approach for counting Writing dimensions: Count WE and WKL only*

Figures 6.3 to 6.14 show test characteristic curves (TCCs) for anchor items, corresponding 2019 LEAP 2025 estimated anchor items (EQ\_ANC), 2018 LEAP 2025 operational items (LEAP 2018), and all 2019 LEAP 2025 estimated items (EQ\_ALL) for ELA and mathematics after applying the Stocking & Lord equating procedure. The **blue** solid line illustrates the anchor items, the **red** dotted line is the 2019 LEAP 2025 equated anchor items, the **black** solid line is the 2018 LEAP 2025 operational items, and the **green** dotted line is for all 2019 LEAP 2025 equated items. Anchor items for each anchor set 1, 2, and 3 are different as mentioned above. For most ELA and mathematics grades, the TCCs for anchor items and the corresponding 2018 estimated anchor items were overlapped across most ability levels.

For ELA, the TCC of the 2018 LEAP 2025 and 2019 LEAP 2025 estimated items (EQ\_ALL) overlapped or were close to each other for all grades, except for grade 8. The TCCs show that the anchor item sets for the 2019 LEAP 2025 were easier than the overall tests especially at grades 3 through 5. The ELA Grade 8 TCC shows that the 2019 LEAP 2025 was more difficult than 2018 LEAP 2025. For mathematics, the TCC of the 2018 LEAP 2025 and 2019 LEAP 2025 estimated items (EQ\_ALL) overlapped across most ability levels across all grades. Anchor sets represented the overall test form in most grades. There were some differences at the extreme ranges, such as low ability or high ability.

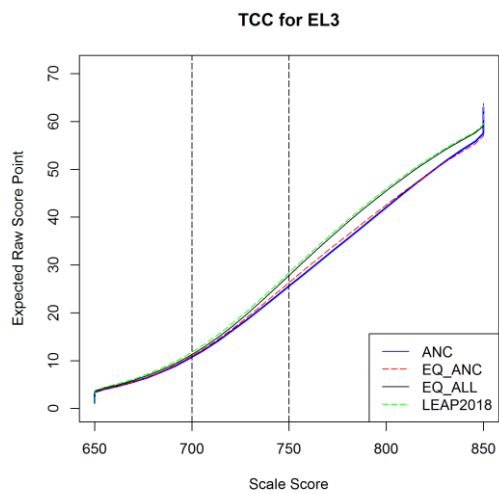
Figures 6.15 to 6.26 present scatter plots of slope item parameters and difficulty item parameters for ELA and mathematics and their correlation after linking 2019 LEAP 2025 to the PARCC 2016 scale.

As can be seen in the ELA slope parameter plots, most parameters were around the identity line. The correlation between anchor item parameters and estimated parameters ranged from 0.96 to 1.00 with Anchor 2. For mathematics, most item slope parameters were around the identity line, and the correlations ranged from 0.96 to 1.00 with Anchor 2.

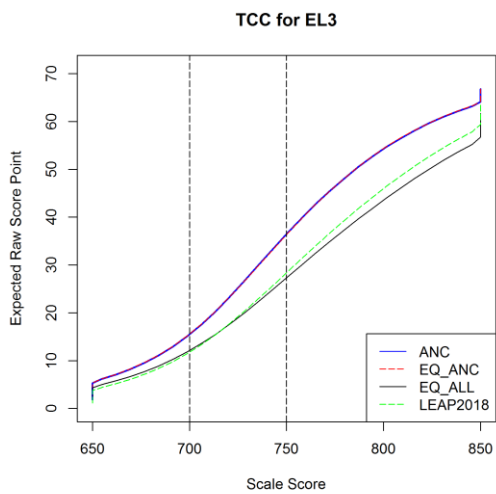
For ELA, most item difficulty parameters were around the identity line, and the correlations ranged from 0.99 to 1.00 with Anchor 2. For mathematics as well, most item difficulty parameters were around the identity line, except for grade 5. Correlations ranged from 0.95 to 1.00 across grades with Anchor 2. It is common to find higher correlations for difficulty parameters than those for slope parameters.

**Figure 6.3 ELA Grade 3 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

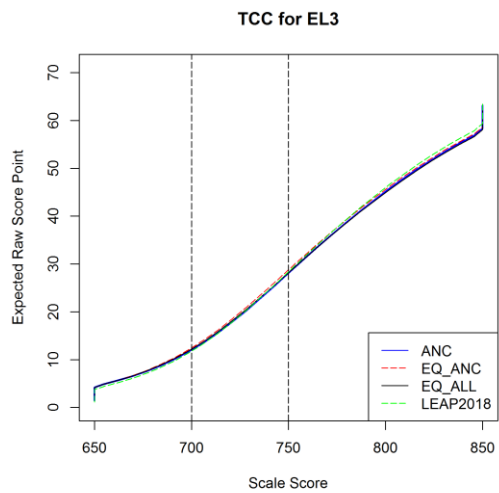
Anchor 1



Anchor 2

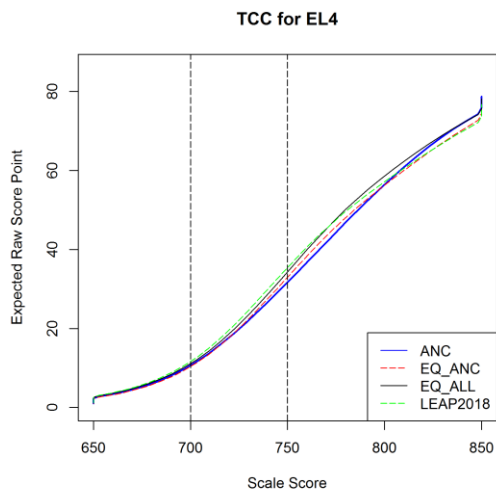


Anchor 3

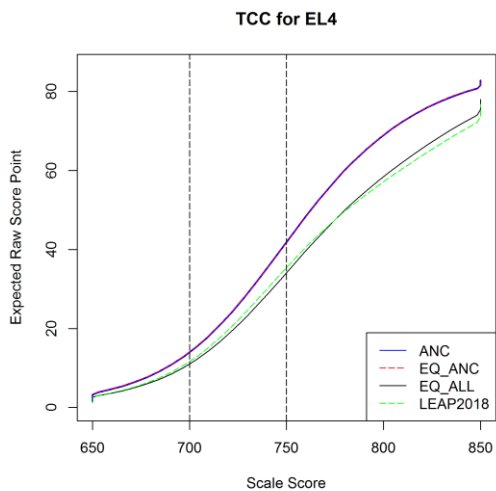


**Figure 6.4 ELA Grade 4 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

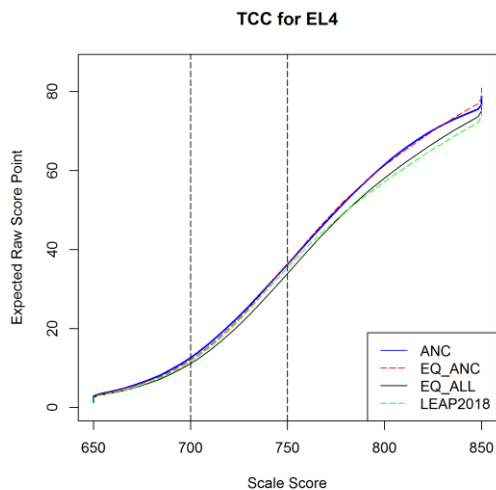
Anchor 1



Anchor 2



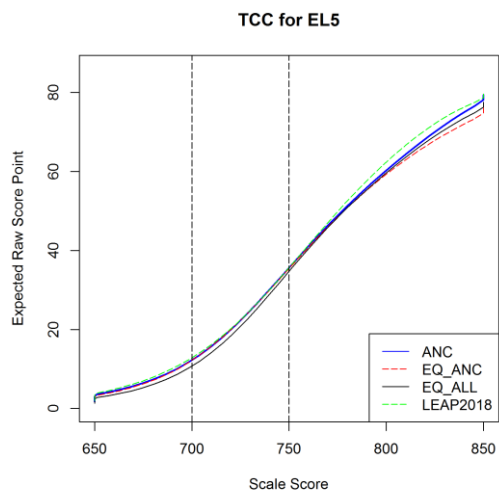
Anchor 3



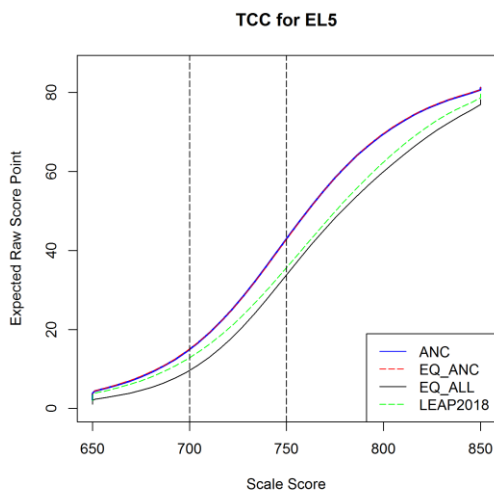


**Figure 6.5 ELA Grade 5 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

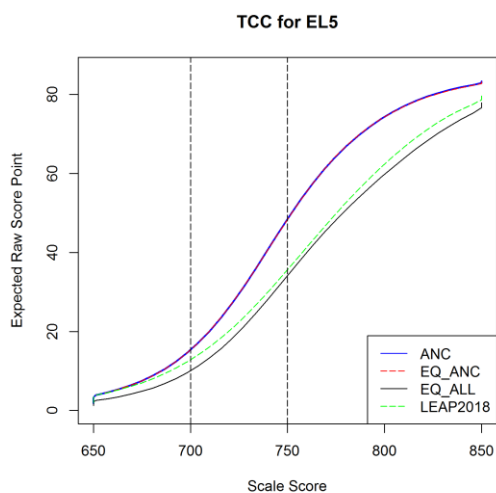
Anchor 1



Anchor 2

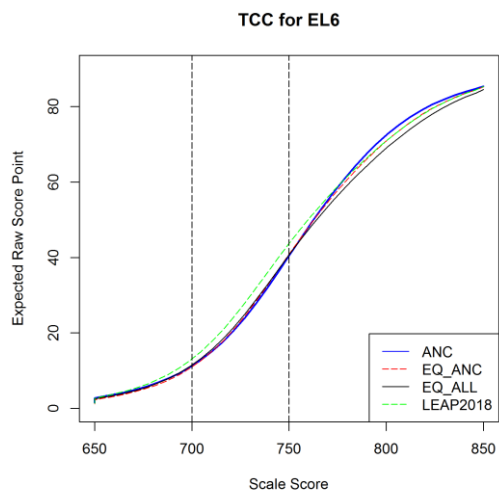


Anchor 3

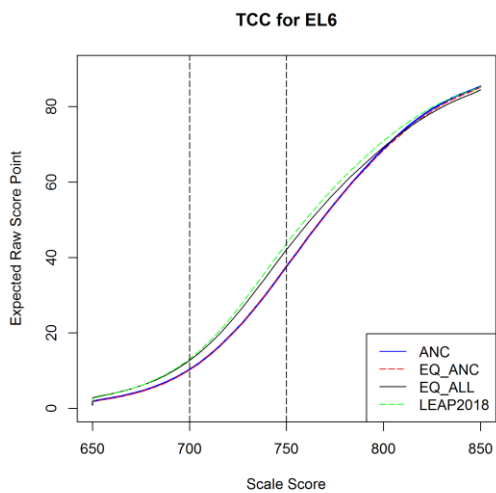


**Figure 6.6 ELA Grade 6 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

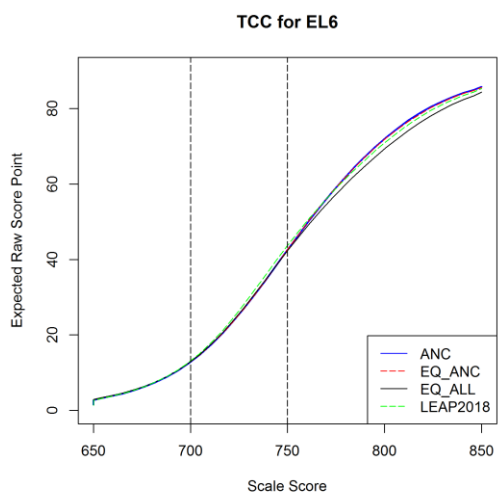
Anchor 1



Anchor 2

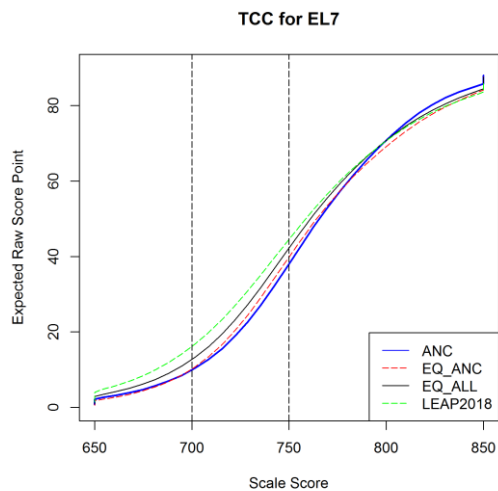


Anchor 3

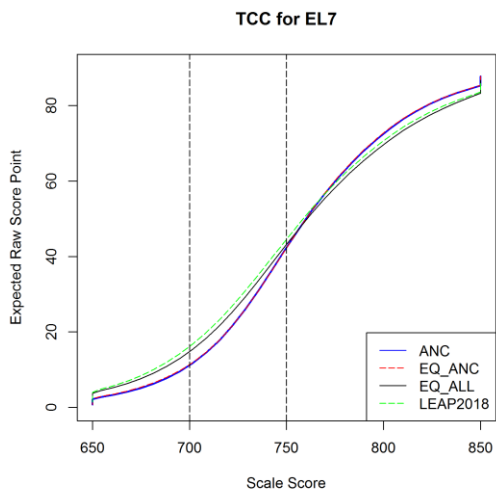


**Figure 6.7 ELA Grade 7 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

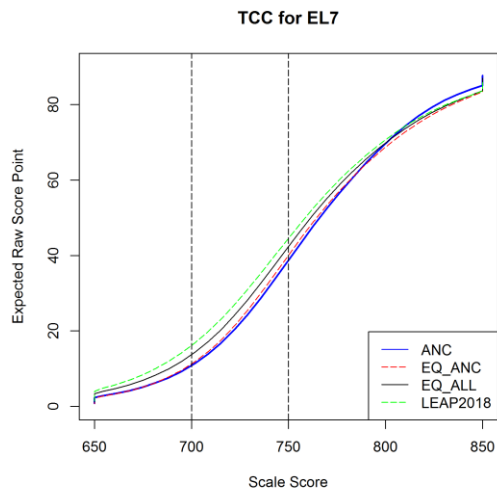
Anchor 1



Anchor 2

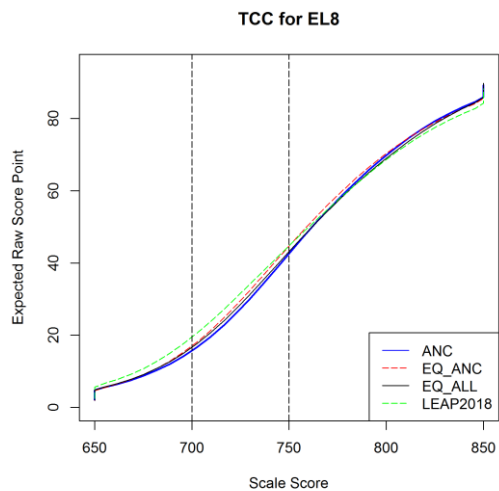


Anchor 3

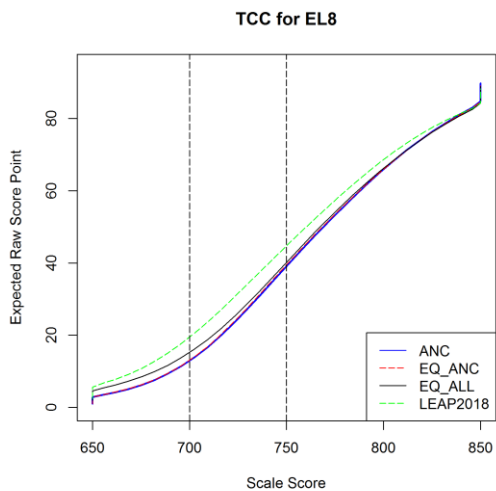


**Figure 6.8 ELA Grade 8 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

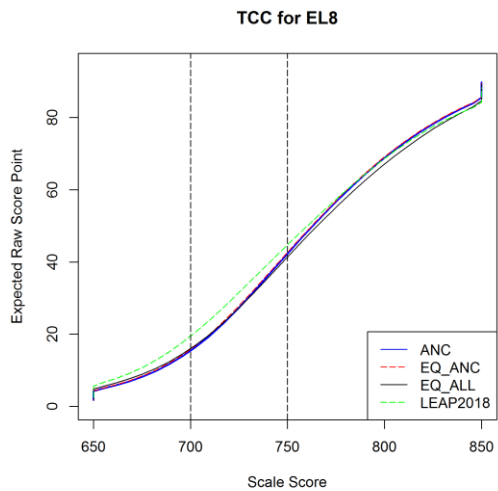
Anchor 1



Anchor 2

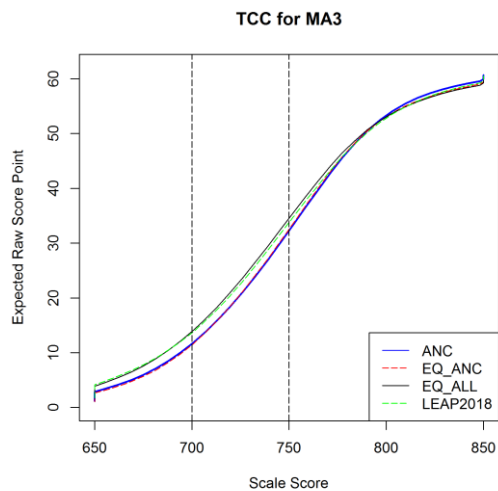


Anchor 3

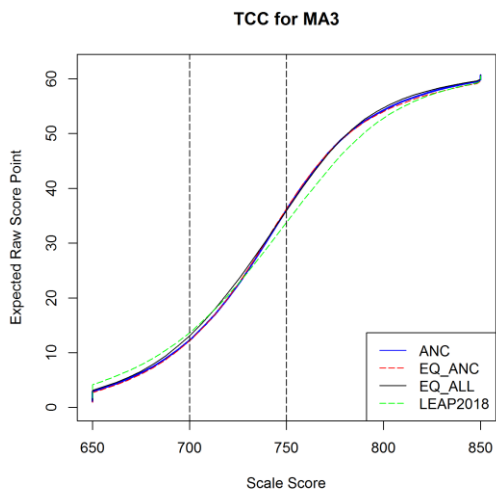


**Figure 6.9 Mathematics Grade 3 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

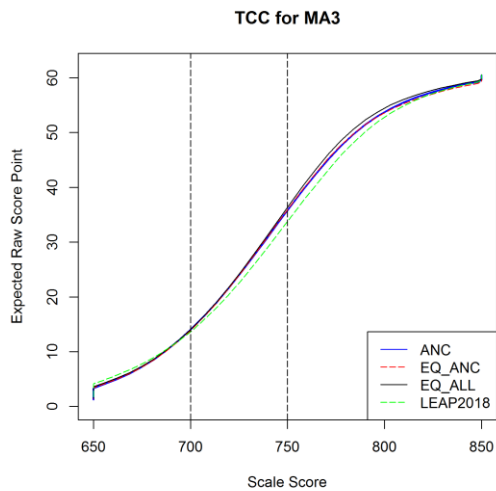
Anchor 1



Anchor 2

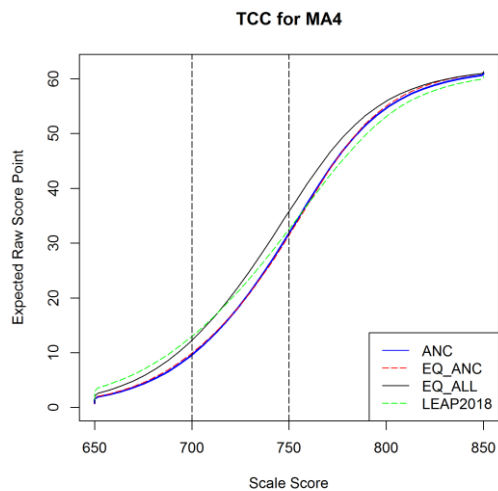


Anchor 3

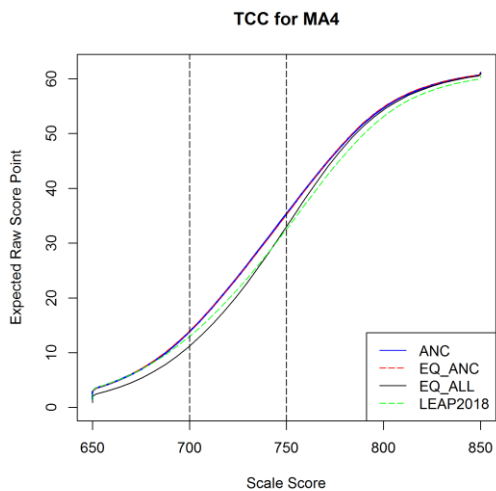


**Figure 6.10 Mathematics Grade 4 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

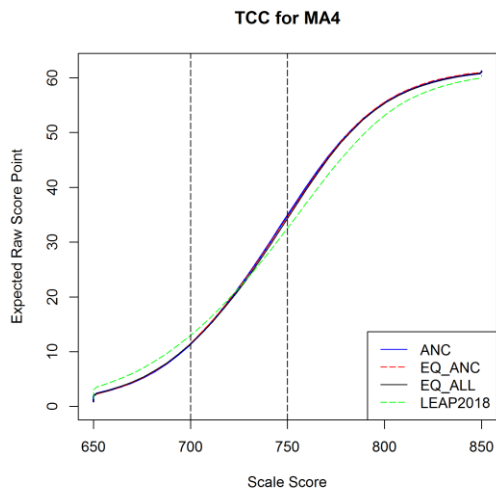
Anchor 1



Anchor 2

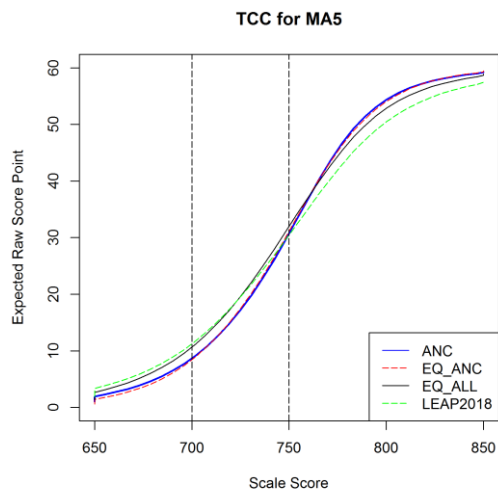


Anchor 3

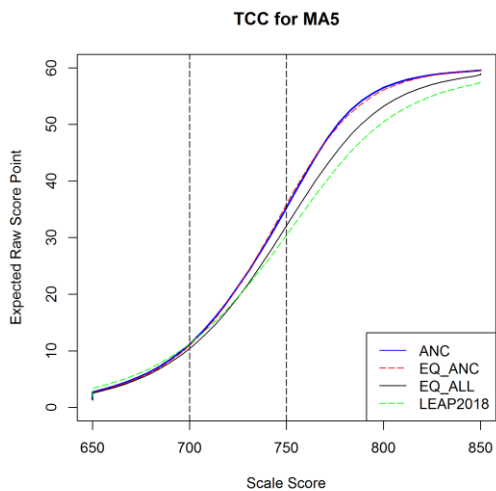


**Figure 6.11 Mathematics Grade 5 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

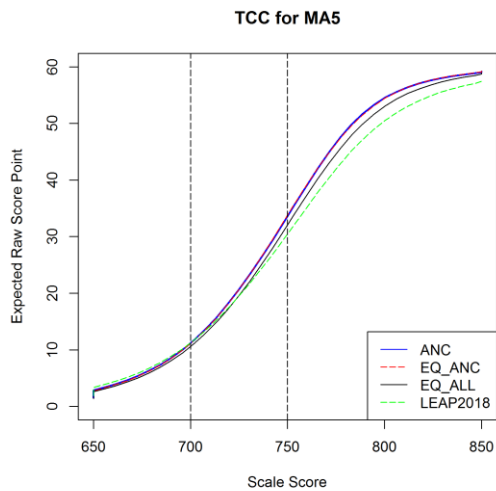
Anchor 1



Anchor 2

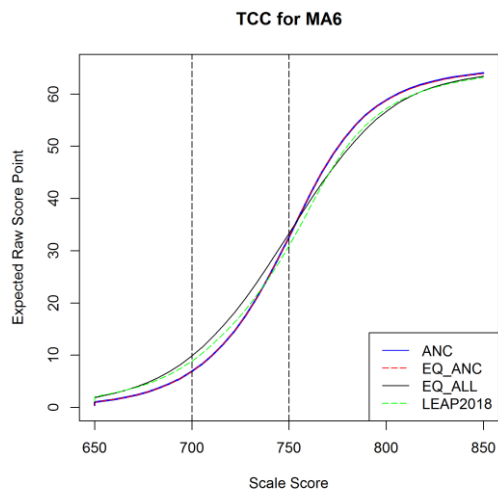


Anchor 3

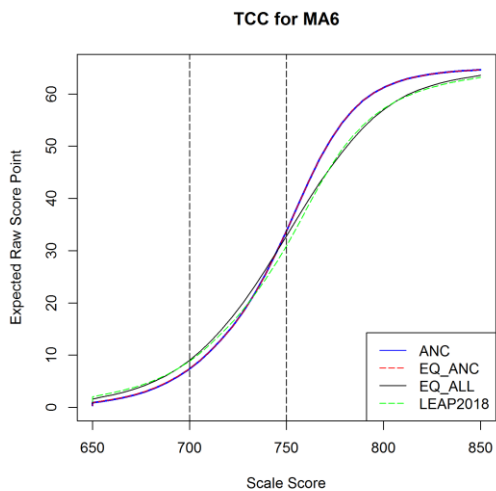


**Figure 6.12 Mathematics Grade 6 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

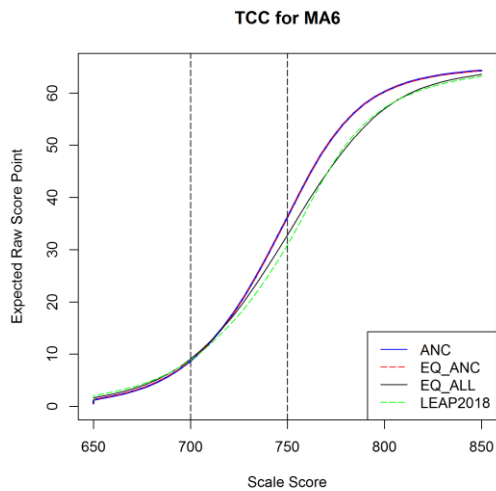
Anchor 1



Anchor 2



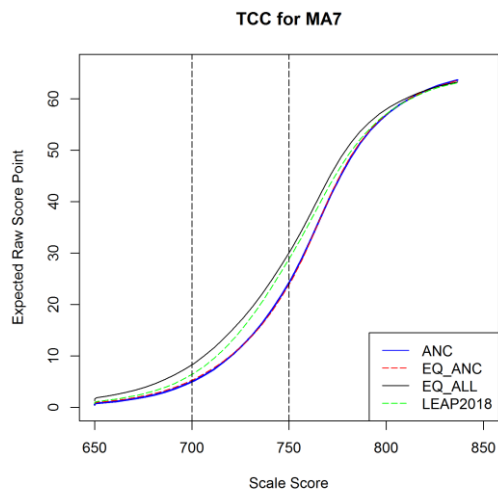
Anchor 3



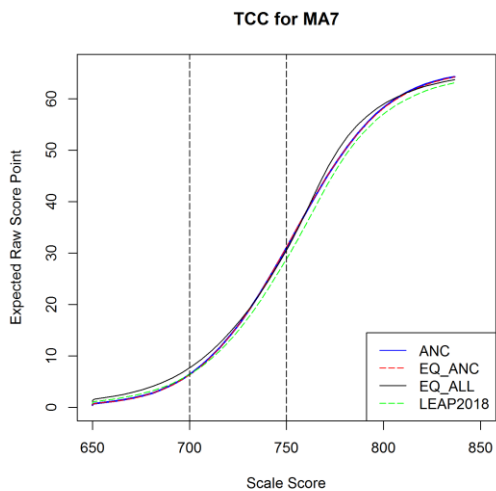


**Figure 6.13 Mathematics Grade 7 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

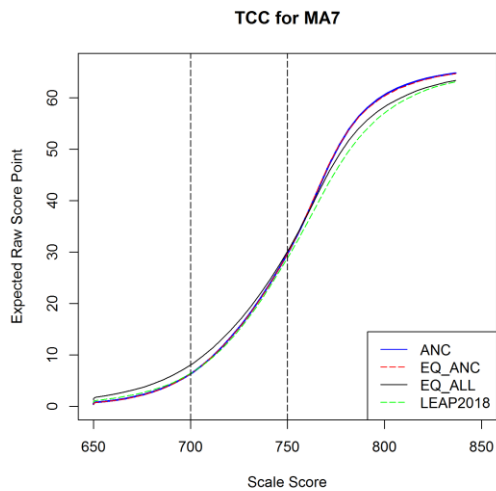
Anchor 1



Anchor 2

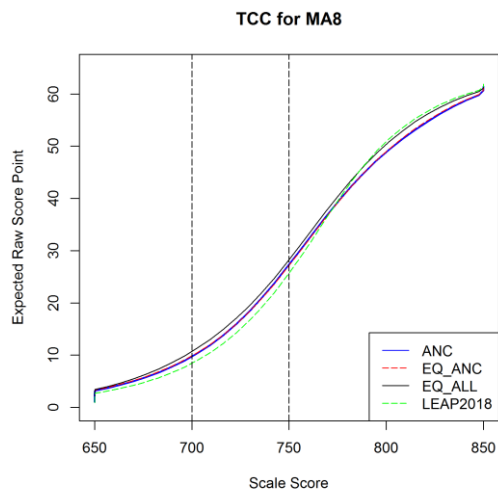


Anchor 3

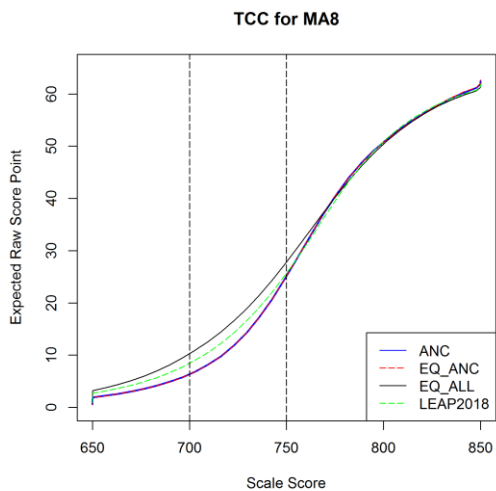


**Figure 6.14 Mathematics Grade 8 TCC between Pre-equated Anchor, Equated Anchor, LEAP 2018, and All 2019 LEAP 2025 Items**

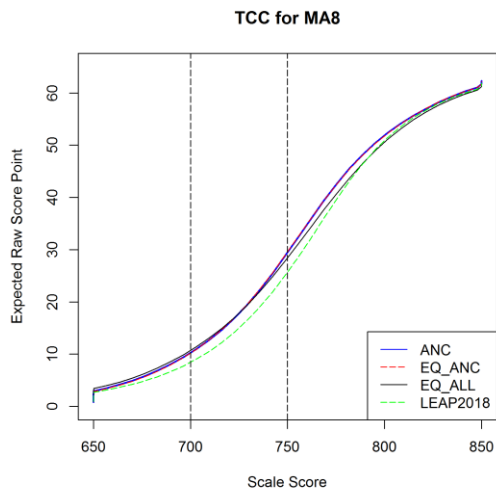
Anchor 1



Anchor 2

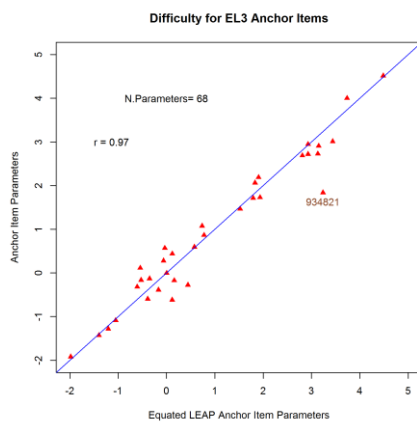
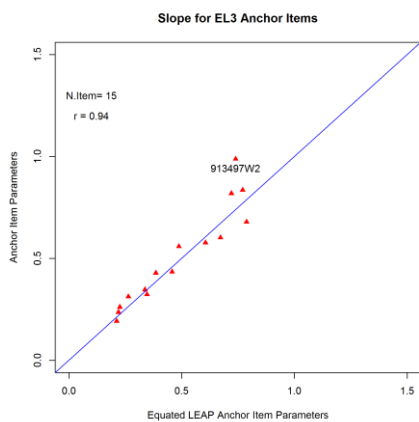


Anchor 3

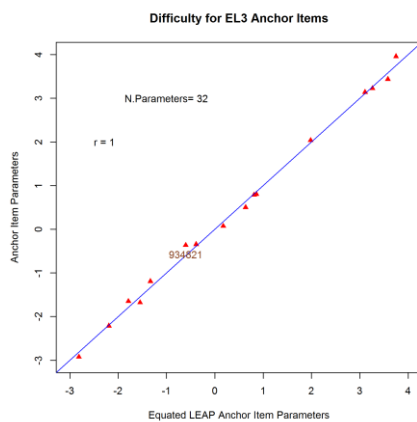
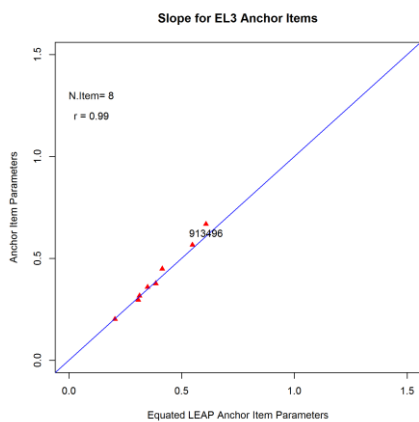


**Figure 6.15 ELA Grade 3 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

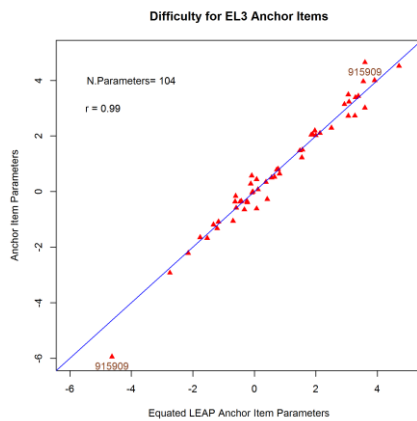
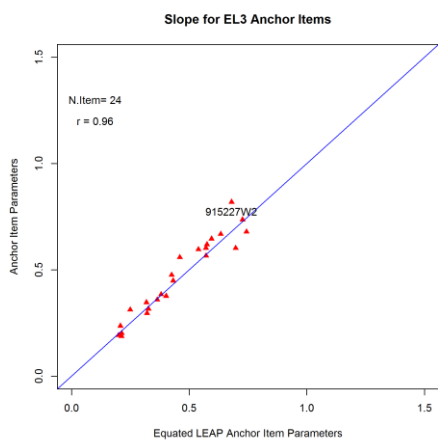
**Anchor 1**



**Anchor 2**

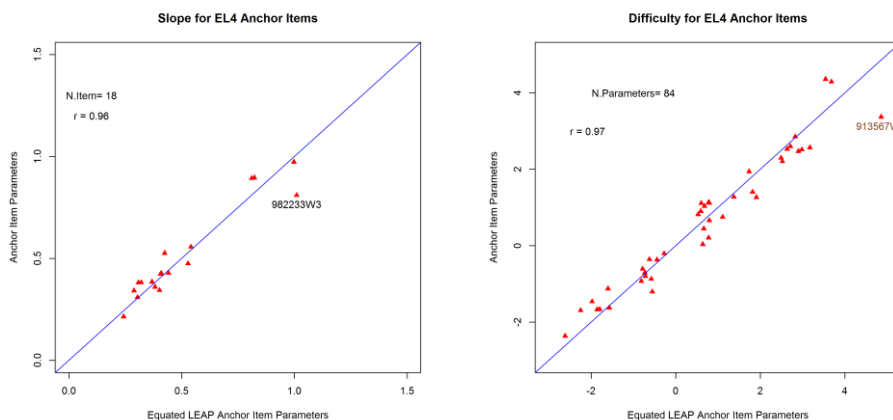


**Anchor 3**

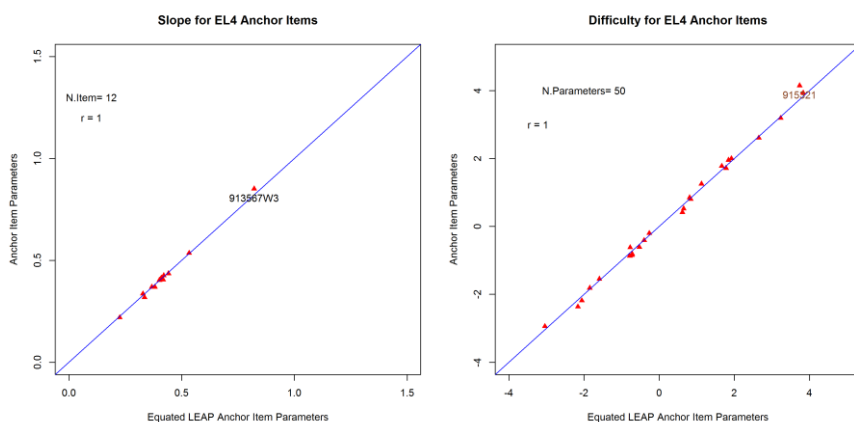


**Figure 6.16 ELA Grade 4 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

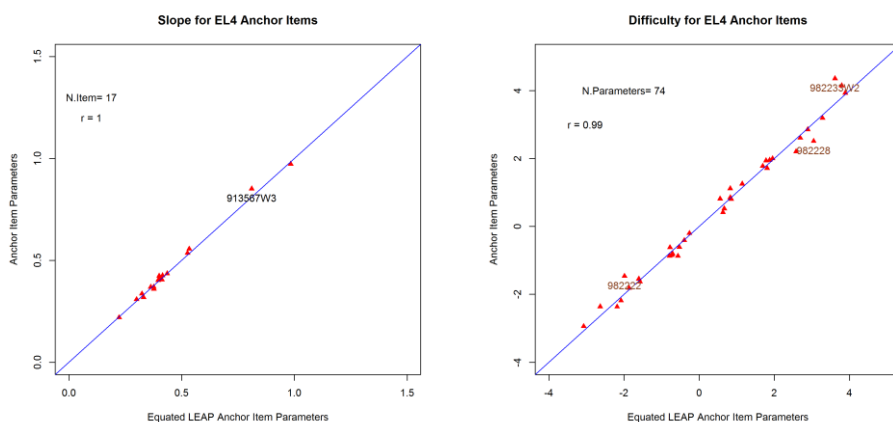
**Anchor 1**



**Anchor 2**

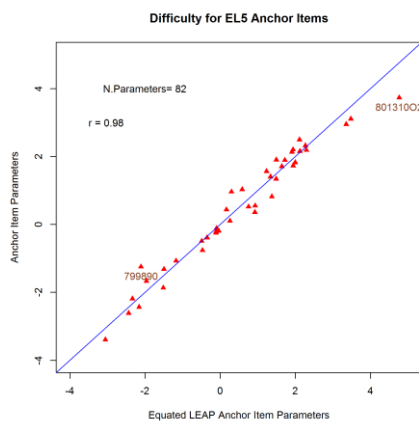
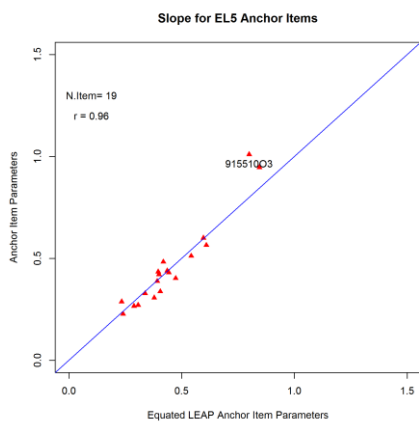


**Anchor 3**

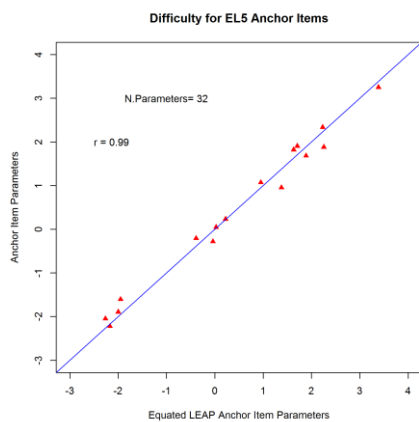
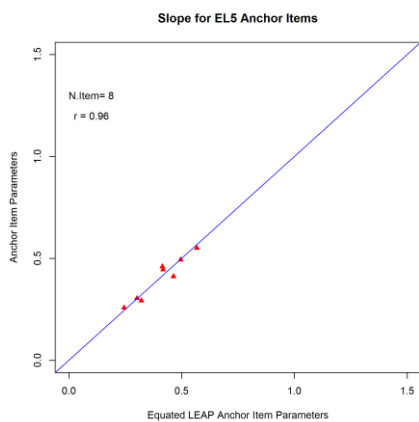


**Figure 6.17 ELA Grade 5 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

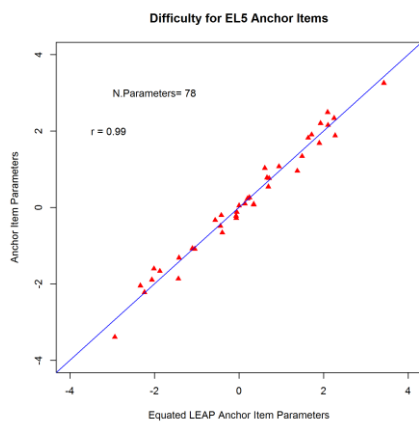
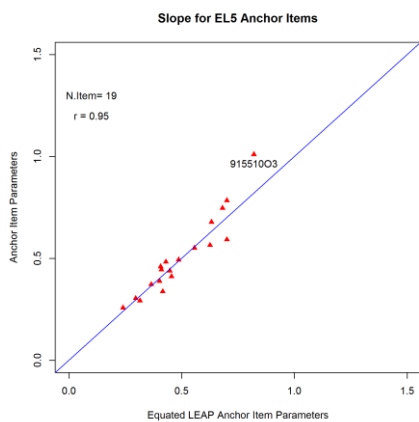
**Anchor 1**



**Anchor 2**

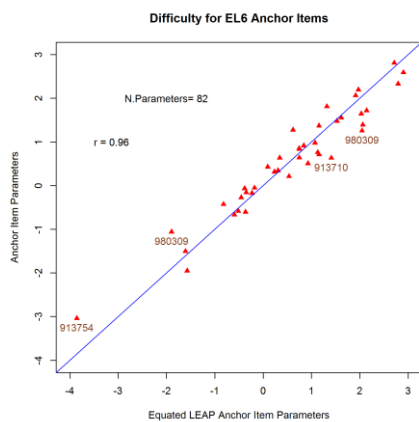
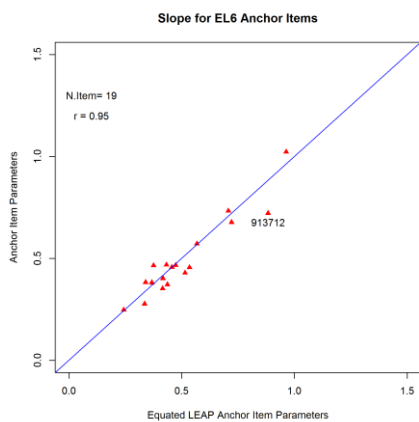


**Anchor 3**

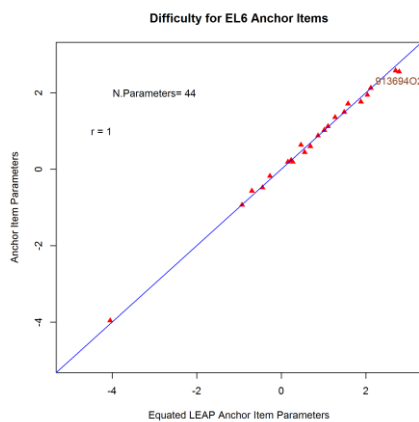
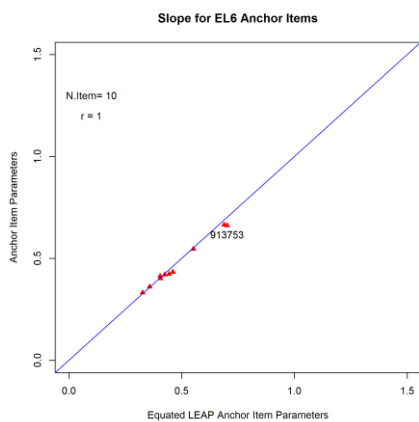


**Figure 6.18 ELA Grade 6 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

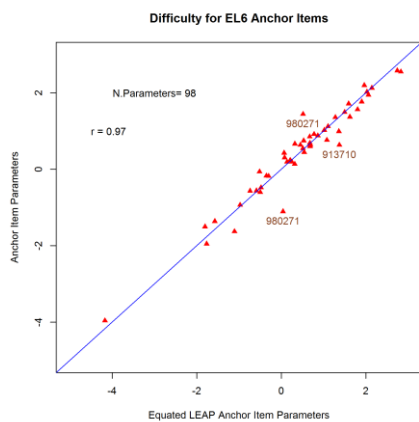
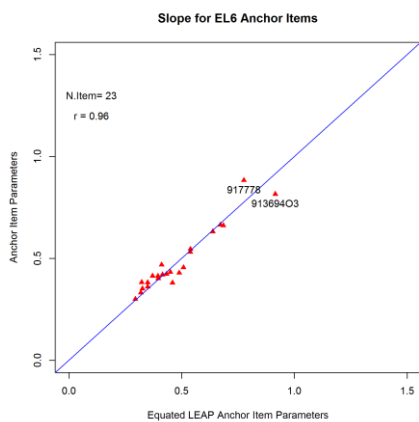
**Anchor 1**



**Anchor 2**

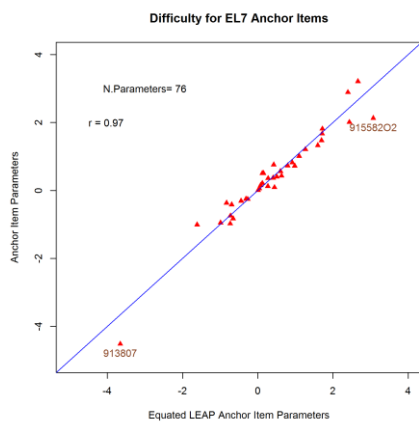
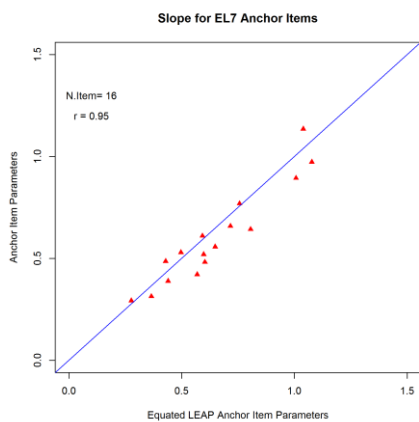


**Anchor 3**

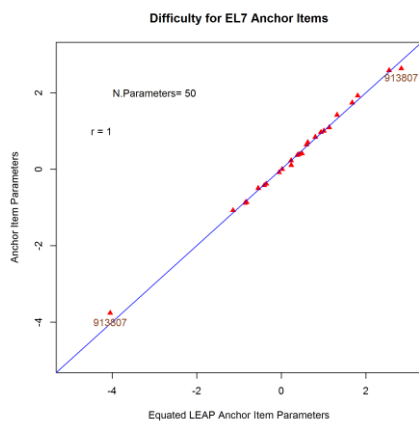
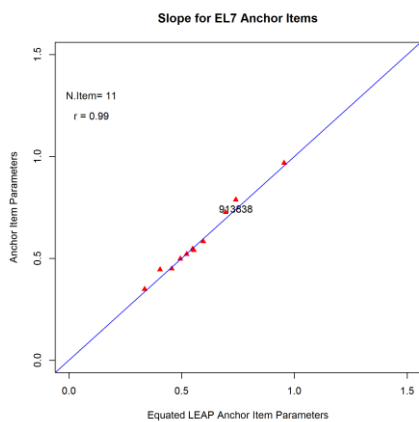


**Figure 6.19 ELA Grade 7 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

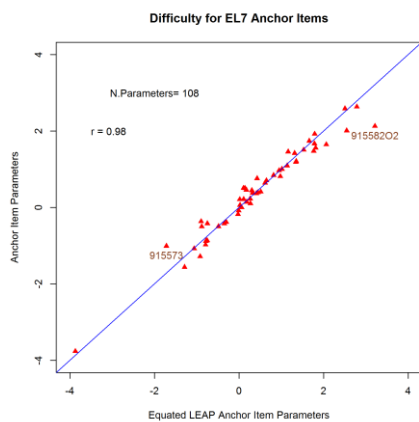
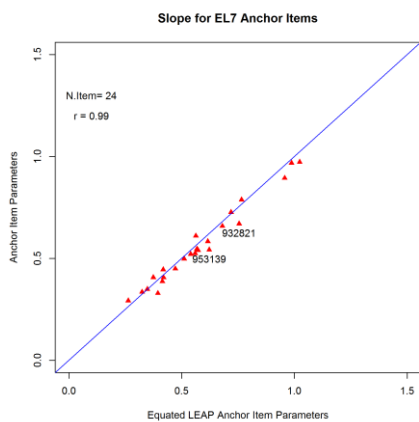
**Anchor 1**



**Anchor 2**

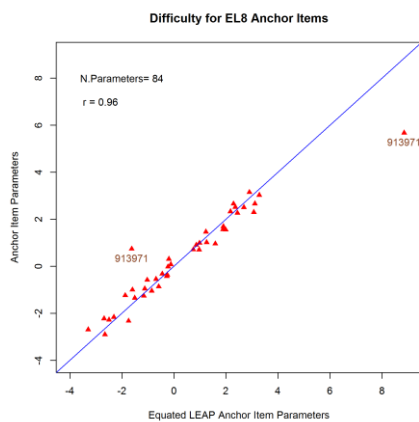
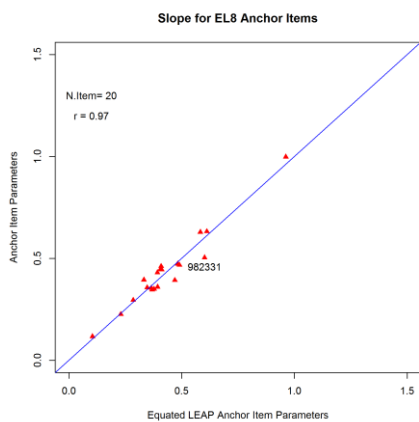


**Anchor 3**

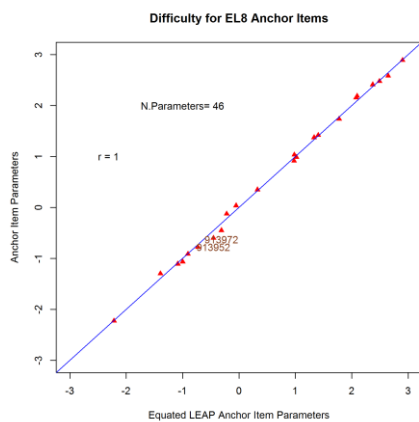
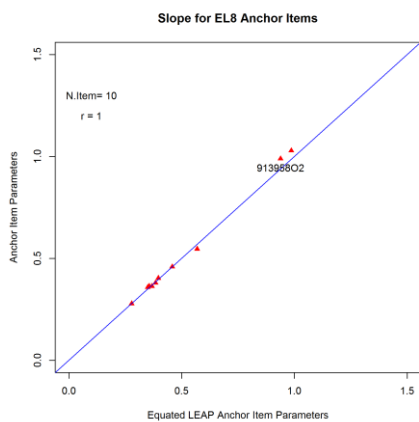


**Figure 6.20 ELA Grade 8 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

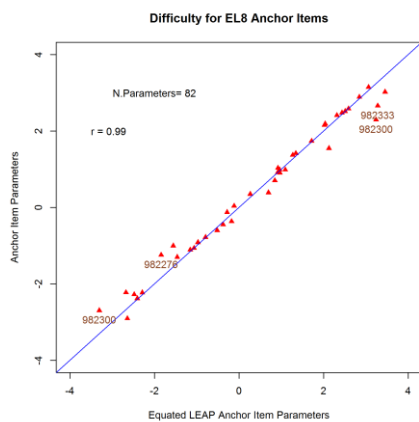
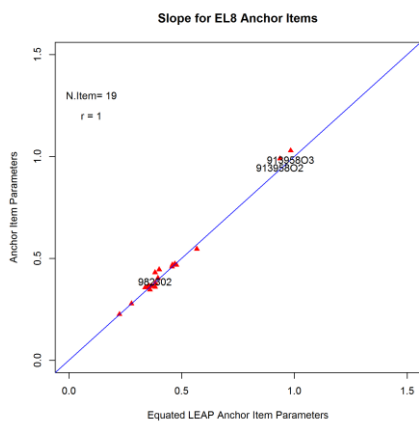
**Anchor 1**



**Anchor 2**



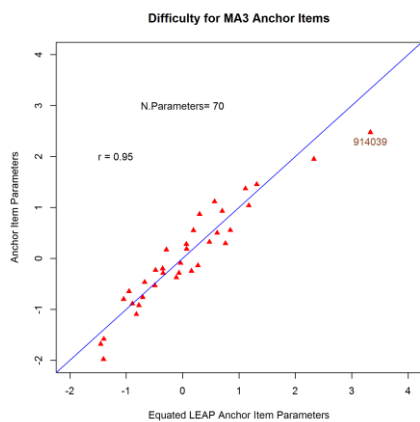
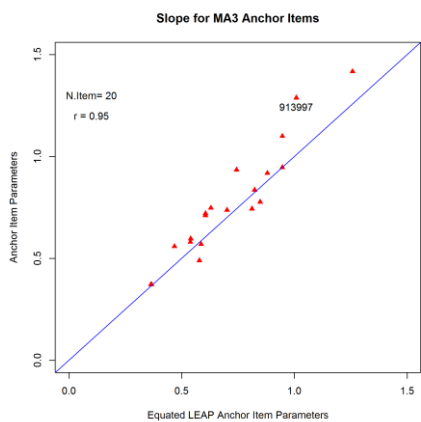
**Anchor 3**



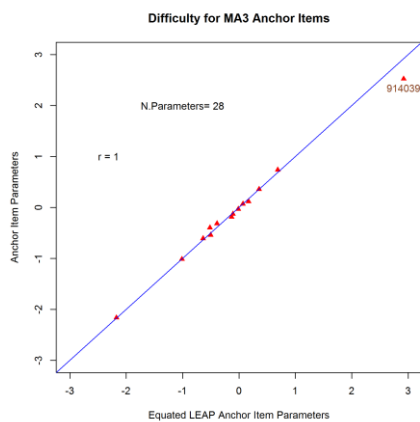
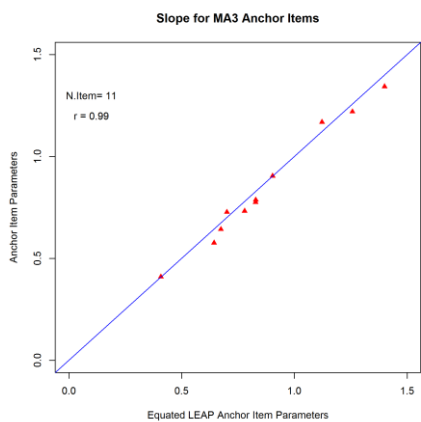


**Figure 6.21 Mathematics Grade 3 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

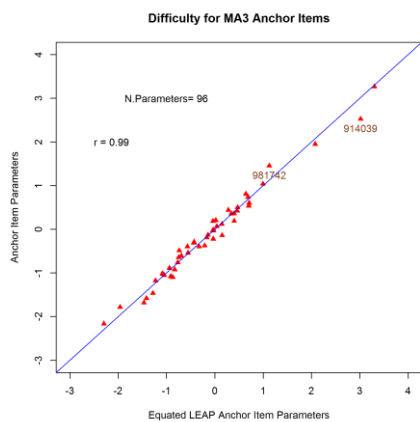
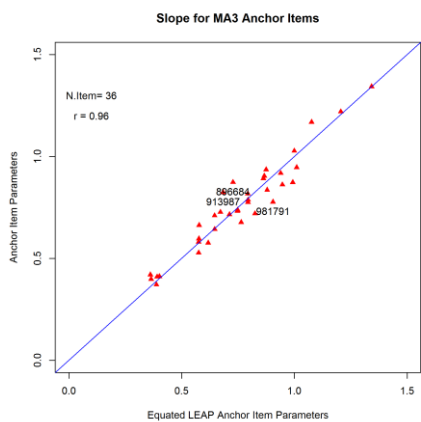
**Anchor 1**



**Anchor 2**

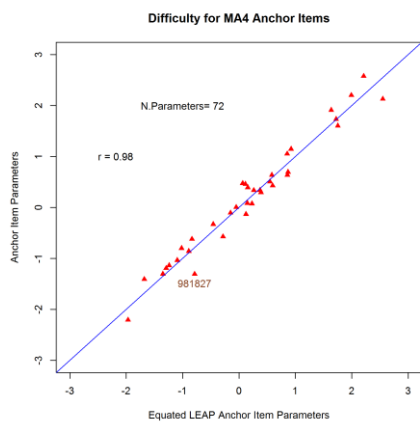
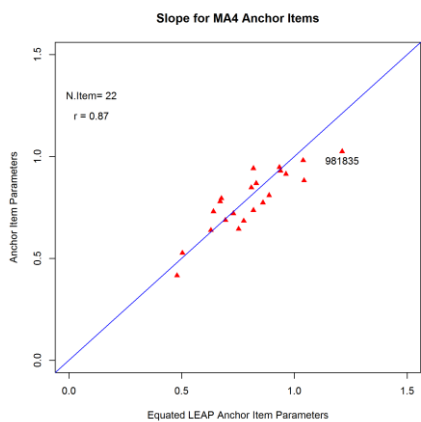


**Anchor 3**

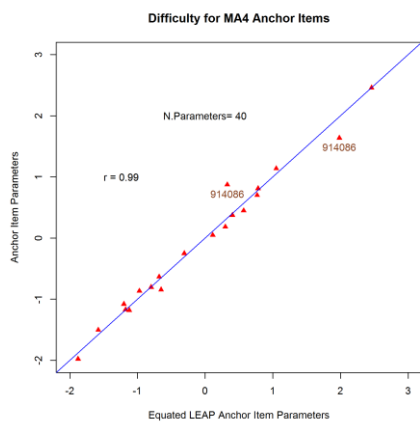
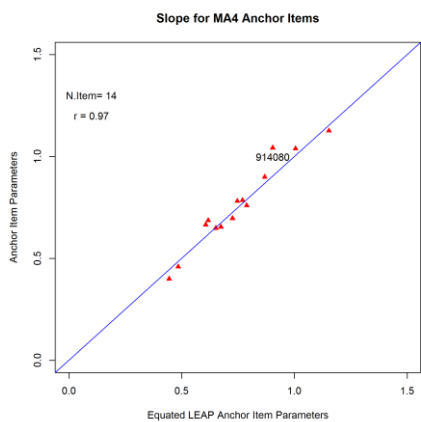


**Figure 6.22 Mathematics Grade 4 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

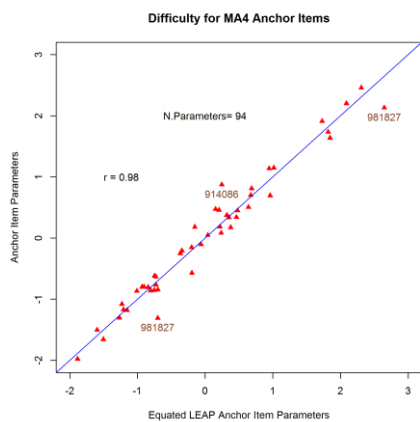
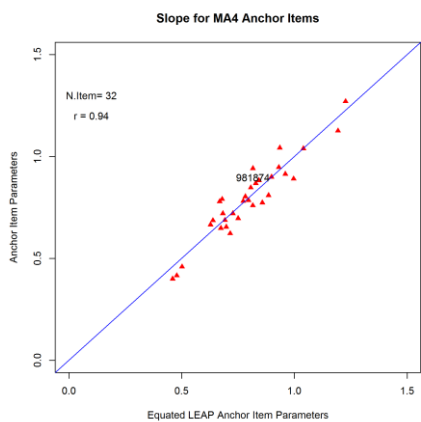
**Anchor 1**



**Anchor 2**

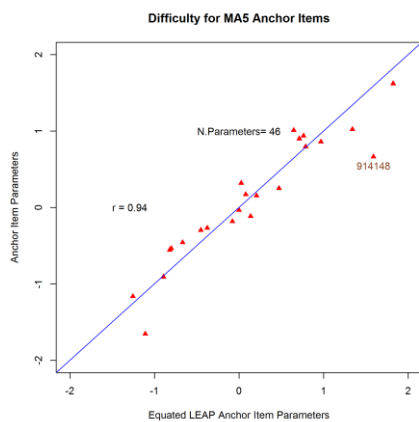
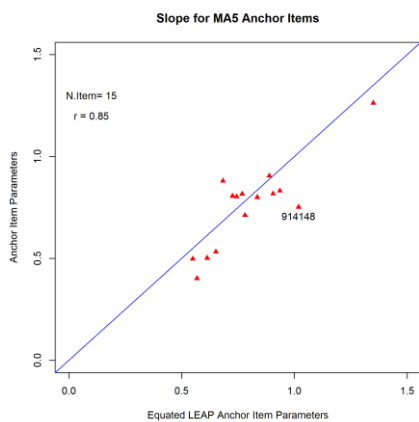


**Anchor 3**

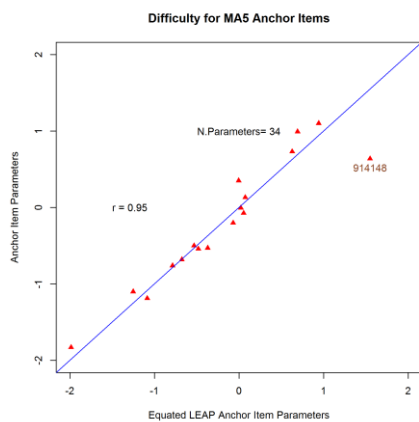
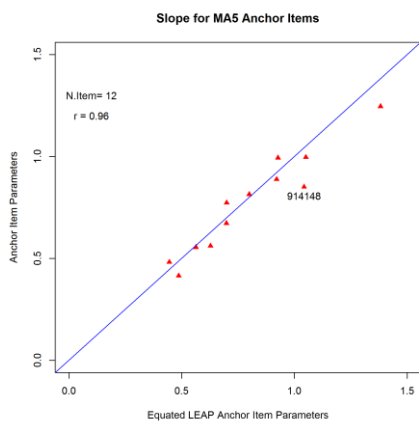


**Figure 6.23 Mathematics Grade 5 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

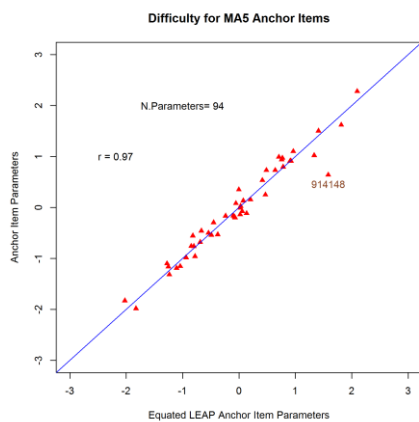
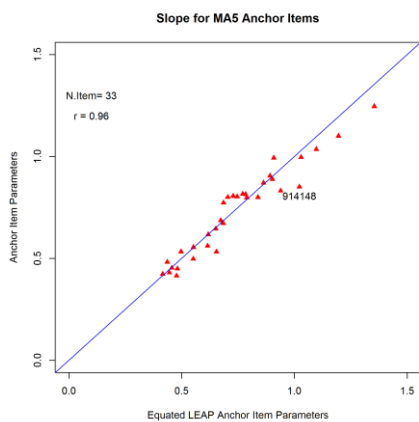
**Anchor 1**



**Anchor 2**

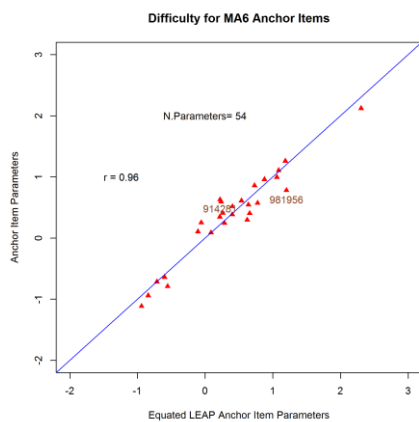
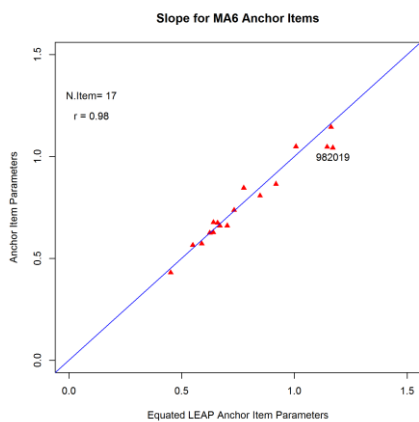


**Anchor 3**

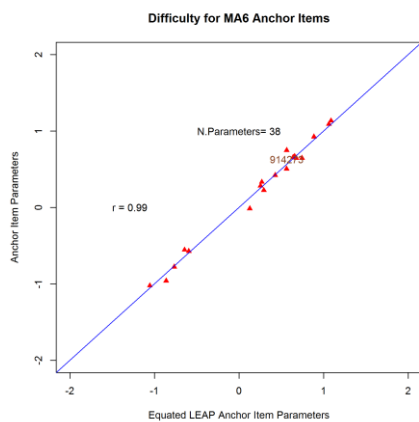
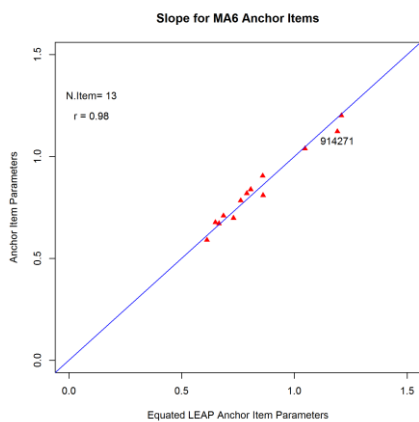


**Figure 6.24 Mathematics Grade 6 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

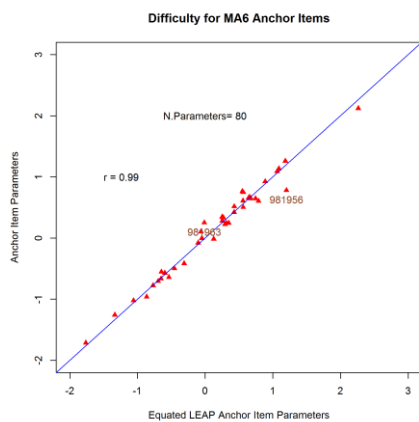
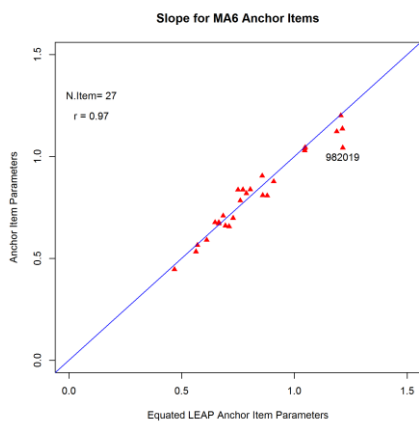
**Anchor 1**



**Anchor 2**

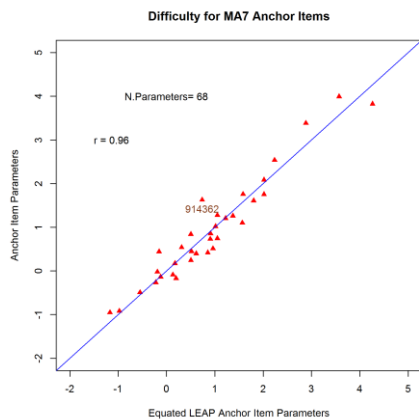
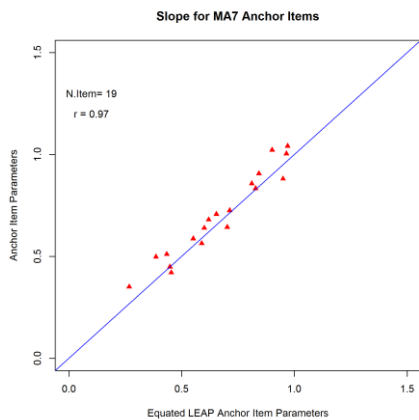


**Anchor 3**

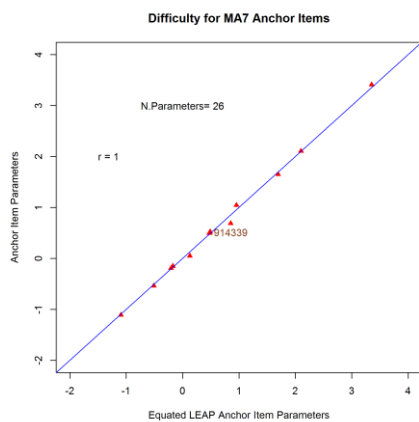
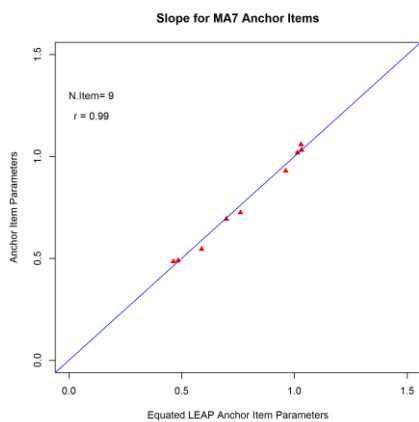


**Figure 6.25 Mathematics Grade 7 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

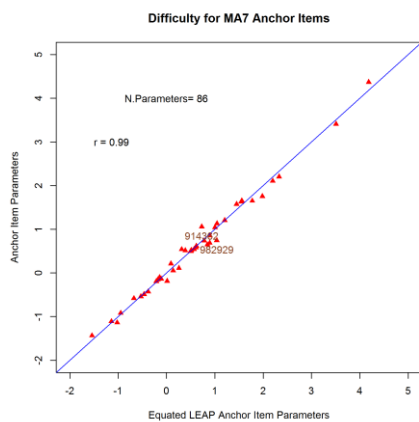
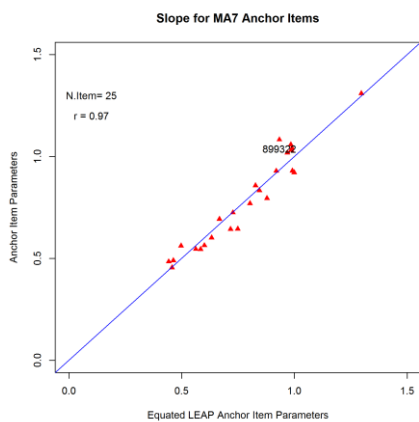
**Anchor 1**



**Anchor 2**

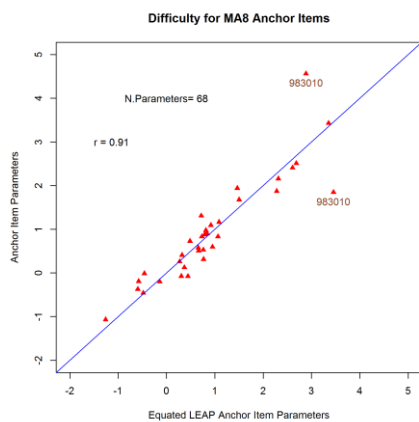
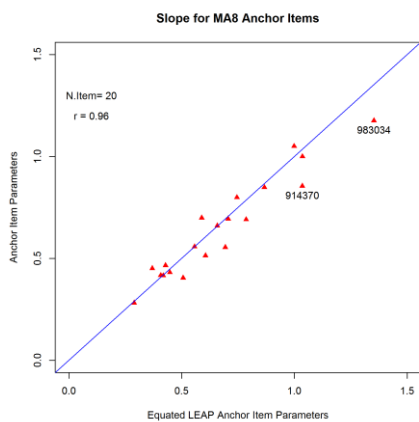


**Anchor 3**

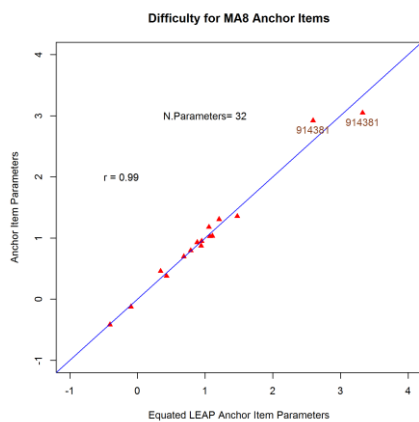
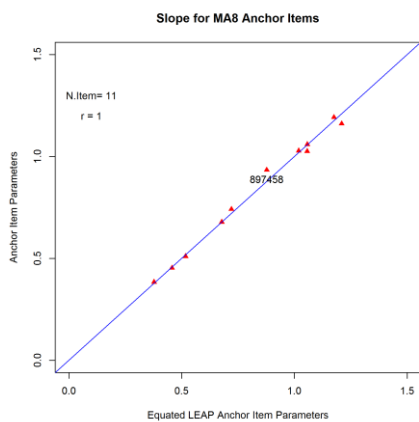


**Figure 6.26 Mathematics Grade 8 Slope and Difficulty Parameters Between Pre-equated and Equated Anchor Item Parameters**

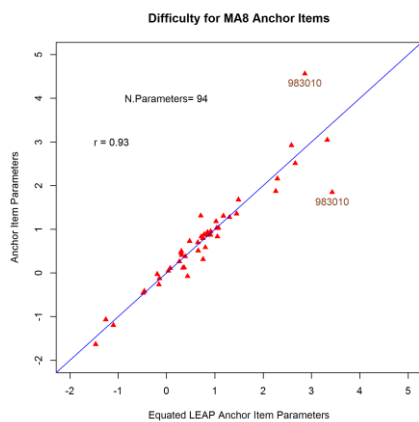
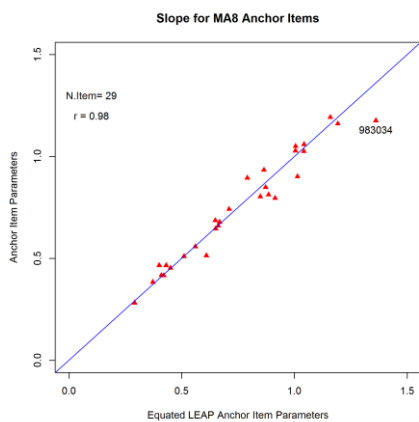
**Anchor 1**



**Anchor 2**



**Anchor 3**



### 6.4.2.1. Evaluation of Anchor Item Stability

Standard 5.15 requires that information about the anchors be presented, stating the following:

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented. (105)

One of the key requirements of anchor items in deriving valid reliable linking results is that the anchor items should form a miniature version of the test in terms of content coverage or test blueprint. Dropping flagged anchor items based solely on statistical criteria may change the content coverage and impact the validity of the results. Before an anchor item may be dropped from an anchor set, the item characteristics, adequacy of the content coverage, and impact to the size of the anchor set should be evaluated.

Outliers of anchor items were reviewed with the Robust Z (Huynh & Meyer, 2010) and the weighted root mean square difference (WRMSD) method in addition to being verified from a content perspective, when reviewers considered aspects of the outliers, such as the number of items and score points for each category and subcategory. If approved by the LDOE, the outliers were dropped from anchor sets and considered to be non-common anchor items during equating. The following evaluation rules were applied in order to check the quality of anchor items and the anchor set.

- Exclude CR items from anchor set if categories were collapsed due to small sample size.
- Exclude items with content or parameter estimation issues.
- Run Robust Z method and remove flagged items from anchor set using the criterion value of  $|1.96|$
- Run STUIRT using the remaining items after Robust Z is applied, and flag items for further inspection if the WRMSD was greater than the values in Table 6.28. If the items are flagged, then they were removed from the anchor set and the ICC was reviewed with the WRMSD (Kim & Kolen, 2004).
- Flag outliers using the plots of slope and difficulty item parameters with their correlations (Kolen & Brennan, 2014).
- Check score points and the numbers of items by reporting category and subcategory before and after dropping an anchor item.

Huynh and Meyer (2010) suggested to applying a z statistic that is robust under the presence of outliers. The robustification is established by replacing mean with median and standard deviation with interquartile range (IQR) for anchor items. A multiplicative constant (0.74) is applied to IQR to emulate the standard deviation of the normal distribution:

$$Z = \frac{(D - Md)}{0.74 \times IQR},$$

where  $D$  is the difference between intact and estimated item parameters of an anchor item and  $Md$  is a median of differences between intact and estimate item parameters for all items. The critical value of  $\pm 1.96$  is often used to evaluate estimated robust z values.

The WRMSD values were calculated to compare to the ICCs using intact and estimated anchor item parameters. WRMSD is defined as

$$SQRT\{\sum_{Q=1}^{41} W_Q [ICC_Q(EST) - ICC_Q(INTACT)]^2\},$$

where  $Q$  represents a quadrature point (i.e., node),  $W$  represents its weight given quadrature point  $Q$  from PARSCALE PH2 output,  $INTACT$  represents intact item parameters, and  $EST$  represents estimated item parameters corresponding to intact item parameters. Table 6.28 summarizes WRMSD flagging criteria for inspection and possible removal of linking items.

**Table 6.28 PARCC WRMSD Flagging Criteria**

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

#### 6.4.2.2. Lowest and Highest Obtainable Scale Scores

A maximum likelihood (MML) procedure cannot produce scale score estimates for students with perfect scores or scores below the level expected when students are guessing. In addition, although MML estimates are available for students with extreme scores other than zero or perfect, occasionally these estimates have standard errors of measurement that are very large, and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessary non-MML procedure. These values, which are set separately by grade, are called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). All grades and content areas in 2019 LEAP 2025 used the same LOSS and HOSS values. The LOSS value was 650, and the HOSS value was 850.

#### 6.4.2.3. Reporting Category and Subcategory Subscores

A student's performance on the ELA reporting categories (i.e., Reading and Writing) and mathematics categories (i.e., Major Content, Additional & Supporting Content, Expressing Mathematical Reasoning, and Modeling & Application) is reported in one of three ratings: *Weak*, *Moderate*, or *Strong*.

Additionally, subcategory ratings are reported at the student level for ELA and mathematics. ELA has three subcategories for reading (i.e., literary text, informational text, and vocabulary) and two subcategories for writing (i.e., written expression and knowledge and use of language conventions). Mathematics has four subcategories and they differ by grade. Subcategory performance is reported in one of three ratings of achievement: *Strong*, *Moderate*, or *Weak*. The 2019 LEAP 2025 reporting categories are summarized in chapter 3. Please see Table 3.1 for ELA and Table 3.8 and 3.9 for mathematics.

Although the performance ratings are determined only by the items included within a category or subcategory, the level of knowledge and ability needed to achieve a performance rating is connected to the level of knowledge and ability required to reach the subject-level achievement levels in the overall tests: a *Weak* rating requires similar knowledge and ability as the *Unsatisfactory* and *Approaching Basic* achievement levels, a *Moderate* rating requires similar knowledge and ability as the *Basic* achievement level, and a *Strong* rating requires similar knowledge and ability as the *Mastery* or *Advanced* achievement levels.



Reading and writing reporting category scores were produced for ELA assessments only. The reading category score range was 10–90 and the writing category score range was 10–60. The method for scaling categories followed the PARCC methodology (Pearson, 2017). For the reading category, two theta score points corresponding to ELA scale scores of 700 and 750 were used for scaling. Linear transformation constants mapping the two theta points to scale score points of 30 and 50 were calculated. After these transformation values were applied to item parameters belonging to the reading category, a scoring table was generated using the TCC inverse method. A similar approach was applied to scale the writing category, using two scale score points of 30 and 35. Two cut scores, 40 and 50 for reading and 30 and 35 for writing, were used to produce three performance-level ratings for each category (see Table 6.29 for cut scores for summatives, categories, and subcategories).

For reporting categories in mathematics and subcategories in ELA and mathematics, only performance-level ratings were reported. Therefore, there is no need to scale these scores. Using the item parameters belonging to a given category (mathematics) or subcategory (ELA), a raw-score-to-theta scoring table is generated by applying the TCC inverse method. PARCC estimated  $\theta_{L3}$  and  $\theta_{L4}$  corresponding to scale scores of 725 and 750 for each content/grade using PARCC 2016 operational items by the TCC inverse method, and these values are the same across years. The two raw scores corresponding to  $\theta_{L3}$  and  $\theta_{L4}$  are cut scores for the category (mathematics) and subcategory (ELA).

This is also illustrated in Table 6.29.

**Table 6.29 Cut Scores for Summative, Reporting Categories, and Subcategories**

Performance Level	Summative Test	Category (ELA)		Category (Mathematics)/Subcategory (Mathematics and ELA)
		Reading	Writing	
1				
2	700	30	25	
3	725	40	30	$\theta_{L3}$
4	750	50	35	$\theta_{L4}$
5	Around 800			

\*Subcategory thetas are those from summative tests (i.e., 725 & 750).

\*\*Yellow highlight shows cut scores for category and subcategory.

The primary purpose of form equating is to establish score equivalency between two (or more) forms. Equivalency is established by first building the forms to be equated according to tight content specifications. Then the form scores are placed on the same scale (by equating), such that students performing on an assessment at the same level of (underlying) achievement should receive the same scale score, although they may not receive the same number-correct score (or raw score). The raw-to-scale-score relationship performs this leveling function based on form-equating studies. Theoretically, differences in the raw-to-scale-score relationship between the two forms can be partially due to differences in the samples utilized for calibration and the differences in item difficulty. The LDOE and DRC strive to maintain equivalent samples or use near-census samples over the years, minimizing the potential differences due to the samples. Differences in the raw-to-scale-score relationship, therefore, can be primarily attributed to the differences in item difficulty.

The forms used in the spring 2019 were post-equated forms. Just as in previous years, equating was conducted using the test characteristic transformation function method in the common-item non-equivalent-groups design (Stocking & Lord, 1983). Tables 6.30 through 6.41. provide scale scores at selected percentiles that can be used to compare the distributional characteristics of the Spring 2019 forms to previous administrations. Although these scale scores are rounded values, there were differences in the scale-score values for a given percentile across the forms. These variations could arise for several reasons: (1) differences in the proficiency (i.e., achievement) of students in the samples or growth in student achievement across years; (2) unevenness in the respective distributions that combine with the number-correct-to-scale-score scoring method, leaving “gaps” in the scale; or (3) other sources of equating error. Other sources of equating error can include subtle content differences between forms, handscoring differences, or unusual student samples. Some equating errors will always be present between forms. This means that the forms will not measure identically, even under optimal testing conditions. In general, however, the test characteristic function equating techniques will “level” the equated forms through the raw-to-scale-score adjustment.

**Table 6.30 Comparisons of Scale Scores at Selected Percentiles—Grade 3 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	822	839	842	845
95	796	810	810	816
90	783	793	797	802
85	774	784	788	792
80	768	775	779	782
75	762	770	773	776
70	757	762	768	770
65	751	757	762	764
60	746	752	757	758
55	741	748	752	752
50	738	743	746	746
45	732	739	741	740
40	727	734	736	734
35	721	727	730	728
30	715	723	724	722
25	712	718	715	715
20	706	710	708	708
15	695	701	701	700
10	687	695	692	690
5	676	679	676	679
1	654	655	650	650

**Table 6.31 Comparisons of Scale Scores at Selected Percentiles—Grade 4 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	816	818	821	824
95	794	796	800	801
90	785	785	789	789
85	777	777	778	780
80	769	771	774	774
75	765	765	767	768
70	760	761	763	762
65	755	756	757	758
60	751	752	753	753
55	746	748	749	750
50	744	744	744	744
45	740	741	740	741
40	735	737	736	736
35	731	733	731	731
30	727	728	727	726
25	722	724	721	721
20	715	717	714	714
15	709	711	707	706
10	701	702	698	699
5	691	691	687	688
1	666	670	668	665

**Table 6.32 Comparisons of Scale Scores at Selected Percentiles—Grade 5 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	816	813	817	821
95	792	793	795	798
90	782	782	782	784
85	774	775	777	776
80	767	769	769	770
75	763	763	765	765
70	758	758	760	759
65	754	754	756	754
60	749	750	753	751
55	745	747	749	745
50	740	743	746	742
45	738	739	740	737
40	733	735	736	733
35	728	731	732	729
30	723	727	728	725
25	720	721	724	718
20	714	716	716	713
15	708	709	711	707
10	701	701	702	701
5	692	691	691	693
1	675	673	676	676

**Table 6.33 Comparisons of Scale Scores at Selected Percentiles—Grade 6 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	813	814	808	812
95	792	790	789	791
90	780	779	777	778
85	772	770	770	771
80	765	763	763	766
75	760	759	758	761
70	756	754	753	756
65	752	748	749	751
60	748	745	746	747
55	745	741	742	743
50	741	736	737	740
45	737	733	735	735
40	734	729	730	731
35	730	724	726	728
30	727	721	721	723
25	723	716	718	718
20	718	711	713	714
15	713	705	707	708
10	706	698	700	701
5	696	689	691	692
1	676	671	675	675

**Table 6.34 Comparisons of Scale Scores at Selected Percentiles—Grade 7 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	825	826	831	826
95	800	800	801	804
90	787	786	789	789
85	777	778	780	782
80	771	770	774	775
75	766	765	767	769
70	761	759	762	764
65	756	756	757	759
60	751	751	752	756
55	747	745	749	750
50	742	742	744	747
45	740	737	740	741
40	735	733	735	736
35	730	728	730	731
30	726	723	726	727
25	721	717	719	720
20	714	711	713	714
15	706	702	707	705
10	697	692	697	695
5	683	675	685	681
1	655	654	662	659

**Table 6.35 Comparisons of Scale Scores at Selected Percentiles—Grade 8 ELA**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	825	834	824	831
95	804	806	801	804
90	790	791	789	793
85	781	782	781	785
80	775	776	774	777
75	770	770	768	771
70	764	764	764	766
65	759	758	758	760
60	754	754	754	755
55	752	749	751	750
50	747	745	745	746
45	743	740	741	741
40	739	734	737	736
35	735	731	732	732
30	731	725	726	727
25	727	719	722	721
20	721	714	716	714
15	714	707	708	707
10	706	696	699	696
5	693	681	683	686
1	670	651	657	667



**Table 6.36 Comparisons of Scale Scores at Selected Percentiles—Grade 3 Mathematics**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	824	822	817	815
95	802	796	793	796
90	789	786	783	784
85	781	776	775	776
80	775	772	771	771
75	770	765	764	764
70	765	761	759	760
65	760	756	755	756
60	756	752	750	752
55	751	747	746	748
50	746	743	742	744
45	741	738	740	738
40	738	733	735	735
35	733	728	731	731
30	728	725	726	724
25	722	720	719	720
20	716	715	713	713
15	710	706	708	705
10	703	699	698	700
5	692	689	686	686
1	672	667	664	672

**Table 6.37 Comparisons of Scale Scores at Selected Percentiles—Grade 4 Mathematics**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	819	812	812	813
95	797	792	790	792
90	786	779	780	781
85	777	774	772	774
80	771	767	768	769
75	766	762	762	763
70	761	756	757	759
65	756	752	753	755
60	752	748	749	750
55	747	744	744	746
50	743	740	740	742
45	738	736	735	737
40	732	732	733	732
35	728	727	728	728
30	723	722	723	724
25	718	717	718	719
20	713	712	715	712
15	708	706	710	706
10	703	700	700	699
5	693	693	689	688
1	677	674	670	673

**Table 6.38 Comparisons of Scale Scores at Selected Percentiles—Grade 5 Mathematics**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	819	808	810	809
95	792	784	784	788
90	779	774	774	778
85	771	767	765	769
80	766	760	759	763
75	759	755	755	757
70	754	751	749	753
65	749	747	745	748
60	745	742	743	744
55	740	740	738	740
50	735	735	734	737
45	731	730	729	733
40	728	728	727	728
35	722	723	722	724
30	720	720	720	719
25	714	715	714	714
20	711	709	711	711
15	705	706	705	705
10	699	699	698	699
5	691	691	689	690
1	678	675	672	674

**Table 6.39 Comparisons of Scale Scores at Selected Percentiles—Grade 6 Mathematics**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	803	808	800	804
95	783	781	780	783
90	771	771	770	773
85	765	762	762	765
80	758	757	757	758
75	753	752	752	754
70	747	746	748	750
65	744	742	743	745
60	740	738	739	742
55	735	734	736	739
50	731	732	732	733
45	729	727	728	729
40	724	724	723	725
35	722	719	721	721
30	717	717	716	717
25	714	711	713	714
20	709	708	707	709
15	706	701	704	703
10	699	697	696	696
5	692	688	686	687
1	679	671	672	667

**Table 6.40 Comparisons of Scale Scores at Selected Percentiles—Grade 7 Mathematics**

	<b>2016</b>	<b>2017</b>	<b>2018</b>	<b>2019</b>
<b>Percentile</b>	<b>Form A</b>	<b>Form B</b>	<b>Form C</b>	<b>Form D</b>
99	797	796	797	796
95	779	777	777	776
90	768	766	766	766
85	760	760	759	761
80	754	754	755	756
75	750	749	750	752
70	746	746	745	748
65	742	741	742	743
60	738	737	739	740
55	734	734	735	736
50	730	731	731	732
45	728	727	729	730
40	723	723	725	726
35	721	721	721	722
30	719	717	718	719
25	714	712	713	714
20	712	709	710	711
15	706	706	706	705
10	703	699	702	701
5	695	694	693	692
1	678	673	679	680

**Table 6.41 Comparisons of Scale Scores at Selected Percentiles—Grade 8 Mathematics**

	2016	2017	2018	2019
Percentile	Form A	Form B	Form C	Form D
99	808	809	807	812
95	787	784	784	788
90	775	771	773	775
85	766	763	764	766
80	761	757	757	758
75	753	751	752	752
70	749	746	746	746
65	744	741	742	742
60	737	736	737	737
55	734	730	732	732
50	731	727	727	730
45	727	724	721	724
40	724	718	718	721
35	720	714	715	715
30	712	710	707	711
25	708	706	702	707
20	704	698	697	699
15	699	693	691	694
10	695	687	684	689
5	684	674	676	677
1	663	656	654	659

Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) for the LEAP 2025 across administrations, see figures 6.26 and 6.27. For most content areas and grades, the TCCs for the three years were similar across ability ranges. For ELA grade 5 and 6, the 2018 forms were slightly easier than the 2017 and 2019 forms for high-performing students. For grade 7, the 2018 and 2019 forms were slightly easier than the 2017 forms. Grade 3 forms have been gradually becoming more difficult from 2017 to 2019. For grade 8, the 2019 form was more difficult than the 2017 and 2018 forms across all ability levels.

For mathematics grades 7 and 8, the 2019 and 2017 forms were slightly easier than the 2018 form for low-performing students. For grades 3 and 4, the 2019 forms were slightly more difficult than the 2018 forms for low-performing students. For grade 5, the 2019 form was easier than the 2017 and 2018 forms for high-performing students. Note that this different form difficulty is adjusted by reporting different scale scores for given raw scores; a scale score of a difficult form is higher than that of an easy form given the same raw score.

Figures 6.28 and 6.29 show SEMs for the 2017- 2019 LEAP 2025 assessments. For most content areas and grades, the SEMs were similar across ability ranges, especially in the middle ability ranges.

Figure 6.26 TCCs Across Years: ELA

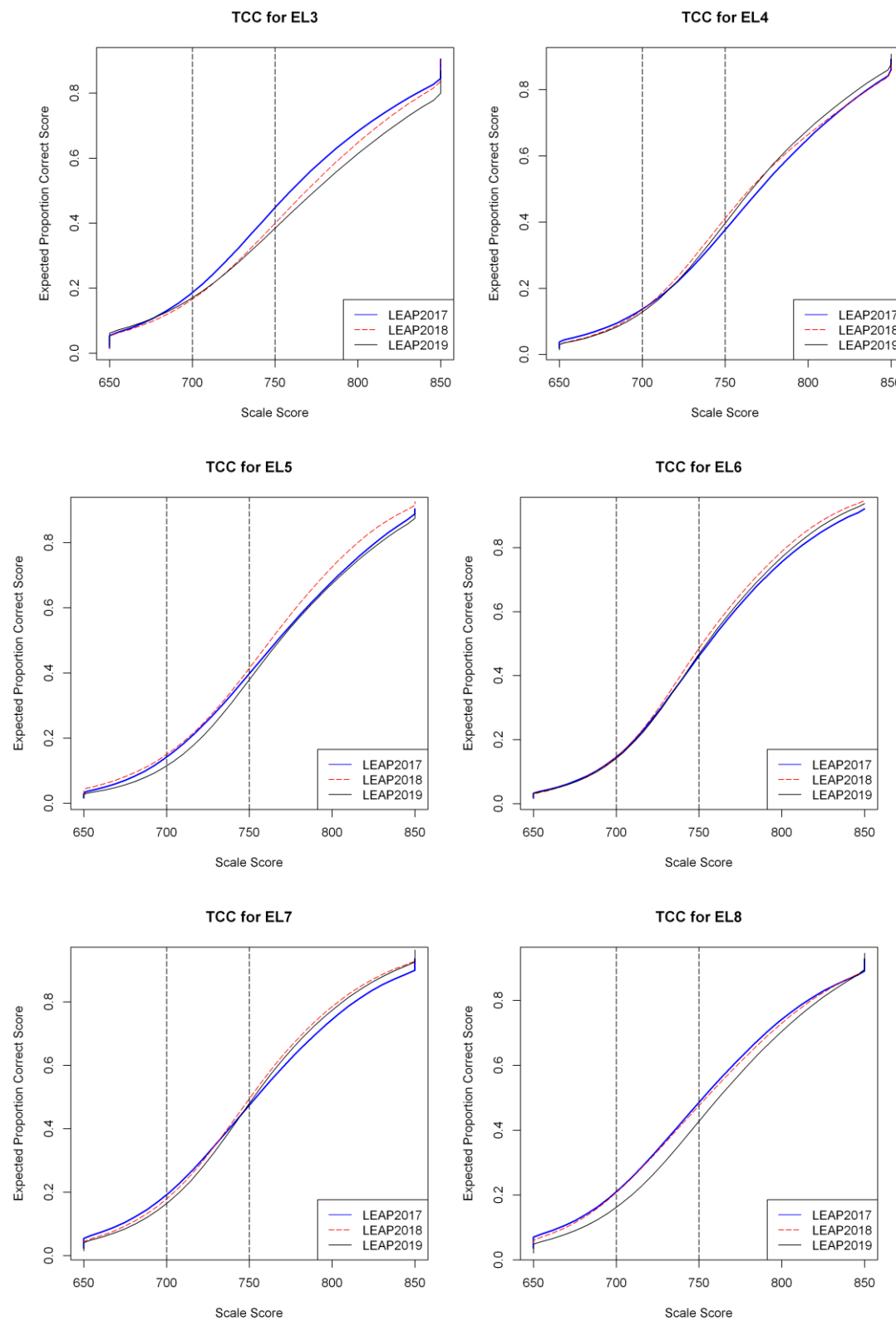


Figure 6.27 TCCs Across Years: Mathematics

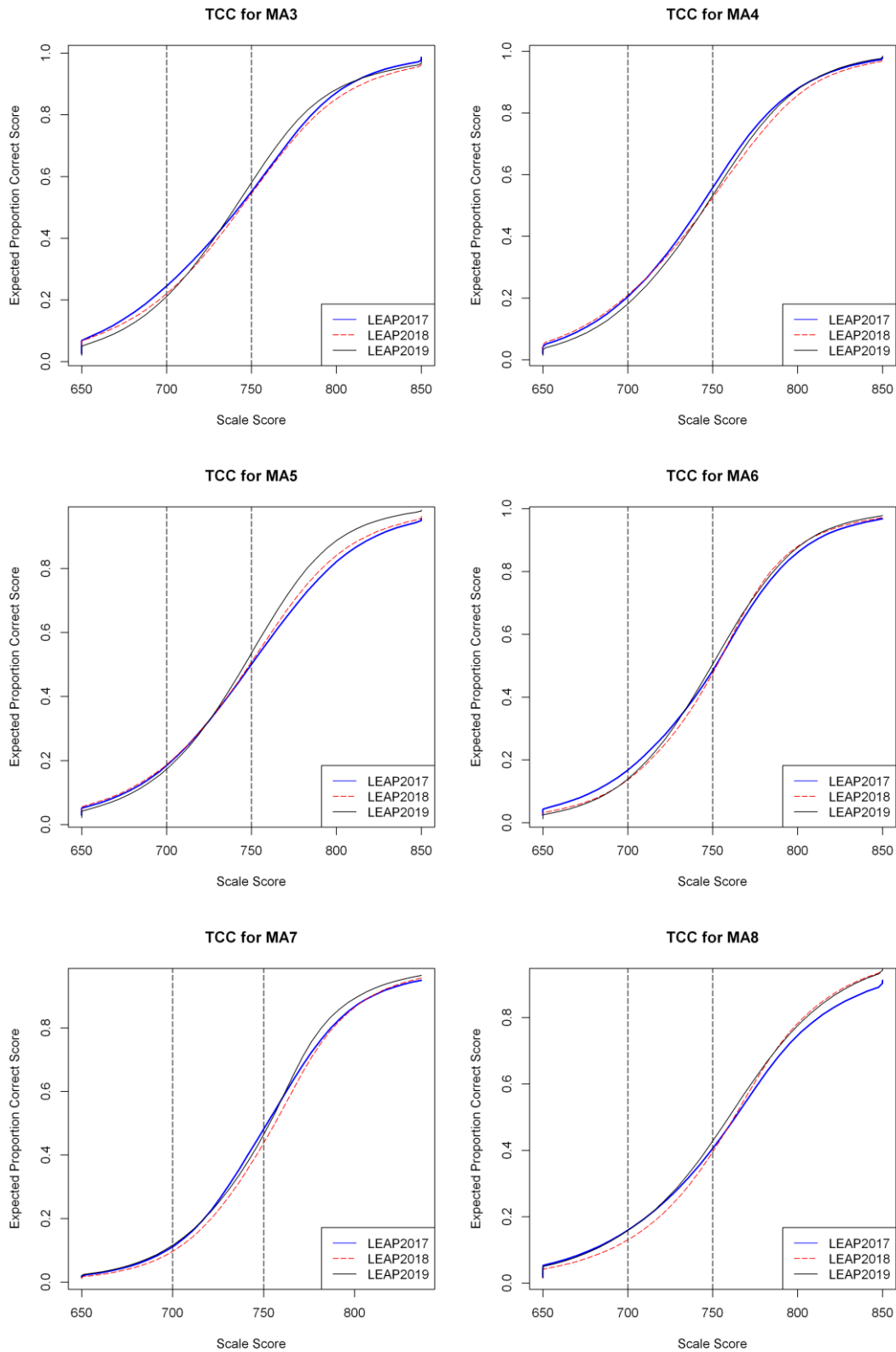




Figure 6.28 SEM Across Years: ELA

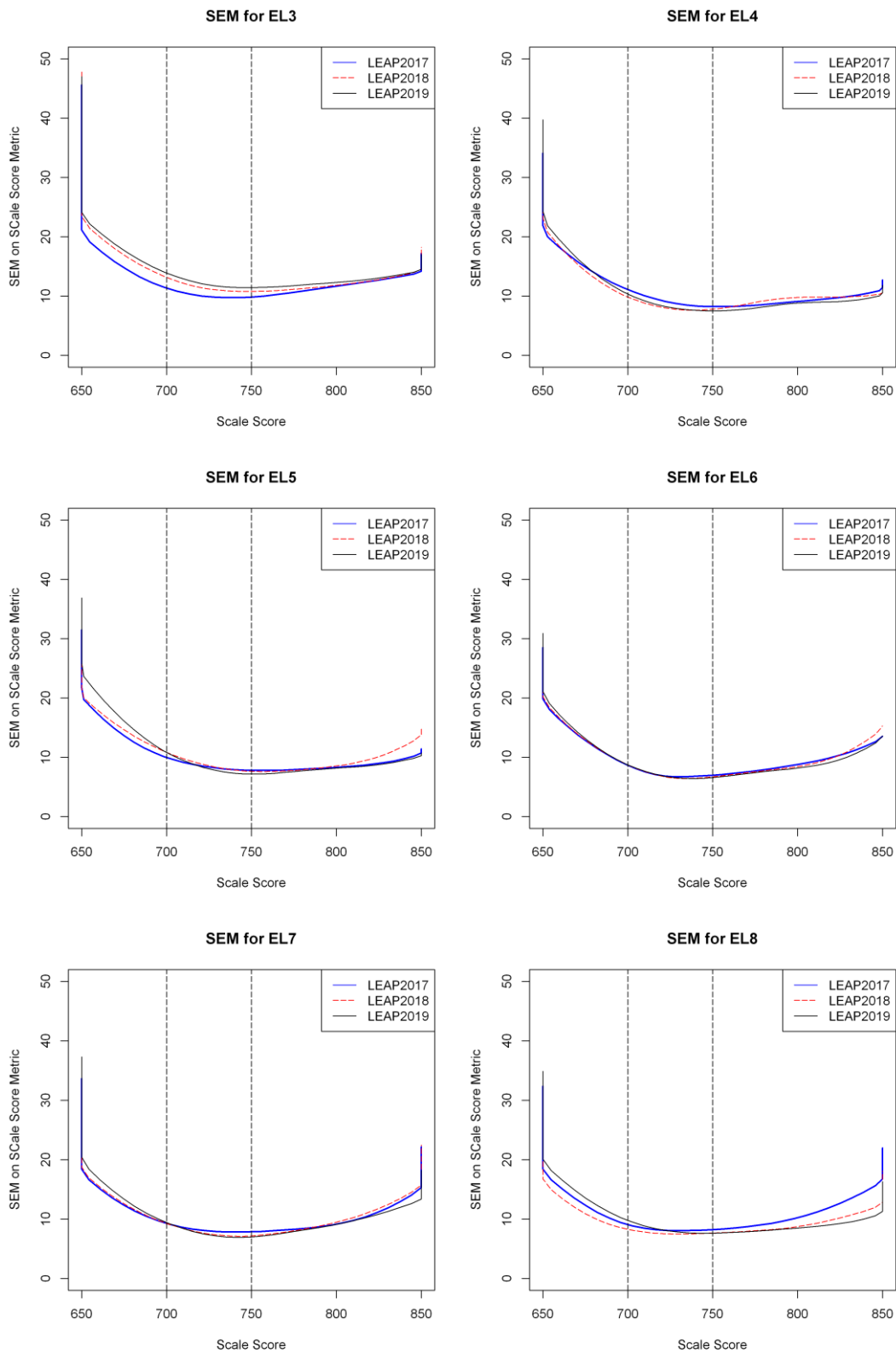
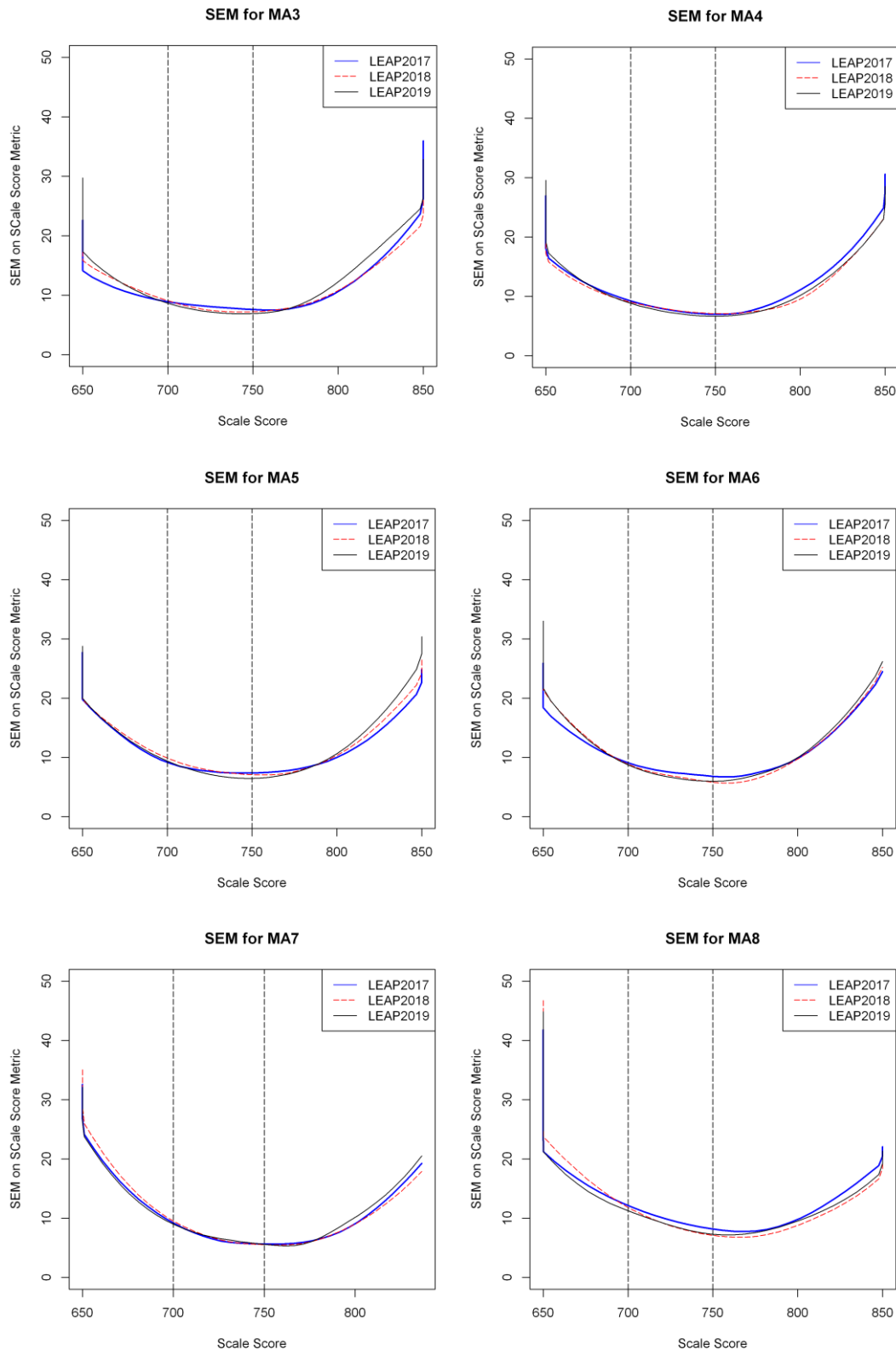


Figure 6.29 SEM Across Years: Mathematics



In summary, the overall purpose of the operational data analyses is to ensure that the test items, as well as the overall test, are functioning appropriately. Operational data analyses also help maintain the test scale so that test results may be appropriately compared across years. The data analyses undertaken by DRC address multiple best practices of the testing industry but are particularly related to the following standards:

**Standard 1.8** The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics. (25)

**Standard 4.14** For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (90)

**Standard 5.2** The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly. (102)

**Standard 5.13** When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions. (105)

**Standard 5.15** In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented. (105)

**Standard 7.2** The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported. (126)

## Chapter 7: Test Results

---

This chapter of the technical report contains information on the results of the Spring 2019 LEAP 2025 ELA and mathematics assessments. The scale score results and achievement level information are presented here. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative description of student achievement. The levels are *Advanced*, *Mastery*, *Basic*, *Approaching Basic*, and *Unsatisfactory*.

While the scale score provides an essential quantitative reference for student achievement, the achievement-level information plainly outlines the meanings of the scores to parents, students, and educators. When combined, scale scores and achievement levels provide a comprehensive set of tools to assess Louisiana student achievement by content and grade level.

This chapter also provides descriptions of the score reports, data structure, and interpretive guide. The American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014) *Standards for Educational & Psychological Testing* addressed in Chapter 7 are 5.1, 6.10, 7.0, and 12.18. Each standard is presented in the pertinent section of this chapter.

The results presented in this chapter are based on census data. The results presented here may differ slightly from the official state summary report of all student populations due to ongoing resolution of test materials and student information. The results in the tables in this chapter are presented as evidence of the reliability and validity of the scores from the LEAP 2025 assessments and should not be used for state accountability purposes.

The following are subgroups reported during the administration of the LEAP 2025 tests:

- Gender: Female and Male
- Race and Ethnicity: Hispanic/Latino, American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, and Two or More Races
- Education Classification
- Economic Status
- English Learner (EL)
- Migrant Status

For the purposes of this report, participation rate is defined as the percentage of students who received a valid scale score given the total number of students who were expected to take the online test or receive a test book. These participation rates are summarized in Table 7.1. Both the percentage of students classified as reportable and the number of students classified as accountable are reported. Reportable students include all students with a valid scale score. The “Accountable” columns shows the total numbers of students who were expected to take the online test or receive a test book. These include students who should have received a LEAP 2025 scale score but who did not take the test and could not be assigned a scale score.

Table 7.1 Participation Rates

Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
3	All Students	≥52,910	99.86%	≥53,440	99.86%
	<b>Gender</b>				
	Female	≥25,670	99.90%	≥25,910	99.92%
	Male	≥27,210	99.85%	≥27,480	99.87%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥4,330	99.95%	≥4,400	99.98%
	American Indian or Alaska Native	≥320	100.00%	≥320	100.00%
	Asian	≥820	99.76%	≥820	99.76%
	Black or African American	≥22,850	99.88%	≥23,120	99.91%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥22,790	99.87%	≥22,940	99.87%
	Two or More Races	≥1,670	99.94%	≥1,680	100.00%
	<b>Education Classification</b>				
	Regular	≥46,560	99.86%	≥47,010	99.86%
	Special	≥6,350	99.80%	≥6,420	99.84%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥37,880	99.90%	≥38,200	99.92%
	Not Economically Disadvantaged	≥14,700	99.84%	≥14,770	99.86%
	<b>English Learner Status</b>				
	Not English Learner	≥50,580	99.86%	≥51,060	99.86%
	English Learner	≥2,330	99.87%	≥2,380	99.87%
	<b>Migrant Status</b>				
	Not Migrant	≥52,760	99.86%	≥53,290	99.86%
	Migrant	≥140	100.00%	≥150	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥48,220	99.85%	≥48,700	99.85%
	Section 504	≥4,690	99.94%	≥4,730	99.89%
	<b>Homeless Status</b>				
	Not Homeless	≥51,830	99.86%	≥52,330	99.86%
	Homeless	≥1,070	99.72%	≥1,110	99.82%
	<b>Foster Care Status</b>				
	Not in Foster Care	≥52,640	99.86%	≥53,160	99.86%
In Foster Care	≥270	100.00%	≥280	100.00%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥51,990	99.86%	≥52,510	99.86%	
Military Affiliated	≥920	99.78%	≥920	99.78%	

Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
4	All Students	≥54,700	99.84%	≥55,170	99.85%
	<b>Gender</b>				
	Female	≥26,840	99.90%	≥27,050	99.92%
	Male	≥27,820	99.87%	≥28,050	99.89%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥4,250	99.86%	≥4,310	99.84%
	American Indian or Alaska Native	≥320	100.00%	≥320	100.00%
	Asian	≥740	100.00%	≥750	100.00%
	Black or African American	≥24,010	99.85%	≥24,250	99.90%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥23,480	99.92%	≥23,590	99.92%
	Two or More Races	≥1,760	100.00%	≥1,770	100.00%
	<b>Education Classification</b>				
	Regular	≥48,210	99.85%	≥48,620	99.86%
	Special	≥6,490	99.77%	≥6,540	99.80%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥39,110	99.88%	≥39,370	99.92%
	Not Economically Disadvantaged	≥15,260	99.91%	≥15,320	99.92%
	<b>English Learner Status</b>				
	Not English Learner	≥52,670	99.84%	≥53,090	99.85%
	English Learner	≥2,020	99.85%	≥2,070	99.71%
	<b>Migrant Status</b>				
	Not Migrant	≥54,580	99.84%	≥55,040	99.85%
	Migrant	≥120	100.00%	≥120	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥48,900	99.83%	≥49,330	99.84%
	Section 504	≥5,790	99.91%	≥5,830	99.95%
	<b>Homeless Status</b>				
	Not Homeless	≥53,590	99.84%	≥54,020	99.85%
	Homeless	≥1,100	99.73%	≥1,140	99.83%
<b>Foster Care Status</b>					
Not in Foster Care	≥54,440	99.84%	≥54,900	99.85%	
In Foster Care	≥260	100.00%	≥260	100.00%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥53,820	99.84%	≥54,280	99.85%	
Military Affiliated	≥880	100.00%	≥890	100.00%	

Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
5	All Students	≥54,780	99.92%	≥54,790	99.92%
	<b>Gender</b>				
	Female	≥26,870	99.93%	≥26,870	99.93%
	Male	≥27,910	99.91%	≥27,910	99.91%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥4,270	99.91%	≥4,270	99.91%
	American Indian or Alaska Native	≥340	100.00%	≥340	100.00%
	Asian	≥840	99.88%	≥840	99.88%
	Black or African American	≥23,850	99.92%	≥23,850	99.92%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥23,690	99.92%	≥23,690	99.92%
	Two or More Races	≥1,720	100.00%	≥1,720	100.00%
	<b>Education Classification</b>				
	Regular	≥48,420	99.93%	≥48,420	99.93%
	Special	≥6,360	99.87%	≥6,360	99.87%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥38,790	99.93%	≥38,790	99.93%
	Not Economically Disadvantaged	≥15,580	99.92%	≥15,580	99.92%
	<b>English Learner Status</b>				
	Not English Learner	≥53,050	99.92%	≥53,050	99.92%
	English Learner	≥1,730	99.94%	≥1,730	99.94%
	<b>Migrant Status</b>				
	Not Migrant	≥54,700	99.92%	≥54,710	99.92%
	Migrant	≥80	100.00%	≥80	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥49,390	99.91%	≥49,390	99.91%
	Section 504	≥5,390	100.00%	≥5,390	100.00%
	<b>Homeless Status</b>				
	Not Homeless	≥53,820	99.92%	≥53,820	99.92%
	Homeless	≥960	99.90%	≥960	99.90%
<b>Foster Care Status</b>					
Not in Foster Care	≥54,560	99.92%	≥54,560	99.92%	
In Foster Care	≥220	100.00%	≥220	100.00%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥53,890	99.92%	≥53,890	99.92%	
Military Affiliated	≥890	100.00%	≥890	100.00%	

Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
6	All Students	≥54,800	99.84%	≥54,800	99.85%
	<b>Gender</b>				
	Female	≥26,950	99.84%	≥26,950	99.85%
	Male	≥27,850	99.83%	≥27,850	99.85%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥3,910	99.85%	≥3,910	99.85%
	American Indian or Alaska Native	≥350	100.00%	≥350	100.00%
	Asian	≥790	99.87%	≥790	99.87%
	Black or African American	≥23,860	99.82%	≥23,870	99.84%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥24,220	99.85%	≥24,220	99.86%
	Two or More Races	≥1,590	99.87%	≥1,590	99.87%
	<b>Education Classification</b>				
	Regular	≥48,920	99.85%	≥48,920	99.86%
	Special	≥5,870	99.73%	≥5,870	99.74%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥38,310	99.81%	≥38,320	99.83%
	Not Economically Disadvantaged	≥16,090	99.90%	≥16,090	99.90%
	<b>English Learner Status</b>				
	Not English Learner	≥53,410	99.84%	≥53,420	99.85%
	English Learner	≥1,380	99.93%	≥1,380	99.93%
	<b>Migrant Status</b>				
	Not Migrant	≥54,730	99.84%	≥54,740	99.85%
	Migrant	≥60	100.00%	≥60	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥49,340	99.84%	≥49,350	99.85%
	Section 504	≥5,450	99.84%	≥5,450	99.85%
	<b>Homeless Status</b>				
	Not Homeless	≥53,800	99.85%	≥53,800	99.86%
	Homeless	≥990	99.50%	≥990	99.50%
<b>Foster Care Status</b>					
Not in Foster Care	≥54,620	99.84%	≥54,630	99.85%	
In Foster Care	≥170	99.43%	≥170	99.43%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥53,940	99.84%	≥53,950	99.85%	
Military Affiliated	≥850	100.00%	≥850	100.00%	



Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
7	All Students	≥52,290	99.80%	≥52,300	99.81%
	<b>Gender</b>				
	Female	≥25,510	99.84%	≥25,510	99.85%
	Male	≥26,780	99.76%	≥26,790	99.78%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥3,560	99.89%	≥3,560	99.89%
	American Indian or Alaska Native	≥340	99.71%	≥340	99.71%
	Asian	≥790	99.75%	≥790	99.75%
	Black or African American	≥23,030	99.74%	≥23,030	99.77%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥23,130	99.85%	≥23,140	99.86%
	Two or More Races	≥1,370	99.78%	≥1,370	99.78%
	<b>Education Classification</b>				
	Regular	≥46,870	99.81%	≥46,870	99.83%
	Special	≥5,420	99.67%	≥5,420	99.69%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥36,090	99.75%	≥36,100	99.77%
	Not Economically Disadvantaged	≥15,800	99.93%	≥15,800	99.94%
	<b>English Learner Status</b>				
	Not English Learner	≥51,050	99.80%	≥51,060	99.82%
	English Learner	≥1,240	99.68%	≥1,240	99.76%
	<b>Migrant Status</b>				
	Not Migrant	≥52,220	99.80%	≥52,230	99.81%
	Migrant	≥60	100.00%	≥60	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥47,060	99.80%	≥47,070	99.81%
	Section 504	≥5,230	99.81%	≥5,230	99.83%
	<b>Homeless Status</b>				
	Not Homeless	≥51,380	99.80%	≥51,390	99.82%
	Homeless	≥910	99.67%	≥910	99.67%
<b>Foster Care Status</b>					
Not in Foster Care	≥52,100	99.80%	≥52,110	99.81%	
In Foster Care	≥190	100.00%	≥190	100.00%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥51,480	99.80%	≥51,480	99.81%	
Military Affiliated	≥810	100.00%	≥810	100.00%	

Participation Rates by Grade and Subgroup					
Grade	Group	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics
8	All Students	≥50,780	99.66%	≥50,800	99.70%
	<b>Gender</b>				
	Female	≥24,790	99.70%	≥24,800	99.75%
	Male	≥25,990	99.62%	≥25,990	99.67%
	<b>Ethnicity</b>				
	Hispanic/Latino	≥3,340	99.79%	≥3,340	99.82%
	American Indian or Alaska Native	≥350	100.00%	≥350	100.00%
	Asian	≥780	100.00%	≥780	100.00%
	Black or African American	≥22,160	99.50%	≥22,160	99.61%
	Native Hawaiian or Other Pacific	≥20	100.00%	≥20	100.00%
	White	≥22,970	99.80%	≥22,980	99.79%
	Two or More Races	≥1,130	99.30%	≥1,130	99.30%
	<b>Education Classification</b>				
	Regular	≥45,830	99.68%	≥45,850	99.73%
	Special	≥4,940	99.43%	≥4,940	99.43%
	<b>Economic Status</b>				
	Economically Disadvantaged	≥34,120	99.58%	≥34,120	99.65%
	Not Economically Disadvantaged	≥16,150	99.86%	≥16,150	99.89%
	<b>English Learner Status</b>				
	Not English Learner	≥49,540	99.66%	≥49,540	99.71%
	English Learner	≥1,240	99.60%	≥1,250	99.60%
	<b>Migrant Status</b>				
	Not Migrant	≥50,720	99.66%	≥50,740	99.70%
	Migrant	≥50	100.00%	≥50	100.00%
	<b>Section 504 Status</b>				
	Not Section 504	≥45,680	99.65%	≥45,700	99.69%
	Section 504	≥5,090	99.78%	≥5,090	99.84%
	<b>Homeless Status</b>				
	Not Homeless	≥49,890	99.67%	≥49,910	99.71%
	Homeless	≥890	99.10%	≥890	99.33%
<b>Foster Care Status</b>					
Not in Foster Care	≥50,600	99.66%	≥50,620	99.71%	
In Foster Care	≥180	98.36%	≥180	98.36%	
<b>Military Affiliation</b>					
Not Military Affiliated	≥50,030	99.65%	≥50,050	99.70%	
Military Affiliated	≥740	100.00%	≥740	100.00%	

\*Students in grade 8 who enrolled in Algebra I had the option of taking the Algebra LEAP 2025 HS test instead of the LEAP 2025 Mathematics grade 8 test.

## 7.1 Current Administration Data

Tables 7.2 through 7.13 show the percentage of students in each achievement level based on the state population for the 2019 administration of the ELA and mathematics assessments. Results from previous years are presented as well for comparison purposes.

**Table 7.2 Comparison of Percentage of Students in Achievement Levels: ELA Grade 3**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥56,800	13.4	17.8	24.7	38.9	5.1
2018	≥55,390	14.2	18.2	22.3	39.8	5.6
2019	≥52,940	13.2	17.2	23.7	39.5	6.4

**Table 7.3 Comparison of Percentage of Students in Achievement Levels: ELA Grade 4**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥56,230	8.8	18.3	29.3	36.2	7.3
2018	≥55,760	10.8	17.0	28.7	34.8	8.8
2019	≥54,800	10.3	18.1	26.6	36.1	8.9

**Table 7.4 Comparison of Percentage of Students in Achievement Levels: ELA Grade 5**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥53,300	8.7	18.8	31.1	37.9	3.4
2018	≥55,310	8.8	17.7	30.4	39.3	3.7
2019	≥54,910	8.4	21.1	30.0	36.0	4.4

**Table 7.5 Comparison of Percentage of Students in Achievement Levels: ELA Grade 6**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥52,370	10.4	24.9	29.8	29.4	5.5
2018	≥52,810	9.3	24.6	31.5	30.3	4.4
2019	≥54,800	9.2	23.5	29.8	32.2	5.3

**Table 7.6 Comparison of Percentage of Students in Achievement Levels: ELA Grade 7**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥51,930	13.2	19.2	26.5	30.3	10.8
2018	≥51,540	10.7	19.2	26.8	31.4	11.9
2019	≥52,350	11.6	16.7	25.1	33.0	13.7

**Table 7.7 Comparison of Percentage of Students in Achievement Levels: ELA Grade 8**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥50,450	11.4	17.4	27.0	35.1	9.0
2018	≥51,020	10.8	17.4	26.6	36.9	8.4
2019	≥50,720	11.7	16.2	25.4	37.6	9.2

**Table 7.8 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 3**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥56,800	11.1	18.4	27.1	36.2	7.1
2018	≥55,360	10.3	19.7	28.1	34.6	7.3
2019	≥52,820	9.7	20.6	26.4	36.5	6.7

**Table 7.9 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 4**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥56,230	8.2	23.2	29.7	35.0	3.8
2018	≥55,680	8.6	22.8	30.3	34.4	3.9
2019	≥54,690	11.1	20.5	27.1	38.0	3.3

**Table 7.10 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 5**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥53,310	11.1	24.9	32.4	27.7	3.9
2018	≥55,200	10.2	25.8	34.0	25.7	4.2
2019	≥54,730	10.3	26.8	28.3	30.5	4.1

**Table 7.11 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 6**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥52,350	12.6	30.8	29.2	23.7	3.7
2018	≥52,670	11.6	29.0	32.0	24.8	2.6
2019	≥54,710	11.4	26.7	31.7	26.6	3.6

**Table 7.12 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 7**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥51,800	11.2	28.9	35.2	22.6	2.1
2018	≥51,420	9.9	29.0	35.7	22.9	2.4
2019	≥52,090	9.1	29.5	34.7	24.5	2.3

**Table 7.13 Comparison of Percentage of Students in Achievement Levels: Mathematics Grade 8**

Year	N	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
2017	≥44,710	20.3	28.2	25.0	24.7	1.8
2018	≥44,910	20.9	27.4	23.7	26.1	1.9
2019	≥44,520	20.9	25.7	25.4	25.7	2.3

Score reports are the primary means of communicating test scores to appropriate school system personnel (e.g., testing coordinators or superintendents), teachers, and parents. Standard 6.10 of the *Standards* states:

When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (119)

Standard 5.1 is related to Standard 6.10. It states:

Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (102)

Interpretations of test scores are disseminated in two ways: the individual score report and the *LEAP 2025 Interpretive Guide* (2019).

In addition to providing interpretation of the test results, the LODE and DRC must ensure that the information is understandable for the target audience. Standard 7.0 states:

Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (125)

The LDOE and DRC strive to create documents that will be accessible to parents, teachers, and all other stakeholders.

The Individual Student-Level Report (ISR) is the primary means for sharing student test results with parents. As such, it is a stand-alone document from which parents can glean information that is relevant to understanding their children's test scores. For more information about the test, parents are provided [A Parent Guide to the LEAP 2025 Student Reports](#). In the 2019 administration year, student reports for each school were posted by grade, then downloaded and printed from eDIRECT by school systems and schools. eDIRECT is DRC's secure online system that provides schools and districts access to student tests and reports.

### 7.1.1 Description of Each Type of Report

In this section, descriptions of the School Roster Report and the ISR are provided.

In compliance with AERA, APA, & NCME (2014) Standard 12.18, the LEAP 2025 score reports provide clear information about the results of individual students and of specific groups of students. Standard 12.18 states:

In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (200)

#### **School Roster Report**

A School Roster Report, which provides summary information about student performance on the LEAP 2025 ELA and Mathematics tests, is available to school systems and schools through eDIRECT. Total test scores and achievement-level indicators are shown for the content area of interest. Reporting category and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by category and subcategory are summarized. More details can be found in the [LEAP 2025 Interpretive Guide](#).

#### **Individual Student-Level Report**

The ISR is another type of report available through the eDIRECT system. ISRs may be downloaded and printed by schools to be sent home to parents. At the top of the page, overall student performance is reported by scale scores and achievement level. To give context to the student score, the student's school system and state averages are presented to the right of the student information. In the middle of the page, category and subcategory performance indicators are reported. achievement-level descriptors and the percentage of students in each achievement level by school, school system, and the state, which allows comparisons of the student's overall achievement level to those of his or her peers, are found at the bottom of the page. When a student does not receive a scale score, his or her achievement level will be left blank. ISRs for students whose scores were invalidated will display a blank scale score for a given content area.

A data file referred to as Louisiana Department of Education Student File (LDESTD) was provided to the LDOE by DRC. It contains one record for every student tested; each record contains demographic information, responses for multiple-choice (MC) items, scores for items that are not MC items, raw scores, content and process standard raw scores, scale scores, and performance-level data for each content area.

The [LEAP 2025 Interpretive Guide](#) was written to help Louisiana school system and school administrators, teachers, parents, and the general public to better understand the LEAP 2025 ELA and mathematics tests. The *LEAP 2025 Interpretive Guide* was developed collaboratively by DRC and LDOE staff. LDOE staff had opportunities to review the guide, provide feedback, and give final approval.

The *LEAP 2025 Interpretive Guide* has three sections. The first section presents an introduction and an overview of key terms and test-related concepts. The second section discusses assessment terms and types of scores that are presented on the ISRs. Sample ISRs are included in the guide. The third section discusses information that is presented on the School Roster Report and an example of the report.

In summary, the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information developed by DRC address multiple best practices of the testing industry but are particularly related to the following standards:

**Standard 5.1** Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (102)

**Standard 6.10** When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (119)

**Standard 7.0** Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (125)

**Standard 12.18** In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (200)

## Chapter 8: Performance-Level Setting

This chapter briefly describes the LEAP 2025 performance-level setting and presents the cut scores and achievement-level descriptors derived from the performance-level setting. Since the LDOE uses PARCC cut scores for the LEAP 2025 ELA and Mathematics tests, a brief overview of the PARCC performance-level setting procedures is included in this chapter. A more detailed discussion and the results of the PARCC performance-level setting may be found in the *Performance Level Setting Technical Report* (Pearson, 2015).

The AERA, APA, & NCME (2014) Standards addressed by the *Performance Level Setting Technical Report* (Pearson, 2015) are 5.21 and 5.22.

Starting in the spring of 2015, the ELA and mathematics assessments measured different content and constructs than did previous tests were administered to Louisiana students. The new tests were built using the PARCC item bank and were fully aligned to the Louisiana Student Standards. The new tests were reported on new scales, and students were classified by achievement levels based on their knowledge and ability to perform different tasks in relation to the new test content and standards.

In terms of the validity of the LEAP 2025 test scores, it is essential to understand that descriptors and cut scores are established in a collaborative and participatory process. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores.

### 8.1 PARCC Performance-Level Setting Process for English Language Arts and Mathematics

According to the *Performance Level Setting Technical Report* (Pearson, 2015), PARCC used the evidence-based standard setting (EBSS) method (Beimers, Way, McClarty, & Miles, 2012) for the PARCC performance-level setting (PLS) process. The EBSS method is used to combine various considerations into the process for setting performance levels, including policy considerations, content standards, research, and educator judgment about what students should know and be able to demonstrate, and to support PARCC's policy goals related to college- and career-readiness expectations. Additional details about the EBSS method can be found in the *Performance Level Setting Technical Report* (Pearson, 2015).

### 8.2 Cut Scores

This section presents the cut scores for each grade and content area of the LEAP 2025. Tables 8.1 and 8.2 show the ELA and mathematics cut scores for students in grades 3 through 8.

**Table 8.1 English Language Arts Cut Scores**

Grade	Cut Scores			
	<i>Approaching Basic</i>	<i>Basic</i>	<i>Mastery</i>	<i>Advanced</i>
3	700	725	750	810
4	700	725	750	790
5	700	725	750	799
6	700	725	750	790
7	700	725	750	785
8	700	725	750	794



**Table 8.2 Mathematics Cut Scores**

Grade	Cut Scores			
	<i>Approaching Basic</i>	<i>Basic</i>	<i>Mastery</i>	<i>Advanced</i>
3	700	725	750	790
4	700	725	750	796
5	700	725	750	790
6	700	725	750	788
7	700	725	750	786
8	700	725	750	801

### 8.2.1 Reporting Category Cut Scores

As stated in Section 6.4.2.3, student performance on ELA and mathematics reporting categories and subcategories was classified into one of three performance ratings: *Strong*, *Moderate*, and *Weak*. Detailed rules for calculating performance ratings for ELA and mathematics reporting categories and subcategories can be found in that section.

The cut scores divide the continuum of student achievement into the following five achievement levels used by the LDOE for reporting purposes:

- *Advanced*: Students performing at this level have **exceeded** college- and career-readiness expectations and are well prepared for the next level of studies in this content area.
- *Mastery*: Students performing at this level have **met** college- and career-readiness expectations and are prepared for the next level of studies in this content area.
- *Basic*: Students performing at this level have **nearly met** college- and career-readiness expectations and may need additional support to be fully prepared for the next level of studies in this content area.
- *Approaching Basic*: Students performing at this level have **partially met** college- and career-readiness expectations and will need much support to be prepared for the next level of studies in this content area.
- *Unsatisfactory*: Students performing at this level have **not yet met** the college- and career-readiness expectations and will need extensive support to be prepared for the next level of studies in this content area.

Table 8.3 summarizes the LEAP 2025 ELA and mathematics scale score ranges for each level of achievement.

**Table 8.3 Achievement-Level Scale Score Ranges**

ELA						
Achievement Level	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
<i>Advanced</i>	810–850	790–850	799–850	790–850	785–850	794–850
<i>Mastery</i>	750–809	750–789	750–798	750–789	750–784	750–793
<i>Basic</i>	725–749					
<i>Approaching Basic</i>	700–724					
<i>Unsatisfactory</i>	650–699					
MATHEMATICS						
Achievement Level	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
<i>Advanced</i>	790–850	796–850	790–850	788–850	786–850	801–850
<i>Mastery</i>	750–789	750–795	750–789	750–787	750–785	750–800
<i>Basic</i>	725–749					
<i>Approaching Basic</i>	700–724					
<i>Unsatisfactory</i>	650–699					

This chapter presented a brief overview of PARCC’s performance-level setting process, which set the cut scores used by the LDOE for reporting student performance on the LEAP 2025 ELA and mathematics tests. These procedures are addressed in more detail in relevant technical reports.

The performance-level setting process undertaken by PARCC addresses the following standards:

**Standard 5.21** When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (107)

**Standard 5.22** When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way. (108)

## Chapter 9: Evidence of Validity

Evidence for validity—the meaning of test scores and the inferences they support—is the central concept underlying the LEAP 2025 validation process. Validity evidence, from the design of the test to item development and scoring, is created throughout the entire assessment process. Therefore, evidence of validity is described throughout the LEAP 2025 technical report. Table 9.1 summarizes the sources of evidence of validity and indicates where the evidence can be found in the technical report.

**Table 9.1 Summary of Evidence of Validity and the Report Chapter in Which it is Found**

Source of Validity	Related Information	Related Chapter/Source
Evidence Based on Test Content	Item Development Process	Chapter 3 2018–2019 LEAP Grades 3-8 ELA and Mathematics Assessment Frameworks
	Test Blueprint and Item Alignment to Curriculum and Standards	Chapter 3 2018–2019 LEAP Grades 3-8 ELA and Mathematics Assessment Frameworks
	Item Bias, Sensitivity, and Content Appropriateness	Chapter 3
	Accommodations	Chapters 3 and 4
Evidence Based on Response Processes	Data Review	2018–2019 LEAP Grades 3-8 ELA and Mathematics Assessment Frameworks
	Classical Item analysis	Chapter 6
Evidence Based on Internal Structure	Differential Item Functioning	Chapter 10
	Reliability and Standard Errors of Measurement	Chapter 9
Evidence Based on Relationships to Other Variables	Divergent Validity	Chapter 9
	Regression of LEAP 2025 from 2018 to 2019	Chapter 9
Evidence Based on the Consequences of Testing	Scale Score and Performance Level Information	Chapter 7
	Test Interpretive Guide	Chapter 4

In this chapter, DRC presents evidence of construct-related validity through studies of test reliability, convergent validity, and divergent validity. All analyses in this chapter are based on census data.

Chapter 9 of this report demonstrates adherence to the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014) Standards 1.13, 1.21, 2.0, 2.3, 2.13, 2.14, 2.16, and 2.19. Each standard is discussed in the pertinent section of this chapter.

## 9.1 Construct-Irrelevant Variance and Construct Underrepresentation

Minimization of construct-irrelevant variance and construct underrepresentation is addressed in the following steps of the test development process: (1) specification, (2) item writing, (3) review, (4) field testing, (5) test construction, and (6) item calibration (see Chapter 3 for more information on steps 1–5 and Chapter 6 for more information on step 6).

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs measured by the test. For example, when tests are not administered under standardized conditions (e.g., one administration may be timed, but another administration is untimed), differences in student performance related to different administration conditions may result. Careful specification of content and review of the items representing that content are first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data, is used to infer construct irrelevance.

Construct underrepresentation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. Specification and review, a process through which test blueprints are developed and reviewed, are primary steps in the development process designed to ensure that content is appropriately represented.

## 9.2 Reliability

Reliability refers to the consistency of students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Often, however, it is impractical to administer multiple forms of the test, and reliability is estimated on a single administration of the test. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across items within a test during a single test administration (Crocker & Algina, 1986). Reliability is a necessary, but not sufficient, condition of validity.

The 2014 *Standards* indicates the following:

The term *reliability* has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency). (33)

In accordance with the *Standards* in developing and maintaining tests of the highest quality, DRC has calculated the reliability of each LEAP 2025 test in a variety of ways: reliability of raw scores, overall standard error of measurement (SEM), IRT-based conditional SEM, and decision consistency of achievement-level classifications.

There are several specific standards that this chapter addresses. These include Standards 2.0, 2.3, 2.13, and 2.19, each of which is articulated below.

**Standard 2.0** Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use. (42)

**Standard 2.3** For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported. (43)

The total score reliabilities are discussed in Section 9.2.1 of this chapter. The SEMs and subscore reliabilities are presented in Sections 9.4.2 and 9.4.3. The SEM of the total score is discussed in Section 9.2.2.

**Standard 2.13** The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (45)

The SEM based on raw scores is discussed in Section 9.2.2 and is reported in raw score units. The conditional SEM is discussed in Section 9.2.3 and is presented in scale score units.

**Standard 2.19** Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported. (47)

Section 9.2 discusses different ways of measuring test reliability, including reliability of raw scores and test-form SEM, IRT-based conditional SEM, and decision consistency of achievement-level classifications. These statistics were computed based on the census data.

## 9.2.1 Test Reliability

The reliability of raw scores by test form was evaluated using Cronbach's (1951) coefficient alpha, which is a lower-bound estimate of test reliability. The reliability coefficient is a ratio of the variance of true test scores to the variance of the total observed scores, with the values ranging from 0 to 1. The closer the value of the reliability coefficient is to 1, the more consistent the scores, where 1 refers to a perfectly consistent test. In general, reliability coefficients that are equal to or greater than 0.8 are considered acceptable for tests of moderate lengths.

Cronbach's coefficient alpha was computed using the formula

$$\alpha = \frac{n}{n-1} \left[ 1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right], \quad (9.1)$$

where  $n$  is the number of items on the test,  $\sigma_i^2$  is the variance of item  $i$ , and  $\sigma_x^2$  is the variance of the total test score.

Total test reliability measures, such as Cronbach's coefficient alpha and SEM, consider the consistency (i.e., reliability) of performance over all test questions in a given form, the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations. The number of items in the test influences these statistics; for example, a longer test can be expected to be more reliable than a shorter test.

The reliability coefficients for the LEAP 2025 are reported in Table 9.2. These reliability coefficients were computed using the census data. The reliability statistics ranged from 0.86 to 0.92 for all ELA forms. The ELA forms have one writing component (RI or RL) that is the same score of another component (WE); the item score for the RI/RL component was excluded from the reliability computation. For mathematics, the reliabilities ranged from 0.92 to 0.94. These results indicate acceptable reliability coefficients for the LEAP 2025 tests.

**Table 9.2 Reliability in English Language Arts and Mathematics**

Content	Grade	Mode	Number of Items	Number of Score Points	SEM	Cronbach's Alpha	N-Count
ELA	3	CBT	26	71	4.09	0.86	≥1,530
ELA	3	PBT	26	71	4.58	0.87	≥51,410
ELA	4	CBT	28	86	4.90	0.90	≥7,570
ELA	4	PBT	28	86	5.26	0.89	≥47,220
ELA	5	CBT	28	86	4.99	0.90	≥54,910
ELA	6	CBT	32	90	5.21	0.91	≥54,800
ELA	7	CBT	32	90	5.51	0.92	≥52,350
ELA	8	CBT	32	94	5.64	0.90	≥50,720
Mathematics	3	CBT	43	62	3.53	0.92	≥1,520
Mathematics	3	PBT	43	62	3.82	0.92	≥51,300
Mathematics	4	CBT	43	62	3.52	0.93	≥7,540
Mathematics	4	PBT	43	62	3.65	0.93	≥47,140
Mathematics	5	CBT	41	60	3.59	0.92	≥54,730
Mathematics	6	CBT	42	65	3.65	0.93	≥54,710
Mathematics	7	CBT	43	66	3.89	0.92	≥52,090
Mathematics	8	CBT	41	65	3.52	0.92	≥44,520

The reliability statistics by subgroup are reported and discussed in Chapter 10.

### 9.2.2 Standard Error of Measurement

The reliability of reported test scores can be characterized by the standard errors associated with the scores. The SEM may be used to determine the range within which a student's true score is likely to fall. An observed score should be regarded not as a student's true score but as an estimate of a student's true score. It is expected that the score a student obtains from a single test administration would fall within one SEM of the student's true score 68% of the time and within approximately two SEMs of the true score 95% of the time. The SEM is an index of the random variability in test scores and is defined as follows:

$$SEM = SD\sqrt{1 - R_{xx'}}, \quad (9.2)$$

where SD represents standard deviation of the raw score distribution, and  $R_{xx'}$  is estimated by  $\hat{\alpha}$  as expressed in Equation 9.1.

The SEM at the test-form level was computed in raw score metric and is also presented in Table 9.2 for ELA and mathematics.

### 9.2.3 Conditional Standard Error of Measurement

In contrast to SEM, conditional standard error of measurement (CSEM) expresses the degree of measurement error in scale score units and is conditioned on the ability of the student. DRC reports the CSEM in support of Standard 2.14, which states:

When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (46)

In further compliance with Standard 2.14, the CSEM of each cut score is reported in Table 9.3.

The CSEMs are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985). The CSEM is defined in the following equation:

$$\text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}, \quad (9.3)$$

where  $I(\theta_i)$  is the test information function, a sum of item information function 2, obtained as

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)q_{ij}(\theta_i)}, \quad (9.4)$$

where  $p'_{ij}(\theta_i)$  is the derivative of  $p_{ij}(\theta_i)$  and  $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$ .

Note that the CSEMs vary in magnitude across the entire range of student ability estimates (i.e., scale scores) and are smaller in the middle of the score distribution and higher at the tails. This pattern is expected when IRT methods are used. The CSEMs at the four cut scores that define the performance levels are presented in Table 9.3.

**Table 9.3 Conditional Standard Errors of Measurement at the *Approaching Basic, Basic, Mastery, and Advanced* Cut Scores**

Content Area	Grade	Mode	<i>Approaching Basic</i>		<i>Basic</i>		<i>Mastery</i>		<i>Advanced</i>	
			Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
ELA	3	CBT	700	14	725	12	750	11	810	13
ELA	3	PBT	700	13	725	12	750	11	810	12
ELA	4	CBT	700	10	725	8	750	8	790	9
ELA	4	PBT	700	10	725	8	750	7	790	8
ELA	5	CBT	700	11	725	8	750	7	799	8
ELA	6	CBT	700	9	725	7	750	7	790	8
ELA	7	CBT	700	9	725	7	750	7	785	8
ELA	8	CBT	700	10	725	8	750	8	794	8
Mathematics	3	CBT	700	9	725	7	750	7	790	11
Mathematics	3	PBT	700	9	725	7	750	7	790	11
Mathematics	4	CBT	700	9	725	7	750	7	796	10
Mathematics	4	PBT	700	9	725	7	750	7	796	9
Mathematics	5	CBT	700	9	725	7	750	6	790	9
Mathematics	6	CBT	700	9	725	7	750	6	788	8
Mathematics	7	CBT	700	9	725	7	750	6	786	8
Mathematics	8	CBT	700	11	725	9	750	7	801	10

Figures 9.1 and 9.2 display the CSEM (conditional standard error of measurement) curves for each grade and content area by mode. Typically, with fixed-form assessments, the estimates of measurement error tend to be higher at the low and high ends of the scale-score range where few items measure those ability levels. Generally, there are few students with extreme scores, and these score levels cannot be estimated as accurately as levels toward the middle of the ability range. The middle ability range, where cut scores are located, shows lower measurement error than the low and high ends of the ability ranges. Figures 9.1 and 9.2 demonstrate that the tests are designed so that measurement error is minimized in the middle of the scale range, where most students are located.



Figure 9.1 CSEM Curves for ELA Grades 3 through 8

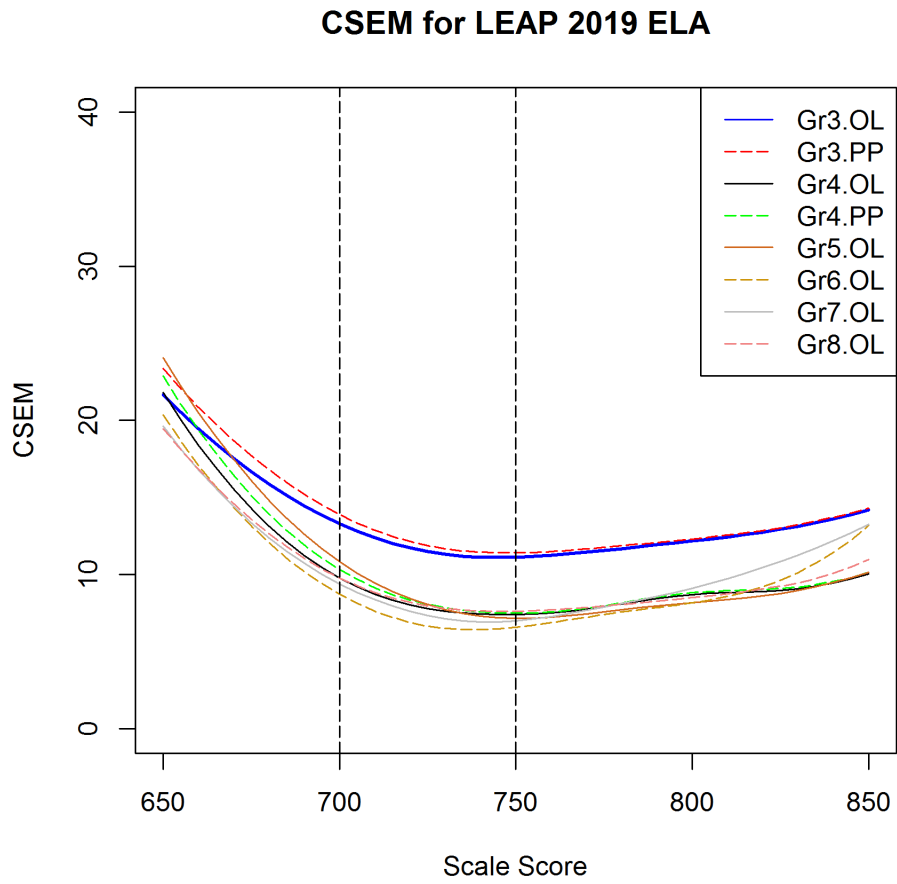
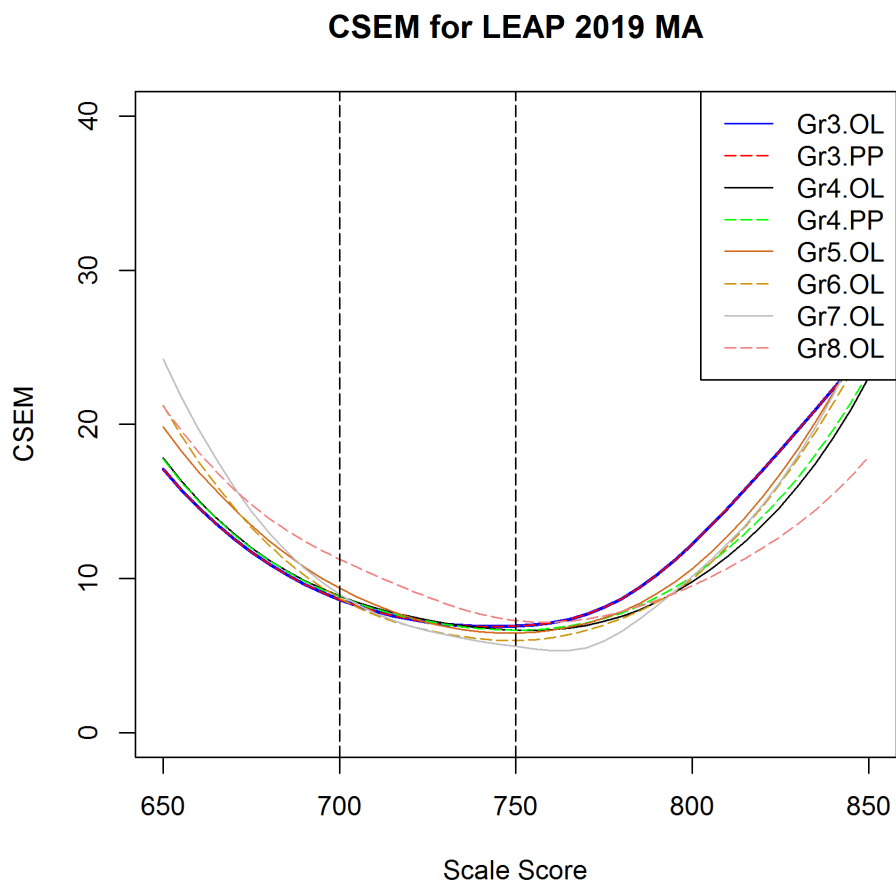


Figure 9.2 CSEM Curves for Mathematics Grades 3 through 8



## 9.2.4 Classification Accuracy and Consistency

### **Classification Accuracy**

Classification accuracy is defined as the extent to which the actual classifications of test takers into various achievement levels match classifications made based on their true scores (Livingston & Lewis, 1995). Classification accuracy refers to the agreement between the observed score and the true score, whereas classification consistency refers to the agreement between two observed scores.

### **Classification Consistency**

Classification consistency is defined as the extent to which the classifications of students in a particular achievement level match based on two independent administrations of the same test form or one administration of two parallel test forms. It is often logistically infeasible, as well as expensive, to obtain data from repeated administrations of a test, be it re-administration of the same test or administration of a parallel form. Therefore, a common practice is to estimate classification consistency from one administration of a test.

The Livingston-Lewis (1995) methodology was used to calculate classification accuracy statistics based on the spring 2019 LEAP 2025 results. The Livingston-Lewis procedure utilizes a beta-binomial model that requires two steps: (1) fitting proportion-correct true scores to a four-parameter beta distribution and (2) using the binomial distribution to estimate classification accuracy and consistency. All calculations for classification accuracy and consistency are based on census data.

Classification consistency and classification accuracy conditioned on achievement level (see Table 9.4 and Table 9.5) and on cut score (see Table 9.6 and Table 9.7) are presented for the 2019 LEAP 2025 in this section of the report. The magnitude of classification consistency and accuracy measures is influenced by several key features of the test design, including the number of items, the location and number of cut scores, the score distribution, and the reliability and associated SEM. As can be seen in Table 9.4, classification accuracy conditioned on achievement level ranged from 0.30 to 0.85 for ELA and 0.41 to 0.89 for mathematics. Classification consistency (see Table 9.5) conditioned on achievement level ranged from 0.36 to 0.76 for ELA and 0.43 to 0.83 for mathematics. Table 9.6 shows that classification accuracy at achievement cut points ranged from 0.88 to 0.99 for ELA and 0.90 to 0.99 for mathematics. Classification consistency (see Table 9.7) conditioned at achievement cut points ranged from 0.83 to 0.98 for ELA and 0.89 to 0.99 for mathematics. Classification consistency and accuracy at achievement cut points tend to be higher values than those conditioned on achievement level. For some tests, classification accuracy and consistency conditioned on the *Advanced* level were lower than 0.50. One reason for these relatively low *Advanced* level values is few highly difficult items to distinguish the *Advanced* level from other achievement levels.

**Table 9.4 Classification Accuracy Conditioned on Level of Achievement**

Content Area	Classification Accuracy						
	Grade	Mode	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
ELA	3	CBT	0.80	0.58	0.63	0.76	0.30
ELA	3	PBT	0.68	0.58	0.60	0.83	0.57
ELA	4	CBT	0.75	0.69	0.70	0.79	0.60
ELA	4	PBT	0.67	0.63	0.72	0.80	0.64
ELA	5	CBT	0.59	0.64	0.72	0.85	0.58
ELA	6	CBT	0.62	0.75	0.75	0.81	0.61
ELA	7	CBT	0.76	0.66	0.73	0.78	0.74
ELA	8	CBT	0.70	0.63	0.70	0.81	0.69
Mathematics	3	CBT	0.80	0.76	0.74	0.82	0.45
Mathematics	3	PBT	0.73	0.74	0.73	0.85	0.57
Mathematics	4	CBT	0.75	0.74	0.75	0.87	0.63
Mathematics	4	PBT	0.73	0.72	0.74	0.89	0.41
Mathematics	5	CBT	0.59	0.72	0.75	0.86	0.60
Mathematics	6	CBT	0.72	0.74	0.80	0.84	0.64
Mathematics	7	CBT	0.41	0.77	0.76	0.84	0.67
Mathematics	8	CBT	0.78	0.64	0.69	0.85	0.69

**Table 9.5 Classification Consistency Conditioned on Level of Achievement**

Content Area	Classification Consistency						
	Grade	Mode	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
ELA	3	CBT	0.74	0.44	0.48	0.68	0.36
ELA	3	PBT	0.64	0.43	0.48	0.72	0.54
ELA	4	CBT	0.71	0.54	0.58	0.70	0.57
ELA	4	PBT	0.63	0.53	0.56	0.71	0.60
ELA	5	CBT	0.53	0.52	0.60	0.76	0.53
ELA	6	CBT	0.62	0.61	0.63	0.74	0.57
ELA	7	CBT	0.71	0.55	0.60	0.68	0.69
ELA	8	CBT	0.66	0.48	0.56	0.73	0.65
Mathematics	3	CBT	0.73	0.64	0.64	0.76	0.44
Mathematics	3	PBT	0.68	0.62	0.62	0.77	0.51
Mathematics	4	CBT	0.67	0.63	0.64	0.81	0.57
Mathematics	4	PBT	0.70	0.58	0.64	0.83	0.43
Mathematics	5	CBT	0.55	0.60	0.64	0.79	0.58
Mathematics	6	CBT	0.66	0.64	0.69	0.78	0.65
Mathematics	7	CBT	0.44	0.61	0.66	0.79	0.64
Mathematics	8	CBT	0.70	0.50	0.58	0.79	0.65

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points, the joint distribution of all the achievement levels is collapsed into a dichotomized distribution around that specific cut point. As an example, for the LEAP 2025 assessments, a dichotomization at the cut point between the *Basic* and *Mastery* classifications was formed. The proportion of correct classifications below this particular cut point is equal to the sum of all the cells at the *Unsatisfactory*, *Approaching Basic*, and *Basic* levels, and the proportion of correct classifications above that particular cut point is equal to the sum of all the cells at the *Mastery* and *Advanced* levels. Table 9.6 shows the classification accuracy and Table 9.7 shows the consistency estimates when conditioned on LEAP 2025 cut scores. The classification accuracy statistics are at or above 0.88, while the classification consistency statistics are at or above 0.83. These results suggest that consistent and accurate achievement-level classifications are being made for students in Louisiana based on the LEAP 2025.

**Table 9.6 Classification Accuracy at Achievement Cut Points**

Content Area	Grade	Mode	Classification Accuracy			
			<i>Unsatisfactory/ Approaching Basic</i>	<i>Approaching Basic/ Basic</i>	<i>Basic/ Mastery</i>	<i>Mastery/ Advanced</i>
ELA	3	CBT	0.88	0.89	0.93	0.99
ELA	3	PBT	0.94	0.90	0.89	0.96
ELA	4	CBT	0.94	0.91	0.91	0.96
ELA	4	PBT	0.95	0.92	0.90	0.95
ELA	5	CBT	0.94	0.91	0.91	0.97
ELA	6	CBT	0.95	0.92	0.92	0.97
ELA	7	CBT	0.95	0.93	0.92	0.94
ELA	8	CBT	0.95	0.92	0.91	0.95
Mathematics	3	CBT	0.91	0.92	0.95	0.99
Mathematics	3	PBT	0.96	0.93	0.92	0.95
Mathematics	4	CBT	0.94	0.93	0.93	0.98
Mathematics	4	PBT	0.95	0.93	0.93	0.97
Mathematics	5	CBT	0.93	0.92	0.93	0.98
Mathematics	6	CBT	0.94	0.92	0.94	0.98
Mathematics	7	CBT	0.92	0.90	0.93	0.99
Mathematics	8	CBT	0.90	0.91	0.94	0.99

**Table 9.7 Classification Consistency at Achievement Cut Points**

Content Area	Grade	Mode	Classification Consistency			
			Unsatisfactory/ Approaching Basic	Approaching Basic/ Basic	Basic/ Mastery	Mastery/ Advanced
ELA	3	CBT	0.83	0.85	0.90	0.98
ELA	3	PBT	0.91	0.87	0.85	0.94
ELA	4	CBT	0.91	0.88	0.88	0.95
ELA	4	PBT	0.93	0.89	0.87	0.92
ELA	5	CBT	0.92	0.87	0.87	0.96
ELA	6	CBT	0.93	0.88	0.88	0.95
ELA	7	CBT	0.93	0.90	0.88	0.91
ELA	8	CBT	0.92	0.89	0.87	0.93
Mathematics	3	CBT	0.87	0.89	0.93	0.99
Mathematics	3	PBT	0.94	0.90	0.89	0.94
Mathematics	4	CBT	0.92	0.89	0.91	0.97
Mathematics	4	PBT	0.93	0.90	0.90	0.97
Mathematics	5	CBT	0.90	0.88	0.91	0.97
Mathematics	6	CBT	0.92	0.89	0.91	0.97
Mathematics	7	CBT	0.89	0.87	0.91	0.98
Mathematics	8	CBT	0.86	0.87	0.91	0.98

### 9.2.1 Convergent Validity

Convergent validity is a subtype of construct validity that can be estimated by the extent to which measures of constructs that theoretically should be related to each other are, in fact, observed as related to each other. Analyses of the internal structure of a test can indicate the extent to which the relationships among test items conform to the construct the test purports to measure. For example, the LEAP 2025 mathematics test is designed to measure a single overall construct—mathematics achievement; therefore, the items comprising the LEAP 2025 mathematics test should measure only mathematics, not language or reading.

This technical report summarizes additional statistics that contribute to construct validity (Cronbach’s coefficient alpha is reported previously in this section, and item fit is reported in Chapter 6). The internal consistency coefficient (i.e., Cronbach’s alpha) reported is typically measured via correlations among the test items and indicates of the degree of the same general construct (Pearson, 2015, page 128). Table 9.2 shows test reliability statistics for ELA and mathematics. The reliability statistics ranged from 0.86 to 0.92 for ELA forms and from 0.92 to 0.94 for mathematics forms, indicating that items on the 2019 LEAP 2025 assessments are homogenous. For a group of items to be homogeneous, the items must measure the same construct (i.e., construct validity) or represent the same content domain (i.e., content validity). Because IRT models were used to calibrate test items and to report student scores, item fit is also relevant to construct validity. The extent to which test items function as the IRT model prescribes is relevant to the validation of test scores. As shown in Chapter 6, no items were flagged for poor model/data fit.

### 9.3 Principal Components Analysis

As another measure of construct validity, DRC examined the unidimensionality of each grade-level LEAP 2025 test. One of the underlying assumptions of the IRT models used to scale the LEAP 2025 tests is that the tests

being calibrated are unidimensional; that is, items in each grade and content area measure a single content domain. For example, mathematics items should measure mathematics ability and not reading skills. Standard 1.13 of the *Standards* states:

If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided. (26–27)

This section examines the internal structure of the LEAP 2025 tests by evaluating the unidimensionality assumption through principal components analysis (PCA). This analysis seeks evidence that there exists a single primary factor, the first principal component, which accounts for much of the relationship between items. The presence of a single or dominant factor suggests that a test is sufficiently unidimensional (i.e., that it measures one underlying construct).

A PCA was conducted for each grade, content area, and mode of the LEAP 2025 assessments. A large first principal component is evident in each analysis. It is common to have additional eigenvalues greater than 1.0, which may suggest the presence of other factors.

For all grades, content areas, and modes of the LEAP 2025 assessments, the ratio of variance accounted for by the first factor to variance accounted for by the second is sufficiently large to indicate that the unidimensionality assumption holds. All the LEAP 2025 content-area tests exhibit first principal components accounting for more than 20% of the test variance for ELA (see Table 9.8) and for mathematics (see Table 9.9). To further investigate the unidimensionality of the ELA and mathematics assessments, the ratio of the first eigenvalue to the second eigenvalue was found (see Tables 9.8 and 9.9). These ratios show that the first eigenvalue is at least four times as large as the second eigenvalue for all the grades, content areas, and modes. This substantial difference in magnitude indicates that one factor appears to be dominant and that the ELA and mathematics tests are essentially unidimensional.

This evidence supports the claim that there is a dominant dimension underlying the items and tasks in each test and that scores from each test represent performance primarily determined by that ability. Construct-irrelevant variance, such as factual knowledge irrelevant to doing well in a subject, does not appear to create significant nuisance factors.

**Table 9.8 Principal Component Analysis for English Language Arts**

<b>Grade</b>	<b>Mode</b>	<b>Components</b>	<b>Eigenvalue</b>	<b>Percentage of Variance Explained</b>	<b>Cumulative Percentage of Variance Explained</b>
3	CBT	First Component	6.19	23.80	23.80
3	CBT	Second Component	1.19	4.58	28.38
3	CBT	Ratio (First/Second)	5.19		
3	PBT	First Component	6.40	24.62	24.62
3	PBT	Second Component	1.12	4.29	28.91
3	PBT	Ratio (First/Second)	5.74		
4	CBT	First Component	7.96	28.41	28.41
4	CBT	Second Component	1.29	4.62	33.04
4	CBT	Ratio (First/Second)	6.15		
4	PBT	First Component	7.46	26.66	26.66
4	PBT	Second Component	1.33	4.73	31.39
4	PBT	Ratio (First/Second)	5.63		
5	CBT	First Component	7.76	27.73	27.73
5	CBT	Second Component	1.33	4.75	32.47
5	CBT	Ratio (First/Second)	5.84		
6	CBT	First Component	8.71	27.21	27.21
6	CBT	Second Component	1.42	4.43	31.63
6	CBT	Ratio (First/Second)	6.15		
7	CBT	First Component	9.32	29.14	29.14
7	CBT	Second Component	1.24	3.87	33.00
7	CBT	Ratio (First/Second)	7.53		
8	CBT	First Component	8.25	25.77	25.77
8	CBT	Second Component	1.30	4.05	29.82
8	CBT	Ratio (First/Second)	6.36		



**Table 9.9 Principal Component Analysis for Mathematics**

Grade	Mode	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
3	CBT	First Component	10.69	24.87	24.87
3	CBT	Second Component	1.47	3.42	28.29
3	CBT	Ratio (First/Second)	7.27		
3	PBT	First Component	11.75	27.32	27.32
3	PBT	Second Component	1.39	3.24	30.56
3	PBT	Ratio (First/Second)	8.43		
4	CBT	First Component	12.72	29.58	29.58
4	CBT	Second Component	1.59	3.69	33.27
4	CBT	Ratio (First/Second)	8.01		
4	PBT	First Component	12.20	28.36	28.36
4	PBT	Second Component	1.56	3.62	31.98
4	PBT	Ratio (First/Second)	7.84		
5	CBT	First Component	11.09	27.04	27.04
5	CBT	Second Component	1.47	3.60	30.64
5	CBT	Ratio (First/Second)	7.52		
6	CBT	First Component	12.19	29.02	29.02
6	CBT	Second Component	1.50	3.58	32.60
6	CBT	Ratio (First/Second)	8.11		
7	CBT	First Component	10.74	24.97	24.97
7	CBT	Second Component	1.77	4.11	29.08
7	CBT	Ratio (First/Second)	6.07		
8	CBT	First Component	10.87	26.52	26.52
8	CBT	Second Component	1.35	3.28	29.81
8	CBT	Ratio (First/Second)	8.08		

## 9.4 Analyses by Reporting Categories and Subcategories

Three sets of analyses were conducted at the reporting category and subcategory levels for ELA and mathematics in another attempt to assess the construct validity of the LEAP 2025 assessments. First, correlation coefficients that measure the relationship between the reporting category scores and subcategory scores in both subjects were computed. Second, the reliability of each reporting category and subcategory was computed. Finally, the SEM was computed for each reportable category and subcategory.

### 9.4.1 Correlations among Reporting Categories and Subcategories

This section reports the strength of the interrelationships among the categories or subcategories by computing the correlation between them. Tables 9.10–9.13 report the uncorrected Pearson product-moment (PPM) correlation coefficients, the PPM corrected for attenuation (CAPPMM), and the reliability coefficients described above. The PPM among the categories and subcategories is presented below the diagonal portion

of the matrix, the CAPPM is presented above the diagonal portion of the matrix, and the reliability coefficients used are shown in Tables 9.10–9.13.

The uncorrected PPM in Tables 9.10–9.13 should be interpreted in the context of the reliability coefficient. In general, lower PPM coefficients are expected between variables that are less reliable. In most cases, the PPM coefficients show that performance on one category or subcategory is moderately to strongly related to performance on another category or subcategory within the same grade and content area. The value of the correlation coefficients will be affected by the limited number of items measuring each category or subcategory. Therefore, caution should be used when comparing the PPM coefficients that measure the relationships between categories or subcategories to those that measure the relationships between content areas. A more modest relationship (i.e., smaller correlation coefficients) is expected to be reported between the categories and subcategories as a consequence of the lower number of items measuring each of the reporting categories. The PPM between two category or subcategory scores may be artificially low because of measurement error.

Standard 1.21 states:

When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates. (29)

The attenuation of the PPM can be corrected statistically using Spearman’s formula:

$$CAPPM = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}, \quad (9.5)$$

where  $r_{xy}$  is the PPM between two claims or GLE strands,  $r_{xx}$  is the reliability of one of those claims or GLE strands, and  $r_{yy}$  is the reliability of the other claim or GLE strand.

ELA shows moderate relationships between the reading and writing reporting categories across all grades, indicating that these two categories measure some different traits. Across all tables, the CAPPM indicates moderate or strong relationships between subcategories. The CAPPM for reading vocabulary, written expression, and knowledge and use of language are moderate. In some cases, the CAPPM is greater than 1.0. “Disattenuated values greater than 1.00 indicate that measurement error is not randomly distributed” (Schumacker, 1996). The moderate or strong relationships suggested by the CAPPM in Tables 9.10–9.13 are further evidence of the validity of the test construct. Since the overall content area is comprised of the category or subcategories subscores and the content area is expected to measure a single dimension, these subscores are expected to be moderately or highly related.

**Table 9.10 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Category: English Language Arts**

Grade	Mode	No.	Category	N Items	1	2
3	CBT	1	Reading	22		0.82
	CBT	2	Writing	4	0.65	
	PBT	1	Reading	22		0.86
	PBT	2	Writing	4	0.69	
4	CBT	1	Reading	24		0.86
	CBT	2	Writing	4	0.76	
	PBT	1	Reading	24		0.86
	PBT	2	Writing	4	0.76	
5	CBT	1	Reading	24		0.84
	CBT	2	Writing	4	0.73	
6	CBT	1	Reading	28		0.80
	CBT	2	Writing	4	0.71	
7	CBT	1	Reading	28		0.86
	CBT	2	Writing	4	0.77	
8	CBT	1	Reading	28		0.86
	CBT	2	Writing	4	0.76	

**Table 9.11 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Subcategories: English Language Arts**

Subcategory Uncorrected and Corrected Correlation Coefficients: English Language Arts									
Grade	Mode	No.	Subcategory	N Items	1	2	3	4	5
3	CBT	1	Reading Literary Text	11	.	0.96	0.91	0.95	0.84
	CBT	2	Reading Information Text	7	0.62	.	0.96	1.07	0.84
	CBT	3	Reading Vocabulary	4	0.56	0.52	.	0.88	0.75
	CBT	4	Written Expression	4	0.55	0.54	0.42	.	1.29
	CBT	5	Knowledge & Use of Language	2	0.57	0.50	0.42	0.68	.
	PBT	1	Reading Literary Text	11	.	1.00	0.85	0.93	0.81
	PBT	2	Reading Information Text	7	0.65	.	0.92	1.16	0.91
	PBT	3	Reading Vocabulary	4	0.52	0.48	.	0.84	0.73
	PBT	4	Written Expression	4	0.58	0.61	0.42	.	1.22
	PBT	5	Knowledge & Use of Language	2	0.56	0.54	0.41	0.69	.
4	CBT	1	Reading Literary Text	7	.	1.05	1.03	1.07	1.03
	CBT	2	Reading Information Text	9	0.69	.	1.02	0.99	0.96
	CBT	3	Reading Vocabulary	8	0.69	0.73	.	0.87	0.85
	CBT	4	Written Expression	4	0.71	0.70	0.63	.	1.31
	CBT	5	Knowledge & Use of Language	2	0.70	0.69	0.63	0.96	.
	PBT	1	Reading Literary Text	7	.	1.06	1.04	1.07	1.04
	PBT	2	Reading Information Text	9	0.67	.	1.03	1.00	0.98
	PBT	3	Reading Vocabulary	8	0.68	0.71	.	0.89	0.88
	PBT	4	Written Expression	4	0.70	0.68	0.64	.	1.36
	PBT	5	Knowledge & Use of Language	2	0.68	0.68	0.63	0.97	.
5	CBT	1	Reading Literary Text	8	.	1.01	0.99	0.96	0.92
	CBT	2	Reading Information Text	10	0.72	.	0.98	1.00	0.96
	CBT	3	Reading Vocabulary	6	0.64	0.66	.	0.86	0.83
	CBT	4	Written Expression	4	0.66	0.70	0.55	.	1.38
	CBT	5	Knowledge & Use of Language	2	0.64	0.69	0.55	0.95	.
6	CBT	1	Reading Literary Text	9	.	0.99	1.00	0.79	0.76
	CBT	2	Reading Information Text	13	0.73	.	1.03	0.91	0.87
	CBT	3	Reading Vocabulary	6	0.64	0.71	.	0.81	0.79
	CBT	4	Written Expression	4	0.57	0.71	0.55	.	1.16
	CBT	5	Knowledge & Use of Language	2	0.57	0.70	0.56	0.91	.
7	CBT	1	Reading Literary Text	10	.	0.99	0.96	0.89	0.85
	CBT	2	Reading Information Text	13	0.76	.	0.97	0.98	0.94
	CBT	3	Reading Vocabulary	5	0.64	0.65	.	0.88	0.85
	CBT	4	Written Expression	4	0.67	0.74	0.58	.	1.20
	CBT	5	Knowledge & Use of Language	2	0.67	0.74	0.58	0.92	.
8	CBT	1	Reading Literary Text	7	.	1.05	1.05	1.03	1.00
	CBT	2	Reading Information Text	13	0.72	.	1.01	0.92	0.91
	CBT	3	Reading Vocabulary	8	0.66	0.70	.	0.80	0.80
	CBT	4	Written Expression	4	0.72	0.71	0.57	.	1.16
	CBT	5	Knowledge & Use of Language	2	0.71	0.71	0.57	0.92	.

**Table 9.12 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Categories: Mathematics**

Grade	Mode	No.	Category	N Items	1	2	3	4
3	CBT	1	Major Content	27	.	1.00	0.96	0.96
	CBT	2	Additional & Supporting Con	10	0.72	.	0.98	1.02
	CBT	3	Expressing Mathematical Rea	3	0.72	0.61	.	1.04
	CBT	4	Modeling & Application	3	0.76	0.66	0.71	.
	PBT	1	Major Content	27	.	1.00	0.99	1.00
	PBT	2	Additional & Supporting Con	10	0.78	.	1.01	1.03
	PBT	3	Expressing Mathematical Rea	3	0.71	0.63	.	1.05
	PBT	4	Modeling & Application	3	0.81	0.72	0.68	.
4	CBT	1	Major Content	29	.	0.98	0.95	0.92
	CBT	2	Additional & Supporting Con	8	0.78	.	0.96	0.94
	CBT	3	Expressing Mathematical Rea	3	0.78	0.69	.	0.98
	CBT	4	Modeling & Application	3	0.74	0.66	0.71	.
	PBT	1	Major Content	29	.	0.97	0.96	0.94
	PBT	2	Additional & Supporting Con	8	0.77	.	0.95	0.94
	PBT	3	Expressing Mathematical Rea	3	0.79	0.69	.	0.94
	PBT	4	Modeling & Application	3	0.73	0.65	0.67	.
5	CBT	1	Major Content	26	.	0.97	0.99	0.93
	CBT	2	Additional & Supporting Con	9	0.77	.	0.98	0.95
	CBT	3	Expressing Mathematical Rea	3	0.77	0.69	.	1.00
	CBT	4	Modeling & Application	3	0.73	0.67	0.70	.
6	CBT	1	Major Content	27	.	1.00	0.95	0.95
	CBT	2	Additional & Supporting Con	8	0.76	.	0.97	0.95
	CBT	3	Expressing Mathematical Rea	4	0.78	0.67	.	1.01
	CBT	4	Modeling & Application	3	0.74	0.63	0.71	.
7	CBT	1	Major Content	27	.	1.00	0.97	0.93
	CBT	2	Additional & Supporting Con	9	0.69	.	0.98	0.95
	CBT	3	Expressing Mathematical Rea	4	0.77	0.60	.	1.01
	CBT	4	Modeling & Application	3	0.72	0.57	0.70	.
8	CBT	1	Major Content	26	.	1.02	0.99	0.94
	CBT	2	Additional & Supporting Con	8	0.79	.	0.99	0.96
	CBT	3	Expressing Mathematical Rea	4	0.75	0.67	.	0.95
	CBT	4	Modeling & Application	3	0.71	0.65	0.62	.

**Table 9.13 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Subcategories: Mathematics**

Grade	Mode	No.	Subcategory	N Items	1	2	3	4
3	CBT	1	A1	9	.	0.89	0.92	0.99
	CBT	2	A2	3	0.58	.	0.88	0.89
	CBT	3	A3	7	0.63	0.51	.	0.89
	CBT	4	A4	8	0.71	0.54	0.57	.
	PBT	1	A1	9	.	0.94	0.89	1.00
	PBT	2	A2	3	0.63	.	0.91	0.97
	PBT	3	A3	7	0.65	0.58	.	0.93
	PBT	4	A4	8	0.72	0.62	0.64	.
4	CBT	1	A1	7	.	0.89	0.96	.
	CBT	2	A2	7	0.65	.	0.83	.
	CBT	3	A3	8	0.70	0.61	.	.
	PBT	1	A1	7	.	0.89	0.97	.
	PBT	2	A2	7	0.64	.	0.84	.
	PBT	3	A3	8	0.69	0.60	.	.
5	CBT	1	A1	5	.	0.95	1.07	0.95
	CBT	2	A2	6	0.55	.	1.00	0.92
	CBT	3	A3	6	0.61	0.62	.	0.93
	CBT	4	A4	7	0.58	0.62	0.61	.
6	CBT	1	A1	8	.	0.95	0.92	.
	CBT	2	A2	7	0.68	.	0.97	.
	CBT	3	A3	12	0.70	0.75	.	.
7	CBT	1	A1	8	.	1.01	1.05	.
	CBT	2	A2	15	0.74	.	1.05	.
	CBT	3	A3	4	0.68	0.73	.	.
8	CBT	1	A1	5	.	0.99	1.00	0.91
	CBT	2	A2	8	0.53	.	1.00	0.96
	CBT	3	A3	5	0.52	0.58	.	0.96
	CBT	4	A4	8	0.55	0.65	0.63	.

### 9.4.2 Reliability of Reporting Categories and Subcategories

Raw score summary statistics (i.e., mean and standard deviation), Cronbach's (1951) coefficient alpha, and SEM were computed for each of the reporting categories or subcategories by grade, content area, and mode using the census data. These statistics are presented in Tables 9.14–9.17 for ELA and mathematics. Reliability indices, such as Cronbach's coefficient alpha (and resulting SEM), are a function of the number of items on a test, the average covariance between item-pairs, and the variance of a test's total score. In general, it is expected that the coefficient alpha would be lower for a reporting category or subcategory assessed by a small number of items than for one assessed by a larger number of items.

### 9.4.3 Standard Error of Measurement of Reporting Categories and Subcategories

This chapter also reports the SEM associated with each of the reporting categories and subcategories in Tables 9.14–9.17 for ELA and mathematics. In these tables the RI/RL writing component was included. These SEMs are reported in the raw score metric.

**Table 9.14 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English Language Arts Reporting Categories**

Grade	Mode	Category	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
3	CBT	Reading	23	47	15.99	8.23	3.43	0.83
	CBT	Writing	4	24	3.47	3.50	1.74	0.75
	PBT	Reading	23	47	20.22	9.17	3.75	0.83
	PBT	Writing	4	24	6.52	4.31	2.07	0.77
4	CBT	Reading	26	56	21.88	10.80	3.95	0.87
	CBT	Writing	4	30	6.89	5.56	1.74	0.90
	PBT	Reading	26	56	23.52	11.01	4.22	0.85
	PBT	Writing	4	30	8.79	5.84	1.86	0.90
5	CBT	Reading	26	56	22.97	10.88	4.01	0.86
	CBT	Writing	4	30	6.44	5.55	1.86	0.89
6	CBT	Reading	29	60	25.99	11.69	4.04	0.88
	CBT	Writing	4	30	9.61	6.85	2.04	0.91
7	CBT	Reading	29	60	28.38	12.29	4.12	0.89
	CBT	Writing	4	30	11.62	7.96	2.48	0.90
8	CBT	Reading	30	64	27.95	12.15	4.45	0.87
	CBT	Writing	4	30	10.36	6.79	1.98	0.92

**Table 9.15 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English Language Arts Reporting Subcategories**

Mean, Standard Deviation, and SEM: English Language Arts								
Grade	Mode	Subcategory	Number of Items	Number of Score Pts.	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
3	CBT	Reading Literary Text	11	22	7.12	4.36	2.25	0.73
	CBT	Reading Information Text	8	17	5.42	3.15	2.07	0.57
	CBT	Reading Vocabulary	4	8	3.45	2.11	1.47	0.52
	CBT	Written Expression	2	18	2.56	2.72	2.01	0.45
	CBT	Knowledge & Use of Language	2	6	0.90	1.02	0.63	0.62
	PBT	Reading Literary Text	11	22	9.49	5.07	2.46	0.77
	PBT	Reading Information Text	8	17	6.22	3.52	2.35	0.55
	PBT	Reading Vocabulary	4	8	4.50	2.11	1.49	0.50
	PBT	Written Expression	2	18	4.80	3.39	2.38	0.50
	PBT	Knowledge & Use of Language	2	6	1.72	1.21	0.74	0.63
4	CBT	Reading Literary Text	8	18	6.42	3.47	2.16	0.61
	CBT	Reading Information Text	10	22	7.54	4.45	2.43	0.70
	CBT	Reading Vocabulary	8	16	7.92	4.13	2.14	0.73
	CBT	Written Expression	2	24	5.16	4.18	2.22	0.72
	CBT	Knowledge & Use of Language	2	6	1.73	1.42	0.71	0.75
	PBT	Reading Literary Text	8	18	6.45	3.76	2.38	0.60
	PBT	Reading Information Text	10	22	8.54	4.55	2.63	0.67
	PBT	Reading Vocabulary	8	16	8.53	4.07	2.17	0.72
	PBT	Written Expression	2	24	6.55	4.42	2.39	0.71
	PBT	Knowledge & Use of Language	2	6	2.25	1.46	0.77	0.72
5	CBT	Reading Literary Text	9	20	7.63	4.11	2.29	0.69
	CBT	Reading Information Text	11	24	9.13	5.10	2.60	0.74
	CBT	Reading Vocabulary	6	12	6.22	3.01	1.88	0.61
	CBT	Written Expression	2	24	4.75	4.18	2.39	0.67
	CBT	Knowledge & Use of Language	2	6	1.69	1.42	0.76	0.71
6	CBT	Reading Literary Text	9	18	7.05	3.77	2.11	0.69
	CBT	Reading Information Text	14	30	13.21	6.18	2.85	0.79
	CBT	Reading Vocabulary	6	12	5.74	3.03	1.90	0.61
	CBT	Written Expression	2	24	7.16	5.28	2.57	0.76
	CBT	Knowledge & Use of Language	2	6	2.46	1.68	0.72	0.82
7	CBT	Reading Literary Text	10	20	8.97	4.70	2.25	0.77
	CBT	Reading Information Text	14	30	12.99	6.27	2.97	0.78
	CBT	Reading Vocabulary	5	10	6.42	2.65	1.71	0.59
	CBT	Written Expression	2	24	8.83	6.22	3.22	0.73
	CBT	Knowledge & Use of Language	2	6	2.79	1.86	0.82	0.80
8	CBT	Reading Literary Text	8	18	7.66	3.86	2.36	0.63
	CBT	Reading Information Text	14	30	12.57	6.12	3.01	0.76
	CBT	Reading Vocabulary	8	16	7.72	3.54	2.12	0.64
	CBT	Written Expression	2	24	7.69	5.17	2.36	0.79
	CBT	Knowledge & Use of Language	2	6	2.67	1.73	0.78	0.80



**Table 9.16 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Mathematics Reporting Categories**

Mean, Standard Deviation, and SEM: Mathematics								
Grade	Mode	Category	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
3	CBT	Major Content	27	30	13.81	6.80	2.37	0.88
	CBT	Additional & Supporting Content	10	10	4.73	2.07	1.32	0.59
	CBT	Expressing Mathematical Reasoning	3	10	2.25	2.10	1.26	0.64
	CBT	Modeling & Application	3	12	2.76	2.77	1.47	0.72
	PBT	Major Content	27	30	17.69	7.12	2.32	0.89
	PBT	Additional & Supporting Content	10	10	5.89	2.28	1.30	0.67
	PBT	Expressing Mathematical Reasoning	3	10	3.68	2.27	1.47	0.58
	PBT	Modeling & Application	3	12	5.04	3.51	1.85	0.72
4	CBT	Major Content	29	30	16.76	7.38	2.23	0.91
	CBT	Additional & Supporting Content	8	10	4.92	2.45	1.35	0.70
	CBT	Expressing Mathematical Reasoning	3	10	2.63	2.31	1.16	0.75
	CBT	Modeling & Application	3	12	2.72	3.01	1.63	0.71
	PBT	Major Content	29	30	17.53	7.11	2.23	0.90
	PBT	Additional & Supporting Content	8	10	5.13	2.43	1.35	0.69
	PBT	Expressing Mathematical Reasoning	3	10	3.38	2.55	1.26	0.76
	PBT	Modeling & Application	3	12	3.35	3.10	1.76	0.68
5	CBT	Major Content	26	28	14.58	6.49	2.28	0.88
	CBT	Additional & Supporting Content	9	10	5.33	2.66	1.41	0.72
	CBT	Expressing Mathematical Reasoning	3	10	3.47	2.59	1.42	0.70
	CBT	Modeling & Application	3	12	2.98	2.71	1.49	0.70
6	CBT	Major Content	27	30	14.90	7.40	2.33	0.90
	CBT	Additional & Supporting Content	8	9	3.73	2.22	1.31	0.65
	CBT	Expressing Mathematical Reasoning	4	14	3.92	3.16	1.60	0.74
	CBT	Modeling & Application	3	12	2.62	2.66	1.53	0.67
7	CBT	Major Content	27	30	13.86	6.86	2.38	0.88
	CBT	Additional & Supporting Content	9	10	4.34	2.11	1.45	0.53
	CBT	Expressing Mathematical Reasoning	4	14	2.99	3.07	1.64	0.71
	CBT	Modeling & Application	3	12	1.90	3.04	1.71	0.68
8	CBT	Major Content	26	29	11.57	6.19	2.28	0.86
	CBT	Additional & Supporting Content	8	10	4.52	2.48	1.35	0.70
	CBT	Expressing Mathematical Reasoning	4	14	2.59	2.67	1.56	0.66
	CBT	Modeling & Application	3	12	2.64	2.43	1.44	0.65

**Table 9.17 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Mathematics Reporting Subcategories**

Mean, Standard Deviation, and SEM: Mathematics								
Grade	Mode	Subcategory	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
3	CBT	A1	9	9	4.88	2.64	1.28	0.76
	CBT	A2	3	4	1.23	1.18	0.79	0.55
	CBT	A3	7	8	3.34	1.98	1.23	0.61
	CBT	A4	8	9	4.37	2.25	1.29	0.67
	PBT	A1	9	9	6.02	2.46	1.20	0.76
	PBT	A2	3	4	1.79	1.34	0.86	0.59
	PBT	A3	7	8	4.37	2.18	1.22	0.69
	PBT	A4	8	9	5.51	2.29	1.27	0.69
4	CBT	A1	7	8	4.82	2.19	1.16	0.72
	CBT	A2	7	7	3.38	2.06	1.05	0.74
	CBT	A3	8	8	4.42	2.20	1.14	0.73
	PBT	A1	7	8	5.02	2.12	1.16	0.70
	PBT	A2	7	7	3.55	2.02	1.06	0.72
	PBT	A3	8	8	4.53	2.12	1.13	0.72
5	CBT	A1	5	5	2.92	1.41	0.97	0.53
	CBT	A2	6	6	3.27	1.73	1.04	0.64
	CBT	A3	6	7	3.14	1.88	1.18	0.61
	CBT	A4	7	8	4.13	2.16	1.18	0.70
6	CBT	A1	8	9	5.46	2.31	1.25	0.71
	CBT	A2	7	8	3.61	2.41	1.25	0.73
	CBT	A3	12	13	5.84	3.49	1.49	0.82
7	CBT	A1	8	9	4.28	2.45	1.36	0.69
	CBT	A2	15	16	7.67	3.59	1.72	0.77
	CBT	A3	4	5	1.91	1.50	0.93	0.61
8	CBT	A1	5	5	2.07	1.31	0.95	0.47
	CBT	A2	8	8	2.94	1.94	1.23	0.60
	CBT	A3	5	6	2.80	1.49	0.98	0.57
	CBT	A4	8	10	3.77	2.67	1.31	0.76

## 9.5 Divergent (Discriminant) Validity

Measures of different constructs should not be highly correlated with each other. Divergent validity is a subtype of construct validity that can be assessed by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures of unrelated or distantly related constructs are examined in support of divergent validity.

To assess the divergent validity of the LEAP 2025 assessments, correlations were computed between the ELA, mathematics, social studies, and science scale scores for students who took more than one LEAP 2025 content-area test in 2019. These correlations are based on the census data, and the results are shown in Table 9.18. The correlation coefficients ranged from 0.70 (between mathematics and social studies in grades 3 and 5) to 0.84 (between ELA and social studies in grade 8 and between social studies and science in grade 8). The correlation coefficients suggest that individual student scores across subjects are moderately related, indicating that these tests measure a similar knowledge base or general underlying ability while still measuring some different traits as planned.

**Table 9.18 Inter-Correlation of English Language Arts and Mathematics Scale Scores**

Grade	ELA/ Mathematics	ELA/ Social Studies	ELA/ Science	Mathematics/ Social Studies	Mathematics/ Science	Social Studies/ Science
3	0.74	0.76	0.80	0.70	0.76	0.79
4	0.74	0.78	0.78	0.73	0.76	0.80
5	0.74	0.79	0.80	0.70	0.77	0.77
6	0.79	0.82	0.79	0.77	0.79	0.81
7	0.79	0.81	0.79	0.76	0.81	0.81
8	0.74	0.84	0.79	0.75	0.76	0.84

## 9.6 Regression of LEAP 2025 from 2018 to 2019

The LEAP 2025 assessments were designed to support an integrated educational system where the scope and sequence of each grade's curriculum will support student readiness for and achievement in the next education level. Effective measurement is expected to result in assessments that produce scores that consistently measure each grade's content and produce data that provide strong evidence of preparedness for the content measured by assessments at the education level.

This study required the collection of data from adjacent grades for each content area. For this purpose, matched longitudinal LEAP 2025 test data from spring 2018 and spring 2019 were used. For example, grade 3 students were matched with grade 4 students, and only matched students were used to estimate correlation and perform linear regression from 2018 to 2019.

Table 9.19 summarizes the correlation and regression results for 2018 and 2019 LEAP 2025. For ELA, the correlation ranged from 0.78 to 0.84, and for mathematics, the correlation ranged from 0.81 to 0.86. Correlations for mathematics were slightly higher than those for ELA. Correlations for both content areas can be considered moderate, which can often be found in state assessments.  $R^2$  indicates how much of the 2018 performance can explain the 2019 performance. For example, 0.60 for ELA 2018 grade 3 and 2019 grade 4 means that 2018's grade 3 performance can explain (predict) about 60% of 2019's grade 4 performance. This  $R^2$  value is generally the power of 2 for the matching correlation. The  $R^2$  values for ELA range from 0.60 to 0.71, and those for mathematics range from 0.66 to 0.73. These also show the moderate relationships between adjacent grades for both ELA and mathematics.

**Table 9.19 Correlation and Regression Summary for 2018 and 2019 LEAP 2025**

Content	2018 Grade	2019 Grade	N	Correlation	R <sup>2</sup>
ELA	3	4	≥51,080	0.78	0.60
	4	5	≥51,920	0.78	0.61
	5	6	≥51,610	0.80	0.64
	6	7	≥49,040	0.84	0.70
	7	8	≥47,720	0.84	0.71
Mathematics	3	4	≥51,090	0.81	0.66
	4	5	≥51,810	0.82	0.67
	5	6	≥51,490	0.81	0.66
	6	7	≥48,870	0.86	0.73
	7	8	≥41,710	0.82	0.67

Figures 9.3 and 9.4 show regression line and scatter plots for ELA and mathematics. The linear lines in the plots are linear regression lines from 2018 to 2019. In general, the length of band given the linear regression line shows the strength of correlation. If the band is narrow, the correlation is high, and if the band is large, the correlation is low. Every plot shows some moderate linear relationships between 2018 and 2019 adjacent grades for both ELA and mathematics.

Figure 9.3 Regression Line and Scatter Plots: ELA

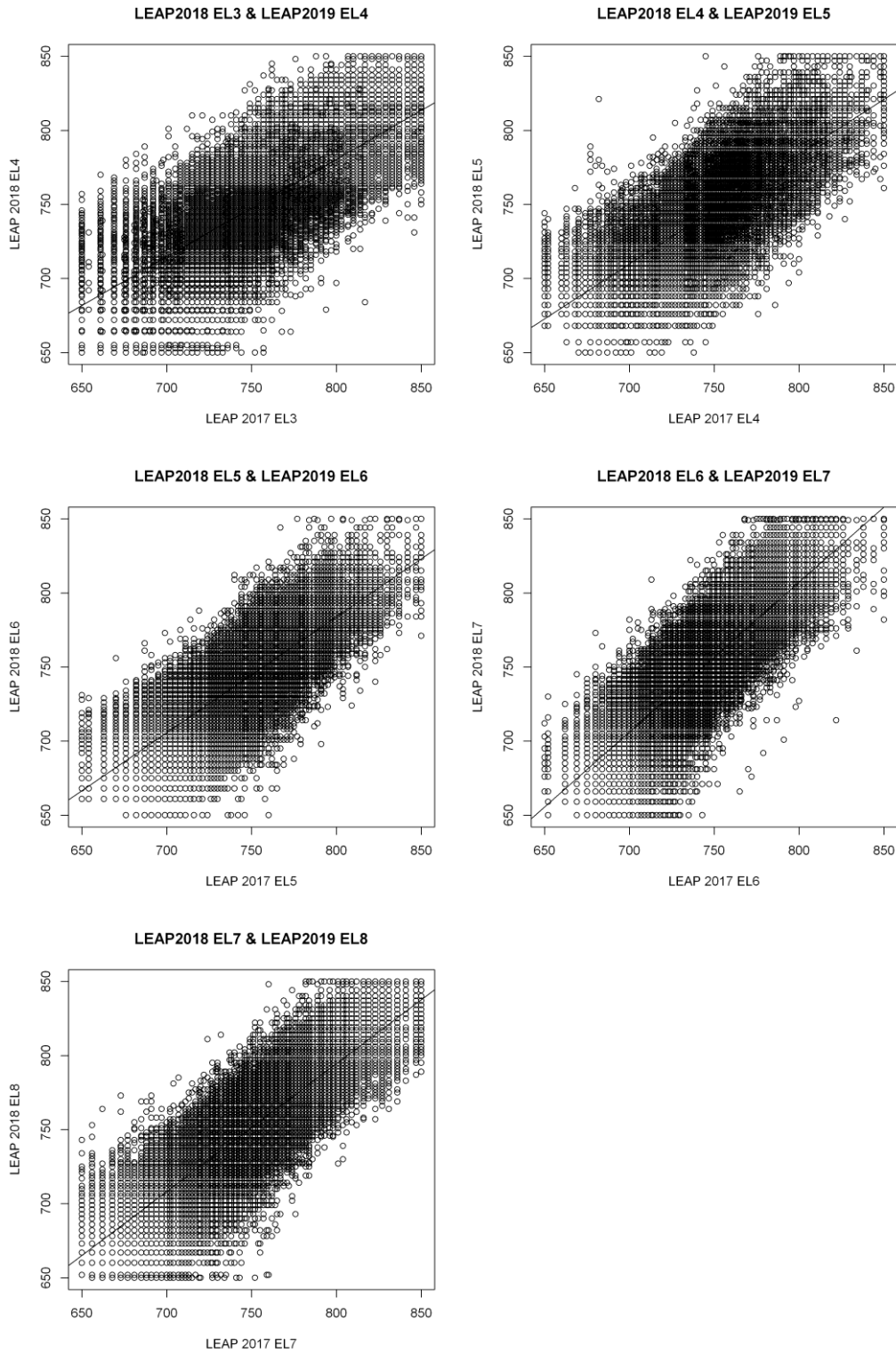
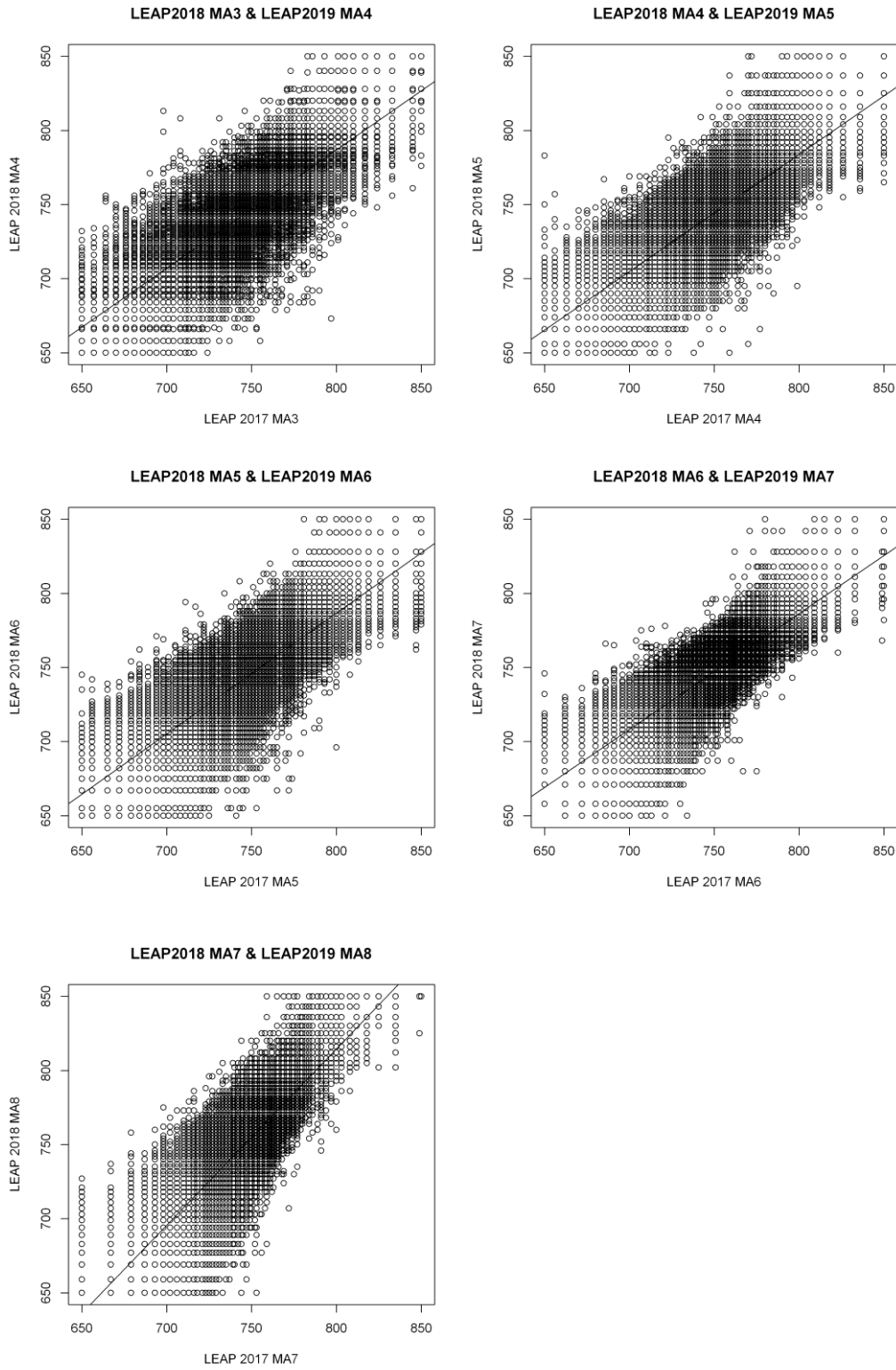


Figure 9.4 Regression Line and Scatter Plots: Mathematics



## 9.7 Summary

In summary, the overall purpose of establishing construct validity is to ensure that the interpretation of test scores is supported. Evidence of validity is necessary to justify the use of the LEAP 2025 test scores. This evidence addresses multiple best practices of the testing industry but particularly relates to the following standards.

**Standard 1.13** If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided. (26)

**Standard 1.21** When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates. (29)

**Standard 2.0** Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use. (42)

**Standard 2.3** For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported. (43)

**Standard 2.13** The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (45)

**Standard 2.14** When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (46)

**Standard 2.16** When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure. (46)

**Standard 2.19** Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported. (47)

## Chapter 10: Fairness

---

As noted in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), there are varying definitions of fairness. This chapter examines fairness as it relates to minimizing bias on a test. This chapter also discusses test performance among varying subgroups assessed by LEAP 2025 assessments. It should be noted that having differences in test performance among subgroups does not mean that a test is unfair—it simply means that groups perform differently on a test. Even when a test is carefully and properly constructed, differences may exist among subgroups as a result of differences in curriculum or learning by students in the subgroup.

This chapter demonstrates for the LEAP 2025 assessments adhere to AERA, APA, & NCME Standards 3.1–3.6. These standards are from Chapter 3 of the *Standards*, which is titled “Fairness in Testing.” Each of these standards is presented in this chapter.

Standard 3.6 states:

Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws. (65)

Test scores of examinee subgroups that differ in meaning are an ongoing concern in any large-scale testing program. To lessen the possibility of differences in test score meaning, DRC follows several steps in the item development and item selection processes, as is explained in Section 10.1 of this chapter. In addition, the LDOE assessment research and development experts, and Louisiana educators, conduct content and bias reviews on items during the selection process, as explained in Chapter 3. These practices adhere to Standard 3.3, which states, “Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test” (64).

The PARCC consortium, as well as DRC, conducted differential item functioning (DIF) studies of their items prior to operational administrations. Items are typically evaluated for possible DIF in the field test phase of the test development process, and any items flagged for DIF are further examined to determine possible bias. During the ELA and mathematics test development process, DRC content experts tried to avoid including operational items flagged for DIF. Section 10.2 of this chapter explains the steps taken to evaluate LEAP 2025 items using DIF to adhere to Standard 3.3.

In addition, the standardized test administration practices and the extensive training process for test score interpretation for LEAP 2025 comply with Standards 3.4 and 3.5, which state:

**Standard 3.4** Test takers should receive comparable treatment during the test administration and scoring process. (65)

**Standard 3.5** Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population. (65)

Section 10.1 of this chapter is also directly relevant to Standards 3.1 and 3.2.



**Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

**Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

This chapter explains the steps taken by DRC to minimize words, phrases, and content that may be regarded as offensive by members of particular demographic subgroups. Section 3.2 of Chapter 3 discusses the content and bias review conducted for LEAP 2025. This review is also critical in fulfilling Standards 3.1 and 3.2. The PARCC operational items used in the 2018 LEAP 2025 forms were critical to the forms construction process. Refer to the PARCC website for the bias and sensitivity guidelines used and the processes and procedures followed by [PARCC](#) pertaining to these items.

## 10.1 Minimizing Bias through Careful Test Development

The construction of a test that is fair for all examinees begins in the early stages of planning and development. The item and test development processes that were used to minimize bias are summarized below.

First, careful attention was paid to content validity during the item development and item selection processes. Bias can occur only if the test is measuring different things for different groups. The possibility of bias is reduced by eliminating irrelevant skills or knowledge from the items.

Second, item writers and test developers followed PARCC Fairness and Sensitivity Guidelines for reducing or eliminating bias. DRC test development staff reviewed all items and other testing materials with these guidelines in mind. Internal editorial reviews were conducted by at least three different people: a content editor who directly supervised the item writers, a style editor, and a content supervisor. The final test was again reviewed by people in these same roles and was also subjected to an independent review by the LDOE assessment research and development specialists.

Third, careful attention was given to item statistics throughout the test development process. As part of the test assembly process, attempts were made to avoid using or reusing items with poor statistical fit or distractors with positive point biserial correlations, since this may indicate that an item is testing an ability that is irrelevant to the construct being measured. DIF statistics were also examined during test construction. Items that had exhibited significant DIF against one or more subgroups were removed from further consideration unless it was essential to include them to meet content specifications.

## 10.2 Evaluating Bias through Differential Item Functioning (DIF) Statistics

After administering the test, an empirical approach known as DIF was used to examine the items. The DIF statistics indicate the degree to which members of a particular focus group perform better or worse than expected on each item as compared to the reference group. The DIF procedures used and the results of these analyses are detailed in this section. It should be noted, however, that all items included in LEAP 2025 were thoroughly reviewed for content and bias by the LDOE and DRC content experts to ensure the items do not test knowledge or ability irrelevant to the construct the test intends to measure. Therefore, DIF flags do not necessarily indicate that an item is biased; rather, DIF flags indicate that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). Items are not necessarily suppressed from operational scoring if they are flagged for DIF.

The position of DRC concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, languages, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test. Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are expected of all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements into the performances on which the measurement is based. If these tests require that students have culturally specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

To lessen this bias, DRC strives to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above and in Chapter 3 of this report, careful attention is given during the test development and test construction processes to lessen the influence of these elements for large numbers of students. Unfortunately, in some cases these elements may continue to play a substantial role in some cases. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted after each operational test administration.

DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Two DIF statistics that are commonly used for this purpose are the Mantel-Haenszel (MH) statistic (1959) and the standardized mean difference (SMD) between the reference and focal groups, proposed by Dorans and Schmitt (1991).

The MH statistic is computed as follows (Zwick, Donoghue, & Grima, 1993):

$$\text{Mantel } \chi^2 = \frac{\left( \sum_k F_k - \sum_k E(F_k) \right)^2}{\sum_k \text{Var}(F_k)},$$

where  $F_k$  is the sum of scores for the focal group at the  $k$ th level of the matching variable. Note that the MH statistic is sensitive to  $N$  such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. Educational Testing Service (ETS) first developed the MH-D DIF statistic. To compute delta, alpha (the odds ratio) is first computed as follows:

$$\alpha_{MH} = \frac{\sum_{k=1}^K N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^K N_{f1k} N_{r0k} / N_k},$$

where  $N_{r1k}$  is the number of correct responses in the reference group at ability level  $k$ ,  $N_{f0k}$  is the number of incorrect responses in the focal group at ability level  $k$ ,  $N_k$  is the total number of responses,  $N_{f1k}$  is the number of correct responses in the focal group at ability level  $k$ , and  $N_{r0k}$  is the number of incorrect responses in the reference group at ability level  $k$ . MH-D DIF is then computed as follows:

$$\text{MH-D DIF} = -2.35 \ln(\alpha_{MH})$$

For selected-response items, the MH ( $\chi^2_{MH}$ ) statistic was used to evaluate potential DIF items. In the MH procedure, subgroups are matched by their raw total test score, using a contingency table with  $K$  ability levels. When applying the MH procedure, the log-odds ratio  $\alpha$  is assumed to be constant across the  $K$  matched levels. The  $\chi^2_{MH}$ , then, estimates a pooled common-odds ratio. Taking the natural logarithm of the common-odds ratio and its confidence limits and multiplying these with the constant  $-2.35$  may then allow the resulting values to be placed on the MH delta metric ( $\Delta_{MH}$ ) for interpretive purposes. Items were flagged for DIF using the following criteria:

- Moderate DIF: Significant MH chi-square statistic ( $p < 0.05$ ) and  $1.0 \leq |\text{MH D-DIF}| < 1.5$
- Large DIF: Significant MH chi-square statistic ( $p < 0.05$ ) and  $|\text{MH D-DIF}| \geq 1.5$

For constructed-response items, an effect size (ES) statistic based on the MH chi-square will be used. The ES is obtained by dividing the SMD statistics by the standard deviation of the item. The SMD is an effect size index of DIF, which is relatively easy to interpret. The SMD compares the mean of the reference and focal group, adjusting for the distribution of reference and focal group members on the conditioning variable, which for these analyses is the LEAP 2025 raw score. The SMD is computed as follows (Zwick et al., 1993):

$$SMD = p_{Fk} \left( \sum_k m_{Fk} - \sum_k m_{Rk} \right),$$

where  $p_{Fk}$  = the proportion of the focal group members at the  $k$ th level of the matching variable,  $m_{Fk} = 1/N_{F1k}$ , and  $m_{Rk} = 1/N_{R1k}$ . Items are flagged using the same rules that are used in NAEP:

- Moderate DIF: If the MH statistic is significant ( $p < .05$ ) and  $|\text{ES}|$  is between 0.17 and 0.25
- Large DIF: If the MH statistic is significant ( $p < .05$ ) and  $|\text{ES}| \geq 0.25$

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group.

### 10.2.1 DIF Statistics for Demographic Groups

DIF analyses were conducted for groups defined by demographic characteristics. Tables 10.1 and 10.2 show the DIF results for the following subgroups:

**Gender:** Focal group is females; reference group is males.

**Ethnicity:** Focal groups are Hispanic/Latino, American Indian or Alaska Native, Asian, Black or African American, and two or more races; reference group is white.

**Education Classification:** Focal group is students who are classified as special education; reference group is all others.

**EL Status:** Focal group is students who are classified as EL; reference group is all others.

**Economic Status:** Focal group is students who are classified as economically disadvantaged; reference group is all others.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

The minimum case count for the focal group was set at 200, and the minimum case count for the reference group was set at 400. The DIF analyses are not performed for subgroups of less than 200. In these cases, the statistical procedures do not have sufficient power to detect potential differences.

Tables 10.1 and 10.2 summarize the number of DIF flags by content area, grade, and test form for each focal group that included at least 200 students. Results are not reported (NR) for groups with an insufficient number of students. The analyses were conducted by test form.

The PBT form for ELA students in grade 3 (see Table 10.1) can be considered as an example. In this form, two items exhibited moderate gender DIF for the female group, one was negative and one positive; one item was flagged for large negative DIF for the Hispanic/Latino group; and one item showed large negative DIF for the EL group.

Table 10.1 2019 LEAP 2025 DIF Statistics: Number of Flagged Items, English Language Arts

DIF Statistics: English Language Arts					Count of Items at DIF Magnitude			
					Moderate		Large	
Grade	Mode	Number of Items	Category	Group	B-	B+	C-	C+
3	CBT	26	Gender	Female	0	0	0	0
			Ethnicity	Hispanic/Latino	NR	NR	NR	NR
			Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
			Ethnicity	Asian	NR	NR	NR	NR
			Ethnicity	Black or African American	NR	NR	NR	NR
			Ethnicity	Two or More Races	NR	NR	NR	NR
			Education Classification	Special	NR	NR	NR	NR
			EL Status	EL	NR	NR	NR	NR
			Economic Status	Economically Disadvantaged	NR	NR	NR	NR
			Section 504 Status	Section 504	NR	NR	NR	NR
3	PBT	26	Gender	Female	1	1	0	0
			Ethnicity	Hispanic/Latino	0	0	1	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	0	0	0	0
			EL Status	EL	0	0	1	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
4	CBT	28	Gender	Female	0	0	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
			Ethnicity	Asian	NR	NR	NR	NR
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	0	0	0	0
			EL Status	EL	1	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
	PBT	28	Gender	Female	0	1	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	1	0	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	2	0	0	0
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0

DIF Statistics: English Language Arts					Count of Items at DIF Magnitude			
					Moderate		Large	
Grade	Mode	Number of Items	Category	Group	B-	B+	C-	C+
5	CBT	28	Gender	Female	1	0	0	0
			Ethnicity	Hispanic/Latino	1	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	0	0	0	0
			EL Status	EL	1	0	1	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
6	CBT	32	Gender	Female	0	3	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	1	0	0	0
			EL Status	EL	1	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
7	CBT	32	Gender	Female	0	0	0	2
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	1	0	0	0
			EL Status	EL	3	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
8	CBT	32	Gender	Female	2	2	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	1	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Education Classification	Special	0	0	0	0
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0

Table 10.2 2019 LEAP 2025 DIF Statistics: Number of Flagged Items, Mathematics

DIF Statistics: Mathematics					Count of Items at DIF Magnitude			
					Moderate		Large	
Grade	Mode	Number of Items	Category	Group	B-	B+	C-	C+
3	CBT	43	Gender	Female	1	1	0	0
			Ethnicity	Hispanic/Latino	NR	NR	NR	NR
			Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
			Ethnicity	Asian	NR	NR	NR	NR
			Ethnicity	Black or African American	1	1	0	0
			Ethnicity	Two or More Races	NR	NR	NR	NR
			Education Classification	Special	NR	NR	NR	NR
			EL Status	EL	NR	NR	NR	NR
			Economic Status	Economically Disadvantaged	NR	NR	NR	NR
			Section 504 Status	Section 504	NR	NR	NR	NR
3	PBT	43	Gender	Female	1	0	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	1	0	0
			Ethnicity	Black or African American	0	1	1	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	4	2	0	0
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	1	0	0	0
			Section 504 Status	Section 504	0	0	0	0
4	CBT	43	Gender	Female	1	0	0	0
			Ethnicity	Hispanic/Latino	0	1	0	0
			Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
			Ethnicity	Asian	NR	NR	NR	NR
			Ethnicity	Black or African American	1	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	0	2	0	2
			EL Status	EL	2	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
	PBT	43	Gender	Female	0	0	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	1	0	0
			Ethnicity	Black or African American	1	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	1	2	0	2
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0

DIF Statistics: Mathematics					Count of Items at DIF Magnitude			
					Moderate		Large	
Grade	Mode	Number of Items	Category	Group	B-	B+	C-	C+
5	CBT	41	Gender	Female	0	1	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	0	0	0
			Ethnicity	Black or African American	0	1	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	1	0	0	3
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
6	CBT	42	Gender	Female	0	1	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	1	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	0	0	0	2
			EL Status	EL	0	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
7	CBT	43	Gender	Female	0	0	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	2	0	0
			Ethnicity	Black or African American	1	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	0	0	0	0
			EL Status	EL	1	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
8	CBT	41	Gender	Female	0	0	0	0
			Ethnicity	Hispanic/Latino	0	0	0	0
			Ethnicity	American Indian or Alaska Native	0	0	0	0
			Ethnicity	Asian	0	1	0	0
			Ethnicity	Black or African American	0	0	0	0
			Ethnicity	Two or More Races	0	0	0	0
			Ed. Classification	Special	2	1	0	0
			EL Status	EL	1	0	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0



## 10.2.2 DIF Statistics for Test Language

All items on one CBT and one PBT form of the mathematics test at each grade are transadapted from English into Spanish. Transadaptation takes into consideration linguistic and cultural differences and grade-level appropriate words. By accounting for these differences, the achievement of Spanish speakers can be measured in the same way as the achievement of English speakers. Please refer to Appendix F for more information about the transadaptation of Spanish mathematics forms. To help confirm that the test items can be measured similarly regardless of the language in which the items are published, a DIF set of analyses was performed. Two DIF analyses were performed using the 2019 LEAP 2025 mathematics operational items, regardless of student count in the reference or focal group. Smaller counts for the groups needed to be tolerated since the overall count for those being administered the Spanish form was low.

For the first analysis, student responses for the shared operational items between 2018 and 2019 LEAP 2025 mathematics were combined. This approach increased the number of students who took the Spanish versions of the items. The Mantel-Haenszel (MH) and the Standardized Mean Difference (SMD) DIF procedures were performed on these shared items and DIF flags applied. The second analysis focused on the items that were not common between the 2018 and 2019 administrations. The MH and the SMD DIF procedures were performed on all 2019 LEAP 2025 operational items, including items that were unique to the 2019 administration in addition to those in common with the 2018 administration. However, DIF flags were applied to only the items that were not shared between 2018 and 2019.

For both analyses, DIF results were carefully reviewed whenever sample sizes were smaller than the required minimum sample size and when an item showed large (C) DIF. All items were determined by the LDOE to be suitable for scoring. Table 10.3 summarizes how many items overall exhibited moderate or large DIF in mathematics.

**Table 10.3 2019 LEAP 2025 DIF Statistics: Number of Flagged Items, Mathematics**

DIF Statistics: Mathematics				Count of Items at DIF Magnitude			
				Moderate		Large	
Grade	Number of Items	Category	Group	B-	B+	C-	C+
3	43	Test Language	Spanish	1	0	5	1
4	41	Test Language	Spanish	1	2	3	0
5	42	Test Language	Spanish	1	0	0	0
6	43	Test Language	Spanish	1	0	0	1
7	43	Test Language	Spanish	2	3	1	0
8	41	Test Language	Spanish	1	0	2	0

## 10.3 Evaluating Bias through Impact Analysis

The impact of achievement testing on subgroups can be determined and reported in the form of average scores and also in terms of test score reliability. Tables 10.4–10.19 present the number of students, test form reliability statistics (i.e., coefficient alpha; see Chapter 9), scale score means and standard deviations, and effect size (i.e., Cohen’s *d*) for the various subgroups of interest by form.

### 10.3.1 Reliability

Tables 10.4–10.11 show the test form reliability coefficients and SEM by student gender, ethnicity, education classification, EL status, economic status, and Section 504 status. The reliability coefficients for English Language Arts forms ranged from 0.82 to 0.93. For mathematics the reliability coefficients ranged from 0.84 to 0.93. These analyses show that the test reliability is of acceptable magnitude for all the subgroups. Note that the reliability coefficients are NR for subgroups with fewer than 10 students.

**Table 10.4 Grade 3 Computer-Based Test Administration Reliability and SEM by Subgroup**

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥1,530	0.86	4.07	≥1,520	0.92	3.52
<b>Gender</b>						
Female	≥730	0.86	4.14	≥720	0.91	3.58
Male	≥800	0.86	3.99	≥790	0.92	3.45
<b>Ethnicity</b>						
Hispanic/Latino	≥160	0.85	4.00	≥140	0.92	3.40
American Indian or Alaska Native	<10	NR	NR	<10	NR	NR
Asian	≥10	0.87	4.27	≥10	0.90	3.82
Black or African American	≥920	0.83	3.96	≥920	0.91	3.39
Native Hawaiian or Other Pacific	<10	NR	NR	<10	NR	NR
White	≥390	0.87	4.15	≥390	0.92	3.70
Two or More Races	≥30	0.84	4.20	≥30	0.91	3.66
<b>Education Classification</b>						
Regular	≥1350	0.86	4.12	≥1330	0.92	3.54
Special	≥180	0.80	3.73	≥180	0.90	3.26
<b>English Learner Status</b>						
Non-EL	≥1420	0.86	4.09	≥1420	0.92	3.53
EL	≥110	0.78	3.75	≥90	0.91	3.29
<b>Economic Status</b>						
Economically Disadvantaged	≥1280	0.84	4.01	≥1270	0.92	3.42
Not Economically Disadvantaged	≥210	0.88	4.38	≥210	0.92	3.83
<b>Section 504 Status</b>						
Non-Section 504	≥1460	0.86	4.09	≥1440	0.92	3.53
Section 504	≥70	0.85	3.72	≥70	0.88	3.15

**Table 10.5 Grade 3 Paper-Based Test Administration Reliability and SEM by Subgroup**

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥51,410	0.87	4.58	≥51,300	0.92	3.82
<b>Gender</b>						
Female	≥24,940	0.86	4.64	≥24,890	0.92	3.83
Male	≥26,440	0.87	4.50	≥26,390	0.93	3.80
<b>Ethnicity</b>						
Hispanic/Latino	≥4,460	0.87	4.54	≥4,390	0.92	3.75
American Indian or Alaska Native	≥300	0.85	4.56	≥300	0.92	3.85
Asian	≥820	0.89	4.72	≥820	0.92	3.65
Black or African American	≥21,770	0.85	4.46	≥21,750	0.92	3.72
Native Hawaiian or Other Pacific	≥40	0.89	4.54	≥40	0.91	3.83
White	≥22,310	0.85	4.66	≥22,300	0.91	3.81
Two or More Races	≥1,620	0.85	4.67	≥1,620	0.92	3.81
<b>Education Classification</b>						
Regular	≥45,250	0.86	4.60	≥45,150	0.92	3.82
Special	≥6,150	0.85	4.31	≥6,150	0.92	3.67
<b>English Learner Status</b>						
Non-EL	≥48,830	0.86	4.59	≥48,810	0.92	3.82
EL	≥2,570	0.82	4.34	≥2,490	0.92	3.68
<b>Economic Status</b>						
Economically Disadvantaged	≥36,850	0.85	4.51	≥36,760	0.92	3.78
Not Economically Disadvantaged	≥14,330	0.85	4.72	≥14,320	0.91	3.74
<b>Section 504 Status</b>						
Non-Section 504	≥46,780	0.87	4.59	≥46,680	0.92	3.82
Section 504	≥4,620	0.83	4.40	≥4,620	0.91	3.72

Table 10.6 Grade 4 Computer-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥7,570	0.90	4.90	≥7,540	0.94	3.51
<b>Gender</b>						
Female	≥3,690	0.90	4.99	≥3,680	0.93	3.52
Male	≥3,870	0.90	4.78	≥3,860	0.94	3.50
<b>Ethnicity</b>						
Hispanic/Latino	≥800	0.90	4.83	≥770	0.94	3.43
American Indian or Alaska Native	≥30	0.90	4.98	≥30	0.91	3.63
Asian	≥100	0.91	4.86	≥100	0.92	3.69
Black or African American	≥3,120	0.87	4.78	≥3,120	0.92	3.27
Native Hawaiian or Other Pacific	<10	NR	NR	<10	NR	NR
White	≥3,270	0.89	5.00	≥3,270	0.93	3.64
Two or More Races	≥220	0.89	5.02	≥220	0.94	3.57
<b>Education Classification</b>						
Regular	≥6,680	0.89	4.94	≥6,650	0.93	3.54
Special	≥890	0.89	4.34	≥890	0.93	3.17
<b>English Learner Status</b>						
Non-EL	≥7,150	0.90	4.92	≥7,150	0.93	3.53
EL	≥410	0.84	4.62	≥390	0.92	3.22
<b>Economic Status</b>						
Economically Disadvantaged	≥5,400	0.88	4.82	≥5,370	0.93	3.39
Not Economically Disadvantaged	≥2,080	0.88	5.02	≥2,080	0.92	3.64
<b>Section 504 Status</b>						
Non-Section 504	≥7,050	0.90	4.91	≥7,020	0.93	3.53
Section 504	≥380	0.83	4.80	≥380	0.88	3.17

Table 10.7 Grade 4 Paper-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥47,220	0.89	5.26	≥47,140	0.93	3.64
<b>Gender</b>						
Female	≥23,190	0.89	5.34	≥23,150	0.93	3.66
Male	≥24,020	0.89	5.14	≥23,970	0.93	3.61
<b>Ethnicity</b>						
Hispanic/Latino	≥3,730	0.90	5.21	≥3,640	0.93	3.57
American Indian or Alaska Native	≥280	0.87	5.29	≥280	0.92	3.66
Asian	≥660	0.91	5.26	≥660	0.93	3.67
Black or African American	≥20,800	0.87	5.24	≥20,790	0.92	3.47
Native Hawaiian or Other Pacific	≥30	0.87	5.34	≥30	0.94	3.45
White	≥20,130	0.88	5.28	≥20,140	0.92	3.72
Two or More Races	≥1,530	0.88	5.32	≥1,520	0.92	3.66
<b>Education Classification</b>						
Regular	≥41,640	0.89	5.26	≥41,550	0.93	3.65
Special	≥5,580	0.87	4.90	≥5,580	0.92	3.35
<b>English Learner Status</b>						
Non-EL	≥45,280	0.89	5.27	≥45,290	0.93	3.65
EL	≥1,940	0.85	5.03	≥1,840	0.92	3.39
<b>Economic Status</b>						
Economically Disadvantaged	≥33,920	0.88	5.24	≥33,830	0.93	3.56
Not Economically Disadvantaged	≥13,120	0.88	5.26	≥13,100	0.91	3.70
<b>Section 504 Status</b>						
Non-Section 504	≥41,950	0.89	5.27	≥41,870	0.93	3.65
Section 504	≥5,270	0.86	5.12	≥5,27	0.92	3.46

Table 10.8 Grade 5 Computer-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥54,910	0.90	4.98	≥54,730	0.92	3.58
<b>Gender</b>						
Female	≥26,910	0.89	5.09	≥26,820	0.92	3.61
Male	≥27,990	0.90	4.85	≥27,900	0.93	3.54
<b>Ethnicity</b>						
Hispanic/Latino	≥4,490	0.90	4.90	≥4,360	0.93	3.51
American Indian or Alaska Native	≥330	0.87	5.20	≥330	0.92	3.64
Asian	≥850	0.91	5.16	≥850	0.92	3.66
Black or African American	≥23,820	0.87	4.87	≥23,770	0.91	3.39
Native Hawaiian or Other Pacific	≥50	0.88	5.27	≥50	0.92	3.63
White	≥23,640	0.89	5.11	≥23,620	0.92	3.69
Two or More Races	≥1,720	0.89	5.09	≥1,720	0.92	3.60
<b>Education Classification</b>						
Regular	≥48,540	0.89	5.02	≥48,360	0.92	3.61
Special	≥6,370	0.87	4.39	≥6,360	0.90	3.14
<b>English Learner Status</b>						
Non-EL	≥52,910	0.89	5.00	≥52,850	0.92	3.60
EL	≥2,000	0.81	4.51	≥1,870	0.91	3.17
<b>Economic Status</b>						
Economically Disadvantaged	≥39,040	0.88	4.90	≥38,900	0.91	3.48
Not Economically Disadvantaged	≥15,480	0.88	5.14	≥15,470	0.91	3.70
<b>Section 504 Status</b>						
Non-Section 504	≥49,510	0.90	5.00	≥49,330	0.92	3.60
Section 504	≥5,400	0.87	4.72	≥5,390	0.91	3.34

Table 10.9 Grade 6 Computer-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥54,800	0.91	5.21	≥54,710	0.93	3.65
<b>Gender</b>						
Female	≥26,960	0.91	5.23	≥26,920	0.93	3.69
Male	≥27,830	0.91	5.10	≥27,790	0.93	3.59
<b>Ethnicity</b>						
Hispanic/Latino	≥4,170	0.91	5.17	≥4,020	0.93	3.53
American Indian or Alaska Native	≥340	0.91	5.30	≥340	0.92	3.71
Asian	≥810	0.92	5.23	≥810	0.93	3.98
Black or African American	≥23,740	0.89	5.13	≥23,730	0.92	3.39
Native Hawaiian or Other Pacific	≥40	0.89	5.56	≥40	0.93	3.72
White	≥24,070	0.90	5.26	≥24,130	0.92	3.82
Two or More Races	≥1,590	0.91	5.25	≥1,590	0.93	3.70
<b>Education Classification</b>						
Regular	≥48,940	0.90	5.22	≥48,850	0.93	3.69
Special	≥5,850	0.87	4.59	≥5,860	0.91	2.95
<b>English Learner Status</b>						
Non-EL	≥53,090	0.91	5.22	≥53,150	0.93	3.66
EL	≥1,700	0.84	4.78	≥1,550	0.92	3.12
<b>Economic Status</b>						
Economically Disadvantaged	≥38,490	0.90	5.15	≥38,390	0.92	3.50
Not Economically Disadvantaged	≥15,940	0.90	5.27	≥15,970	0.92	3.88
<b>Section 504 Status</b>						
Non-Section 504	≥49,350	0.91	5.22	≥49,260	0.93	3.67
Section 504	≥5,440	0.89	5.01	≥5,450	0.92	3.36

Table 10.10 Grade 7 Computer-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥52,350	0.92	5.50	≥52,090	0.92	3.88
<b>Gender</b>						
Female	≥25,530	0.91	5.54	≥25,400	0.91	3.94
Male	≥26,810	0.92	5.36	≥26,680	0.92	3.81
<b>Ethnicity</b>						
Hispanic/Latino	≥3,800	0.93	5.42	≥3,590	0.92	3.79
American Indian or Alaska Native	≥330	0.91	5.60	≥330	0.91	3.88
Asian	≥800	0.93	5.52	≥800	0.93	4.49
Black or African American	≥22,890	0.91	5.43	≥22,870	0.90	3.43
Native Hawaiian or Other Pacific	≥40	0.92	5.42	≥40	0.92	3.95
White	≥23,070	0.90	5.53	≥23,040	0.91	4.17
Two or More Races	≥1,370	0.91	5.59	≥1,370	0.91	3.92
<b>Education Classification</b>						
Regular	≥46,940	0.91	5.52	≥46,700	0.91	3.95
Special	≥5,400	0.89	4.83	≥5,380	0.89	2.95
<b>English Learner Status</b>						
Non-EL	≥50,830	0.92	5.51	≥50,770	0.92	3.90
EL	≥1,510	0.87	4.80	≥1,320	0.87	3.12
<b>Economic Status</b>						
Economically Disadvantaged	≥36,250	0.91	5.46	≥36,040	0.91	3.61
Not Economically Disadvantaged	≥15,720	0.90	5.51	≥15,710	0.91	4.29
<b>Section 504 Status</b>						
Non-Section 504	≥47,120	0.92	5.51	≥46,870	0.92	3.92
Section 504	≥5,220	0.90	5.29	≥5,210	0.90	3.46



Table 10.11 Grade 8 Computer-Based Test Administration Reliability and SEM by Subgroup

Group	ELA			Mathematics		
	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
<b>All Students</b>	≥50,720	0.90	5.63	≥44,520	0.92	3.51
<b>Gender</b>						
Female	≥24,760	0.90	5.63	≥21,430	0.92	3.59
Male	≥25,960	0.90	5.53	≥23,090	0.92	3.43
<b>Ethnicity</b>						
Hispanic/Latino	≥3,640	0.92	5.58	≥3,130	0.92	3.42
American Indian or Alaska Native	≥330	0.90	5.69	≥310	0.91	3.57
Asian	≥800	0.92	5.69	≥570	0.94	3.89
Black or African American	≥22,010	0.88	5.54	≥20,660	0.90	3.30
Native Hawaiian or Other Pacific	≥20	0.87	6.58	≥20	0.94	3.81
White	≥22,760	0.89	5.71	≥18,840	0.92	3.67
Two or More Races	≥1,120	0.89	5.72	≥980	0.92	3.63
<b>Education Classification</b>						
Regular	≥45,800	0.89	5.66	≥39,670	0.92	3.56
Special	≥4,920	0.84	5.02	≥4,850	0.87	2.87
<b>English Learner Status</b>						
Non-EL	≥49,140	0.90	5.64	≥43,130	0.92	3.52
EL	≥1,570	0.84	4.84	≥1,390	0.89	3.08
<b>Economic Status</b>						
Economically Disadvantaged	≥34,340	0.89	5.58	≥31,780	0.91	3.39
Not Economically Disadvantaged	≥16,020	0.89	5.67	≥12,420	0.92	3.73
<b>Section 504 Status</b>						
Non-Section 504	≥45,640	0.90	5.64	≥39,680	0.92	3.54
Section 504	≥5,070	0.87	5.48	≥4,840	0.90	3.25

### 10.3.2 Effect Size

One way to evaluate the magnitude of the standardized mean difference (SMD) is to calculate the ES. Cohen's  $d$  was used to calculate the ES. Cohen's  $d$  is given by the following formula:

$$d = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}},$$

where  $\bar{x}_a$  is the mean score of group A,  $\bar{x}_b$  is the mean score of group B,  $s_a^2$  is the variance of group A,  $s_b^2$  is the variance of group B,  $n_a$  is the number of students in group A, and  $n_b$  is the number of students in group B.

Cohen's  $d$ , then, expresses the difference in group means in terms of the standard deviation. For example, if  $d = .34$  for two groups, then it may be interpreted that the SMD between the two groups is .34 of the pooled standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the  $d$  statistic:  $d = .20$  is a small ES,  $d = .50$  is a medium ES, and  $d = .80$  is a large ES.

Using Cohen's (1988) guidelines, certain trends become apparent in Tables 10.12–10.19. Results are NR for subgroups with fewer than 10 students. On the ELA test in most grades, there are medium differences in mean test scores between females and males where females outperform males. Although there were no ESs larger than a small ES,  $|0.20|$ , for mathematics, females tend to perform better than males in general. For most ELA and mathematics tests, mean scale scores and ES show that Asian and white students tend to outperform other ethnicity groups across grades. For most ELA and mathematics tests, there were clear performance differences between regular education and special education students in Education Classification, between not economically disadvantaged and economically disadvantaged in economic status, non-EL and EL students in EL status, and non-migrant and migrant students in migrant status.

Table 10.12 Impact Analysis, Grade 3 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥1,530	719.32	38.10		≥1,920	722.23	29.28	
<b>Gender</b>								
Male	≥800	714.87	37.92		≥800	720.60	29.78	
Female	≥730	724.19	37.73	-0.25	≥730	724.01	28.64	-0.12
<b>Ethnicity</b>								
White	≥390	737.95	39.79		≥390	733.83	30.16	
Hispanic/Latino	≥160	711.23	37.79	0.68	≥160	721.16	27.99	0.43
American Indian or Alaska Native	<10	NR	NR	NR	<10	NR	NR	NR
Asian	≥10	756.92	39.02	-0.48	≥10	756.15	25.72	-0.74
Black or African American	≥920	711.82	34.24	0.73	≥920	717.07	27.58	0.59
Native Hawaiian or Other Pacific	<10	NR	NR	NR	<10	NR	NR	NR
Two or More Races	≥30	729.86	34.25	0.21	≥30	726.97	27.42	0.23
<b>Education Classification</b>								
Regular	≥1,350	721.97	38.12		≥1,350	724.25	29.16	
Special	≥180	699.83	31.91	0.59	≥180	707.36	25.78	0.59
<b>Economic Status</b>								
Not Economically Disadvantaged	≥210	741.40	43.79		≥210	734.92	31.89	
Economically Disadvantaged	≥1,280	715.51	35.72	0.70	≥1,280	720.04	28.21	0.52
<b>English Learner Status</b>								
Non-EL	≥1,420	720.78	38.24		≥1,420	722.74	29.44	
EL	≥110	700.65	30.90	0.53	≥110	715.61	26.37	0.24
<b>Migrant Status</b>								
Nonmigrant	≥1,530	719.39	38.11		≥1,530	722.25	29.29	
Migrant	<10	NR	NR	NR	<10	NR	NR	NR
<b>Section 504 Status</b>								
Non-Section 504	≥1,460	720.00	38.12		≥1,460	722.95	29.41	
Section 504	≥70	706.12	35.49	0.37	≥70	708.12	22.72	0.51

Table 10.13 Impact Analysis, Grade 3 Paper-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥51,410	746.65	41.86		≥51,020	743.01	32.54	
<b>Gender</b>								
Male	≥26,440	743.58	41.97		≥26,130	741.78	33.14	
Female	≥24,940	749.91	41.51	-0.15	≥24,690	744.33	31.84	-0.08
<b>Ethnicity</b>								
White	≥22,310	760.51	39.28		≥22,150	753.40	30.80	
Hispanic/Latino	≥4,460	736.64	43.36	0.60	≥4,370	740.00	31.69	0.43
American Indian or Alaska Native	≥300	748.21	39.20	0.31	≥300	743.46	31.12	0.32
Asian	≥820	769.95	46.59	-0.24	≥820	769.10	34.87	-0.51
Black or African American	≥21,770	733.05	38.94	0.70	≥21,480	731.71	30.28	0.71
Native Hawaiian or Other Pacific	≥40	756.13	45.70	0.11	≥40	754.95	32.59	-0.05
Two or More Races	≥1,620	753.64	39.43	0.17	≥1,610	745.68	31.36	0.25
<b>Education Classification</b>								
Regular	≥45,250	749.87	41.15		≥44,750	745.60	31.76	
Special	≥6,150	722.97	39.37	0.66	≥6,070	724.00	31.90	0.68
<b>Economic Status</b>								
Not Economically Disadvantaged	≥14,330	768.51	39.07		≥14,260	760.12	30.02	
Economically Disadvantaged	≥36,850	738.24	39.79	0.76	≥36,470	736.37	30.99	0.77
<b>English Learner Status</b>								
Non-EL	≥48,830	748.25	41.46		≥48,330	743.68	32.48	
EL	≥2,570	716.23	37.68	0.78	≥2,490	730.06	30.97	0.42
<b>Migrant Status</b>								
Nonmigrant	≥51,260	746.71	41.85		≥50,680	743.05	32.53	
Migrant	≥150	724.31	41.49	0.54	≥140	731.64	33.56	0.35
<b>Section 504 Status</b>								
Non-Section 504	≥46,780	748.28	41.97		≥46,240	744.40	32.51	
Section 504	≥4,620	730.09	36.88	0.44	≥4,580	729.00	29.38	0.48

Table 10.14 Impact Analysis, Grade 4 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥7,570	736.92	34.27		≥7,930	736.85	32.28	
<b>Gender</b>								
Male	≥3,870	733.98	33.99		≥3,870	736.86	32.96	
Female	≥3,690	740.00	34.29	-0.18	≥3,690	736.84	31.56	0.00
<b>Ethnicity</b>								
White	≥3,270	749.37	32.05		≥3,270	749.05	30.49	
Hispanic/Latino	≥800	732.65	34.71	0.51	≥800	734.36	32.73	0.47
American Indian or Alaska Native	≥30	738.90	34.54	0.33	≥30	737.32	29.21	0.38
Asian	≥100	765.84	37.21	-0.51	≥100	765.85	32.18	-0.55
Black or African American	≥3,120	723.54	30.77	0.82	≥3,120	723.44	28.07	0.87
Native Hawaiian or Other Pacific	<10	NR	NR	NR	<10	NR	NR	NR
Two or More Races	≥220	743.96	32.81	0.17	≥220	741.97	33.01	0.23
<b>Education Classification</b>								
Regular	≥6,680	740.17	33.29		≥6,680	739.39	31.69	
Special	≥890	712.57	31.62	0.83	≥890	717.81	30.28	0.68
<b>Economic Status</b>								
Not Economically Disadvantaged	≥2,080	757.48	31.33		≥2,080	756.82	29.69	
Economically Disadvantaged	≥5,400	729.07	31.98	0.89	≥5,390	729.30	29.79	0.92
<b>English Learner Status</b>								
Non-EL	≥7,150	738.20	34.11		≥7,150	737.75	32.21	
EL	≥410	715.06	29.22	0.68	≥410	721.55	29.64	0.51
<b>Migrant Status</b>								
Nonmigrant	≥7,560	736.93	34.27		≥7,560	736.86	32.26	
Migrant	≥10	729.45	32.26	0.22	≥10	732.73	46.19	0.13
<b>Section 504 Status</b>								
Non-Section 504	≥7,050	738.03	34.43		≥7,050	737.96	32.33	
Section 504	≥520	722.00	28.10	0.47	≥520	721.90	27.63	0.50

Table 10.15 Impact Analysis, Grade 4 Paper-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥47,220	745.27	34.24		≥46,860	741.45	31.62	
<b>Gender</b>								
Male	≥24,020	742.13	34.00		≥23,770	740.79	32.32	
Female	≥23,190	748.53	34.18	-0.19	≥22,970	742.13	30.87	-0.04
<b>Ethnicity</b>								
White	≥20,130	757.67	31.39		≥20,020	753.40	29.28	
Hispanic/Latino	≥3,730	737.11	36.28	0.64	≥3,670	736.91	31.66	0.56
American Indian or Alaska Native	≥280	745.39	30.94	0.39	≥280	742.69	29.21	0.37
Asian	≥660	765.81	38.76	-0.26	≥660	766.20	32.20	-0.44
Black or African American	≥20,800	733.66	31.92	0.76	≥20,530	729.56	29.02	0.82
Native Hawaiian or Other Pacific	≥30	751.03	31.08	0.21	≥30	745.77	30.82	0.26
Two or More Races	≥1,530	750.92	32.33	0.21	≥1,510	744.39	30.06	0.31
<b>Education Classification</b>								
Regular	≥41,640	748.32	33.42		≥41,210	744.26	30.89	
Special	≥5,580	722.49	31.63	0.78	≥5,530	720.52	29.06	0.77
<b>Economic Status</b>								
Not Economically Disadvantaged	≥13,120	763.70	31.93		≥13,050	758.83	28.80	
Economically Disadvantaged	≥33,920	738.24	32.39	0.79	≥33,600	734.75	30.05	0.81
<b>English Learner Status</b>								
Non-EL	≥45,280	746.48	33.84		≥44,850	742.18	31.48	
EL	≥1,940	716.99	31.18	0.87	≥1,890	724.08	29.95	0.58
<b>Migrant Status</b>								
Nonmigrant	≥47,110	745.32	34.23		≥46,630	741.48	31.61	
Migrant	≥110	723.87	34.72	0.63	≥110	729.81	33.31	0.37
<b>Section 504 Status</b>								
Non-Section 504	≥41,950	747.00	34.33		≥41,520	743.06	31.65	
Section 504	≥5,270	731.50	30.18	0.46	≥5,220	728.64	28.30	0.46

Table 10.16 Impact Analysis, Grade 5 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥54,910	742.46	32.08		≥54,910	736.98	29.75	
<b>Gender</b>								
Male	≥27,990	740.00	31.81		≥27,960	735.83	30.24	
Female	≥26,910	745.01	32.16	-0.16	≥26,880	738.18	29.19	-0.08
<b>Ethnicity</b>								
White	≥23,640	753.66	30.69		≥23,610	747.28	28.58	
Hispanic/Latino	≥4,490	737.77	32.66	0.51	≥4,490	733.45	30.57	0.48
American Indian or Alaska Native	≥330	746.05	30.18	0.25	≥330	739.01	29.54	0.29
Asian	≥850	766.29	35.74	-0.41	≥850	766.40	32.30	-0.67
Black or African American	≥23,820	731.02	28.71	0.76	≥23,780	726.10	26.18	0.77
Native Hawaiian or Other Pacific	≥50	759.62	31.03	-0.19	≥50	752.14	28.34	-0.17
Two or More Races	≥1,720	746.15	31.07	0.24	≥1,720	739.90	28.72	0.26
<b>Education Classification</b>								
Regular	≥48,540	745.57	31.28		≥48,480	739.63	29.32	
Special	≥6,370	718.72	27.88	0.87	≥6,360	716.85	24.90	0.79
<b>Economic Status</b>								
Not Economically Disadvantaged	≥15,480	760.33	30.32		≥15,470	754.20	28.33	
Economically Disadvantaged	≥39,040	735.56	29.91	0.82	≥38,990	730.37	27.39	0.86
<b>English Learner Status</b>								
Non-EL	≥52,910	743.45	31.90		≥52,840	737.77	29.59	
EL	≥2,000	716.20	24.61	0.86	≥2,000	716.33	26.48	0.73
<b>Migrant Status</b>								
Nonmigrant	≥54,830	742.47	32.08		≥54,760	737.00	29.75	
Migrant	≥80	733.89	29.76	0.27	≥80	728.25	30.01	0.29
<b>Section 504 Status</b>								
Non-Section 504	≥49,510	743.92	32.18		≥49,450	738.35	29.89	
Section 504	≥5,400	729.05	27.70	0.47	≥5,390	724.45	25.23	0.47

Table 10.17 Impact Analysis, Grade 6 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥54,800	739.70	30.18		≥54,800	733.77	29.60	
<b>Gender</b>								
Male	≥27,830	734.84	29.69		≥27,820	731.87	30.04	
Female	≥26,960	744.73	29.86	-0.33	≥26,940	735.74	29.00	-0.13
<b>Ethnicity</b>								
White	≥24,070	749.67	28.78		≥24,070	744.40	27.64	
Hispanic/Latino	≥4,170	733.42	31.78	0.56	≥4,170	728.52	30.67	0.56
American Indian or Alaska Native	≥340	741.45	30.14	0.29	≥340	735.13	28.20	0.34
Asian	≥810	763.88	34.01	-0.49	≥810	763.32	32.58	-0.68
Black or African American	≥23,740	729.47	27.12	0.72	≥23,720	722.61	26.46	0.81
Native Hawaiian or Other Pacific	≥40	749.83	27.79	-0.01	≥40	740.77	30.98	0.13
Two or More Races	≥1,590	744.97	30.05	0.16	≥1,590	737.79	29.11	0.24
<b>Education Classification</b>								
Regular	≥48,940	742.96	29.07		≥48,910	736.79	28.68	
Special	≥5,850	712.46	25.08	1.06	≥5,850	708.57	24.71	1.00
<b>Economic Status</b>								
Not Economically Disadvantaged	≥15,940	755.83	28.44		≥15,940	750.53	27.40	
Economically Disadvantaged	≥38,490	733.24	28.26	0.80	≥38,470	727.03	27.60	0.85
<b>English Learner Status</b>								
Non-EL	≥53,090	740.65	29.87		≥53,070	734.54	29.36	
EL	≥1,700	710.10	24.24	1.03	≥1,700	710.02	26.96	0.84
<b>Migrant Status</b>								
Nonmigrant	≥54,720	739.71	30.17		≥54,700	733.78	29.59	
Migrant	≥70	730.86	33.34	0.29	≥70	725.26	34.03	0.29
<b>Section 504 Status</b>								
Non-Section 504	≥49,350	741.21	30.18		≥49,320	735.12	29.67	
Section 504	≥5,440	726.06	26.56	0.51	≥5,440	721.62	26.00	0.46



Table 10.18 Impact Analysis, Grade 7 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥52,350	744.57	35.94		≥52,350	733.26	26.11	
<b>Gender</b>								
Male	≥26,810	737.90	35.24		≥26,780	731.97	26.49	
Female	≥25,530	751.57	35.33	-0.39	≥25,500	734.62	25.65	-0.10
<b>Ethnicity</b>								
White	≥23,070	755.82	33.13		≥23,040	742.69	24.77	
Hispanic/Latino	≥3,800	736.28	39.49	0.57	≥3,790	729.57	26.75	0.52
American Indian or Alaska Native	≥330	748.71	34.02	0.21	≥330	734.94	23.83	0.31
Asian	≥800	768.11	40.51	-0.37	≥800	757.52	31.26	-0.59
Black or African American	≥22,890	733.44	34.00	0.67	≥22,870	723.38	23.00	0.81
Native Hawaiian or Other Pacific	≥40	753.10	37.53	0.08	≥40	736.85	26.46	0.24
Two or More Races	≥1,370	748.88	34.13	0.21	≥1,370	735.15	25.33	0.30
<b>Education Classification</b>								
Regular	≥46,940	748.69	34.20		≥46,890	736.04	25.08	
Special	≥5,400	708.71	30.25	1.18	≥5,390	709.10	22.11	1.09
<b>Economic Status</b>								
Not Economically Disadvantaged	≥15,720	763.02	32.34		≥15,710	747.85	24.72	
Economically Disadvantaged	≥36,250	736.82	34.38	0.78	≥36,210	727.11	24.08	0.85
<b>English Learner Status</b>								
Non-EL	≥50,830	745.78	35.41		≥50,770	733.88	25.98	
EL	≥1,510	704.07	29.61	1.18	≥1,510	712.50	21.65	0.83
<b>Migrant Status</b>								
Nonmigrant	≥52,270	744.59	35.93		≥52,210	733.27	26.12	
Migrant	≥70	727.34	37.18	0.48	≥70	725.91	24.30	0.28
<b>Section 504 Status</b>								
Non-Section 504	≥47,120	746.40	35.94		≥47,070	734.45	26.17	
Section 504	≥5,220	728.06	31.47	0.52	≥5,220	722.60	23.06	0.46

Table 10.19 Impact Analysis, Grade 8 Computer-Based Test Administration

Group	ELA				Mathematics			
	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
<b>All Students</b>	≥50,720	745.65	36.32		≥50,720	729.42	34.09	
<b>Gender</b>								
Male	≥25,960	738.16	35.78		≥23,170	726.42	34.64	
Female	≥24,760	753.51	35.21	-0.43	≥21,490	732.66	33.20	-0.18
<b>Ethnicity</b>								
White	≥22,760	756.67	34.58		≥18,840	741.48	32.55	
Hispanic/Latino	≥3,640	737.10	40.41	0.55	≥3,300	723.72	34.20	0.54
American Indian or Alaska Native	≥330	747.77	35.62	0.26	≥310	731.73	33.12	0.30
Asian	≥800	770.67	40.89	-0.40	≥570	759.60	42.13	-0.55
Black or African American	≥22,010	734.25	33.10	0.66	≥20,640	718.09	30.76	0.74
Native Hawaiian or Other Pacific	≥20	754.83	38.49	0.05	≥20	739.50	43.57	0.06
Two or More Races	≥1,120	754.89	34.27	0.05	≥980	737.09	33.62	0.13
<b>Education Classification</b>								
Regular	≥45,800	749.47	34.92		≥39,810	732.85	33.23	
Special	≥4,920	710.13	29.13	1.14	≥4,850	701.31	27.40	0.97
<b>Economic Status</b>								
Not Economically Disadvantaged	≥16,020	764.21	33.40		≥12,430	747.33	32.65	
Economically Disadvantaged	≥34,340	737.26	34.25	0.79	≥31,900	722.66	32.00	0.77
<b>English Learner Status</b>								
Non-EL	≥49,140	746.93	35.80		≥43,110	730.27	33.95	
EL	≥1,570	705.78	29.08	1.16	≥1,560	706.04	29.50	0.72
<b>Migrant Status</b>								
Nonmigrant	≥50,660	745.67	36.32		≥44,610	729.44	34.09	
Migrant	≥60	733.39	36.67	0.34	≥60	718.97	32.66	0.31
<b>Section 504 Status</b>								
Non-Section 504	≥45,640	747.46	36.35		≥39,830	730.83	34.28	
Section 504	≥5,070	729.36	31.68	0.50	≥4,840	717.88	30.14	0.38

Additional data for mean scale scores are provided in Tables 10.20 and 10.21. These tables report the number of students, mean scale scores, and standard deviations for special education classification. Groups that have fewer than 50 students are NR.

**Table 10.20 Special Education Classification Scale-Score Means and Standard Deviations: English Language Arts**

Special Education Classification Scale-Score Means and Standard Deviations: English Language Arts							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
3	Gifted	≥960	810.12	28.18	≥51,980	744.67	41.29
	Talented	≥560	784.56	35.85	≥52,370	745.44	41.88
	Autism	≥320	714.23	37.60	≥52,610	746.05	41.96
	Deaf-Blindness	<50	NR	NR	≥52,940	745.85	42.01
	Developmental Delay	≥590	713.19	36.12	≥52,350	746.22	41.93
	Emotional Disturbance	≥80	714.52	36.55	≥52,860	745.91	42.00
	HI—Deaf	<50	NR	NR	≥52,920	745.87	42.00
	HI—Hard-of-Hearing	<50	NR	NR	≥52,900	745.88	42.01
	Mild Mental Disability	≥370	691.51	25.72	≥52,560	746.25	41.85
	Moderate Mental Disability	<50	NR	NR	≥52,920	745.88	42.00
	Orthopedic Impairment	<50	NR	NR	≥52,890	745.86	42.01
	Other Health Impairment	≥780	715.38	37.46	≥52,160	746.31	41.91
	Specific Learning Disability	≥2,210	715.43	31.52	≥50,730	747.18	41.91
	Speech or Language Impairment	≥1,780	745.23	41.67	≥51,160	745.88	42.02
	Traumatic Brain Injury	<50	NR	NR	≥52,930	745.87	42.01
	Visual Impairment	<50	NR	NR	≥52,910	745.85	42.01
	Other	<50	NR	NR	≥52,940	745.86	42.01
	HI—Hearing Impairment	<50	NR	NR	≥52,940	745.85	42.01
Unknown	<50	NR	NR	≥52,940	745.85	42.01	
4	Gifted	≥1,170	797.46	23.85	≥53,630	742.95	33.63
	Talented	≥890	772.27	27.39	≥53,900	743.64	34.28
	Autism	≥290	717.92	33.66	≥54,500	744.26	34.32
	Deaf-Blindness	<50	NR	NR	≥54,800	744.11	34.37
	Developmental Delay	<50	NR	NR	≥54,750	744.13	34.37
	Emotional Disturbance	≥140	715.84	34.22	≥54,660	744.19	34.34
	HI—Deaf	<50	NR	NR	≥54,770	744.13	34.36
	HI—Hard-of-Hearing	≥60	725.23	30.99	≥54,740	744.13	34.37
	Mild Mental Disability	≥420	694.06	19.58	≥54,380	744.50	34.17
	Moderate Mental Disability	<50	NR	NR	≥54,790	744.12	34.37
	Orthopedic Impairment	≥50	727.74	34.22	≥54,750	744.13	34.36
	Other Health Impairment	≥1,150	717.34	28.75	≥53,650	744.69	34.25
	Specific Learning Disability	≥2,820	716.06	26.93	≥51,980	745.63	34.07
	Speech or Language Impairment	≥1,400	743.02	33.16	≥53,400	744.14	34.40
	Traumatic Brain Injury	<50	NR	NR	≥54,790	744.12	34.37
	Visual Impairment	<50	NR	NR	≥54,760	744.12	34.37
	Other	<50	NR	NR	≥54,800	744.12	34.37
	HI—Hearing Impairment	<50	NR	NR	≥54,800	744.11	34.37
Unknown	<50	NR	NR	≥54,800	744.11	34.37	

Special Education Classification Scale-Score Means and Standard Deviations: English Language Arts							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
5	Gifted	≥1,370	792.25	23.78	≥53,540	741.18	31.23
	Talented	≥1,260	769.26	27.56	≥53,650	741.83	31.90
	Autism	≥330	725.02	33.89	≥54,580	742.57	32.04
	Deaf-Blindness	<50	NR	NR	≥54,910	742.46	32.07
	Developmental Delay	≥50	714.68	27.63	≥54,860	742.48	32.07
	Emotional Disturbance	≥150	716.55	25.73	≥54,760	742.53	32.06
	HI—Deaf	<50	NR	NR	≥54,890	742.47	32.07
	HI—Hard-of-Hearing	≥50	725.31	31.93	≥54,850	742.48	32.07
	Mild Mental Disability	≥390	699.09	16.91	≥54,520	742.77	31.95
	Moderate Mental Disability	<50	NR	NR	≥54,900	742.46	32.07
	Orthopedic Impairment	≥80	728.59	34.66	≥54,830	742.48	32.07
	Other Health Impairment	≥1,220	718.06	25.82	≥53,690	743.02	31.99
	Specific Learning Disability	≥3,010	713.63	22.54	≥51,900	744.13	31.75
	Speech or Language Impairment	≥980	739.91	32.64	≥53,920	742.51	32.06
	Traumatic Brain Injury	<50	NR	NR	≥54,900	742.46	32.07
	Visual Impairment	<50	NR	NR	≥54,880	742.46	32.07
	Other	<50	NR	NR	≥54,910	742.46	32.08
	HI—Hearing Impairment	<50	NR	NR	≥54,910	742.46	32.08
	Unknown	<50	NR	NR	≥54,910	742.46	32.08
6	Gifted	≥1,540	785.55	22.89	≥53,250	738.37	29.31
	Talented	≥1,560	764.29	27.14	≥53,230	738.98	29.96
	Autism	≥270	717.58	27.10	≥54,520	739.81	30.15
	Deaf-Blindness	<50	NR	NR	≥54,790	739.70	30.18
	Developmental Delay	<50	NR	NR	≥54,760	739.73	30.17
	Emotional Disturbance	≥180	709.67	24.50	≥54,610	739.80	30.15
	HI—Deaf	≥10	NR	NR	≥54,780	739.71	30.18
	HI—Hard-of-Hearing	≥70	724.99	26.48	≥54,720	739.72	30.18
	Mild Mental Disability	≥220	691.05	15.93	≥54,570	739.91	30.06
	Moderate Mental Disability	<50	NR	NR	≥54,790	739.71	30.18
	Orthopedic Impairment	≥50	733.60	28.67	≥54,750	739.71	30.18
	Other Health Impairment	≥1,150	711.96	24.68	≥53,640	740.30	30.01
	Specific Learning Disability	≥3,130	708.50	20.39	≥51,660	741.59	29.63
	Speech or Language Impairment	≥670	734.88	30.19	≥54,120	739.76	30.18
	Traumatic Brain Injury	<50	NR	NR	≥54,790	739.71	30.18
	Visual Impairment	<50	NR	NR	≥54,770	739.70	30.18
	Other	<50	NR	NR	≥54,800	739.70	30.18
	HI—Hearing Impairment	<50	NR	NR	≥54,800	739.70	30.18
	Unknown	<50	NR	NR	≥54,800	739.70	30.18

Special Education Classification Scale-Score Means and Standard Deviations: English Language Arts							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
7	Gifted	≥1,510	798.16	24.49	≥50,830	742.97	34.99
	Talented	≥1,630	771.74	30.29	≥50,710	743.69	35.76
	Autism	≥230	713.72	36.98	≥52,110	744.71	35.87
	Deaf-Blindness	<50	NR	NR	≥52,350	744.57	35.94
	Developmental Delay	<50	NR	NR	≥52,340	744.57	35.94
	Emotional Disturbance	≥210	706.87	30.15	≥52,130	744.72	35.88
	HI—Deaf	<50	NR	NR	≥52,320	744.59	35.93
	HI—Hard-of-Hearing	≥50	719.93	29.21	≥52,290	744.59	35.94
	Mild Mental Disability	≥180	684.74	18.24	≥52,160	744.78	35.81
	Moderate Mental Disability	<50	NR	NR	≥52,350	744.57	35.94
	Orthopedic Impairment	≥60	736.54	31.95	≥52,290	744.58	35.94
	Other Health Impairment	≥1,140	710.98	29.82	≥51,200	745.32	35.70
	Specific Learning Disability	≥2,980	703.82	25.91	≥49,360	747.03	34.97
	Speech or Language Impairment	≥440	737.48	34.73	≥51,900	744.63	35.94
	Traumatic Brain Injury	<50	NR	NR	≥52,330	744.58	35.94
	Visual Impairment	<50	NR	NR	≥52,320	744.58	35.94
	Other	<50	NR	NR	≥52,340	744.57	35.94
	HI—Hearing Impairment	<50	NR	NR	≥52,350	744.57	35.94
Unknown	<50	NR	NR	≥52,350	744.57	35.94	
8	Gifted	≥1,560	798.59	26.27	≥49,150	743.97	35.31
	Talented	≥1,690	772.80	30.71	≥49,020	744.71	36.14
	Autism	≥260	720.32	37.45	≥50,460	745.78	36.27
	Deaf-Blindness	<50	NR	NR	≥50,720	745.65	36.32
	Developmental Delay	<50	NR	NR	≥50,720	745.65	36.32
	Emotional Disturbance	≥200	707.79	33.13	≥50,520	745.80	36.25
	HI—Deaf	<50	NR	NR	≥50,700	745.66	36.32
	HI—Hard-of-Hearing	≥50	731.02	30.74	≥50,660	745.67	36.32
	Mild Mental Disability	≥190	689.87	17.38	≥50,520	745.87	36.21
	Moderate Mental Disability	<50	NR	NR	≥50,720	745.65	36.32
	Orthopedic Impairment	≥50	730.44	41.87	≥50,670	745.67	36.31
	Other Health Impairment	≥1,050	710.89	30.06	≥49,670	746.39	36.08
	Specific Learning Disability	≥2,760	707.26	25.06	≥47,950	747.87	35.63
	Speech or Language Impairment	≥270	732.47	34.48	≥50,440	745.72	36.32
	Traumatic Brain Injury	<50	NR	NR	≥50,710	745.66	36.32
	Visual Impairment	≥30	NR	NR	≥50,690	745.66	36.32
	Other	<50	NR	NR	≥50,720	745.65	36.32
	HI—Hearing Impairment	<50	NR	NR	≥50,720	745.65	36.32
Unknown	<50	NR	NR	≥50,720	745.65	36.32	

Table 10.21 Special Education Classification Scale-Score Means and Standard Deviations: Mathematics

Special Education Classification Scale-Score Means and Standard Deviations: Mathematics							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
3	Gifted	≥960	793.89	24.35	≥51,860	741.32	32.02
	Talented	≥560	766.59	28.75	≥52,260	742.01	32.60
	Autism	≥320	720.30	34.21	≥52,500	742.41	32.60
	Deaf-Blindness	<50	NR	NR	≥52,820	742.28	32.66
	Developmental Delay	≥590	716.48	29.61	≥52,230	742.57	32.58
	Emotional Disturbance	≥80	715.05	28.94	≥52,740	742.32	32.65
	HI—Deaf	<50	NR	NR	≥52,800	742.29	32.66
	HI—Hard-of-Hearing	<50	NR	NR	≥52,780	742.29	32.66
	Mild Mental Disability	≥380	695.62	20.80	≥52,440	742.61	32.49
	Moderate Mental Disability	<50	NR	NR	≥52,800	742.30	32.65
	Orthopedic Impairment	<50	NR	NR	≥52,770	742.29	32.66
	Other Health Impairment	≥780	718.17	29.55	≥52,040	742.64	32.57
	Specific Learning Disability	≥2,210	716.06	25.38	≥50,610	743.42	32.46
	Speech or Language Impairment	≥1,770	743.90	31.59	≥51,050	742.22	32.70
	Traumatic Brain Injury	<50	NR	NR	≥52,810	742.28	32.66
	Visual Impairment	<50	NR	NR	≥52,790	742.27	32.66
	Other	<50	NR	NR	≥52,820	742.28	32.66
	HI—Hearing Impairment	<50	NR	NR	≥52,820	742.28	32.66
	Unknown	<50	NR	NR	≥52,820	742.28	32.66
4	Gifted	≥1,170	790.14	21.57	≥53,520	739.66	31.07
	Talented	≥900	762.18	25.18	≥53,790	740.38	31.72
	Autism	≥300	720.87	32.68	≥54,390	740.85	31.71
	Deaf-Blindness	<50	NR	NR	≥54,690	740.74	31.75
	Developmental Delay	<50	NR	NR	≥54,640	740.75	31.74
	Emotional Disturbance	≥140	711.99	29.16	≥54,550	740.81	31.72
	HI—Deaf	<50	NR	NR	≥54,660	740.75	31.74
	HI—Hard-of-Hearing	≥60	734.82	28.78	≥54,630	740.75	31.75
	Mild Mental Disability	≥420	695.59	17.16	≥54,260	741.09	31.58
	Moderate Mental Disability	<50	NR	NR	≥54,680	740.74	31.75
	Orthopedic Impairment	≥50	727.38	29.10	≥54,640	740.75	31.75
	Other Health Impairment	≥1,150	715.77	26.39	≥53,540	741.28	31.64
	Specific Learning Disability	≥2,820	714.24	23.41	≥51,870	742.18	31.51
	Speech or Language Impairment	≥1,390	741.89	30.83	≥53,290	740.71	31.77
	Traumatic Brain Injury	<50	NR	NR	≥54,680	740.75	31.75
	Visual Impairment	<50	NR	NR	≥54,650	740.75	31.75
	Other	<50	NR	NR	≥54,680	740.74	31.75
	HI—Hearing Impairment	<50	NR	NR	≥54,690	740.74	31.75
	Unknown	<50	NR	NR	≥54,680	740.74	31.75

Special Education Classification Scale-Score Means and Standard Deviations: Mathematics							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
5	Gifted	≥1,370	785.57	21.58	≥53,350	735.85	28.80
	Talented	≥1,260	757.72	25.87	≥53,460	736.61	29.59
	Autism	≥330	722.10	31.22	≥54,390	737.19	29.65
	Deaf-Blindness	<50	NR	NR	≥54,720	737.10	29.68
	Developmental Delay	≥50	717.04	24.59	≥54,680	737.11	29.68
	Emotional Disturbance	≥150	714.79	24.50	≥54,580	737.16	29.67
	HI—Deaf	<50	NR	NR	≥54,700	737.11	29.68
	HI—Hard-of-Hearing	≥50	729.09	25.56	≥54,670	737.10	29.68
	Mild Mental Disability	≥380	696.43	16.29	≥54,340	737.39	29.55
	Moderate Mental Disability	<50	NR	NR	≥54,720	737.10	29.68
	Orthopedic Impairment	≥80	723.54	31.69	≥54,650	737.12	29.67
	Other Health Impairment	≥1,220	715.34	22.61	≥53,500	737.59	29.64
	Specific Learning Disability	≥3,010	712.55	18.92	≥51,710	738.53	29.57
	Speech or Language Impairment	≥980	737.02	29.75	≥53,740	737.10	29.68
	Traumatic Brain Injury	<50	NR	NR	≥54,710	737.10	29.68
	Visual Impairment	<50	NR	NR	≥54,690	737.10	29.68
	Other	<50	NR	NR	≥54,720	737.10	29.68
	HI—Hearing Impairment	<50	NR	NR	≥54,730	737.10	29.68
	Unknown	<50	NR	NR	≥54,730	737.10	29.68
6	Gifted	≥1,540	782.99	21.85	≥53,160	732.49	28.49
	Talented	≥1,560	753.84	25.50	≥53,140	733.33	29.44
	Autism	≥270	716.69	28.71	≥54,440	734.00	29.51
	Deaf-Blindness	<50	NR	NR	≥54,710	733.92	29.53
	Developmental Delay	<50	NR	NR	≥54,680	733.94	29.52
	Emotional Disturbance	≥180	704.44	25.73	≥54,520	734.02	29.49
	HI—Deaf	<50	NR	NR	≥54,700	733.92	29.53
	HI—Hard-of-Hearing	≥70	718.33	26.12	≥54,640	733.94	29.53
	Mild Mental Disability	≥220	686.54	15.07	≥54,480	734.11	29.41
	Moderate Mental Disability	<50	NR	NR	≥54,710	733.92	29.53
	Orthopedic Impairment	≥50	722.34	33.86	≥54,660	733.93	29.52
	Other Health Impairment	≥1,150	707.91	24.08	≥53,550	734.48	29.38
	Specific Learning Disability	≥3,130	704.67	19.44	≥51,570	735.69	29.10
	Speech or Language Impairment	≥670	730.75	29.87	≥54,030	733.95	29.52
	Traumatic Brain Injury	<50	NR	NR	≥54,700	733.92	29.53
	Visual Impairment	<50	NR	NR	≥54,680	733.92	29.52
	Other	<50	NR	NR	≥54,710	733.92	29.53
	HI—Hearing Impairment	<50	NR	NR	≥54,710	733.92	29.53
	Unknown	<50	NR	NR	≥54,710	733.92	29.53

Special Education Classification Scale-Score Means and Standard Deviations: Mathematics							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
7	Gifted	≥1,510	776.41	20.55	≥50,570	732.12	25.09
	Talented	≥1,630	750.58	22.57	≥50,450	732.85	25.97
	Autism	≥230	717.64	28.28	≥51,850	733.48	26.02
	Deaf-Blindness	<50	NR	NR	≥52,090	733.40	26.05
	Developmental Delay	<50	NR	NR	≥52,080	733.40	26.05
	Emotional Disturbance	≥210	710.09	22.24	≥51,870	733.50	26.02
	HI—Deaf	<50	NR	NR	≥52,060	733.41	26.05
	HI—Hard-of-Hearing	≥50	718.95	22.32	≥52,030	733.42	26.05
	Mild Mental Disability	≥180	691.84	14.33	≥51,900	733.55	25.97
	Moderate Mental Disability	<50	NR	NR	≥52,090	733.40	26.05
	Orthopedic Impairment	≥60	724.48	23.39	≥52,030	733.41	26.05
	Other Health Impairment	≥1,140	709.69	22.14	≥50,940	733.94	25.88
	Specific Learning Disability	≥2,970	705.78	18.53	≥49,110	735.08	25.49
	Speech or Language Impairment	≥440	728.70	26.01	≥51,640	733.44	26.05
	Traumatic Brain Injury	<50	NR	NR	≥52,070	733.41	26.05
	Visual Impairment	<50	NR	NR	≥52,060	733.41	26.05
	Other	<50	NR	NR	≥52,080	733.41	26.05
	HI—Hearing Impairment	<50	NR	NR	≥52,090	733.40	26.05
	Unknown	<50	NR	NR	≥52,090	733.40	26.05
8	Gifted	≥650	787.31	31.40	≥43,870	728.73	33.32
	Talented	≥1,250	750.71	30.07	≥43,270	728.97	33.94
	Autism	≥250	711.20	33.65	≥44,260	729.69	34.00
	Deaf-Blindness	<50	NR	NR	≥44,520	729.58	34.03
	Developmental Delay	<50	NR	NR	≥44,520	729.58	34.03
	Emotional Disturbance	≥190	696.42	29.30	≥44,330	729.73	33.97
	HI—Deaf	<50	NR	NR	≥44,500	729.59	34.03
	HI—Hard-of-Hearing	≥50	716.72	34.71	≥44,460	729.60	34.02
	Mild Mental Disability	≥190	680.83	18.34	≥44,320	729.80	33.92
	Moderate Mental Disability	<50	NR	NR	≥44,520	729.58	34.03
	Orthopedic Impairment	<50	NR	NR	≥44,470	729.60	34.02
	Other Health Impairment	≥1,030	702.70	28.02	≥43,480	730.22	33.90
	Specific Learning Disability	≥2,750	699.31	24.27	≥41,770	731.58	33.63
	Speech or Language Impairment	≥260	719.81	33.96	≥44,260	729.64	34.02
	Traumatic Brain Injury	<50	NR	NR	≥44,510	729.59	34.02
	Visual Impairment	<50	NR	NR	≥44,490	729.60	34.03
	Other	<50	NR	NR	≥44,520	729.58	34.03
	HI—Hearing Impairment	<50	NR	NR	≥44,520	729.58	34.03
	Unknown	<50	NR	NR	≥44,520	729.58	34.03



## 10.4 Mode Effect Study

It is also important to evaluate fairness in test administration in addition to evaluating fairness by examining performance among subgroups. The 2019 LEAP 2025 ELA and mathematics tests were administered as both paper-based tests (PBTs) and computer-based tests (CBTs) for grades 3 and 4. The *Standards* indicate that results across different testing modes should be comparable. The mode comparability for the 2019 LEAP 2025 CBT and PBT in grades 3 and 4 was investigated using the following steps:

- The mode effect study was performed using the CBT as the focal group and the PBT as the reference group.
- The study was based on equivalent groups design. Equivalent PBT students that match CBT students were selected using propensity score matching (PSM).
- At the item level, DIF analysis was performed using the PSM samples.
- At the test level, ESs based on difference scores of scale scores between the CBT and the PBT were used to examine the mode effect.
- Similar to PARCC's decision to not apply a mode adjustment, the LDOE also decided to not apply any mode adjustment to the LEAP 2025.

### 10.4.1 Sampling Using Propensity Score Matching

The CBT was administered to a smaller number of students than the PBT in grades 3 and 4; therefore, the CBT was designated as the focal group for PSM (Rosenbaum & Rubin, 1983) and the PBT was considered the reference group. That is, all CBT students and their matching PBT students were selected using covariates (matching variables), such as the 2018 LEAP 2025 ELA and mathematics scale scores and the 2019 bio-demographic information, such as gender, ethnicity, economically disadvantaged, accommodations, and ELL. Only scale scores of the grade 3 students who took the 2018 PBT were used in this study as there are no LEAP 2025 grade 2 tests. Therefore, school means from the 2018 grade 3 tests were used to match with 2019 LEAP 2025 grade 3 school means. All grade 4 students who took the 2019 LEAP 2025 CBT were included, and a sample of matching PBT students was drawn using the R package, MatchIt for PSM.

Table 10.22 shows the number of equivalent CBT and PBT students matched by the PSM method. Only 2019 grade 4 students who took 2018 grade 3 PBT and 2019 grade 3 students whose schools took 2018 PBT were included in PSM. There were more 2019 Grade 4 students who took 2019 CBT and 2018 CBT than the 2019 Grade 4 students who took 2019 PBT and 2018 CBT. There were more 2019 Grade 3 students who took 2019 CBT and whose schools took 2018 CBT than the 2019 Grade 3 students who took 2019 PBT and whose schools took 2018 CBT. PSM cannot be applied to these students. Grade 3 had a small number of CBT students, making its matching PBT student count small. For mathematics grade 4, there were 6,201 CBT students and 43,410 PBT students who had all PSM covariate information, such as bio-demographics, 2018 ELA and mathematics performance information, and 2019 mode information. Of the 43,410 PBT students, 6,201 were selected (a number equivalent to the number of CBT students) by considering all covariates.

**Table 10.22 Number of Students Used for Propensity Score Matching**

Content	Grade	CBT	PBT	
		Total*	Total*	Selected
Mathematics	3	≥840	≥50,610	≥840
Mathematics	4	≥6,200	≥43,410	≥6,200
ELA	3	≥830	≥50,610	≥830
ELA	4	≥6,200	≥43,370	≥6,200

*\*Total: Number of students who have information for all covariates*

At the item level, DIF analysis was performed using the MH statistic by Holland and Thayer (1988). There were unique items in each ELA CBT and PBT forms, and these items were dropped from analysis. Table 10.23 shows the number of mode DIF items flagged using the same rules that are used in NAEP. For mathematics, there was one C- and one B- item in grade 3 and one item each flagged for C- and C+ for grade 4. For ELA, grade 3 had two B+ items while no items were flagged for grade 4. The negative sign indicates the CBT item was more difficult than the same PBT item.

**Table 10.23 2018 LEAP 2025 Mode DIF Statistics: Number of Flagged Items**

Content	Grade	N of Items	DIF			
			-C	C	-B	B
Mathematics	3	43	1		1	
Mathematics	4	43	1	1		
ELA	3	20				2
ELA	4	22				

Scale scores of the CBT and PBT administrations were estimated using the item parameters for score reporting, and their difference scores were calculated. ESs of the difference scores were calculated as follows:

$$ES = (\text{CBT Mean} - \text{PBT Mean}) / \sqrt{(\text{CBT VAR} + \text{PBT VAR})/2}, \text{ where } \text{VAR} = \text{SD}^2.$$

Table 10.24 shows the mean scale scores and standard deviations for the CBT and PBT administrations. When the mean scale scores were compared, the CBT appeared more difficult than the PBT for mathematics and ELA. A flag criterion of  $|0.2|$ , which can be considered a small difference criterion, was applied and grade 3 ELA was flagged. Factors considered when making the decision to apply a mode adjustment or not included that the student count available for grade 3 made the results somewhat unreliable and that a consistent pattern of which mode performed better was not established.

**Table 10.24 Mode Study Scale Score Differences and Effect Size**

Content	Grade	N of Students	PBT		CBT		Mean Diff PBS-CBT	ES	Flag > $ 0.2 $
			Mean	Std. Dev.	Mean	Std. Dev.			
Mathematics	3*	≥840	722.05	30.12	718.33	27.84	6.62	0.13	
Mathematics	4	≥6200	737.93	31.33	738.21	32.11	-0.28	-0.01	
ELA	3*	≥830	720.34	38.18	712.57	35.21	9.62	0.21	YES
ELA	4	≥6200	743.39	33.98	738.10	34.32	5.29	0.15	

*\*Less reliable due to small sample size; using 2018 school scale score mean as independent variable*

## 10.5 Summary

In summary, the overall purpose of this chapter is to address fairness concerns that are relevant to the administration of LEAP 2025 assessments. The information in this chapter addresses multiple best practices of the testing industry and is particularly related to the following standards:

**Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

**Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

**Standard 3.3** Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test. (64)

**Standard 3.4** Test takers should receive comparable treatment during the test administration and scoring process. (65)

**Standard 3.5** Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population. (65)

**Standard 3.6** Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws. (65)

**Standard 3.16** When credible research indicates that test scores for some relevant subgroups are differentially affected by construct-irrelevant characteristics of the test or of the examinees, when legally permissible, test users should use the test only for those subgroups for which there is sufficient evidence of validity to support score interpretations for the intended uses. (70)

## Appendix A—Text Complexity Placemat Template

### Worksheet: Text Complexity Analysis

Title	Author	Text Description

### Recommended Placement for Assessment: Grade 4

--



Qualitative Measures	Quantitative Measures												
<p><b>PURPOSE:</b></p> <p><b>TEXT STRUCTURE</b></p> <p><b>Organization of Main Ideas:</b></p> <p><b>Text Features:</b></p> <p><b>Use of Images:</b></p> <p><b>LANGUAGE FEATURES</b></p> <p><b>Conventionality:</b></p> <p><b>Vocabulary:</b></p> <p><b>Sentence Structure:</b></p> <p><b>KNOWLEDGE DEMANDS</b></p> <p><b>Subject Matter Knowledge:</b></p> <p><b>Intertextuality:</b></p>	<p><b>Common Core State Standards Appendix A Complexity Band Level</b> (if applicable):</p> <p><b>Lexile or Other Quantitative Measure of the Text:</b></p>												
	<p align="center"><b>Considerations for Passage Selection</b></p>												
	<p>Passage selection should be based on the ELA Content Specifications targets and the cognitive demands of the assessment tasks.</p> <p><b>Potential Challenges This Text May Pose (check all that apply):</b></p> <table border="1"> <tbody> <tr> <td><input type="checkbox"/></td> <td>Accessibility</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Sentence and text structures</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Archaic language, slang, idioms, or other language challenges</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Background knowledge</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Bias and sensitivity issues</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Word count</td> </tr> </tbody> </table>	<input type="checkbox"/>	Accessibility	<input type="checkbox"/>	Sentence and text structures	<input type="checkbox"/>	Archaic language, slang, idioms, or other language challenges	<input type="checkbox"/>	Background knowledge	<input type="checkbox"/>	Bias and sensitivity issues	<input type="checkbox"/>	Word count
<input type="checkbox"/>	Accessibility												
<input type="checkbox"/>	Sentence and text structures												
<input type="checkbox"/>	Archaic language, slang, idioms, or other language challenges												
<input type="checkbox"/>	Background knowledge												
<input type="checkbox"/>	Bias and sensitivity issues												
<input type="checkbox"/>	Word count												

Adapted from Smarter Balanced and the 2012 ELA SCASS work

## Appendix B—Item Content and Bias Review

---

### ***English Language Arts:***

Educators reviewed items for passage set quality and overall grade-level appropriateness. Item-specific characteristics reviewed included content alignment; cognitive complexity; bias, fairness, and sensitivity; and technical design. For content alignment, educators reviewed items to determine if the item was aligned to the content and skills indicated in the associated standard(s). With regard to cognitive complexity, educators reviewed the items for grade-level appropriateness and appropriate range of difficulty. Educators also assigned a depth-of-knowledge (DOK) level based on Webb’s DOK scale and analyzed each item for appropriate source of challenge, which indicates that the most difficult part of the item is indeed the skill defined in the standard the item is purported to measure. For bias, fairness, and sensitivity, educators reviewed items to ensure that barriers to successful performance on the test items were nonexistent or were removed via suggested revisions. For technical design the different parts of the items were analyzed to ensure that each functioned as it should.

### ***Mathematics:***

Educators reviewed items for content alignment, cognitive complexity, difficulty, bias, fairness, sensitivity, and technical design. For content alignment, educators reviewed the items to determine if the item was aligned to the LSSM and/or the LEAP 2025 Evidence Statements and appropriate for the grade level. With regard to cognitive complexity, educators assigned a depth-of-knowledge (DOK) level based on Webb’s DOK scale and analyzed each item for appropriate source of challenge to estimate a difficulty level (low, medium, high), to verify that the item only measured student mastery of the aligned standard and not some other irrelevant concept or skill. For bias, fairness, and sensitivity, educators reviewed items to ensure that barriers to successful performance on the test items were nonexistent or were removed via suggested revisions. Technical design referred to analyzing the different parts of the item to ensure that each functioned as it should.

### ***Summer/Fall 2018 Item Content and Bias Review Process***


The newly-developed items available for field-testing are aligned to the Louisiana Student Standards for ELA and Mathematics and/or LEAP 2025 Mathematics Evidence Statements as determined by committees of Louisiana educators during the Summer/Fall 2018 LEAP 2025 Grades 3-8 Item Content and Bias Review. The process used to train educators on their role at the Item Review is outlined below.

1. Committee members attended a general session and received an overview of the process, which included:
  - a. Background of the review process that took place prior to acquired items being field-tested
  - b. Overview of purpose of the review:
    - i. To confirm content alignment of each item to designated Louisiana Student Standard
    - ii. To confirm grade appropriateness of each item
    - iii. To confirm cognitive complexity of each item
    - iv. To confirm the correct key(s) or response for each item
    - v. To confirm items are free of issues of bias, fairness, or sensitivity that could impact student responses to item
  - c. Overview of review process:
    - i. Committee members review each item individually and decide status of each item: accepted, accepted with revisions, or rejected.
    - ii. Group discusses and comes to consensus regarding status of each item.
    - iii. Items are either accepted or edited by group as needed.
    - iv. Items that are accepted or accepted with revisions are considered appropriate for future field-testing.
2. After the general session, committee members reported to their assigned committees, based on grade level.
3. Each group facilitator went over the review process again by first walking the group through the use of the review spreadsheet.
  - a. The ELA spreadsheet contained prepopulated columns of the following information for each item: item ID, passage, keys, primary, secondary, and tertiary standards (as applicable), and depth of knowledge.
  - b. The mathematics spreadsheet contained prepopulated columns of the following information for each item: ABBI ID, IDEAS ID, max points, Louisiana Student Standards for Mathematics (LSSM)/LEAP2025 evidence statement, secondary and tertiary standards (as applicable), estimated difficulty level (low, medium, high), cognitive complexity (depth of knowledge 1, 2, or 3), key, and item type.
4. The facilitator reviewed each column in the spreadsheet so committee members could better understand what information needed to be filled in.
5. The facilitator went over what each column represented to make sure all participants understood the information being presented.
6. The facilitator went over the columns to be filled in individually by each committee member as they reviewed an item (accept, accept with revisions, reject), and any comments they might have about the item.
7. The facilitator then reviewed the process to be used with each item.

- a. For the first couple of items, the facilitator and committee members worked through reviewing the item(s) together. They then came to group consensus as to the status of the items. Items were edited as necessary by the group.
- b. When ready, committee members reviewed each item individually (facilitator indicated how many items to do before getting back together for group discussion of each item).
- c. Once every member finished reviewing the assigned set of items, individuals shared the status they gave each item. Items were discussed one at a time until a status was assigned through group consensus to each item in a given set.
- d. This process continued for all items taken to the item review.
- e. Committee members were provided with an evaluation at the end of the meeting so DRC and the LDOE can make improvements for future item reviews.

***Item Content and Bias Review Training Slides Covering Universal Design and Bias and Sensitivity***

## Universal Design



- Universal Design Principles make test items accessible for the widest range of the population possible. Some of these include:
  - Use simple, common words instead of low-frequency words when possible.
  - Avoid irregularly-spelled words, words with ambiguous or multiple meanings, and technical terms unless they are defined and integral to the meaning.
  - Ensure clarity of noun-pronoun relationships



## Bias and Sensitivity



- In order to have fairness in assessment, it is critical to ensure test materials are free of possible barriers to success among diverse groups of test takers
- Barriers can be reduced by ensuring items:
  - do not measure irrelevant knowledge or skill
  - do not anger, upset or distract test takers
  - treat all groups of people with respect

## Most Common Forms of Bias



- stereotypical representations
- geographical bias
- socioeconomic bias
- religious bias
- gender bias
- linguistic bias
- exclusion or underrepresentation of groups

**Table B.1 LEAP 2025 2018 Item Content and Bias Review Totals**

<b>Content Area</b>	<b>Grade</b>	<b>Total Items Reviewed</b>	<b>Total Accepted Items at DOK 1</b>	<b>Total Accepted Items at DOK 2</b>	<b>Total Accepted Items at DOK 3</b>	<b>Accepted as Is</b>	<b>Accepted with Revision</b>	<b>Rejected</b>
ELA	3	32	0	27	5	23	9	0
ELA	4	39	0	30	9	21	18	0
ELA	5	32	0	22	10	12	20	0
ELA	6	34	0	17	17	10	24	0
ELA	7	32	0	19	13	19	13	0
ELA	8	33	0	20	13	18	15	0
Math	3	4	0	1	3	2	2	0
Math	4	4	0	2	2	1	3	0
Math	5	4	0	3	1	4	0	0
Math	6	4	0	0	4	0	4	0
Math	7	4	0	2	2	0	4	0
Math	8	4	0	2	2	2	2	0

## Appendix C—Item Alignment Review Process

---

### June 2018 Item Alignment Review Process

The acquired items available for selection are aligned to the Louisiana Student Standards for ELA and Mathematics and/or LEAP 2025 Mathematics Evidence Statements. Their alignment was determined by committees of Louisiana educators during the June 2018 LEAP 2025 Grades 3-8 Item Alignment Review. The process used to train educators at the Item Alignment Review is outlined below.

1. Committee members attended a general session and received an overview of the process, which included the following:
  - a. Background on items and the review process that took place prior to items being field-tested
  - b. Overview of the purpose of the review:
    - i. To confirm the content alignment of each item to a designated Louisiana State Standard
    - ii. To confirm the grade appropriateness of each item
    - iii. To confirm the cognitive complexity of each item
    - iv. To confirm that items were free of issues of bias, fairness, or sensitivity that could impact student responses to each item
  - c. Overview of the review process:
    - i. Committee members reviewed each item individually and decided the status of each item: accepted with current alignment, accepted with realignment, or rejected.
    - ii. The group discussed the items and came to a consensus regarding the status of each item.
    - iii. Items that appropriately measured the intended standard and/or evidence statement and are free of issues of bias, fairness, or sensitivity that could impact student responses were accepted and considered appropriate for inclusion in the LDOE item bank.
2. After the general session, committee members reported to their committees, which were assigned based on grade level.
3. Each group facilitator went over the review process again by first walking the group through the use of the review spreadsheet.
  - a. The ELA spreadsheet contained prepopulated columns of the following information for each item: item ID, passage, keys, primary, secondary, and tertiary standards (as applicable), and depth of knowledge.
  - b. The mathematics spreadsheet contained prepopulated columns of the following information for each item: session, ABBI ID, IDEAS ID, maximum points, Louisiana Student Standards for Mathematics (LSSM)/LEAP 2025 evidence statement, secondary and tertiary standards (as applicable), estimated difficulty level (low, medium, high), cognitive complexity (depth of knowledge 1, 2, or 3), key, and item type.
4. The facilitator reviewed each column in the spreadsheet.
  - a. The facilitator went over what each column represented to make sure all participants understood the information being presented.
  - b. The facilitator went over the columns to be filled in individually by each committee member as they reviewed an item (accept with current alignment, accept with realignment, reject),

and the place in the spreadsheet for any comments they might have about the item. If committee members did not agree with an alignment, they were asked to propose a new alignment, if possible.

5. The facilitator then reviewed the process to be used with each item.
  - a. For the first couple of items, the facilitator and committee members reviewed the item and its alignment(s) together. They then came to group consensus as to the status of the items.
  - b. When ready, committee members reviewed each item individually (the facilitator indicated how many items to do before getting back together for the group discussion of each item). This process continued for all items taken to the item alignment review.
  - c. Committee members were provided with an evaluation at the end of the meeting to help DRC and the LDOE could make improvements with future item alignment reviews.

**Table C.1 LEAP 2025 June 2018 Item Alignment Review Totals**

Content Area	Grade	Total Items Reviewed	Total Accepted Items at DOK 1	Total Accepted Items at DOK 2	Total Accepted Items at DOK 3	Accepted with Current Alignment	Accepted with Realignment	Rejected
ELA	3	169	0	105	60	153	12	4
ELA	4	110	0	77	33	105	5	0
ELA	5	168	0	105	62	146	23	1
ELA	6	190	0	113	76	167	22	1
ELA	7	150	0	104	45	136	13	1
ELA	8	165	0	128	31	132	27	6
Math	3	88	48	29	6	65	18	5
Math	4	90	68	16	3	75	12	3
Math	5	90	69	12	7	74	14	2
Math	6	86	63	12	11	84	2	0
Math	7	91	36	34	16	71	15	5
Math	8	94	47	31	16	89	5	0

**Table C.2 LEAP 2025 June 2018 ELA Grades 3-4 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	Not Specified	Teacher of English Learners (EL)
2	Female	White	Teacher
3	Female	White	Teacher
4	Female	White	Administrator (Visually Impaired [VI] students)
5	Female	White	Teacher
6	Female	White	Teacher
7	Female	African American	Teacher
8	Female	White	Teacher

**Table C.3 LEAP 2025 June 2018 ELA Grades 5-6 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	White	Special Education (SPED) Teacher
2	Female	Asian	Teacher (EL)
3	Male	Hispanic	Teacher (EL)
4	Female	White	Teacher
5	Female	African American	Supervisor
6	Female	African American	Teacher
7	Female	African American	Teacher
8	Female	White	Teacher

**Table C.4 LEAP 2025 June 2018 ELA Grades 7- 8 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	White	Teacher (SPED)
2	Female	White	Teacher (VI)
3	Female	White	Teacher (EL)
4	Female	White	Teacher
5	Male	White	District Supervisor
6	Male	African American	Teacher
7	Female	White	Instructional Coordinator
8	Female	African American	Teacher
9	Female	African American	Teacher

**Table C.5 LEAP 2025 June 2018 Mathematics Grades 3-4 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	White	Teacher (EL)
2	Female	White	Teacher (SPED – Gifted)
3	Male	White	Instructional Supervisor
4	Female	White	Instructional Supervisor
5	Female	White	Teacher
6	Female	White	Teacher (SPED – LD)
7	Female	White	Teacher
8	Female	White	Teacher
9	Female	White	Instructional Supervisor

**Table C.6 LEAP 2025 June 2018 Mathematics Grades 5-6 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	African American	Teacher (SPED – Gifted)
2	Female	White	Teacher (SPED – VI/EL)
3	Female	White	Instructional Supervisor
4	Female	Hispanic/Latino	Teacher
5	Female	White	Teacher
6	Female	White	Teacher
7	Female	White	Teacher
8	Female	Not Specified	Instructional Supervisor
9	Female	African American	Teacher
10	Female	African American	Teacher

**Table C.7 LEAP 2025 June 2018 Mathematics Grades 7-8 Item Alignment Review Committee Make-Up**

<b>Member #</b>	<b>Gender</b>	<b>Race/Ethnicity</b>	<b>Background</b>
1	Female	White	Teacher
2	Female	White	Instructional Supervisor
3	Female	African American	Teacher
4	Female	African American	Teacher
5	Female	African American	Instructional Supervisor
6	Female	African American	Teacher
7	Female	White	Teacher
8	Female	White	Teacher
9	Female	African American	Teacher
10	Female	African American	Teacher

## Appendix D—Accommodated Print Form Creation

---

### ***Guidelines for Building Accommodated Print Forms***

Careful consideration is given to all items that are used for accommodated print (AP) forms and/or braille forms. Fairness for all populations, item integrity, and student-item interaction for technology-enhanced (TE) items are factors when selecting items that will appear on an AP form. TE items used for AP are modified as described below to allow the student to interact with the item in a way similar to the online interaction, thereby maintaining both the rigor and the content being assessed.

- Drag-and-drop items in the online environment require a student to place the answer options in an interactive table. For the AP form, the student is presented with a table with the same information as the interactive table (column or row headers, any completed cells, and blank spaces) and the answer options are listed below the table (similar to the online form in which the options are listed either below or to the right of the table). For ELA drag-and-drop items, a number or letter is added in front of each of the draggers and the directions are modified to ask the student to write only the corresponding letters and/or numbers in the table rather than having a student write out long answers. In mathematics, the directions are modified to ask the student to write the correct answer in its corresponding box. Students are also able to circle the text and draw arrows to indicate where it should be placed or add labels to the answer choices and write only the label in the box, as long as the intended response is clear to the test administrator who will transcribe the answers into the online system.
- Match interaction table items in the online environment require a student to select a checkbox in one or more columns for each of multiple rows. In the AP form, the student is provided with a table and asked to mark an X in the correct places.
- Highlight-text items or item parts in the online environment require a student to click on the selected text, which highlights the selected word, phrase, or sentence. In the AP form, the text is presented in the same format and the student is asked to circle the answer. Where only certain words or phrases are selectable in the online system, those options are underlined in the AP form to indicate which words and/or phrases the student should select from.
- Drop-down menu items in the online environment have answer options in a drop-down menu format, oftentimes as part of a complete sentence. The AP form displays the item with a blank line in place of the drop-down menu in the sentence, with all the answer options for the drop-down menu presented vertically below the sentence. The directions are then modified to ask the student to circle the word/phrase that belongs in the blank.
- Short answer items in the online environment require a student to type the answer in a box. In the AP form, a box is provided for the student to write the response.
- Keypad input items in the online environment require a student to enter a numeric response including all rational and irrational numbers as well as expressions and equations. In the AP form, a box is provided for the student to write the response.
- Graphing items, including coordinate planes, number lines, line plots, and bar graphs, in the online environment require a student to complete a graph by plotting points, adding Xs to create a line plot, or raising/lowering bars to create a bar graph or histogram. In the AP form, the student is provided with the same coordinate plane, number line, line plot, or bar graph as in the online item, including titles, axis labels, and keys, and is asked to complete the graph.

Displaying items similarly in both print and online, and allowing the student to interact with the item in a similar manner, maintain the item integrity by assessing a similar construct in a similar manner, providing students who are unable to access the assessment online with an assessment at the same level of rigor as the online test.



AP forms are thoroughly reviewed by DRC and LDOE content experts to ensure a valid and reliable assessment for students who are unable to participate in the online assessment. These forms are also used as the source files for the creation of braille forms for students in grades 5–8 in ELA and mathematics.

## Appendix E—Transadaptation Process for Spanish Mathematics Forms

---

For English Learners, the LDOE offers the mathematics assessments in Spanish for both computer-based tests (CBT) in all grades and paper-based tests (PBT) in grades 3 and 4 only to mirror the English language forms, the text-to-speech (TTS) for CBT and large print and human voice audio CDs for PBT forms. The Spanish language versions of the test were developed through transadaptation. Transadaptation takes into consideration the grade-level appropriateness of the words and sentence structures used and the linguistic and cultural differences that exist between speakers of two different languages. Accounting for these differences allows experts to ensure that a Spanish language version of an item will measure the same construct as the English-language version of the item at the same level of rigor. The item is therefore expected to measure the achievement of English learners in the same way that the English version of the item does for native speakers of English.

Once the operational form was approved in English, DRC provided item IDs for acquired items to New Meridian, who then identified which of those items had previously appeared on a Spanish transadapted form. Once New Meridian identified the items that had previously been transadapted and provided the transadaptations of those items, DRC identified the English version of all items that had not been previously transadapted (either because they were Louisiana-owned items that would appear in field-test positions or because they were acquired items that had not been previously used on a Spanish-language form by PARCC). These items were then provided to the Spanish transadaptation subcontractor for initial transadaptation. DRC's Spanish Test Development Team reviewed the previously transadapted items to ensure consistency between those items transadapted as part of the PARCC assessments and those transadapted specifically for Louisiana. The team provided guidance to the translator conducting the initial transadaptation in grade-level and culturally appropriate ways. Upon completion of the transadaptation by the subcontractor, DRC's Spanish Test Development team conducted reviews by native Spanish speakers for content and grade-level appropriateness of the transadaptation. The team also conducted an editorial review. At least two members of DRC's Spanish Test Development team compared each English item to the Spanish transadaptation to ensure that the transadaptation:

- was accurate;
- contained grade-appropriate wording;
- contained answer choices that were reasonably parallel;
- did not introduce ambiguity into the Spanish version;
- contained graphics that were clearly transadapted;
- did not alter current teaching and learning practices in the content area; and
- remained free of gender, ethnic, cultural, socioeconomic, and regional bias.

The Spanish Test Development team then reconciled any discrepancies and submitted the transadaptations to a senior Spanish Test Development team member for resolution. After approval by the senior Spanish Test Development team member, the item moved forward to be imported into DRC's item banking system.

Both previously transadapted items and newly transadapted items were imported into DRC's item banking system and formatted for online use. Each Spanish item was paired with the corresponding English item in the item bank, and the Spanish item was formatted. Graphics for the item were then finalized for review. The

finalized transadaptation was then compared to the Spanish version of the item in the DRC assessment system and the English version of the item, and all changes were verified.

DRC's Spanish Test Development team then used the final, approved communication assistance scripts in English to transadapt descriptions of graphics as necessary. These descriptions were used when preparing the TTS forms for review. Scripting the TTS forms and reviewing the finalized Spanish forms were conducted by native Spanish speakers at DRC prior to submitting the forms to the LDOE for a translation review by a third-party translation vendor. The vendor reviewed the transadapted forms and provided feedback to the LDOE and DRC. Experienced DRC Spanish Test Development team members and the translation vendor resolved any issues, and DRC made modifications as necessary. The forms were then approved by both DRC and the LDOE translation vendor.

## Appendix F—LEAP 2025 Spring 2018 Handscoring/AI Documentation

---

# LEAP 2025 SPRING 2019

## HANDSCORING/AI DOCUMENTATION

### LEAP 2025 GRADES 3-8

ELA, Math, Science, and Social Studies

### LEAP 2025 HIGH SCHOOL AND EOC

Algebra I, Geometry, English I, English II, English III, Biology,  
and U. S. History

## Contents

Schedule, Locations, and Staffing .....	1
Training and Scoring Schedule .....	1
Scorer Degree Requirements .....	2
Training .....	3
Training Materials .....	4
EOC Biology, LEAP 2025 Biology, LEAP 205 U.S. History, and LEAP 2025 Grades 3-8 Science and Social Studies .....	4
EOC English III.....	4
LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math (Items and Materials Developed by DRC).....	5
LEAP 2025 Algebra I, Geometry, English I, English II, and Grades 3-8 ELA and Math (Items and Materials Developed by PARCC) .....	6
Algebra I, Geometry, and Grades 3-8 Math Training Set Composition .....	7
English I, English II, and Grades 3-8 ELA Training Set Composition .....	8
English I Constructed Response (CR) Items and Associated Training Materials.....	10
English II Prose Constructed Response (CR) Items and Associated Training Materials.....	10
<i>Grades 3-8 ELA CR Items and Associated Training Materials</i> .....	11
Algebra I Items and Associated Training Materials.....	12
Geometry Items and Associated Training Materials.....	13
<i>Grade 3 Math Items and Associated Training Materials</i> .....	14
<i>Grade 4 Math Items and Associated Training Materials</i> .....	14
<i>Grade 5 Math Items and Associated Training Materials</i> .....	14
<i>Grade 6 Math Items and Associated Training Materials</i> .....	15
<i>Grade 7 Math Items and Associated Training Materials</i> .....	15
<i>Grade 8 Math Items and Associated Training Materials</i> .....	15
Qualifying .....	16
EOC Biology and EOC English III .....	16
EOC Biology .....	16
EOC English III.....	16

LEAP 2025 Constructed-Response and Extended-Response Items .....	17
LEAP 2025 English I, English II, and Grades 3-8 ELA.....	17
LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math.....	17
LEAP 2025 U.S. History and Grades 3-8 Social Studies .....	18
LEAP 2025 Biology and Grades 3-8 Science .....	18
Spring 2019 Scoring Plan.....	19
LEAP 2025 High School and EOC.....	19
LEAP 2025 Grades 3-8 .....	19
Handscoring Rules.....	20
AI Scoring .....	20
Handscoring .....	20
Reader Monitoring Procedures.....	21
Team Leader Read-Behinds .....	21
Validity Sets and Inter-Rater Reliability .....	21
Calibration Sets .....	24
Handscoring Quality Control Reports .....	25
Scoring Summary Report Sample .....	25
Scoring Summary Report Sample with AI (Reader ID #3) .....	25
Reader Feedback Logs.....	26
Handling Unusual Responses.....	26
Nonscore Codes and Definitions.....	26
Nonscore Code Definitions .....	27
Nonscore Codes by Course .....	27
Alerts.....	27
Artificial Intelligence Scoring .....	28
AI Scoring – Measurement, Inc. ....	28
Model Building.....	29
Evaluation Metric.....	30
Scoring Responses with the AI Engine .....	31

Quality Control of the AI Engine .....	32
Identifying Responses for Human Review .....	32
Alert Detection System .....	33
Identification of Nonscorable Responses .....	34
Identifying Copied Text and Plagiarism with the AI Engine .....	34
AI Scoring – Pearson .....	37
The Intelligent Essay Assessor.....	37
How the Intelligent Essay Assessor was Trained .....	38
Quality Monitoring.....	40
Scoring (DRC) .....	41
Rescores .....	41
Appendix A.....	42
DRC-MI Streaming Scoring Documentation.....	42
SECTION 1 – General Information.....	43
SECTION 2 – SCHEMA SUPPLEMENT.....	44
SECTION 3 – STATUS CODE INFORMATION .....	47
Appendix B.....	48
AI Model Data – EOC English III (Spring 2019) .....	48
Quadratic Weighted Kappa (QWK) and Inter-rater Reliability (IRR).....	48
Score Point Distribution (SPD) .....	48
AI Model Data – LEAP 2025 U.S. History ER (Spring 2019) .....	49
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 49	
AI Model Building – Social Studies Grades 5-8 ERs (Spring 2019) .....	50
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 50	
AI Model CR Performance – ELA Grades 6-8, English I, and English II (Spring 2019) .....	51
Spring 2019 LEAP 2025 and EOC Items – IRR and SPD from Previous Administrations .....	52
<i>Algebra I</i> .....	52
<i>Algebra I (continued)</i> .....	53
<i>Geometry</i> .....	54



<i>Math Grade 3</i> .....	56
<i>Math Grade 4</i> .....	57
Math Grade 5 .....	58
<i>Math Grade 6</i> .....	59
<i>Math Grade 7</i> .....	60
<i>Math Grade 8</i> .....	61
English I .....	62
English II .....	63
EOC English III (All Report Data Scored by DRC and MI [AI]) .....	64
<i>ELA Grade 3</i> .....	65
<i>ELA Grade 4</i> .....	65
<i>ELA Grade 5</i> .....	66
<i>ELA Grade 6</i> .....	66
<i>ELA Grade 7</i> .....	67
<i>ELA Grade 8</i> .....	67
Biology (EOC).....	68
Biology ERs and CRs (LEAP 2025) .....	69
Grade 3 Science.....	70
Grade 4 Science.....	70
Grade 5 Science.....	70
Grade 6 Science.....	71
Grade 7 Science.....	71
Grade 8 Science.....	71
<i>U.S. History ERs and CRs (LEAP 2025)</i> .....	72
<i>Social Studies Grade 3</i> .....	73
<i>Social Studies Grade 4</i> .....	73
<i>Social Studies Grade 5</i> .....	73
<i>Social Studies Grade 6</i> .....	74
<i>Social Studies Grade 7</i> .....	74

*Social Studies Grade 8* ..... 74

# Schedule, Locations, and Staffing

All reader training and handscoring for the spring 2019 administration of LEAP 2025 grades 3-8 and high school assessments and End-of-Course (EOC) high school assessments will take place at the DRC scoring center locations noted in the table below.

## Training and Scoring Schedule

DRC's reader training and scoring schedule is based on the spring testing windows of April 1, 2019 – May 3, 2019 (LEAP 2025 grades 3-8) and April 15, 2019 – May 17, 2019 (LEAP 2025 and EOC high school).

Reader training and scoring locations and the anticipated dates for each are noted below.

Grade/Content Area or Course	DRC Scoring Center Location	Anticipated Staffing	2019 Reader Training and Scoring Window
3 ELA	Columbus, OH	2 Scoring Directors, 5 Team Leaders, 55 Readers	May 9 – June 5
4 ELA	Madison, WI	2 Scoring Directors, 6 Team Leaders, 60 Readers	May 9 – June 5
5 ELA	Madison, WI	2 Scoring Directors, 4 Team Leaders, 45 Readers	April 4 – May 10
6 ELA	Plymouth, MN	1 Scoring Director, 1 Team Leader, 13 Readers	April 1 – May 10
7 ELA	Plymouth, MN	2 Scoring Directors, 4 Team Leaders, 30 Readers	April 1 – May 10
8 ELA	Atlanta, GA	2 Scoring Directors, 4 Team Leaders, 30 Readers	April 4 – May 10
3 Math	Woodbury, MN	2 Scoring Directors, 6 Team Leaders, 60 Readers	May 9 – June 5
4 Math	Plymouth, MN	2 Scoring Directors, 4 Team Leaders, 40 Readers	May 13 – June 5
5 Math	Sharonville, OH	2 Scoring Directors, 4 Team Leaders, 40 Readers	April 4 – May 10
6 Math	Lake Mary, FL	2 Scoring Directors, 5 Team Leaders, 50 Readers	April 4 – May 10
7 Math	Woodbury, MN	2 Scoring Directors, 5 Team Leaders, 50 Readers	April 4 – May 10
8 Math	Sharonville, OH	2 Scoring Directors, 5 Team Leaders, 50 Readers	April 4 – May 10
3 Science	Indianapolis, IN	2 Scoring Directors, 5 Team Leaders, 55 Readers	May 9 – June 5
4 Science	Indianapolis, IN	2 Scoring Directors, 9 Team Leaders, 90 Readers	May 9 – June 5
5 Science	Atlanta, GA	2 Scoring Directors, 6 Team Leaders, 55 Readers	April 4 – May 10
6 Science (CRs)	Indianapolis, IN	1 Scoring Director, 4 Team Leaders, 40 Readers	April 4 – May 10
6 Science (ERs)	Indianapolis, IN	1 Scoring Director, 3 Team Leaders, 25 Readers	April 1 – May 10
7 Science	Indianapolis, IN	2 Scoring Directors, 7 Team Leaders, 65 Readers	April 4 – May 10
8 Science (CRs)	Atlanta, GA	1 Scoring Director, 3 Team Leaders, 30 Readers	April 4 – May 10
8 Science (ERs)	Plymouth, MN	1 Scoring Director, 3 Team Leaders, 30 Readers	April 3 – May 10
3 SS	Atlanta, GA	1 Scoring Director, 2 Team Leaders, 20 Readers	May 13 – June 5
4 SS	Atlanta, GA	1 Scoring Director, 2 Team Leaders, 20 Readers	May 13 – June 5
5 & 6 SS (CRs only)	Atlanta, GA	1 Scoring Director, 4 Team Leaders, 40 Readers	April 1 – May 10
7 & 8 SS (CRs only)	Atlanta, GA	1 Scoring Director, 4 Team Leaders, 40 Readers	April 1 – May 10
5, 6, 7, & 8 SS (ERs only)	Plymouth, MN	1 Scoring Director, 4 Team Leaders, 40 Readers	April 3 – May 10

<b>Grade/Content Area or Course</b>	<b>DRC Scoring Center Location</b>	<b>Anticipated Staffing</b>	<b>2019 Reader Training and Scoring Window</b>
<b>LEAP 2025 Algebra I</b>	Sharonville, OH	2 Scoring Directors, 4 Team Leaders, 40 Readers	April 11 – May 22
<b>LEAP 2025 Geometry</b>	Sharonville, OH	2 Scoring Directors, 2 Team Leaders, 25 Readers	April 11 – May 22
<b>LEAP 2025 English I</b>	Plymouth, MN	1 Scoring Director, 1 Team Leader, 15 Readers	April 11 – May 22
<b>LEAP 2025 English II</b>	Plymouth, MN	1 Scoring Director, 1 Team Leader, 23 Readers	April 11 – May 22
<b>EOC English III</b>	Plymouth, MN	1 Scoring Director, 1 Team Leader, 5 Readers	April 11 – May 22
<b>EOC Biology</b>	Plymouth, MN	1 Scoring Director, 5 Readers	April 11 – May 22
<b>LEAP 2025 Biology</b>	Plymouth, MN	1 Scoring Director, 3 Team Leaders, 30 Readers	April 15 – May 22
<b>LEAP 2025 U.S. History</b>	Plymouth, MN	1 Scoring Director, 4 Team Leaders, 40 Readers	April 10 – May 22

Each DRC scoring center is a secure facility. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials may leave the scoring center and must sign legally-binding confidentiality agreements before work begins. DRC will retain these agreements for the duration of the contract. To prevent the unauthorized duplication of secured materials, cell phone/camera use within the scoring rooms is strictly forbidden. Readers only have access to student responses they are qualified to score. Each scorer is assigned a unique username and password to access DRC’s imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

Scorers will be divided by content area or course as detailed in the previous table. Depending on the overall progress of the project, more scorers may be added to some groups. Additionally, depending on the overall progress of the project, some groups may subdivide and work on different items.

## Scorer Degree Requirements

DRC readers scoring for Louisiana have at least a four-year college degree. DRC has a Human Resources Director dedicated solely to recruiting and retaining our handscoring staff. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. During personal interviews, reader candidates are asked to demonstrate their own proficiency in writing by responding to a DRC writing topic and in mathematics by solving word problems with correct work shown. All of this results in a highly educated and diverse workforce. Our personnel files for readers and Team Leaders include evaluations for each project completed. We use these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performance on similar projects at DRC.

# Training

In preparation for the scoring of all LEAP 2025 and EOC items, DRC scoring supervisors will train readers using the same training materials that were used by previous vendors for prior administrations of the same items. These training materials originated from the sources noted below.

Reader training materials for the following were provided to DRC by Pacific Metrics and previously approved by LDOE:

- EOC Biology and English III

Reader training materials for the following were developed by DRC in conjunction with LDOE:

- LEAP 2025 grades 3-8 Science and Social Studies, as well as select items for grades 3-8 Math (noted as DRC Material Type on pages 14-15) originating with the spring 2018 DRC field test
- LEAP 2025 Biology and U.S. History, as well as select items for Algebra I and Geometry (noted as DRC Material Type on pages 12-13) originating with the spring 2018 DRC field test

Reader training materials for the following were provided to DRC by New Meridian and were approved by the Partnership for Assessment of Readiness for College and Careers (PARCC):

- LEAP 2025 grades 3-8 ELA and Math items developed by PARCC
- LEAP 2025 Algebra I, Geometry, English I, and English II items developed by PARCC

The materials include:

- Passages, items/prompts, associated stimuli for applicable content areas/courses and item types;
- Rubrics;
- Anchor Sets;
- Training Sets (or Practice Sets); and
- Qualifying Sets.

DRC will start the training with a review of passages, items/prompts, rubrics, and anchor responses, followed by the scoring and discussion of Training/Practice Sets and the scoring and discussion of Qualifying Sets. Once this process has been completed for an item or prompt, qualified readers will be able to start scoring live student responses. A group of scorers will score responses for a particular item until the scoring for that item is complete. Then they may move on to score a different item. Depending on the overall progress of the project and the current quantity of responses available to score for each item, some groups may subdivide and work on different items. Additionally, depending on the overall progress of the project, more scorers may be added to some groups when the groups are ready to score new items.

The following tables detail the composition of the training materials for all of the spring 2019 administration of the LEAP 2025 grades 3-8 and high school and EOC assessments.

## Training Materials

### *EOC Biology, LEAP 2025 Biology, LEAP 205 U.S. History, and LEAP 2025 Grades 3-8 Science and Social Studies*

Reader training for the EOC biology task is conducted using item-specific anchor sets, training sets, and qualifying sets provided by Pacific Metrics. The LEAP 2025 biology, U.S. history, and grades 3-8 science and social studies item-specific training materials were developed by DRC.

<b>Set Type*</b>	<b>Biology, U.S. History, and Grades 3-8 Science and Social Studies Training Materials</b>	<b>Annotated</b>
Anchor Set	Most item-specific anchor sets contain at least two responses per score point (with at least one example of each of the top scores).*	Yes
Training Sets	There are at least two training sets for each item <ul style="list-style-type: none"> <li>● 10 responses per training set</li> <li>● All numeric score points are represented*</li> </ul>	No
Qualifying Sets	There are two qualifying sets for each item <ul style="list-style-type: none"> <li>● 10 responses per qualifying set</li> <li>● All numeric score points are represented*</li> </ul>	No
*Examples of responses at the top score points or for all score-point combinations may not be present in some anchor, training, and qualifying sets as there may have been few or no examples found during rangefinding or subsequent field test scoring. In such cases, DRC Scoring Directors will identify examples of these scores during live scoring to supplement reader training.		

### *EOC English III*

For English III, the Content and Style dimensions will be trained using prompt-specific scoring guides, training sets, and qualifying sets provided by Pacific Metrics.

<b>Set Type</b>	<b>Content and Style Dimension Training Materials</b>	<b>Annotated</b>
Content and Style Anchor Set*	20 responses representing all numeric score points and including nonscore condition codes	Yes
Content and Style Training Sets 1 – 3	30 responses across the three training sets All numeric score points are represented in each set. Readers will score each response for both Content and Style.	No
Content and Style Qualifying Sets 1 – 3	40 responses across the three qualifying sets All numeric score points are represented in each set. Readers will score each response for both Content and Style.	No
* Some responses appear more than once in the Anchor Set to illustrate both a Content and a Style score.		

EOC English III training materials covering the four elements of the Conventions dimension (Sentence Formation [F], Usage [U], Mechanics [M], and Spelling [S]) are made up of sets that include examples from multiple writing prompts. Roughly half of the responses in each set are a score point 0 for each element and a score point 1 for each element.

<b>Set Type</b>	<b>Conventions (F, U, M, S) Element Training Materials</b>	<b>Annotated</b>
Conventions Anchor Set*	6 responses for each Conventions element representing both numeric score points (three 0s and three 1s for each of the four elements)	Yes
Conventions Training Sets 1 – 2	20 responses across both training sets Both numeric score points are represented in each set. Readers will score each response for all four elements.	No
Conventions Qualifying Sets 1 – 2	20 responses across both training sets Both numeric score points are represented in each set. Readers will score each response for all four elements.	No
* Some responses appear more than once in the Conventions Anchor Set to illustrate scores in more than one element.		

*LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math (Items and Materials Developed by DRC)*

Training materials for math items developed by DRC and field tested in spring of 2018 are made up of item-specific anchor sets, training sets, and qualifying sets developed by DRC.

<b>Set Type*</b>	<b>Algebra I, Geometry, and Grades 3-8 Math Training Materials</b>	<b>Annotated</b>
Anchor Set	Each item-specific anchor set contains at least two responses per score point (with at least one of each of the top score points).	Yes
Training Sets	There are two training sets for each item representing the range of responses <ul style="list-style-type: none"> <li>● 10 responses per training set</li> <li>● All numeric score points are represented*</li> </ul>	No
Qualifying Sets	There are three qualifying sets for each item <ul style="list-style-type: none"> <li>● 10 responses per qualifying set</li> <li>● All numeric score points are represented*</li> </ul>	No
*Examples of responses at the top score points may not be present in some anchor, training, and qualifying sets as there may have been few or no examples found during rangefinding or subsequent field test scoring. In such cases, DRC Scoring Directors will identify examples of these scores during live scoring to supplement reader training.		

## *LEAP 2025 Algebra I, Geometry, English I, English II, and Grades 3-8 ELA and Math (Items and Materials Developed by PARCC)*

For all LEAP 2025 English I, English II, and grades 3-8 ELA items and a portion of the LEAP 2025 Algebra I, Geometry, and grades 3-8 math items (exceptions are referred to on page 5 in the *LEAP 2025 Algebra I, Geometry, and grades 3-8 Math [Items Developed by DRC]* section and specifically noted on pages 11-15), DRC will use the PARCC-approved training and qualifying materials provided by New Meridian. Training materials for each item can be grouped into one of two categories: “prototype” item materials or “abbreviated” item materials.

### **Prototype Item Materials**

Some PARCC items included in the Louisiana forms are prototype items, meaning they have full sets of associated training materials, including Anchor Sets, Practice Sets, and Qualifying Sets. DRC will start the training with a review of passages/items, rubrics, and anchor responses, followed by the scoring and discussion of Practice Sets and the scoring and discussion of Qualifying Sets. Once this process has been completed for a prototype item included on the Louisiana form, qualified readers will start scoring live student responses for that item.

### **Abbreviated Item Materials**

Unlike prototype items, abbreviated PARCC item training materials have only two item-specific Practice Sets and no Qualifying Sets; therefore, abbreviated items require a two-step training/qualifying process. First, scorers will train and qualify as described in the Prototype Item Materials section above using PARCC-approved training materials for an associated prototype item that is similar to the abbreviated one they will be scoring on the Louisiana form.<sup>1</sup> Readers who do not qualify on the prototype item will not be allowed to continue the training.

After qualifying on the associated prototype item, readers receive additional item-specific training on the abbreviated item they are going to score. This consists of an item-specific Anchor Set and two item-specific Practice Sets. After completing the abbreviated item’s training, readers may begin scoring live responses for the abbreviated item.

---

<sup>1</sup> Item associations were determined by PARCC and Pearson with the understanding that aspects of training are generalizable across similar items. For mathematics, the determination of prototype versus abbreviated items was made by PARCC and Pearson based on similar item types and by evidence statements. For ELA items, this determination by PARCC and Pearson was based on grade/course and task type.



The following tables detail the composition of the training materials provided by New Meridian for math and ELA:

*Algebra I, Geometry, and Grades 3-8 Math Training Set Composition*

<b>Set Type</b>	<b>Mathematics Prototype Item Training</b>	<b>Mathematics Abbreviated Item Training</b>	<b>Annotated</b>
Anchor Set	3 responses per score point (Composite items will have 3 responses per composite score)	3 responses per score point (Composite items will have 3 responses per composite score)	Yes
Practice Set 1	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Practice Set 2	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Qualifying Set 1	10 responses comparable to the anchor set responses		No
Qualifying Set 2	10 responses comparable to the anchor set responses		No
Qualifying Set 3	10 responses comparable to the anchor set responses		No

*English I, English II, and Grades 3-8 ELA Training Set Composition*

<b>Set Type</b>	<b>English Prototype Item Training</b>	<b>English Abbreviated Item Training**</b>	<b>Annotated</b>
Anchor Set (for the RCWE and WE traits)	3 responses per score point <ul style="list-style-type: none"> <li>Anchor Sets for prototype RST and LAT item training include scores for the combined trait Reading Comprehension and Written Expression (RCWE).</li> <li>Anchor sets for prototype NWT item training include scores for Written Expression (WE).</li> </ul>	3 responses per score point <ul style="list-style-type: none"> <li>Anchor Sets for abbreviated RST and LAT item training include scores for the combined trait Reading Comprehension and Written Expression (RCWE).</li> <li>Anchor Sets for abbreviated NWT item training include scores for Written Expression (WE).</li> </ul>	Yes
Practice Set 1	5 responses representing the range of responses for <ul style="list-style-type: none"> <li>the RCWE trait (for LAT and RST items)</li> <li>the WE trait (for NWT items)</li> </ul>	10 responses representing the range of responses for both traits appropriate to the task type	Yes
Practice Set 2	5 responses representing the range of responses for the Knowledge of Language and Conventions trait	10 responses representing the range of responses for both traits appropriate to the task type	Yes
Practice Set 3	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Practice Set 4	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Qualifying Set 1	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Qualifying Set 2	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Qualifying Set 3	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Direct Copy Set*	3-5 responses composed entirely or partially of text copied from passage or passages (includes both traits appropriate to the task type)	3-5 responses composed entirely or partially of text copied from passage or passages (includes both traits appropriate to the task type)	Yes

\*The PARCC-approved Direct Copy sets provide additional annotated sample responses that explain the scoring rationale for responses composed entirely or partially of text copied from the source passage(s) associated with an item. DRC scoring supervisors review these item-specific sets with the readers prior to scoring the associated item.

\*\*Some English abbreviated item training sets approved by PARCC were for items that have previously been field tested only. The abbreviated (FT) training materials that were provided to DRC for these ELA CRs consist of a full-length anchor set with some annotations and a five-response practice set (unannotated). The full range of score points may not be represented in some anchor sets or practice sets.

Set Type	English Prototype and Abbreviated Item Training	Annotated
Anchor Set (for the Knowledge of Language and Conventions trait)	<ul style="list-style-type: none"> <li>● There are 3 responses per score point in each set.</li> <li>● There are two mixed-prompt Anchor Sets per grade level (one set for NWT item training, another set for LAT/RST item training). These sets are not exclusive to specific prototype or abbreviated items; they are intended to familiarize readers with the conventions features appropriate to each task type.</li> <li>● Subsequent Practice Sets for prototype and abbreviated items will require readers to practice scoring the Knowledge of Language and Conventions trait along with the RCWE trait (for LAT or RST) or with the WE trait (for NWT).</li> <li>● In addition, readers will be required to qualify on the Knowledge of Language and Conventions trait during each prototype item qualifying session.</li> </ul>	Yes

Some items selected for use on the spring 2019 administration were previously only field tested by PARCC. Consequently, the abbreviated training materials available for use with these items are abridged versions of typical abbreviated sets of materials. They consist of:

- Anchor Set (for ELA, some have annotations and some lack examples of the top scores)
- One Practice Set of 5 responses (scored but unannotated in the case of ELA)
- Approximately 10 validity responses

Since these materials are somewhat limited compared to typical abbreviated materials, DRC will bolster the training by using the PARCC-approved field test validity responses provided by New Meridian as additional practice responses. DRC Scoring Directors will then pull additional validity responses from operational Louisiana responses to use during the scoring window. The Scoring Directors will also find examples of higher-scoring responses that might be missing from the field test anchors. The validity and additional exemplar responses, along with the DRC Scoring Directors’ notes for all papers used during the training of the abbreviated (FT) items, will be submitted to LDOE for approval.

While the field test-only abbreviated item materials are somewhat limited compared to regular abbreviated materials (the main difference being a lack of formal written annotations and fewer practice responses), using the PARCC-approved validity responses provided by New Meridian as additional practice responses is intended to help fill that gap. It is important to note that readers still must qualify via standard qualification procedures on the prototype items for all items by first going through full training with the appropriate prototype Anchor Set, Practice Sets 1-4, and Qualifying Sets 1-3 (as well as the Conventions sets).

*English I Constructed Response (CR) Items and Associated Training Materials*

Question	Form	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
7	D	LAT	902152	VH017536_2T	Abbreviated	VH037763_2T
20	D	RST	914552	GG431834057	Abbreviated	VH017542 2T
9	E	RST	914552	GG431834057	Abbreviated	VH017542 2T
14	E	NWT	983215	GG604245591	Abbreviated (FT)	6139
9	A (SR)	RST	902161	VH017542_2T	Prototype	N/A
14	A (SR)	NWT	906152	VH084830	Abbreviated	6139
9	C (AE)	RST	902194	VH017614_2T	Abbreviated	VH017542 2T
14	C (AE)	NWT	902203	6139	Prototype	N/A
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.						
Abbreviated (FT) – Item has previously only been field tested by Pearson/PARCC. Abbreviated (FT) training materials for ELA consist of a full-length anchor set with some annotations and a five-response practice set (unannotated).						

*English II Prose Constructed Response (CR) Items and Associated Training Materials*

Question	Form	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
7	D	LAT	906197	HH428127697	Abbreviated	7233 2T
20	D	RST	983688	HH607742252	Abbreviated	7121 2T
9	E	RST	983688	HH607742252	Abbreviated	7121 2T
14	E	NWT	983642	HH432845949	Abbreviated	VF908613
9	A (SR)	RST	902331	VH004490	Abbreviated	7121_2T
14	A (SR)	NWT	902354	7064	Abbreviated	VF908613
7	C (AE)	LAT	906181	HH431436431	Abbreviated	7233 2T
20	C (AE)	RST	906190	HH433954866	Abbreviated	7121 2T
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.						

*Grades 3-8 ELA CR Items and Associated Training Materials*

Grade	Question	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
3	7	RST	915227	A1598	Abbreviated (FT)	VF906000
	12	NWT	913497	AA431426588	Abbreviated	VF910093
4	7	LAT	913567	VH170170	Abbreviated	VF925727
	20	RST	982233	VH060330	Abbreviated (FT)	VF653524
5	7	LAT	801310	VF821667	Abbreviated	VF882724
	20	RST	915510	VH198972	Abbreviated (FT)	2208
6	9	RST	913715	DD502035970	Abbreviated	3538
	14	NWT	913694	D1466	Abbreviated	VH000592
7	9	RST	915582	E1567	Abbreviated (FT)	VH014400
	14	NWT	913842	EE430133306	Abbreviated	4284
8	7	LAT	913958	F1460	Abbreviated	5271
	20	RST	982327	FF506834510	Abbreviated (FT)	VH007336
<p>*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.</p>						
<p>Abbreviated (FT) – Item has previously only been field tested by Pearson/PARCC. Abbreviated (FT) training materials for ELA consist of a full-length anchor set with some annotations and a five-response practice set (unannotated).</p>						

*Algebra I Items and Associated Training Materials*

Question	Form	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
13	D	901832	3031-M44083P	Abbreviated	3003_M43111
15	D	938741	MA10144 (DRC ID)	DRC	N/A
28	D	980927	VH251952	Abbreviated	VH046614
29	D	938735	MA10137 (DRC ID)	DRC	N/A
43	D	938744	MA10147 (DRC ID)	DRC	N/A
44	D	938737	MA10139 (DRC ID)	DRC	N/A
45	D	938769	MA10178 (DRC ID)	DRC	N/A
13	E	980924	M44463	Abbreviated	VH046614
15	E	980909	M43216	Abbreviated	VH046614
28	E	980927	VH251952	Abbreviated	VH046614
29	E	980911	2679-M43312	Abbreviated	3003-M43111
43	E	901851	M41726	Abbreviated	3003-M43111
44	E	938737	MA10139 (DRC ID)	DRC	N/A
45	E	980923	M000312	Abbreviated	3003-M43111
13	A (SR)	901836	M43318	Abbreviated	3003-M43111
15	A (SR)	901882	VH196970	Abbreviated	VH046614
28	A (SR)	901859	3003-M43111	Prototype	N/A
29	A (SR)	901814	M47147	Abbreviated	2407-M41752
43	A (SR)	938769	MA10178 (DRC ID)	DRC	N/A
44	A (SR)	901848	M47287	Abbreviated	M41686
45	A (SR)	901857	VH046479	Abbreviated	2407-M41752
13	B (AE)	901832	3031 M44083P	Abbreviated	3003_M43111
15	B (AE)	901882	VH196970	Abbreviated	VH046614
28	B (AE)	901687	2407_M41752_AT	Prototype	N/A
29	B (AE)	938737	MA10139 (DRC ID)	DRC	N/A
43	B (AE)	901851	M41726	Abbreviated	3003_M43111
44	B (AE)	901705	VF883359_AT	Abbreviated	VH046614
45	B (AE)	901857	VH046479	Abbreviated	2407_M41752
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.					
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.					

*Geometry Items and Associated Training Materials*

Question	Form	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
13	D	902012	M41169	Abbreviated	VF935309
15	D	980937	M43798	Abbreviated	2904-M43021
27	D	939083	MGM0141 (DRC ID)	DRC	N/A
28	D	980942	VH236248	Abbreviated	2904-M43021
43	D	939077	MGM0135 (DRC ID)	DRC	N/A
44	D	980938	M100106	Abbreviated	VF935309
45	D	980936	VH239429	Abbreviated	2904-M43021
13	E	902012	M41169	Abbreviated	VF935309
15	E	980937	M43798	Abbreviated	2904-M43021
25	E	980929	M1000516	Abbreviated (FT)	2904-M43021
28	E	902042	3020-M44058	Abbreviated	3042-M44133
43	E	980930	M1000518	Abbreviated (FT)	2904-M43021
44	E	980938	M100106	Abbreviated	VF935309
45	E	980936	VH239429	Abbreviated	2904-M43021
13	A (SR)	901939	M43794	Abbreviated	V935309
15	A (SR)	902046	M46668	Abbreviated	3042-M44133
27	A (SR)	902027	M43233	Abbreviated	VH001716
28	A (SR)	902036	2904-M43021	Prototype	N/A
43	A (SR)	902047	VH150404	Abbreviated	V935309
44	A (SR)	939101	MGM0160 (DRC ID)	DRC	N/A
13	B (AE)	902012	M41169	Abbreviated	VF935309
15	B (AE)	902046	M46668	Abbreviated	3042_M44133
27	B (AE)	902027	M43233	Abbreviated	VH001716
28	B (AE)	902042	3020-M44058	Abbreviated	3042-M44133
43	B (AE)	902062	VH150384	Abbreviated	VF613786
44	B (AE)	939101	MGM0160 (DRC ID)	DRC	N/A
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.					
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.					
Abbreviated (FT) Material Type – Item has previously only been field tested by Pearson/PARCC. Abbreviated (FT) training materials consist of a full-length Anchor Set and a five-response Practice Set (both are annotated).					

### Grade 3 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	981736	VH054794	Abbreviated	VH093931
18	914048	M05158	Abbreviated	M00848
32	898001	N/A	DRC	N/A
33	981742	M300388PD	Abbreviated	M00848
48	914039	M02527	Abbreviated	M00848
49	981747	4127-M03599P	Abbreviated	M01883
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				

### Grade 4 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	914084	4112-M03491P	Abbreviated	0081_M00445
18	914086	M04133	Abbreviated	M03436
32	981831	M400526	Abbreviated	M03436
33	899959	N/A	DRC	N/A
48	899955	N/A	DRC	N/A
49	981927	0318-M01475	Abbreviated (FT)	M03436
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				
Abbreviated (FT) Material Type – Item has previously only been field tested by Pearson/PARCC. Abbreviated (FT) training materials consist of a full-length Anchor Set and a five response Practice Set (both are annotated).				

### Grade 5 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	914152	M03820	Abbreviated	M03555
18	914148	M03888	Abbreviated	VH141466
32	902410	N/A	DRC	N/A
33	902414	N/A	DRC	N/A
48	914195	0154-M00796	Abbreviated	VH084803
49	934015	N/A	DRC	N/A
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				



### Grade 6 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
30	981963	M25151	Abbreviated	VH122131
34	981961	VH082639	Abbreviated	VH122131
35	981954	VH139067	Abbreviated	M21577
36	981956	VH220482	Abbreviated	M21577
47	914231	1740-M23030	Abbreviated	VH122131
48	903511	N/A	DRC	N/A
49	914281	M25152	Abbreviated	VF655921
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				

### Grade 7 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
31	914362	VH083535	Abbreviated	VF643181
34	982922	M25544	Abbreviated	M20598
36	868848	M25578	Abbreviated	M20598
37	900539	N/A	DRC	N/A
47	900520	N/A	DRC	N/A
48	914339	VH151385	Prototype	N/A
49	982929	M22009	Abbreviated	M22018
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				

### Grade 8 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
31	983010	VH097312	Abbreviated	M21063
34	982987	M800114	Abbreviated (FT)	M21063
35	982999	M22203	Abbreviated	M21063
36	870899	1282-M21381	Abbreviated	M20198
42	899312	N/A	DRC	N/A
46	914381	M25425	Abbreviated	M21063
48	899329	N/A	DRC	N/A
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				
Abbreviated (FT) Material Type – Item has previously only been field tested by Pearson/PARCC. Abbreviated (FT) training materials consist of a full-length Anchor Set and a five response Practice Set (both are annotated).				
DRC Material Type – Training materials built by DRC using 2018 field test responses. These materials consist of an annotated Anchor Set, two Practice Sets, and three Qualifying Sets specific to each CR.				

# Qualifying

Scorers must demonstrate their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with true scores on qualifying sets). After each qualifying set has been scored, the DRC Scoring Director responsible for training the item will lead the scorers in a discussion of the set.

Any scorer who does not qualify by the end of the qualifying process for an item will not be allowed to score actual student work for that item.

In order to maintain scoring comparability with prior administrations of the same items, DRC will use the same qualifying standards for the spring 2019 administration of the LEAP 2025 and EOC items as were established by the vendors who scored these items previously. Qualifying standards for LEAP 2025 biology and grades 3-8 science were approved by the LDOE in February 2019.

## EOC Biology and EOC English III

Descriptions of the qualifying standards for the EOC Biology and EOC English III item types are below. These standards were established by Pacific Metrics.

### *EOC Biology*

Course	Qualifying Standard
EOC Biology	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.

### *EOC English III*

Course	Qualifying Standard
EOC English III (Content and Style)	EOC English III scorers will first qualify on the Content and Style dimensions before moving on to qualify in the Conventions dimension. Each reader will complete at least two qualifying sets, and a score of 70% exact agreement or higher is required in each dimension in order to qualify. Since readers complete two sets, they may qualify on one dimension in the first set and the other dimension in the second set.
EOC English III (Conventions)	Once qualified for Content and Style, readers must then qualify for each of the four elements (F, U, M, and S) that make up the Conventions dimension. An exact agreement rate of 80% or higher is required once on each of the individual Conventions elements. A scorer may qualify on some elements in the first qualifying set and the remaining elements in the second qualifying set.

## LEAP 2025 Constructed-Response and Extended-Response Items

For all LEAP 2025 ELA and math CR items, DRC will follow the same qualification standards followed by Pearson for PARCC. A description of these qualifying standards is below.

### LEAP 2025 English I, English II, and Grades 3-8 ELA

Course	Qualifying Standard	
English I, English II, and Grades 3-8 ELA	Perfect Agreement	Perfect Plus Adjacent Agreement
	70% average for both traits on two of three qualifying sets	96% across the three qualifying sets combined on both traits
	70% on each trait at least once across three qualifying sets	

Readers of English I, English II, and grades 3-8 ELA responses are required to meet all three of the qualifications listed in the table. Perfect Plus Adjacent Agreement of 96% means that out of the entire pool of scores that a reader gives across the three qualifying sets for an item, no more than 4% of those scores can be non-adjacent. In other words, no more than 2 of the 60 applied scores can be non-adjacent (3 sets x 10 responses/set x 2 traits = 60 applied scores).

### LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math

Course	Qualifying Standard		
Algebra I, Geometry, and Grades 3-8 Math	Comprehensive	Perfect Agreement	Perfect Plus Adjacent Agreement
	0, 1, 2, 3 Rubric	70% on two of three sets	96% on two of three sets
	0, 1, 2, 3, 4 Rubric	70% on two of three sets	95% on two of three sets

Course	Qualifying Standard		
Algebra I, Geometry, and Grades 3-8 Math	Composite (multi-part) Items*	Perfect Agreement	Perfect Plus Adjacent Agreement
	0, 1 Rubric	90% on two of three sets	100% on two of three sets
	0, 1, 2 Rubric	80% on two of three sets	96% on two of three sets
	0, 1, 2, 3 Rubric	70% on two of three sets	96% on two of three sets
	0, 1, 2, 3, 4 Rubric	70% on two of three sets	95% on two of three sets

\*For mathematics composite items, the appropriate qualifying standard should be achieved on each part of the item. For example, if an item has Part A with a top score of 1, Part B with a top score of 2, and Part C with a top score of 3, a scorer/supervisor would need to achieve 90% perfect agreement on Part A, 80% perfect agreement on Part B, and 70% perfect agreement on Part C, with no more than one nonadjacent score per part across all three qualifying sets.

*LEAP 2025 U.S. History and Grades 3-8 Social Studies*

Course and Item Type	Qualifying Standard
<b>U.S. History and Grades 3-8 Social Studies</b> 0-2 point CRs	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
<b>U.S. History and Grades 5-8 Social Studies</b> 0-8 point, 2-dimension ERs (Content, 0-4; Claims, 0-4)	Scorers must qualify with 70% exact agreement or higher in both the Content trait and the Claims trait on one or more of the qualifying sets in order to score student responses. Since scorers complete two sets, they may qualify on one trait in the first set and the other trait in the second set.

*LEAP 2025 Biology and Grades 3-8 Science*

Course and Item Type	Qualifying Standard	
<b>Biology and Grades 3-8 Science</b> 0-2 point CRs	0-2 Rubric	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
<b>Biology and Grades 3-8 Science</b> Composite (multi-part) ER items*	0-1 Rubric	Scorers must qualify with 90% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-2 Rubric	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-3 Rubric	Scorers must qualify with 70% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-4 Rubric	Scorers must qualify with 70% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-5 Rubric	Scorers must qualify with 70% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-6 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-7 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0-8 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
<b>Grades 3 and 4 Science</b> Comprehensive (single part) ER items	0-6 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
<b>Biology and Grades 5-8 Science</b> Comprehensive (single part) ER items	0-9 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.

\*Qualifying Sets are made up of 10 responses comparable to the Anchor Set responses. For multi-part (composite) Biology and Grades 3-8 Science ERs, the appropriate qualifying standard should be achieved on each part of the item. For example, if an item has Part A with a top score of 6 and Part B with a top score of 3, a scorer would need to achieve 60% perfect agreement on Part A and 70% perfect agreement on Part B on one or more of the qualifying sets. A scorer may qualify on one part in the first qualifying set and the other part in the second qualifying set.

# Spring 2019 Scoring Plan

The charts below provide an overview of the Spring 2019 LEAP 2025 and EOC scoring plan, detailing the types of scoring that will be done for each course/grade.

## LEAP 2025 High School and EOC

Course	Handscoring Only	AI Scoring	AI Vendor
LEAP 2025 English I	NWT_GG604245591 (Form E) RST_VH017614_2T (Form C – AE) NWT_6139 (Form C – AE) RST_VH017542_2T (Form A – SR) NWT_VH084830 (Form A –SR)	LAT_VH017536_2T (Form D) RST_GG431834057 (Forms D & E)	Pearson
LEAP 2025 English II	RST_HH607742252 (Forms D & E) LAT_HH431436431 (Form C – AE) RST_HH433954866 (Form C – AE) NWT_7064 (Form A – SR) RST_VH004490 (Form A – SR)	LAT_HH428127697 (Form D) NWT_HH432845949 (Form E)	Pearson
EOC English III	Writing Prompt (Form X – AE)	Writing Prompt (Form W)	Measurement Inc.
LEAP 2025 Algebra I	All CRs	N/A	
LEAP 2025 Geometry	All CRs	N/A	
LEAP 2025 Biology	All CRs and ERs	N/A	
EOC Biology	ERs (operational and AE form)	N/A	
LEAP 2025 U.S. History	All CRs, ER (AE form)	ER (operational)	Measurement Inc.
Note: All Administrative Error [AE] form items are handscored by DRC scoring supervisors.			

## LEAP 2025 Grades 3-8

Course	Handscoring Only	AI Scoring*	AI Vendor
ELA grade 3	Both CRs	N/A	
ELA grade 4	Both CRs	N/A	
ELA grade 5	Both CRs	N/A	
ELA grade 6	N/A	Both CRs	Pearson
ELA grade 7	RST_E1567	NWT_EE430133306	Pearson
ELA grade 8	RST_FF506834510	LAT_F1460	Pearson
Math grades 3-8	All CRs	N/A	
Science grades 3-8	All CRs and ERs	N/A	
Social Studies grades 3 and 4	All CRs	N/A	
Social Studies grades 5-8	All CRs	All ERs	Measurement Inc.
*DRC's handscoring teams will provide a second read for at least ten percent of all AI-scored responses.			

# Handscoring Rules

## AI Scoring

For EOC English III and the LEAP 2025 U.S. History and grades 5-8 Social Studies ER items, Measurement Incorporated's (MI) Project Essay Grade (PEG) AI scoring system will provide the first score (the score of record). For select CRs in LEAP 2025 English I, English II, and grades 6-8 ELA (see Spring 2019 Scoring Plan on page 19), Pearson's Intelligent Essay Assessor (IEA) will provide the first score (the score of record). DRC's handscoring teams will provide a second read for at least ten percent of these responses in order to capture the inter-rater reliability statistics that will be used to manage scoring consistency of both the AI scoring systems and the handscoring teams. Scoring Directors will also review nonscores, alerts, and flagged responses as required. (For additional information about the nonscore, alert, and flagged response review process, please see the Handling Unusual Responses section starting on page 26.) The AI scoring process is discussed in-depth later in this document.

## Handscoring

All scores for handscored items (noted as Handscoring Only in the Spring 2019 Scoring Plan) will be provided by DRC's handscoring team. The first score will be the score of record. Ten percent of the responses will be scored twice to monitor and maintain inter-rater reliability. Scoring Directors will review all nonscores and alerts.

In addition, per PARCC/Pearson rules for ELA and math, if the first two scores are nonadjacent (e.g., 0, 2), a third, independent reading by a Team Leader or Scoring Director will be conducted for additional quality control monitoring. In the unlikely event that a response receives three nonadjacent scores (i.e., 0, 2, 4), a Scoring Director or Project Manager will review the response and provide retraining as needed.

Calculating the Final Score:

- The score associated with the first scorer is always the score of record, regardless of how many subsequent scores are applied.
- After handscoring, when the final score-processing for the ELA items takes place, the Written Expression trait score is multiplied by 3 (for the Narrative Writing Task). The Reading Comprehension and Written Expression (RCWE) trait score is multiplied by 4 (for the Literary Analysis and Research Simulation tasks), and one fourth of this weighted score will be assigned as the Reading Comprehension score, and three fourths of this weighted score will be assigned as the Written Expression score. The Knowledge of Language and Conventions score is not weighted. (For more information, please see the Scoring Rules found in the '/Scoring Documentation' folder posted on the Reporting SFTP site.)

# Reader Monitoring Procedures

## Team Leader Read-Behinds

Throughout the handscoring process, DRC Project Managers, Scoring Directors, and Team Leaders will review the statistics that are generated on a daily basis. DRC will assign one Team Leader for every 10–12 readers. (When test numbers are low and smaller groups of 10 or fewer readers are used, these groups may be supervised directly by the Scoring Director.) If scoring patterns are apparent among individual scorers, Team Leaders or Scoring Directors will handle these issues on an individual basis. If a scorer appears to need clarification of the scoring rules, DRC supervisors typically monitor one out of five of the scorer’s readings, making adjustments to that ratio as needed. If a supervisor disagrees with a reader’s scores during monitoring, he or she will correct the score and provide retraining in the form of direct feedback to the reader, using rubric language and applicable training responses. The supervisor’s corrected score becomes the score of record; it is not a second read.

DRC will also monitor the inter-rater reliability, which is to be based on the 10% of responses that receive second reads. If a scorer falls below the expected rate of agreement, the Team Leader or Scoring Director will retrain the scorer. If a scorer fails to improve after retraining and feedback, DRC will remove the scorer from the project. In this situation, DRC will remove all unreported scores that were assigned by the scorer during the period in question. These unreported responses with dropped scores will then be re-dealt and rescored.

## Validity Sets and Inter-Rater Reliability

In addition to the feedback that supervisors provide to readers based on regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC will also conduct validity scoring. For LEAP 2025 Algebra I, Geometry, English I, and English II, and grades 3-8 Math and ELA items from New Meridian, DRC will utilize the same validity responses that Pearson used. These validity responses were approved by PARCC and supplied by New Meridian.

DRC scoring supervisors will identify validity responses during live scoring for all newly operational LEAP 2025 items in Algebra I, Geometry, Biology, U.S. History, and grades 3-8 Math, Science, and Social Studies. DRC will post these validity responses and their scores to the secure LDOE SFTP site for LDOE content staff to review and approve. Validity responses previously identified and approved for EOC Biology and EOC English III will be reused.

The validity responses will be added to DRC’s image handscoring system prior to the beginning of scoring (with the exception of the previously noted LEAP 2025 validity responses, which will be identified and added during live scoring). The distribution of validity responses will be more frequent at the beginning of the scoring window and will decrease as agreement levels reveal a strong understanding and application of the scoring guidelines by the scorers. Validity reports compare scorers’ scores to pre-determined scores and can help detect potential room drift as well as individual scorer drift. This data will be used to make decisions regarding the retraining and/or release of scorers, as well as the rescoring of responses.

To monitor inter-rater reliability, DRC will produce handscoring quality control reports on a daily basis (see the sample on page 25) that provide exact, adjacent, and nonadjacent agreement rates for each reader and item on a daily and cumulative basis. These rates are calculated based on responses that are scored by two readers (or PEG or IEA—the AI scoring systems—and one reader). MI’s PEG AI scoring system will provide the first scores (the scores of record) for EOC English III and the LEAP 2025 U.S. History and grades 5-8 Social Studies ERs. For LEAP 2025 English I and English II, and select CRs in grades 5-8 ELA (see Spring 2019 Scoring Plan), Pearson’s IEA will provide the first score (the score of record). This data will be used in conjunction with scores from human-conducted second reads to calculate inter-rater reliability statistics in these content areas. Metrics and standards associated with the two AI scoring systems and their processes are described in the AI Scoring section starting on page 28. AI scores will be attributed to reader ID number 3 in the appropriate scoring reports. The calculations on these reports are:

- **Percent Exact (%EX)**—total number of responses by reader where scores are the same, divided by the number of responses that were scored twice.
- **Percent Adjacent (%AD)**—total number of responses by reader where scores are one point apart, divided by the number of responses that were scored twice.
- **Percent Non-Adjacent (%NA)**—total number of responses by reader where scores are more than one score point apart, divided by the number of responses that were scored twice.

DRC will strive to maintain the inter-rater and validity exact agreement rates at or above the percentages noted below. When a reader’s validity or inter-rater agreement falls 5% or more below these expectations, or if Perfect Agreement + Adjacent percentages fall below the rates noted, the reader will be flagged for additional monitoring and/or retraining by their Team Leader or Scoring Director. Additionally, for all items which will be AI scored, low inter-rater reliability will be investigated to see if it is an indication that the handscorers need retraining or if the AI needs retraining (see the AI Scoring section for details about AI training).



The validity and inter-rater reliability expectations for EOC and LEAP 2025 items are shown below.

<b>Agreement Rate Expectations for Validity and Inter-Rater Reliability – LEAP 2025 and EOC</b>			
<b>Content Area/Course</b>	<b>Score Point Range</b>	<b>Perfect Agreement</b>	<b>Perfect Agreement + Adjacent</b>
<b>Grades 3-8 ELA, English I, English II</b>	0-3 or 0-4 Rubric, Multi-trait	65% (each trait)	96% (each trait)
<b>EOC English III</b> (Style and Content)	1-4 (each domain)	70%	95%
<b>EOC English III</b> (F, U, M, S)	0-1 (each domain)	80%	95%
<b>Algebra I, Geometry, Grades 3-8 Math</b>	0-1 Rubric	90%	95%
<b>Algebra I, Geometry, Grades 3-8 Math</b>	0-2 Rubric	80%	95%
<b>Algebra I, Geometry, Grades 3-8 Math</b>	0-3 Rubric	70%	95%
<b>Algebra I, Geometry, Grades 3-8 Math</b>	0-4 Rubric	65%	95%
<b>EOC Biology</b>	0-4 Rubric	80%	95%
<b>LEAP 2025 Biology and Grades 3-8 Science</b> CR items	0-2 Rubric	80%	95%
<b>LEAP 2025 Biology and Grades 3-8 Science</b> Composite (multi-part) ER items	0-1 Rubric	90%	100%
	0-2 Rubric	80%	95%
	0-3 Rubric	70%	95%
	0-4 Rubric	70%	95%
	0-5 Rubric	70%	95%
	0-6 Rubric	60%	93%
	0-7 Rubric	60%	93%
	0-8 Rubric	60%	90%
<b>LEAP 2025 Grades 3 and 4 Science</b> Comprehensive (single part) ER items	0-6 Rubric	60%	93%
<b>LEAP 2025 Biology and Grades 5-8 Science</b> Comprehensive (single part) ER items	0-9 Rubric	60%	90%
<b>LEAP 2025 U.S. History and Grades 3-8 Social Studies</b> CR items	0-2	80%	95%
<b>LEAP 2025 U.S. History and Grades 5-8 Social Studies</b> 0-8 point, 2-dimension ER items (Content 0-4; Claims 0-4)	0-4 (each trait)	70%	95%

Each reader will be expected to maintain an acceptable level of exact agreement on validity responses and on inter-rater reliability as described above. Additionally, readers will be expected to maintain an acceptably low rate of nonadjacent agreement for validity and inter-rater agreement. To monitor this, we will sum each reader's percentages of exact and adjacent agreement rates and require each reader to maintain the levels displayed under "Perfect Agreement + Adjacent" in the tables on the previous page.

## Calibration Sets

Calibration sets are another means of ensuring consistency in scoring. DRC will use these sets to maintain calibration across the entire scorer population after breaks from scoring (e.g. weekends; down time between scoring periods; when moving between items/prompts). Calibration sets will also be used for an item if trends occur (e.g., low agreement between certain score points, if a certain type of response is missing from initial training).

The responses in these targeted sets help illustrate particular points and familiarize readers with the types of responses commonly seen during operational scoring. They are chosen by DRC scoring supervisors during live scoring or supplied by New Meridian (for Algebra I, Geometry, English I, English II, and grades 3-8 ELA and math). After the readers take one of these calibration sets (usually 5-10 responses), the Scoring Director will review the set from the front of the room using rubric language and the anchor responses to explain the reasoning behind each response's score. These sets do not have a passing requirement but are designed to help refocus readers on how to properly use the scoring guidelines to score responses. The Scoring Director or Team Leaders will provide individual feedback to any scorers in need of additional clarification based on their performance.

## Handscoring Quality Control Reports

### Scoring Summary Report Sample

#### EOC Biology Q000

Totals	Inter-rater Reliability				Score Point Distribution									
	2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%4	%B	%F	%R	%U
<b>Current Handscore</b>	<b>11,834</b>	<b>87</b>	<b>12</b>	<b>1</b>	<b>58,111</b>	<b>47</b>	<b>32</b>	<b>12</b>	<b>6</b>	<b>3</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
7435	500	84	14	1	2,437	42	34	11	7	6	0	0	0	0
7443	512	87	13	0	2,751	50	37	10	3	0	0	0	0	0
7762	624	88	11	1	2,968	51	28	7	4	2	0	0	0	0
8103	812	87	13	0	4,237	47	30	10	7	6	0	0	0	0
8339	447	88	11	1	2,410	52	32	9	6	1	0	0	0	0

### Scoring Summary Report Sample with AI (Reader ID #3)

#### Grade 10 English II Q000

#### Conventions

Totals	Inter-rater Reliability				Score Point Distribution										
	2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%B	%F	%N	%R	%T	%U
<b>Current Handscore</b>	<b>636</b>	<b>87</b>	<b>13</b>	<b>0</b>	<b>2,440</b>	<b>62</b>	<b>26</b>	<b>3</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>3</b>	<b>0</b>	<b>0</b>	<b>3</b>
3	318	87	13	0	2,122	63	27	3	0	0	0	2	0	0	2
11775	18	86	14	0	18	72	28	0	0	0	0	0	0	0	0
13021	36	81	19	0	36	64	31	0	0	0	0	6	0	0	0
16132	76	83	17	0	76	64	33	0	0	0	3	0	0	0	0
<b>18887</b>	<b>107</b>	<b>91</b>	<b>9</b>	<b>0</b>	<b>107</b>	<b>16</b>	<b>2</b>	<b>0</b>	<b>0</b>	<b>3</b>	<b>2</b>	<b>34</b>	<b>3</b>	<b>2</b>	<b>39</b>

## Reader Feedback Logs

Reader performance and intervention information will be tracked and updated in bi-weekly Reader Feedback Logs. These Reader Feedback Logs provide at-a-glance information about retraining actions taken with individual readers to ensure scoring consistency in regard to reliability, score point distribution, and validity performance. The logs address the following possible actions:

- Action 1—Includes one or more of the following: increase monitor rate, show and discuss examples of errant scores, pair scorer with a supervisor or stronger reader, provide additional review or training materials/recalibration
- Action 2—Rescoring of responses for which scores have not been handed off for reporting
- Action 3—Removal from scoring item

Below is an example of a Reader Feedback log:

Algebra I Q000									M/D/Yr
Reader	%EX Low	%NA High	Score Point Distribution Skewed	Validity %EX Low	Validity %NA High	Comments	Action 1	Action 2	Action 3
3782				●			●		
12860			●				●		
13296				●			●		
16070	●						●		
18961				●			●		

In addition to the Reader Feedback Logs, DRC will continue to provide the LDOE with handscoring quality control reports (the same cumulative scoring reports that we have provided in the past; samples are provided on page 25). The Scoring Summary reports show inter-rater reliability data and score point distribution information for each item (by part where appropriate).

## Handling Unusual Responses

### Nonscore Codes and Definitions

Handscored responses that cannot be assigned a score based on the rubric will be assigned a nonscore code. When readers apply nonscore codes, the responses are automatically routed to DRC handscoring supervisors for validation. Responses that receive a nonscore code count as zero points toward student scores that display on reports. The nonscore code will display in the response string that is included in the file provided to the LDOE.

If readers suspect plagiarism but have no concrete evidence, they score the response and alert it for suspected plagiarism. These responses are sent to supervisors for additional investigation. When supervisors find evidence of student-student plagiarism, each of the associated responses is scored according to rubric requirements and processed as an alert. Responses with proven student-internet plagiarism receive a score of 0 and are also processed as alerts. If supervisors cannot find definitive

proof of plagiarism in a response but suspect it to be likely, the response is scored using the rubric and processed as an alert. All responses with a possible plagiarism alert are sent to LDOE for final determination. (For additional information on final alert processing, see *Alerts* section below).

The non-score codes and the courses to which they apply are described below:

### Nonscore Code Definitions

Nonscore Code	Explanation
B	Blank/no response
F	Response is not written in English (Math responses from Spanish forms will be scored by a Spanish-qualified math scorer.)
I	Response does not contain enough original writing to evaluate. There is an insufficient amount of original writing to score and/or the response is composed of copied text. (Insufficient also means copied text that may have slight changes but does not introduce original ideas/thoughts.)
N	Don't understand/know
R	Refusal to respond
T	Off-topic
U	Incoherent, unintelligible, or undecipherable

### Nonscore Codes by Course

Course	B	F	I	N	R	T	U
LEAP 2025 Algebra I, English I, English II, Geometry, 3-8 ELA, and 3-8 math	✓	✓	N/A	✓	✓	✓	✓
EOC English III	✓	✓	✓	N/A	N/A	N/A	✓
EOC Biology	✓	✓	N/A	N/A	✓	N/A	✓
LEAP 2025 Biology, 3-8 Science, U.S. History, and 3-8 Social Studies ERs and CRs	✓	✓	✓	✓	✓	N/A	✓

## Alerts

Scorers have the ability to apply an alert flag to specific student responses. These are responses that may indicate the possibility of teacher interference, plagiarism, or disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). After setting the alert flag, which states the reason for the alert, and providing a brief description (as necessary), the reader will score the response according to the specific scoring guidelines for that item.

Likewise, PEG and IEA have the ability to detect specific alerts (described in detail later in the *Artificial Intelligence Scoring* section of this document). All alerted responses (whether identified by a human reader or by AI) are automatically routed to the Scoring Director who reviews the score and forwards appropriate responses (including grade, content area/course, lithocode, item number, and reason for alert) to senior project staff and DRC's Project Management Team for review.

If it is concluded that a response warrants an alert, DRC Project Management will contact the LDOE with the student’s LASID and post to the SFTP site the response information provided by the scoring staff for LDOE to review. If it is determined that a void is required due to plagiarism, the LDOE applies an invalidation to the record in eDIRECT. At no point during this process do scorers, Team Leaders, or Scoring Directors have access to demographic information for any students participating in the assessment. Note that the alert status of responses is not passed on in data files.

## Artificial Intelligence Scoring

As part of our comprehensive scoring solution, DRC uses two artificial intelligence (AI) scoring systems. Measurement Incorporated’s (MI) Project Essay Grade (PEG) is used to score students’ responses to the writing prompt for EOC English III and the extended-response items (ER) for LEAP 2025 U.S. History and grades 5-8 Social Studies. Pearson’s Intelligent Essay Assessor (IEA) is used to score student responses to selected constructed-response (CR) items in grades 6-8 ELA, English I, and English II.

### AI Scoring – Measurement, Inc.

The items in the following table will be AI scored by MI during the Spring 2019 administration. AI scoring models were built for each of these items by MI in the fall of 2016 or fall of 2017 and followed the model-building process described below. (Model-building data for all items included on the spring 2019 test may be found in the Appendix.)

Course	Item Type	IDEAS ID	Model Built
EOC English III	Writing Prompt	851370	Fall 2016
LEAP 2025 U.S. History	ER	892955	Fall 2017
LEAP 2025 U.S. History	ER	894104	Fall 2017
LEAP 2025 Grade 5 Social Studies	ER	807773**	Fall 2016
LEAP 2025 Grade 6 Social Studies	ER	804889*	Fall 2016
LEAP 2025 Grade 7 Social Studies	ER	805627*	Fall 2016
LEAP 2025 Grade 8 Social Studies	ER	808905*	Fall 2016
<p>*In spring 2017, human scored targeted samples of ≈ 500 responses per item used to augment and retrain the original AI models built in 2016. These samples were intended to find high score points to add to the existing AI models for the purpose of retraining the models prior to operational scoring in spring 2017.  **Similarly, the original 2016 model for grade 5 ER 807773 will be augmented prior to operational scoring in spring 2019 using a targeted sample of spring 2019 responses.</p>			

## Model Building

For each model built, PEG analyzed a set of inputs that were randomly pulled from the training set itself, which is made up of approximately 2,500 examples of student field test responses scored by expert human scorers. Specifically, the training set was divided into two independent pieces:

- One set of response data was used to train the AI engine and produce the scoring model. This attributed to 85% of the training set (~2,125 responses).
- The remaining 15% of the training set (~375 responses) was then used to validate the resulting model.

A regression model was built by choosing a set of variables (e.g., grammar, punctuation, style, etc.) and using least squares Linear Regression to find a best-fit relationship based on the training set. An algorithm chose the initial set of variables and added to the set as needed to produce a good fit, by taking into account correlation statistics and multicollinearity. Once the model was built, it was then run against the validation set, so that it could be evaluated for accuracy. Training was complete once PEG's validation set scores agreed with the human scores; however, if this level of accuracy was not met, then further iterations of training (which may involve new parameterizations or new algorithms) were used to produce a different model with higher accuracy. This process was completed for each trait that needed to be scored.

To further understand the importance of the validation set, consider that one of the risks inherent in machine learning is over-fitting the data. This means that it is possible to home in on particular elements of the responses in training data in such a way that the model does not generalize well to unseen data. To mitigate this risk, PEG uses a hold-out validation strategy<sup>2</sup> in which a randomly chosen subset of the initial training data is set aside, never used in training, but used only to evaluate the generalizability of models trained from the remainder of the set.

Validation is implicit in PEG's model training and, so, is complete for any model in production. The essential element of the process is that the models are trained on a larger subset of the training sample (approximately 85%), then validated against an entirely separate smaller subset of the training sample (approximately 15%). What is critical about this process and all validation schemes used in PEG training is that the AI's agreement is always based upon samples the AI has not encountered during training. Put another way, the samples used to train are never the same as the samples used to validate. This maximizes generalizability and minimizes the chance for over-fitting.

---

<sup>2</sup> PEG's agreements are based on a hold-out validation set pattern, as opposed to a cross-validation pattern. Cross-validation was evaluated in the past, but MI has since learned that hold-out validation provides (1) equally valid models with a massive improvement in training time, as well as (2) an easy way to ensure that the validation set remains partitioned from the rest of the training set at all times.

## Evaluation Metric

When PEG builds a model, it selects the model elements that maximize scoring accuracy for the data in question. Therefore, it is important to choose an agreement statistic on which PEG can optimize its models in such a way that the final model will exhibit reliable, accurate scoring. The inter-rater reliability of two human raters is often measured via perfect/adjacent agreement or the Pearson product-moment correlation coefficient (Pearson's  $r$ ). However, these two metrics each have significant disadvantages. Perfect/adjacent agreement is highly influenced by the overall scale and underlying distribution of the "true" scores (Williamson & Breyer, 2012), while Pearson's  $r$  is insensitive to mean difference between raters (Schuster, 2004).

MI has found that using quadratic weighted kappa, which has become the industry standard for AI scoring, as the optimization and evaluation metric leads to the most reliable and accurate scoring. Quadratic weighted kappa as a metric can detect changes in mean difference and variance between raters and is therefore well suited for comparing the accuracy of AI scoring with that of human scoring, as well as measuring the agreement of two independent human raters. For the sake of clarity in the discussion below, the quadratic weighted kappa between PEG and Reader 1 is referred to as  $\kappa\omega(\text{PEG}, \text{R1})$  and quadratic weighted kappa between Reader 1 and Reader 2 is referred to as  $\kappa\omega(\text{R1}, \text{R2})$ .

Even though quadratic weighted kappa performs well as an optimization metric, there are still some deficiencies in using it as an evaluation metric. Quadratic weighted kappa is far less influenced by the overall scale and underlying distribution of the "true" scores than perfect/adjacent agreement, but it does still display some sensitivity to those aspects of the data. In addition, while AI scoring can outperform human scoring with regard to scoring accuracy, the quality of the human scoring data has a significant impact on PEG's ability to accurately model the data. That is, a low  $\kappa\omega(\text{R1}, \text{R2})$  will usually lead to a low  $\kappa\omega(\text{PEG}, \text{R1})$ . Because of these issues with sensitivity to scale and distribution of scores and being bound by the quality of the training data scores themselves, it is difficult to give a fixed number in all scales for what an acceptable value would be for  $\kappa\omega(\text{PEG}, \text{R1})$ . In cases of four or more levels (e.g. a score ranging from 1-4, or broader) a  $\kappa\omega(\text{PEG}, \text{R1})$  of 0.7 has become a rule of thumb as a go-no-go metric. In these broader scales, a  $\kappa\omega(\text{PEG}, \text{R1})$  that is less than 0.7 to any significant degree is typically grounds for rejecting the item for AI scoring. In cases where this metric is 0.7 or above, the performance is usually considered satisfactory for AI scoring; however, other metrics such as those discussed in the next paragraph are often considered for additional information.

For instance, where the score range is smaller, such as binary (0-1) or ternary (0-2) ranges, the QWK is of more limited use, as QWK subtracts the rate of chance agreement which is quite high in the binary and ternary cases. In binary and ternary cases, the percent-exact and percent-adjacent agreements can be valuable additional metrics as they are exhaustive in these extremely-limited-range cases. Also useful in such extreme cases is to compare the human-machine agreement with the human-human agreement. In these cases the difference between  $\kappa\omega(\text{PEG}, \text{R1})$  and  $\kappa\omega(\text{R1}, \text{R2})$  can be used as an additional evaluation metric. MI defines that value as follows:

$$\Delta\kappa = \kappa\omega(\text{PEG}, \text{R1}) - \kappa\omega(\text{R1}, \text{R2})$$



When  $\Delta\kappa$  is positive, PEG's scores are more in agreement with Reader 1 than Reader 1's scores are in agreement with Reader 2. When  $\Delta\kappa$  is negative, the opposite is true; Reader 1 and Reader 2 show higher agreement levels than PEG and Reader 1. Of course, in both cases the absolute value of  $\Delta\kappa$  maintains its weight as a relative value between the two kappa values. That is, a larger  $\Delta\kappa$  means more separation between the two kappa values being compared.

The first phase of training is to maximize agreement between the PEG (machine) score and the final expert human score. If high agreement can be reached in this phase (for instance, a quadratic weighted kappa of  $\geq 0.7$ ), then the model is considered fit. The PEG team conducts secondary analysis such as this R1 vs. R2 analysis in cases where there is some question as to the fitness of the model – for instance, in a case in which PEG's quadratic weighted kappas are quite low, R1 vs. R2 analysis may be conducted to determine if the lack of agreement is a shortcoming of PEG's training, or if it is implicit in the data. This was not necessary in the current set, with the exception of the binary (i.e. zero-or-one) scores for some English traits. Analysis in this case showed not only that human-human quadratic weighted kappas in the training set were low, but, more to point, that random sets of such binary scoring showed similarly low quadratic weighted kappas. In this case, the low quadratic weighted kappa was simply an artifact of the definition of quadratic weighted kappa itself and no further R1 vs. R2 analysis was necessary.

$\Delta\kappa$  is a good metric to quickly show how accurately PEG was able to score a set of data with respect to how accurate human raters are on the same data, but MI also reports other metrics that its clients may be more familiar with, such as perfect/adjacent agreement, Pearson's  $r$ , and standard mean difference. However, since PEG was optimized on quadratic weighted kappa,  $\kappa_w$  and  $\Delta\kappa$  are the best reflections of actual performance.

### *Scoring Responses with the AI Engine*

The PEG AI scoring engine extracts and uses a large and proprietary set of linguistic feature metrics both during training and during production scoring. During training, PEG's models "learn" to represent the many complex and almost always non-linear relationships found between these linguistic features and the score points assigned by human experts. During production scoring, these same features are extracted from submitted responses. The previously trained models related to the item in question are then used to map these features to their predicted score points.

After PEG has been trained on a scored training set provided by DRC, it is available to receive batches of student responses in a mutually agreed upon format (XML or plain-text). The current preferred scoring method is to exchange XML documents via a web service. No static files are exchanged during this process. The web service supports discovery via Web Service Description Language (WSDL). The file transfer will be encrypted and will satisfy FERPA security requirements. Each record in the batch provides PEG with the student's response and a number of identifiers. The identifiers typically consist of a test ID that uniquely identifies the test, an item ID that uniquely identifies the prompt/item, and a FERPA-compliant student ID that uniquely identifies either the student or the student-test combination. The tables in Section 2 of the "DRC – Streaming Scoring" document (see Appendix) also contain information on identifiers.

When PEG receives the file, it processes the batch of responses and records the scores. Each record is specific to a student-test-item combination and will contain the item's score or a reason why it could not be scored (most commonly because the response is too short, or does not contain English). After the batch is processed, the scored records will be returned to DRC for reporting.

DRC will send files to MI daily. Scored files will typically be returned to DRC in 2 to 3 days; however, these timeframes are not definite, because they are dependent on numerous variables involved (e.g. number of responses submitted, number of different items, number of traits per item, the average response length, the standard deviation of response lengths, number of unique words submitted in each response, etc.).

Regardless of whether responses are scored by humans or machines, it is inevitable that scoring anomalies requiring human intervention will occur. Built into MI's automated scoring engine are a variety of triggers for identifying alert papers and responses in which it has low confidence. This is detailed later under "Identifying Responses for Human Review."

### *Quality Control of the AI Engine*

The guidelines below are purposefully general as they have proven to be the best practice for training the PEG engine. The PEG team followed this standard procedure in the DRC/Louisiana project and attempted to maximize human-machine quadratic weighted kappa among all holdout sets.

PEG holds out a 15% set of training data for use in validation. This holdout set is not seen by the AI during training. Instead, once training is complete, the holdout set is submitted for test evaluation and PEG's output is compared to the known, human-expert scores. As discussed in "Evaluation Metric" above, the quadratic weighted kappa has proven to be the most valuable agreement metric in PEG's recent history; however, others (e.g., exact, adjacent, and any host of others) are also applicable.

This evaluation was performed along with model building prior to operational scoring, and the results were shared with LDOE and the TAC to demonstrate sufficient scoring accuracy by PEG. For details on these results, please see pages 51-53 in the Appendix.

Once training and model building is complete, the performance of any given model is essentially deterministic (so, for a precise, given input, the output is expected to be identical). The PEG team monitors the services for unexpected events (for instance physical damage to its cloud infrastructure), and handles any data flow issues (for instance, if the client was using a different item number during live scoring than was used during training) but the AI itself does not change during live scoring. When read-behind data becomes available to the PEG team (typically this is on an annual basis), it can be used to re-evaluate and, if necessary, retrain the existing models prior to the next season of use, but such changes do not happen during live scoring. As part of our continuous improvement cycle, the analysis of this data is on-going with no current end date (i.e., items are being reviewed on a rolling basis).

### *Identifying Responses for Human Review*

Built into MI's automated scoring engine are a variety of triggers for identifying responses that require human review, including potential alerts (suspected plagiarism included) and potential nonscorable

responses (e.g., responses that are primarily copied text, lack proper development, lack enough content to be scored, or are written in an unsupported language). Many of these triggers have client-configurable thresholds. These can be set to standard defaults and then modified as needed. Thresholds are generally deliberately conservative. DRC will work with LDOE content staff and MI to look at the responses that PEG identifies for human review to make sure the high and low copied text and minimum word count settings are set appropriately. (See pages 35-36 for detailed information about these custom thresholds.)

Please note that all responses that are identified in the sections below for human review will be automatically forwarded to a DRC Scoring Director who will determine the correct score or nonscore code to apply to the response. The Scoring Director will provide the final, reported score (or nonscore) for these responses. If the Scoring Director needs assistance in determining the correct score or nonscore, DRC will work with LDOE content staff to ensure that the response is scored correctly.

### *Alert Detection System*

PEG has a robust system for detecting potential alerts, which is described in detail in this section. When PEG detects the presence of alert language, this alone does not indicate that a response is unscorable. Therefore, unless the response is unscorable for some other reason, PEG will return scores as well as the alert status code of 500 (in cases of unscorable alerts, the status code is in the range of 501-599, inclusive). Regardless of the alert flag, any responses returned with a flag to DRC will be evaluated by the handscoring supervisory team, who will determine if the response needs to be processed as an alert as described previously in this document (see *Handling Unusual Responses – Alerts*). When it is concluded that a response does warrant an alert, DRC Project Management will contact the LDOE with the student's LASID and post the response information to the SFTP site for LDOE's review.

PEG's Alert flagging system is a pattern-matching system, targeting phrases suggestive of violence towards self or others, drug or alcohol abuse, feelings of anxiety or depression or the use of weapons. This system is rules-based. It responds to concentrations of "alert language" detected within submissions. Typically, these are word counts of particularly violent or profane language often found in actionable alerts. (Such language may also be found in non-alert submissions, but PEG does not attempt to determine "intent" in these cases, rather it flags only the presence of detected verbiage.) PEG currently tracks two types of alert language that differ only in severity (e.g., a statement regarding a person "killing" is considered more severe than a statement regarding a person "beating up," but both are counted as forms of alert language). By default, PEG issues an alert flag if it encounters one instance of severe alert language or two instances of less-severe language. PEG may also issue an alert flag if high counts of profanity are found. By default, this is three instances of severely profane or five instances of less profane verbiage. Although this means that non-actionable alerts may also certainly be flagged, PEG's default settings are purposefully kept highly sensitive to alert language. These levels are configurable, however, so if the rate of return is too high or too low, adjustments can be made. For the responses that it cannot score, PEG returns a condition code to the test delivery system indicating why the response could not be scored (i.e., the response receives a tentative nonscore code that is reviewed by a Scoring Director and corrected if needed). The test delivery system can then route the flagged responses to DRC's performance assessment handscoring system. DRC will perform human handscoring for the limited number of responses that cannot be scored by AI.

With regards to the process and timing, the alerts detection is typically run in series with other essay analysis, so it is no slower (or faster) than a regular scoring. A batch of individually identified extended responses are posted to PEG's Streaming Scoring service, and at that point a response may be flagged as a potential alert. This flag takes the form of a "status code."

The rules are purposefully over-sensitive (they are more likely to give false positives than false negatives), so it is likely that the great majority of ER's flagged with a "5##" status code will not require actual intervention with the student; however, PEG is in no way capable of diagnosing this. Instead PEG just follows rules designed to sense and flag the use of language which has, in the past, been associated with alerts.

### *Identification of Nonscorable Responses*

PEG's nonscorable configurability includes the settings listed below, which can flag responses so that they are sent to DRC Scoring Directors who will determine the correct score or nonscore code to apply. These can be set to any threshold, with extreme values effectively disabling any given setting. These are the only nonscorable parameters which can be configured in this way. Each nonscorable setting relates to status codes and general rules surrounding of insufficiency and indecipherability as described below.

1. MIN\_WORDS: this controls status code 200 and may correspond to the business concept of "Insufficient" (i.e., too-short response)
2. MIN\_CORRECT\_WORD: this controls the status code 220 and is similar to the business concept of "Indecipherable" (i.e., foreign words and non-words)
3. Copied Text Low: this controls status code 605
4. Copied Text High: this controls status code 610

By adjusting each setting, PEG may impose a reasonable approximation of the scoring rules regarding Insufficiency and/or Indecipherability.

Once the scoring in the cloud is complete, the scores and statuses are sent back to the MI Delivery Service which then returns these scores and codes to DRC.

That entire process typically requires less than 100 hours (~4 days), and quite often takes less than a single day).

### *Identifying Copied Text and Plagiarism with the AI Engine*

Prior to describing the functionality PEG uses to detect copied text and plagiarized responses, an important distinction must be made between what is considered copied and what is considered plagiarized. Copied text is that which a student copies from the directions, prompt, passage(s), or reference sources supplied with an item. A response composed predominantly of text copied from item sources will not be alerted for any sort of suspected testing violation, but in most cases, it will receive a lower score (or a nonscore of "I") depending on the amount of original student writing in the response and/or how much text is copied. Responses flagged by PEG for this condition are sent to DRC scoring supervisors for review. Based on this review, EOC English III responses having an insufficient amount of

original writing to score will receive a nonscore of “I.” For LEAP 2025 U.S. History and grades 5-8 Social Studies ERs, any response having an insufficient amount of original writing to score, because it is made up entirely or almost entirely of text copied from the directions or reference sources, will also receive a score of “I” (unless the item-specific rubric makes exceptions for the use of relevant copied text).

Text that a student extracts and uses from a source external to the test itself is considered plagiarized. When PEG detects these responses (this process is explained in the next paragraph), they are also sent to DRC scoring supervisors for review, and if they are deemed to warrant an alert for suspected plagiarism, DRC’s supervisors route the responses through the same alert process described in an earlier section of this document (Handling Unusual Responses – Alerts).

PEG’s copied text and plagiarism detection functionality compares student responses to texts that students may have copied or plagiarized. To do this, per-item reference texts must be provided. For EOC English III, this is the prompt and any associated reading material provided with each test item. For the LEAP 2025 U.S. History and grades 5-8 Social Studies ERs, this includes the prompt and any associated source material (including MC/MS items) provided with each test item. In addition to external sources of plagiarism previously provided by LDOE based on results from past administrations, DRC will pre-identify other websites that may be likely sources of external plagiarism. These may include Wikipedia’s pages relevant to the topic and/or other “top hit” websites. These external sources will be used by the AI engine to identify potentially plagiarized responses. All of these text references will be added to the appropriate scoring models for each related item.

Upon receiving a response, PEG conducts a high-speed sequence scan of both the reference text and the response. Each sequence is evaluated for both the length and density of copied/plagiarized text. Length is a direct character count, and density is a measure of similarity between sequences. A verbatim copy has a density of 1.0, and a copy that contains some substitutions, additions, or deletions would likely have a density in the ~0.6 - 0.4 range. The product of these two numbers provides a value that is used to flag responses requiring human review due to large amounts of copied/plagiarized text. Clients can configure two thresholds for a low and high flag. For example, the default values for these are 50 and 100 respectively. So, a verbatim copy of 72 characters (~12 prompt words) would be reported as a low match, whereas a verbatim copy of 100 characters (roughly 16 words) would be flagged as a high match. Similarly, a copy (even with some substitutions) of 40 words would still be reported as a high match in the default setting example. The low and high matches will be flagged with status codes. This is similar to the alert flagging above. There will be a three-digit code for low-match (status code 605) and a three-digit code for high-match (status code 610).

Custom thresholds for copied text, plagiarism, and insufficient responses have been established by DRC in consultation with LDOE and were based on recommendations from MI. They are described below:

1. When PEG scans responses for copied text/plagiarism, any text copied from the supplied reference texts (regardless of whether it is contained within quotations marks) will be considered when determining if a response meets or exceeds the thresholds required for it to be routed to DRC for human review. These configurations are noted in 2a–4b on page 36.

2. EOC English III
  - a. Copied text thresholds
    - i. Low flag (status 605) – 125 characters
    - ii. High flag (status 610) – 200 characters
  - b. MIN\_WORDS (status 200) – 45 words or fewer
3. LEAP 2025 Grades 5-8 Social Studies
  - a. Copied text thresholds
    - i. Low flag (status 605) – 125 characters
    - ii. High flag (status 610) – 200 characters
  - b. MIN\_WORDS (status 200) – 25 words or fewer
4. LEAP 2025 U.S. History
  - a. Copied/plagiarized text thresholds
    - i. Low flag (status 605) – 85 characters
    - ii. High flag (status 610) – 170 characters
  - b. MIN\_WORDS (status 200) – 25 words or fewer

These settings are deliberately conservative. While some flagged responses are composed exclusively of text copied directly from source/passage material, the majority of responses that PEG flags with status codes 605 and 610 contain a combination of copied text, relevant information cited or paraphrased from the sources, and some amount of original student writing. They are flagged because they meet or exceed the copied text thresholds noted above and need to be checked by DRC scoring supervisors to determine whether they contain a sufficient amount of original student writing to evaluate. Upon review, most will be found to contain enough original writing to be considered scorable. When the supervisor determines that there is sufficient original student writing to score, and there is no evidence of plagiarism, he or she validates the original numeric scores returned by PEG and they are submitted as final scores for that response. On the other hand, if the supervisor determines that the response contains insufficient original student writing to evaluate, he or she will override PEG's scores and apply the appropriate scores or nonscores as necessary, depending on the content area scoring rules. For EOC English III and LEAP 2025 U.S. History and Social Studies, flagged responses composed entirely of text copied from item source material (or copied text combined with an insufficient amount of original student work) are given a nonscore of "I" (Insufficient).

Less frequently, responses will be flagged as potential nonscores for having too little written to be evaluated at all (status code 200). Just as DRC requires all nonscores given by human readers to be reviewed by scoring supervisors, this same requirement holds true when PEG flags responses as potential nonscores. For example, if the DRC supervisor reviews a response flagged by PEG and agrees with PEG's assessment that the response has too little writing to be assessed, the supervisor will validate the AI score of "I," and this nonscore code will be submitted as the final score for that response. On the other hand, if DRC's supervisor reviews the response, and based on the training responses provided in the handscoring training materials, he or she feels that there is enough original student writing to score, the supervisor scores the response and also overrides PEG's original nonscore, changing PEG's nonscore of "I" to the correct numeric scores. These become the scores of record.

## AI Scoring – Pearson

The items in the following table will be AI scored by Pearson during the Spring 2019 LEAP 2025 administration. AI scoring models for each of these items were previously built and used by Pearson during PARCC operational scoring. (Model-building data for all items included on the Spring 2019 test may be found in the Appendix.)

Course	Task Type	IDEAS ID	PARCC UIN	Model Built
English I	LAT	902152	VH017536_2T	2017
English I	RST	914552	GG431834057	2018
English II	LAT	906197	HH428127697	2017
English II	NWT	983642	HH432845949	2017
Grade 6 ELA	RST	913715	DD502035970	2017
Grade 6 ELA	NWT	913694	D1466	2017
Grade 7 ELA	NWT	913842	EE430133306	2017
Grade 8 ELA	LAT	913958	F1460	2017

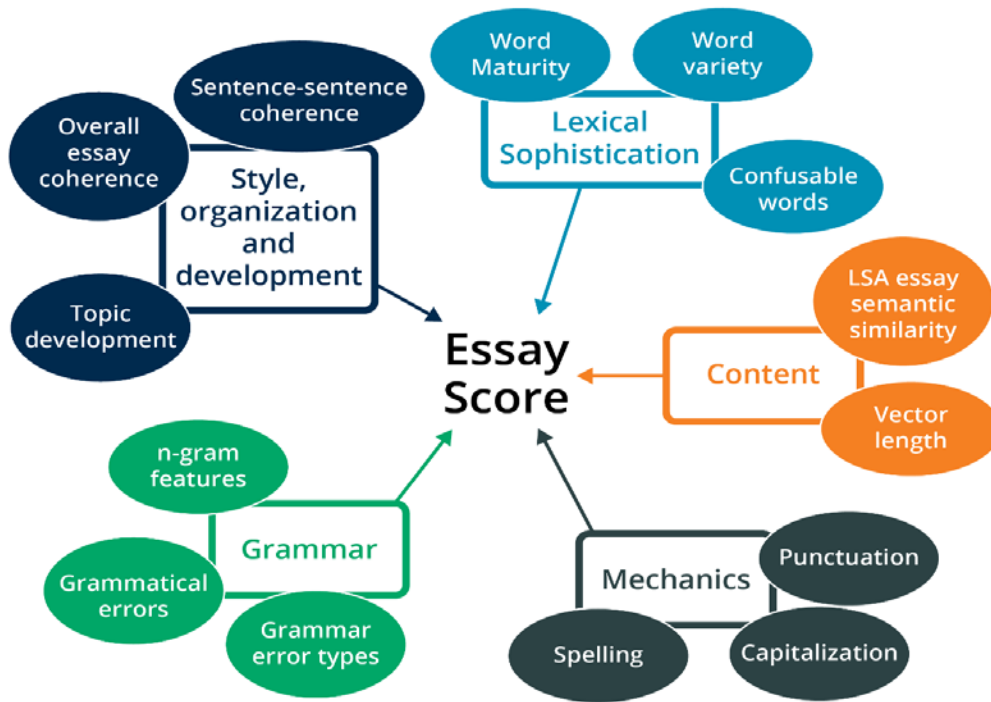
### *The Intelligent Essay Assessor*

Pearson's Intelligent Essay Assessor (IEA) uses a range of machine learning and natural language processing technologies to learn to score based on human-scored responses. One of the hallmarks of IEA is its ability to score constructed responses in content domains using Pearson's unique implementation of Latent Semantic Analysis (LSA), an approach that generates semantic similarity of words and passages by analyzing large bodies of relevant text. LSA can then "understand" the meaning of text much the same as a human scorer.

IEA's background knowledge of English is derived from a collection of texts equivalent to what students are likely to have encountered over the course of their academic career (about 12 million words). Because LSA operates over the semantic representation of texts, rather than at the individual word level, it can evaluate similarity even when texts have few words in common. For example, LSA finds the following two sentences to have a high degree of semantic similarity:

- Surgery is often performed by a team of doctors.
- On many occasions, several physicians are involved in an operation.

The following figure illustrates some of the features used in IEA and how they relate to specific constructs of student writing performance.



**Example features used in the Intelligent Essay Assessor.** Like human scorers, IEA evaluates essays for ideas, organization, development, and various grammatical and mechanics errors.

IEA is trained to associate features extracted from each essay with scores assigned by human scorers. A machine learning-based approach is used to determine the optimal set of features, and the weights for each of those features, to best model the scores for each essay. From these comparisons, IEA derives a prompt- and trait-specific scoring model that predicts the scores human scorers would assign to any new responses.

The automated scoring process mimics the approach that human scorers take when evaluating essays. Human scorers train based on anchors of annotated student responses with agreed-upon scores. Human scorers compare new responses against the anchor set of two to three examples per score point to determine the appropriate score. IEA scores essays similarly, but makes comparisons against a much larger set of examples. Rather than comparing a new essay against the 16-24 examples in an anchor set, it compares against the set of hundreds or thousands of responses on which it was trained.

### *How the Intelligent Essay Assessor was Trained*

For the ELA prompts that will be used by Louisiana, IEA was trained based on operational PARCC responses using Pearson’s Continuous Flow approach to training and scoring. When these prompts were first administered, student responses flowed to IEA even before human scoring started. IEA then selected a sample of responses for humans to score first to expedite the creation of automated scoring models. The sample included responses that represented different demographic subgroups to ensure



equity in scoring, as well as responses that were algorithmically selected to likely span the score range. As the human-scored responses flowed back to IEA, the engine automatically built potential scoring models, evaluating them against the industry standards for performance criteria included in the table below.

<b>Evaluation of Automated Scoring Systems</b>	
<b>Criterion</b>	<b>Threshold</b>
Quadratic weighted kappa (QWK)	Greater than or equal to 0.70
Pearson correlation (r)	Greater than or equal to 0.70
Standardized mean difference (SMD) between human and automated scoring	Less than or equal to  0.15
Difference in QWK or r from human-human rates	Less than or equal to 0.10
Difference in exact agreement from human-human rates	Less than or equal to 0.05

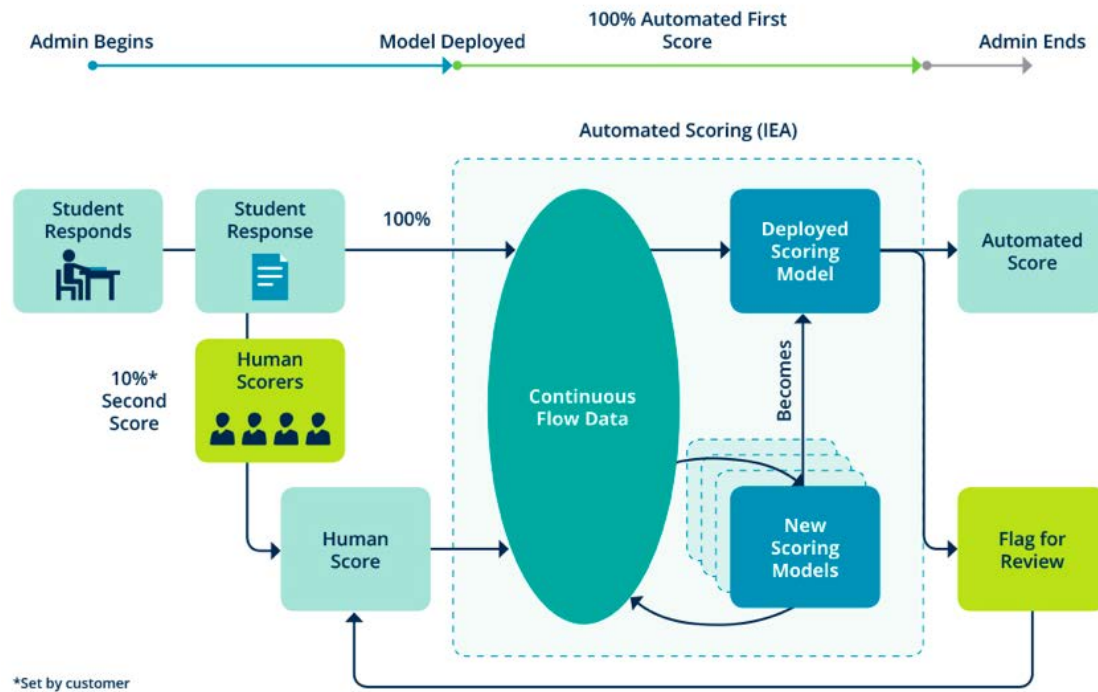
**Evaluating Automated Scoring.** *Statistical Criteria for the Evaluation of Automated Scoring Systems based on those used by Williamson et al, Smarter Balanced, and PARCC.*

While the engine was being trained, scoring and psychometrics teams met daily to review progress, quality, and next steps. When IEA met or exceeded the performance criteria for a given constructed response item, it took over as the first scorer for that item.

Responses for which IEA is less confident in its score are routed for additional human scoring. This “smart routing” of responses by the scoring engine occurs when responses fall in a particular score range for which the engine has lower agreement with human scorers, or for responses that are highly unusual or creative.

The figure on the following page depicts the entire Continuous Flow process.

## Continuous Flow Scoring



**Continuous Flow.** As student responses flowed to IEA, it selected responses for human scorers to score. As the human scores flowed back to IEA, the engine continued to try to build a scoring model that would pass the agreed upon performance criteria. Once the scoring model passed the criteria, it was deployed and began scoring all student responses, with humans applying a second score as a quality check, as well as scoring any responses flagged for review by IEA.

IEA is also trained to recognize a variety of different non-responses (e.g., non-English language, “don’t understand,” refusal to answer, off-topic, unintelligible), assigning corresponding condition codes to them or flagging them for human review when less certain. Detection of copying between students is done out of band and accomplished by using Latent Semantic Analysis to compare each student response to every other student response and flagging highly similar responses for human review. The comparison is cumulative. Every response gets checked against every other response that has been received, as they come in, within that same administration and within that prompt. Child in danger alerts are also scanned for out of band and flagged for human review.

### Quality Monitoring

Human scorers play a key role in maintaining quality throughout the scoring process starting with IEA learning to score based on their scores. Since the models for the 2019 Louisiana items are built and IEA has already established the performance characteristics necessary to accomplish first scoring, DRC human scorers will score 10% of the responses scored by IEA to monitor quality. Should agreement rates between IEA and the human scorers fall below the established agreement rates, the automated scoring model can be examined to determine the appropriate action. This action may include adjusting IEA’s confidence threshold to send more responses for human scoring or retraining the scoring engine and

rescoring student responses.

## Scoring (DRC)

DRC will use human scorers to read behind MI and Pearson's AI engines. Ten percent of the AI-scored student responses will be randomly selected to be read a second time by DRC's handscoring teams. This will provide inter-rater reliability statistics that compare the scores given by PEG and IEA to the scores given by each individual reader. Throughout the handscoring process, DRC Project Managers, Scoring Directors, and Team Leaders will review handscoring reports detailing these results.

If the inter-rater reliability (AI compared to handscoring on the 10% sample) shows exact agreement that is less than desired or nonadjacent agreement that is higher than desired, DRC will investigate and take immediate action. If scoring patterns are apparent among individual readers, scoring supervisors will deal with issues of this sort on an individual basis. If a reader appears to need clarification of the scoring rules, DRC supervisors typically monitor one out of five of the scorer's readings, making adjustments to that ratio as needed. If a supervisor disagrees with a reader's scores during monitoring, he or she will provide retraining in the form of direct feedback to the reader, using rubric language and applicable training responses.

If, however, the agreement rates for either PEG or IEA and for large numbers of readers are not as anticipated, DRC scoring experts will need to review the responses that received different scores from the AI engine(s) and from readers. Based on this, the DRC scoring experts will need to determine if they feel that the readers need to be retrained or if they are disagreeing with scores given by AI. In the unlikely scenario that DRC's scoring experts believe that they have detected unexpected trends in the scores given by PEG or IEA, DRC would take examples to LDOE and the appropriate AI vendor to review. Based on this review, if DRC, LDOE, and the vendor determined that the AI modelling was not resulting in sufficiently accurate scores, corrective measures would be put into place. Depending on the nature and timing of the issue and subsequent related LDOE policy decisions, DRC and the AI vendor will enact measures such as updating the AI modeling, providing LDOE with response information (e.g., Item ID, Student IDs, updated scale scores, updated achievement levels), and/or using expert handscorers to determine the final score for student responses.

## Rescores

The rescoring process includes automatic rescores that occur during the scoring process, as well as parent-requested rescores that take place after the official scoring window. The rescores for all subjects will be performed by expert readers.

Please refer to [LEAP 2025 HS\\_EOC Processing Rules – Scoring.pdf](#) on the LDOE Reporting SFTP site at /<YYYY> - EOC LEAP 2025 HS <Spring>/Processing Rules - Final/ for a complete description of the rescore rules and process.

# Appendix A

## DRC-MI Streaming Scoring Documentation

### **DRC – MI STREAMING SCORING SUBMIT SERVICE DOCUMENTATION**

**NOTICE:** The contents of this document and any references to external resources are intended for review only by representatives of Data Recognition Corporation, Measurement Incorporated, and LDOE, and are considered private. Technical specifications are subject to change.

**REVISED:** 2015-11-23; *created*

#### **CONTENTS:**

SECTION 1 – General Information	43
SECTION 2 – SCHEMA SUPPLEMENT	44-46
SECTION 3 – STATUS CODE INFORMATION	47

## SECTION 1 – General Information

**1.1 PURPOSE:** Submit Service accepts groups (“batches”) of constructed responses for processing by the MI Streaming Scoring product.

**1.2 SERVICE TYPE:** The Submit Service uses a standard SOAP web service interface.

**1.3 INTEGRATION:** Application-generated service definition (WSDL 1.1) document is available; WCF (Windows Community Foundation) client integration is also possible. The WSDL and WCF URLs for each environment are as follows:

### DEVELOPMENT

- WSDL:
- WCF:

### STAGING

- WSDL:
- WCF:

### PRODUCTION:

- WSDL:
- WCF:

**1.4 SERVICE SIGNATURE:** The Submit Service provides a single operation **SubmitBatch**. The operation signature – request and response structure – is defined in the WSDL. The structure of each complex type, with field descriptions and expected value ranges is described below.

## SECTION 2 – SCHEMA SUPPLEMENT

**2.1.1 SUPPLEMENTAL SCHEMA DOCUMENTATION:** The following tables are supplemental to the schema for the Submit Service, but are not, themselves, the schema. The service schema is contained within the WSDL, and may be emitted from that source to an XML schema document (XSD) through various means, though this will likely be unnecessary. To reduce confusion in terminology, the following tables will be referred to as the “supplement” or “schema supplement”.

**2.1.2 TABLE STRUCTURE:** Each table documents a specific complex type defined by the Submit Service WSDL, with each row in a table representing a field of that complex type. Column definitions are provided here.

- **Name:** Name of field; note that for complex type fields, the name of the field and the name of the type may, or may not be the same.
- **Type:** Field type; this may be a simple type (string, integer, etc.) or another complex type, which is described in another table.
- **Min:** Minimum expected occurrences (minOccurs). This value will be either 1 or 0 for all fields. For fields with 0 minOccurs, that field may be omitted from the complex type, and it will still be schema-compliant. Omitting a field may still cause an application-level error due to invalid data, refer to the **Range** column for application-level constraints.
- **Max:** Maximum expected occurrences (maxOccurs). This value will usually be 1 or *unbounded*. Unbounded fields/elements may appear multiple time within the complex type, which allows for list-like data structures within the service. While there is no theoretical upper limit to the number of occurrences, some constraints are enforced at the application level. See the **Range** column for more information.
- **Description:** This column defines the field’s purpose.
- **Range:** Application-enforced constraints on a field’s value are given here. If the field has a minOccurs of 0 in the schema, but is expected to be included by the application, it will be designated *required* in this column. Fields with a maxOccurs of *unbounded* within the schema with an application-enforced limit will be described here. Strings will have their maximum expected length defined here, if any.

### 2.2.1 SubmitBatch (REQUEST ELEMENT)

Name	Type	Min	Max	Description	Range
request	SubmitBatchRequest	0	1	Application-defined request element	<i>Required.</i>

### 2.2.1 SubmitBatchRequest

Name	Type	Min	Max	Description	Range
BatchId	string	1	1	DRC Batch ID; no validation performed by MI	Max length 50; longer values will be truncated.
ClientId	string	1	1	MI-Assigned client/project identifier; other projects sharing the environment will be assigned separate ClientIds.	Only values provided by MI will be accepted.
ConstructedResponses	ConstructedResponseList	0	1	List of constructed response elements to be scored for this batch	<i>Required.</i>

### 2.2.2 ConstructedResponseList

Name	Type	Min	Max	Description	Range
ConstructedResponse	ConstructedResponse	0	<i>unbounded</i>	List of individual CRs to be scored	<i>Required.</i> Missing or zero-length lists will not be entered for scoring. Lists exceeding 2000 CRs will also be rejected.

### 2.2.3 ConstructedResponse

Name	Type	Min	Max	Description	Range
EssayText	string	1	1	Student-generated response text.	This field is technically nillable, though nil or zero-length essays will not be scored. The field also technically has no max length, but essays exceeding 30,000 characters will also not be scored. Description codes will be returned for each of these cases.
ItemId	string	1	1	Identifier for Item/prompt	Responses that do not have a valid ItemId will not be scored; the range and convention for ItemIds are defined by DRC and MI.
ResponseId	string	1	1	DRC constructed response ID; no validation performed by MI	Max length 256; longer values will be truncated.

### 2.3.1 SubmitBatchResponse (RESPONSE ELEMENT)

Name	Type	Min	Max	Description	Range
SubmitBatchResult	SubmitBatchResult	0	1	Application-defined result element	<i>Required.</i>

### 2.3.2 SubmitBatchResult

Name	Type	Min	Max	Description	Range
BatchId	string	1	1	DRC batch ID as stored by MI (same value given in request)	Value may be truncated if it exceeds 50 characters
ClientId	string	1	1	MI-assigned client identifier (same value given in request)	
MIBatchId	ser:guid	1	1	MI-generated Batch ID	ser:guid is an extension of string, bounding the expected value to a Guid data type. It may be treated as a string or parsed to a Guid by the client.
StatusCode	StatusCode	1	1	Application-generated response code indicating success/failure of operation	

### 2.3.3 StatusCode

Name	Type	Min	Max	Description	Range
Code	integer	0	1	Numeric status code	<i>Required.</i> Will fall in the range 0-999. See <b>section 3</b> for more information
Description	string	0	1	Short description of status	<i>Required.</i> See <b>section 3</b> for more information



## SECTION 3 – STATUS CODE INFORMATION

**3.1 STATUS CODES:** Each SubmitBatch response will contain a status code indicating success or failure in adding the batch to the Streaming Scoring system. Individual CRs processed by Streaming Scoring will also receive similarly structured Status Codes upon delivery, albeit with similar values. Note that lower-level errors will not receive application-generated responses, and therefore will not be given status codes. These types of errors include (but are not limited to): malformed requests (which violate the schema), service unavailable, and TCP/HTTP errors. Expected status codes and their description for the SubmitBatch operation can be found in the following table.

### 3.2 SubmitBatch STATUS CODES

Code	Description	Notes
0	SUCCESS	Batch successfully accepted and queued for scoring.
100	INVALID_CLIENT_ID	ClientId value in request is not valid.
120	NO_REQUEST_DATA	request element is nil or missing.
140	NO_ESSAY_DATA	ConstructedResponses element is missing or contains zero CRs.
150	BATCH_TOO_LARGE	ConstructedResponses element contains more than 2000 CRs.
190	INTERNAL_ERROR	An unexpected internal error occurred at the application level.

### 3.3 Individual CR STATUS CODES

Code	Description	Notes
200	too few words (configurable)	blank or extremely short response; response sent to DRC for Supervisor Review
220	not enough correctly spelled words (configurable)	"Indecipherable" (i.e., foreign words and non-words); response sent to DRC for Supervisor Review
400	unexpected item_id	the item_id is not one of the items PEG AI has modeled; potential set-up issue to be resolved between MI and DRC
500	Alert, otherwise same as 0, above	alerted response sent to DRC for Supervisor Review
520	Alert, otherwise same as 200, above	alerted response sent to DRC for Supervisor Review
522	Alert, otherwise same as 220, above	alerted response sent to DRC for Supervisor Review
530	Alert, otherwise same as 300, above	alerted response sent to DRC for Supervisor Review
540	Alert, otherwise same as 400, above	the item_id is not one of the items PEG AI has modeled; potential set-up issue to be resolved between MI and DRC; alerted response sent to DRC for Supervisor Review
605	copied text low threshold (configurable)	sent to DRC for Supervisor Review
610	copied text high threshold (configurable)	sent to DRC for Supervisor Review
900	timeout	unable to complete essay score prediction within time limits; sent to DRC for Supervisor Review
950	system error processing essay	internal PEG error

# Appendix B

## AI Model Data – EOC English III (Spring 2019)

### Quadratic Weighted Kappa (QWK) and Inter-rater Reliability (IRR)

Course	Item #	Rater	Content				Style				Sentence Formation			Usage			Mechanics			Spelling		
			QWK	EX	ADJ	NON-	QWK	EX	ADJ	NON-	QWK	EX	ADJ	QWK	EX	ADJ	QWK	EX	ADJ	QWK	EX	ADJ
EOC English III	851370	H-H*	0.69	61	36	3	0.77	62	37	1	0.27	76	24	0.44	72	28	0.39	77	23	0.49	82	18
		AI-H**		58	40	2		61	39	0		71	29		70	30		78	22		83	17

\*Human to human (H-H) inter-rater metrics are from Pacific Metrics EFT scoring.

\*\*Human to AI (AI-H) inter-rater metrics are from the MI 2016 model-building results.

### Score Point Distribution (SPD)

Course	Item #	Rater	Content				Style				Sentence Formation		Usage		Mechanics		Spelling	
			1%	2%	3%	4%	1%	2%	3%	4%	0%	1%	0%	1%	0%	1%	0%	1%
EOC English III	851370	H	11	43	37	9	6	37	45	11	23	77	38	62	26	74	20	80
		AI	11	42	37	9	7	37	45	12	24	76	39	61	26	74	20	80

## AI Model Data – LEAP 2025 U.S. History ER (Spring 2019)

*Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD)*

Course	IDEAS Item #	# of Responses	Content										Claims											
			QWK	Inter-Rater Agreement %				Score Point Distribution %					QWK	Inter-Rater Agreement %				Score Point Distribution %						
				Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s		4s	Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s	4s
USH	894104	2500	0.86	H to H	62	33	5	Human	31	34	22	9	4	0.84	H to H	61	32	7	Human	39	28	21	9	4
		15%		AI to H	70	29	1	AI	30	38	21	8	3		AI to H	63	36	1	AI	39	28	21	8	3
USH	892955	2500	0.88	H to H	65	32	3	Human	34	29	25	9	3	0.88	H to H	64	32	4	Human	37	26	25	10	3
		15%		AI to H	74	26	0	AI	31	34	24	9	2		AI to H	72	28	0	AI	37	28	22	10	3

Human to human metrics are from DRC EFT scoring in Spring 2017.

AI to human metrics are from the MI 2017 model-building results.

- AI model was built in Fall 2017
- Included 2,500 responses from the Spring 2017 EFT
- Responses scored using DRC developed training materials
- 100% were scored by a second human reader and adjacent scores were resolved

## AI Model Building – Social Studies Grades 5-8 ERs (Spring 2019)

### Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD)

Grade	IDEAS Item #	# of Responses	Content										Claims											
			QWK	Inter-Rater Agreement %				Score Point Distribution %					QWK	Inter-Rater Agreement %				Score Point Distribution %						
				Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s		4s	Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s	4s
5	807773	2599	0.89	H to H <sup>1</sup>	78	21	1	Human	62	25	12	2	0	0.88	H to H <sup>1</sup>	79	20	1	Human	67	23	9	1	0
		≈500		H to H <sup>3</sup>	92	7	1	Human	3	29	48	17	3		H to H <sup>3</sup>	91	8	1	Human	8	33	45	11	2
		15%		AI to H	77	23	1	AI	50	27	18	4	1		AI to H	77	23	1	AI	54	26	16	4	1
6	804889	2975	0.79	H to H <sup>1</sup>	67	32	1	Human	42	44	12	1	0	0.76	H to H <sup>1</sup>	68	31	1	Human	52	38	9	1	0
		≈500		H to H <sup>2</sup>	98	2	0	Human	7	28	50	14	1		H to H <sup>2</sup>	99	1	0	Human	14	47	32	6	1
		15%		AI to H	71	28	0	AI	38	43	16	2	1		AI to H	73	25	2	AI	52	35	11	1	0
7	805627	2610	0.83	H to H <sup>1</sup>	73	25	2	Human	45	41	12	2	0	0.83	H to H <sup>1</sup>	73	25	2	Human	57	31	11	2	0
		≈500		H to H <sup>2</sup>	98	1	0	Human	9	18	39	26	8		H to H <sup>2</sup>	98	1	1	Human	12	20	38	22	8
		15%		AI to H	71	29	1	AI	35	40	16	7	1		AI to H	74	25	2	AI	52	28	14	3	3
8	808905	2543	0.86	H to H	65	33	2	Human	30	36	25	7	2	0.86	H to H	64	34	2	Human	30	37	25	7	2
		≈500		H to H <sup>2</sup>	90	9	0	Human	1	6	34	35	24		H to H <sup>2</sup>	91	8	1	Human	1	7	35	34	23
		15%		AI to H	67	32	1	AI	25	33	24	13	5		AI to H	70	28	2	AI	21	37	26	12	4

H to H<sup>1</sup> – Human scored 2016 Field Test sample of ≈ 2500 responses per item.

H to H<sup>2</sup>, H to H<sup>3</sup> – Human scored targeted samples of ≈ 500 responses per item used to augment and retrain the original AI models from 2016. These samples come from spring operational responses and are intended to find high score points to add to the existing AI models for the purpose of retraining the models prior to operational scoring. H to H<sup>2</sup> augmentation sample was scored in spring 2017. H to H<sup>3</sup> augmentation sample is to be scored in spring 2019.

AI – Data based on holdout subsets chosen by stratified random sampling from the full ≈ 3000 per item response count (2016 FT and 2018 sample) and excluded from the training process.

## AI Model CR Performance – ELA Grades 6-8, English I, and English II (Spring 2019)

Prompt	Grade	Trait	IEA-Human Agreement					
			Exact	SP0	SP1	SP2	SP3	SP4
E06_N_D1466	6	1	Blue	Green	Blue	Blue	Blue	Blue
		2	Blue	Blue	Blue	Blue	Blue	Blue
E06_R_DD502035970	6	1	Blue	Green	Blue	Blue	Blue	Blue
		2	Blue	Blue	Blue	Blue	Blue	Blue
E07_N_EE430133306	7	1	Blue	Blue	Green	Blue	Blue	Blue
		2	Blue	Blue	Orange	Blue	Blue	Blue
E08_L_F1460	8	1	Blue	Blue	Blue	Blue	Blue	Blue
		2	Blue	Blue	Blue	Blue	Blue	Blue
E09_L_VH017536_2T	9	1	Blue	Blue	Blue	Blue	Blue	Blue
		2	Blue	Blue	Blue	Green	Blue	Blue
E09_R_GG431834057	9	1	Blue	Green	Blue	Blue	Blue	Blue
		2	Blue	Green	Blue	Blue	Blue	Blue
E10_L_HH428127697	10	1	Blue	Blue	Blue	Blue	Blue	Blue
		2	Blue	Green	Blue	Blue	Blue	Blue
E10_N_HH432845949	10	1	Blue	Blue	Blue	Blue	Blue	Blue
		2	Blue	Blue	Blue	Blue	Blue	Blue

- Trait 1 = Reading Comprehension and Written Expression or Written Expression
- Trait 2 = Conventions
- Blue indicates IEA-Human performance higher than Human-Human performance
- Green indicates IEA-Human performance is within 5.25% of Human-Human performance
- Orange indicates IEA-Human performance is more than 5.25% below Human-Human performance
- Source – Pearson

## Spring 2019 LEAP 2025 and EOC Items – IRR and SPD from Previous Administrations

### Algebra I

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
901882	A, B	VH196970	Pearson Spring 2016	9,586	Part A	0,1	1,950	97.7%	99.9%	71.1%	12.5%				16.4%
				9,586	Part B	0,1,2	1,950	89.8%	97.1%	66.2%	6.7%	3.9%			23.1%
901882	A, B	VH196970	DRC Fall 2017, Op	8,522	Part A	0,1	1,940	99.0%	100.0%	94.0%	3.0%				4.0%
				8,522	Part B	0,1,2	1,940	99.0%	100.0%	94.0%	2.0%	1.0%			4.0%
901882	A, B	VH196970	DRC Spring 2018, Op	50,072	Part A	0,1	10,654	99.0%	100.0%	90.0%	8.0%				2.0%
				50,072	Part B	0,1,2	10,654	97.0%	100.0%	93.0%	3.0%	2.0%			2.0%
901882	A, B	VH196970	DRC Summer 2018, Op	1,625	Part A	0,1	372	99.0%	100.0%	97.0%	0.0%				3.0%
				1,625	Part B	0,1,2	372	99.0%	100.0%	96.0%	1.0%	0.0%			3.0%
901882	A, B	VH196970	DRC Fall 2018, Op	9,092	Part A	0,1	1,940	99.0%	100.0%	94.0%	3.0%				4.0%
				9,092	Part B	0,1,2	1,940	99.0%	100.0%	94.0%	2.0%	1.0%			4.0%
901836	A	M43318	DRC Fall 2017, Op	8,509	Overall	0,1,2,3	2,084	96.0%	100.0%	71.0%	12.0%	8.0%	3.0%		6.0%
901836	A	M43318	DRC Fall 2018, Op	9,062	Overall	0,1,2,3	2,084	96.0%	100.0%	71.0%	12.0%	8.0%	3.0%		6.0%
901814	A	M47147	DRC Fall 2017, Op	8,780	Part A	0,1,2	2,184	97.0%	100.0%	78.0%	8.0%	7.0%			8.0%
				8,780	Part B	0,1,2	2,184	99.0%	100.0%	88.0%	3.0%	1.0%			8.0%
901814	A	M47147	DRC Summer 2018, Op	1,637	Part A	0,1,2	412	97.0%	99.0%	88.0%	3.0%	1.0%			8.0%
				1,637	Part B	0,1,2	412	99.0%	100.0%	91.0%	1.0%	0.0%			8.0%
901859	A	3003-M43111	Pearson Spring 2016	253,395	Part C	0,1,2,3	48,917	92.1%	99.3%	43.7%	5.4%	14.9%	25.9%		10.1%
901859	A	3003-M43111	DRC Fall 2017, Op	8,485	Part C	0,1,2,3	2,504	98.0%	100.0%	73.0%	4.0%	7.0%	13.0%		2.0%
938769	A, D	MA10178	DRC Spring 2018, FT	1,579	Overall	0,1,2,3	324	94.0%	99.0%	65.0%	14.0%	13.0%	6.0%		2.0%
901848	A	M47287	Pearson Spring 2016	17,146	Overall	0,1,2,3,4	3,335	97.2%	99.6%	70.1%	9.2%	0.9%	0.4%	0.2%	19.1%
901848	A	M47287	DRC Fall 2017, Op	8,445	Overall	0,1,2,3,4	2,796	100.0%	100.0%	78.0%	4.0%	0.0%	0.0%	0.0%	17.0%
901848	A	M47287	DRC Summer 2018, Op	1,580	Overall	0,1,2,3,4	428	100.0%	100.0%	87.0%	1.0%	0.0%	0.0%	0.0%	12.0%
901857	A, B	VH046479	Pearson Spring 2017	78,418	Part A	0,1,2	13,963	88.2%	99.8%	51.2%	36.3%	2.9%			9.6%
				78,418	Part B	0,1	13,963	91.8%	99.7%	68.9%	19.0%			12.1%	
901857	A, B	VH046479	DRC Fall 2017, Op	8,686	Part A	0,1,2	2,258	94.0%	100.0%	77.0%	13.0%	1.0%			9.0%
				8,686	Part B	0,1	2,258	97.0%	100.0%	86.0%	5.0%			9.0%	
901857	A, B	VH046479	DRC Spring 2018, Op	8,686	Part A	0,1,2	2,258	94.0%	100.0%	77.0%	13.0%	1.0%			9.0%
				8,686	Part B	0,1	2,258	97.0%	100.0%	86.0%	5.0%			9.0%	
901857	A, B	VH046479	DRC Summer 2018, Op	49,959	Part A	0,1,2	11,927	88.0%	100.0%	57.0%	33.0%	4.0%			5.0%
				49,959	Part B	0,1	11,927	94.0%	100.0%	80.0%	14.0%			5.0%	
901857	A, B	VH046479	DRC Fall 2018, Op	1,623	Part A	0,1,2	396	92.0%	100.0%	80.0%	14.0%	0.0%			6.0%
				1,623	Part B	0,1	396	99.0%	100.0%	93.0%	1.0%			6.0%	

Form Key: Form A = Seniors only, Form B = Administrative Error (AE), Forms D and E = Operational

Algebra I (continued)

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
901832	B, D	3031-M44083P	Pearson Spring 2016	95,907	Part B	0,1,2	18,835	91.3%	99.9%	29.9%	45.1%	11.9%			13.1%
901832	B, D	3031-M44083P	DRC Spring 2018, Op	55,162	Part B	0,1,2	10,236	91.0%	100.0%	32.0%	47.0%	21.0%			0.0%
901832	B, D	3031-M44083P	DRC Fall 2018, Op	6,329	Part B	0,1,2	1,140	92.0%	100.0%	51.0%	40.0%	9.0%			0.0%
938741	D	MA10144	DRC Spring 2018, FT	1,620	Overall	0,1,2,3	350	95.0%	99.0%	50.0%	10.0%	17.0%	11.0%		2.0%
980927	D, E	VH251952	Pearson Spring 2018	124,433	Part A	0,1,2	23,748	97.3%	99.6%	69.6%	15.0%	4.9%			10.6%
				124,433	Part B	0,1,2	23,748	95.4%	99.3%	72.0%	8.3%	5.7%			14.0%
				124,433	Part C	0,1,2	23,748	90.9%	98.8%	67.5%	11.5%	7.0%			13.9%
938735	D	MA10137	DRC Spring 2018, FT	1,655	Part B	0,1,2,3	316	94.0%	98.0%	79.0%	9.0%	7.0%	5.0%		0.0%
938744	D	MA10147	DRC Spring 2018, FT	1,606	Overall	0,1,2,3	344	90.0%	98.0%	67.0%	18.0%	4.0%	8.0%		1.0%
938737	B, D, E	MA10139	DRC Spring 2018, FT	1,582	Overall	0,1,2,3,4	382	94.0%	100.0%	71.0%	12.0%	4.0%	2.0%	5.0%	7.0%
938769	D	MA10178	DRC Spring 2018, FT	1,579	Overall	0,1,2,3	324	94.0%	99.0%	65.0%	14.0%	13.0%	6.0%		2.0%
980924	E	M44463	Pearson Spring 2017	77,183	Overall	0,1,2,3	14,754	87.6%	99.0%	36.9%	14.7%	30.4%	11.2%		6.8%
980909	E	M43216	Pearson Spring 2018	98,152	Overall	0,1,2,3	18,677	87.5%	99.3%	61.9%	13.9%	10.6%	3.7%		9.9%
980911	E	2679-M43312	Pearson 2015 FT	1,799	Part A	0,1,2	402	95.0%	99.5%	70.9%	12.4%	2.9%			13.9%
				1,799	Part B	0,1,2	402	94.5%	100.0%	19.2%	62.6%	3.3%			15.0%
901851	B, E	M41726	DRC Spring 2018, Op	52,490	Overall	0,1,2,3	11,918	92.0%	100.0%	57.0%	14.0%	15.0%	8.0%		6.0%
901851	B, E	M41726	DRC Fall 2018, Op	6,011	Overall	0,1,2,3	1,556	96.0%	100.0%	66.0%	11.0%	9.0%	4.0%		9.0%
980923	E	M000312	Pearson 2017 FT	1,593	Overall	0,1,2,3	264	89.0%	100.0%	65.1%	15.0%	7.6%	6.3%		6.1%
901687	B	2407-M41752	DRC Spring 2018, OP	53,117	Part A	0,1,2	11,413	98.0%	100.0%	74.0%	3.0%	19.0%			4.0%
				53,117	Part B	0,1,2	11,413	96.0%	100.0%	83.0%	7.0%	6.0%			4.0%
				53,117	Part C	0,1,2	11,413	98.0%	100.0%	89.0%	4.0%	3.0%			4.0%
901687	B	2407-M41752	DRC Spring 2018, OP	6,022	Part A	0,1,2	1,470	99.0%	100.0%	80.0%	2.0%	10.0%			7.0%
				6,022	Part B	0,1,2	1,470	99.0%	100.0%	87.0%	3.0%	2.0%			7.0%
				6,022	Part C	0,1,2	1,470	99.0%	100.0%	90.0%	1.0%	1.0%			7.0%
901705	B	VF883359	DRC Spring 2018, Op	53,281	Part A	0,1,2,3	11,808	98.0%	100.0%	89.0%	4.0%	1.0%	2.0%		5.0%
				53,281	Part B	0,1	11,808	93.0%	100.0%	84.0%	11.0%				5.0%
901705	B	VF883359	DRC Fall 2018, Op	6,097	Part A	0,1,2,3	1,570	100.0%	100.0%	87.0%	2.0%	1.0%			8.0%
				6,097	Part B	0,1	1,570	98.0%	100.0%	84.0%	7.0%				8.0%

Form Key: Form A = Seniors only, Form B = Administrative Error (AE), Forms D and E = Operational

## Geometry

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
902012	B, D, E	M41169	Pearson Spring 2016	90,471	Overall	0,1,2,3	16,723	87.1%	98.7%	46.2%	12.3%	14.6%	7.0%		19.9%
902012	B, D, E	M41169	DRC Spring 2018, Op	38,108	Overall	0,1,2,3	9,066	90.0%	100.0%	45.0%	15.0%	26.0%	9.0%		5.0%
902012	B, D, E	M41169	DRC Fall 2018, Op	5,823	Overall	0,1,2,3	1,424	96.0%	100.0%	47.0%	14.0%	23.0%	9.0%		7.0%
980937	D, E	M43798	Pearson Spring 2017	42,156	Overall	0,1,2,3	7,901	95.2%	99.5%	65.8%	14.1%	3.6%	1.2%		15.3%
939083	D	MGM0141	DRC Spring 2018, FT	1,592	Overall	0,1,2,3,4	354	95.0%	100.0%	70.0%	3.0%	6.0%	5.0%	11.0%	4.0%
980942	D	VH236248	Pearson 2016 FT	1,633	Part A	0,1,2,3	341	84.2%	98.2%	44.4%	25.3%	13.5%	6.1%		10.8%
				1,633	Part B	0,1,2,3	341	79.5%	97.7%	54.4%	16.8%	12.7%	3.4%		12.8%
939077	D	MGM0135	DRC Spring 2018, FT	1,595	Overall	0,1,2,3,4	356	95.0%	100.0%	70.0%	14.0%	7.0%	2.0%	2.0%	5.0%
980938	D, E	M100106	Pearson 2017 FT	1,635	Overall	0,1,2,3,4	314	93.0%	98.7%	73.8%	5.2%	5.7%	3.9%		11.4%
980936	D, E	VH239429	Pearson Spring 2017	42,154	Overall	0,1,2,3	8,173	84.3%	99.1%	71.6%	16.1%	3.6%	2.3%		6.4%
980929	E	M1000516	Pearson 2017 FT	1,612	Overall	0,1,2,3,4	314	87.9%	96.8%	63.1%	7.5%	6.8%	3.9%	6.8%	12.0%
902042	B, E	3020-M44058	Pearson Spring 2016	45,304	Part A	0,1,2,3	8,509	94.5%	99.7%	47.9%	29.7%	7.3%	4.1%		11.0%
				45,304	Part B	0,1	8,509	96.1%	99.8%	61.4%	21.9%				16.7%
				45,304	Part C	0,1,2	8,509	94.8%	97.7%	61.2%	4.7%	12.2%			21.9%
902042	B, E	3020-M44058	DRC Spring 2018, Op	38,085	Part A	0,1,2,3	8,517	96.0%	100.0%	55.0%	34.0%	5.0%	3.0%		4.0%
				38,085	Part B	0,1	8,517	97.0%	100.0%	78.0%	19.0%				4.0%
				38,085	Part C	0,1,2	8,517	97.0%	99.0%	79.0%	5.0%	13.0%			4.0%
902042	B, E	3020-M44058	DRC Fall 2018, Op	5,710	Part A	0,1,2,3	1,318	98.0%	100.0%	56.0%	30.0%	6.0%	2.0%		6.0%
				5,710	Part B	0,1	1,318	98.0%	100.0%	77.0%	17.0%				6.0%
				5,710	Part C	0,1,2	1,318	98.0%	99.0%	76.0%	5.0%	14.0%			6.0%
980930	E	M1000518	Pearson 2017 FT	1,500	Part B	0,1,2,3	298	95.3%	100.0%	60.4%	11.0%	11.9%	1.4%		15.2%
901939	A	M43794	DRC Fall 2017, Op	6,811	Overall	0,1,2,3	1,696	93.0%	100.0%	72.0%	10.0%	11.0%	2.0%		5.0%
901939	A	M43794	DRC Summer 2018, Op	450	Overall	0,1,2,3	162	98.0%	100.0%	74.0%	0.0%	2.0%	4.0%		19.0%

Form Key: Form A = Seniors only, Form B = Administrative Error (AE), Forms D and E = Operational



Geometry (continued)

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
902046	A, B	M46668	Pearson Spring 2016	42,630	Overall	0,1,2,3	7,622	92.9%	98.9%	70.0%	8.5%	5.2%	0.5%		15.8%
902046	A, B	M46668	DRC Fall 2017, Op	6,821	Overall	0,1,2,3	1,880	97.0%	100.0%	78.0%	9.0%	3.0%	0.0%		9.0%
902046	A, B	M46668	DRC Spring 2018, Op	38,108	Overall	0,1,2,3	9,657	95.0%	100.0%	76.0%	10.0%	6.0%	1.0%		7.0%
902046	A, B	M46668	DRC Summer 2018, Op	423	Overall	0,1,2,3	148	99.0%	100.0%	74.0%	3.0%	3.0%	0.0%		19.0%
902046	A, B	M46668	DRC Fall 2018, Op	5,601	Overall	0,1,2,3	1,396	96.0%	100.0%	73.0%	9.0%	7.0%	1.0%		10.0%
902027	A, B	M43233	Pearson Spring 2017	84,614	Overall	0,1,2,3,4	15,944	88.2%	97.7%	51.8%	12.5%	9.5%	5.0%	5.4%	15.9%
902027	A, B	M43233	DRC Spring 2018, Op	38,085	Overall	0,1,2,3,4	9,519	94.0%	100.0%	60.0%	13.0%	10.0%	5.0%	6.0%	7.0%
902027	A, B	M43233	DRC Summer 2018, Op	420	Overall	0,1,2,3,4	156	96.0%	100.0%	70.0%	3.0%	2.0%	1.0%	2.0%	22.0%
902027	A, B	M43233	DRC Fall 2018, Op	5,712	Overall	0,1,2,3,4	1,530	96.0%	100.0%	60.0%	10.0%	8.0%	5.0%	7.0%	9.0%
902036	B	2904-M43021	Pearson Spring 2016	42,708	Part A	0,1,2	8,216	95.9%	99.5%	53.9%	6.8%	33.2%			6.1%
				42,708	Part B	0,1,2	8,216	94.7%	99.2%	61.0%	6.5%	24.1%			8.5%
				42,708	Part C	0,1,2	8,216	94.9%	98.4%	75.2%	3.5%	4.2%			17.2%
902036	B	2904-M43021	DRC Fall 2017, Op	6,800	Part A	0,1,2	1,518	99.0%	100.0%	69.0%	9.0%	20.0%			2.0%
				6,800	Part B	0,1,2	1,518	96.0%	99.0%	68.0%	10.0%	19.0%			2.0%
				6,800	Part C	0,1,2	1,518	97.0%	99.0%	91.0%	3.0%	3.0%			2.0%
902036	B	2904-M43021	DRC Summer 2018, Op	433	Part A	0,1,2	110	100.0%	100.0%	84.0%	3.0%	6.0%			8.0%
				433	Part B	0,1,2	110	100.0%	100.0%	86.0%	1.0%	6.0%			8.0%
				433	Part C	0,1,2	110	100.0%	100.0%	89.0%	1.0%	2.0%			8.0%
902047	B	VH150404	Pearson Spring 2016	47,576	Part A	0,1,2	8,713	97.9%	99.6%	64.1%	6.6%	2.8%			26.5%
				47,576	Part B	0,1,2	8,713	92.1%	99.7%	52.6%	19.2%	7.0%			21.2%
902047	B	VH150404	DRC Fall 2017, Op	6,775	Part A	0,1,2	1,636	98.0%	100.0%	81.0%	10.0%	4.0%			5.0%
				6,775	Part B	0,1,2	1,636	96.0%	100.0%	74.0%	15.0%	6.0%			5.0%
902047	B	VH150404	DRC Summer 2018, Op	430	Part A	0,1,2	134	99.0%	100.0%	80.0%	2.0%	3.0%			14.0%
				430	Part B	0,1,2	134	100.0%	100.0%	77.0%	4.0%	5.0%			14.0%
939101	A, B	MGM0160	DRC Spring 2018, FT	1,665	Part C	0,1,2,3,4	336	80.0%	97.0%	73.0%	15.0%	8.0%	2.0%	1.0%	1.0%
902062	B	VH150384	Pearson Spring 2016	2,581	Overall	0,1,2,3,4	542	88.6%	97.4%	56.6%	6.1%	4.1%	1.5%	0.8%	30.9%
902062	B	VH150384	DRC Spring 2018, Op	38,056	Overall	0,1,2,3,4	9,554	96.0%	100.0%	79.0%	9.0%	4.0%	1.0%	1.0%	7.0%
902062	B	VH150384	DRC Fall 2018, Op	5,747	Overall	0,1,2,3,4	1,452	97.0%	100.0%	76.0%	9.0%	4.0%	2.0%	1.0%	9.0%

Form Key: Form A = Seniors only, Form B = Administrative Error (AE), Forms D and E = Operational

Math Grade 3

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
981736	Op	VH054794	Pearson Spring 2017	52,491	Part A	0,1,2	9,873	76.2%	98.9%	46.9%	33.0%	16.7%			3.4%
				52,491	Part B	0,1,2	9,885	82.8%	98.3%	35.4%	22.5%	37.8%			4.3%
914048	Op	M05158	Pearson Spring 2017	79,640	Overall	0,1,2,3	7,819	92.4%	99.3%	52.8%	18.5%	11.5%	15.7%		1.5%
914048	Op	M05158	DRC Spring 2018, Op	61,502	Overall	0,1,2,3	11,828	90.0%	100.0%	34.0%	30.0%	24.0%	6.0%		6.0%
898001	Op	N/A	DRC Spring 2018, FT	1,659	Part A	0,1,2	318	94.0%	100.0%	41.0%	21.0%	37.0%			1.0%
				1,659	Part B	0,1	318	98.0%	100.0%	95.0%	4.0%			1.0%	
981742	Op	M300388PD	Pearson 2017 FT	1,500	Part B	0,1,2	295	88.1%	98.3%	73.4%	7.3%	17.4%			1.9%
914039	Op	M02527	Pearson Spring 2017	7,113	Overall	0,1,2,3	699	92.7%	98.7%	37.9%	30.2%	23.4%	1.7%		6.8%
914039	Op	M02527	DRC Spring 2018, Op	61,394	Overall	0,1,2,3	11,578	88.0%	100.0%	18.0%	28.0%	45.0%	7.0%		1.0%
981747	Op	4127-M03599P	Pearson Spring 2018	102,233	Part B	0,1,2,3	20,403	90.8%	99.2%	48.2%	25.5%	8.8%	13.4%		4.1%
				102,233	Part C	0,1,2	20,403	92.4%	99.8%	32.9%	28.8%	33.1%			5.1%

Math Grade 4

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
914084	Op	4112-M03491P	Pearson Spring 2017	383,723	Part C	0,1,2	37,737	94.9%	99.9%	65.0%	28.5%	2.5%			4.0%
914084	Op	4112-M03491P	DRC Spring 2018, Op Paper	5,830	Part C	0,1,2	1,238	96.0%	100.0%	67.0%	28.0%	3.0%			1.0%
914084	Op	4112-M03491P	DRC Spring 2018, Op Online	56,155	Part C	0,1,2	10,776	95.0%	100.0%	63.0%	28.0%	5.0%			4.0%
914086	Op	M04133	Pearson Spring 2017	107,359	Overall	0,1,2,3	10,670	91.1%	99.4%	53.1%	23.9%	7.4%	14.9%		0.7%
914086	Op	M04133	DRC Spring 2018, Op	61,742	Overall	0,1,2,3	11,702	95.0%	100.0%	54.0%	24.0%	7.0%	9.0%		5.0%
981831	Op	M400526	Pearson 2017 FT	1,500	Overall	0,1,2,3	288	85.8%	99.3%	47.2%	21.4%	22.1%	9.2%		0.1%
899959	Op	N/A	DRC Spring 2018, FT	1,622	Overall	0,1,2,3	302	82.0%	99.0%	34.0%	24.0%	11.0%	30.0%		0.0%
899955	Op	N/A	DRC Spring 2018, FT	1,651	Part A	0,1,2	306	88.0%	98.0%	39.0%	10.0%	49.0%			1.0%
				1,651	Part B	0,1	306	96.0%	100.0%	88.0%	11.0%				1.0%
981927	Op	0318-M01475	Pearson 2017 FT	1,500	Part A	0,1,2	300	98.7%	100.0%	54.8%	10.8%	33.6%			0.7%
				1,500	Part B	0,1,2	300	99.3%	100.0%	79.8%	3.3%	15.4%			1.6%
				1,500	Part C	0,1,2	300	93.7%	99.7%	64.2%	8.9%	24.5%			2.3%

Math Grade 5

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
914152	Op	M03820	Pearson Spring 2017	216,578	Overall	0,1,2,3,4	43,004	76.0%	98.0%	26.0%	25.6%	22.4%	15.5%	9.3%	1.3%
914148	Op	M03888	Pearson Spring 2017	72,736	Overall	0,1,2,3	7,272	86.7%	98.9%	39.9%	27.5%	13.2%	18.7%		0.8%
914148	Op	M03888	DRC Spring 2018, Op	59,662	Overall	0,1,2,3	11,464	93.0%	99.0%	57.0%	22.0%	8.0%	12.0%		1.0%
902410	Op	N/A	DRC Spring 2018, FT	1,653	Part B	0,1,2	306	87.0%	100.0%	46.0%	20.0%	33.0%			1.0%
902414	Op	N/A	DRC Spring 2018, FT	1,651	Overall	0,1,2,3	318	87.0%	99.0%	63.0%	20.0%	7.0%			0.0%
914195	Op	0154-M00796	Pearson Spring 2017	92,904	Part B	0,1,2	9,282	95.9%	99.8%	80.4%	8.3%	6.4%			4.8%
914195	Op	0154-M00796	DRC Spring 2018, Op	61,037	Part B	0,1,2	11,260	91.0%	100.0%	75.0%	15.0%	10.0%			0.0%
934015	Op	N/A	DRC Spring 2018, FT	1,660	Part B	0,1	320	93.0%	100.0%	85.0%	15.0%				0.0%
				1,660	Part C	0,1,2,3,4	320	89.0%	98.0%	58.0%	19.0%	11.0%	4.0%	7.0%	0.0%

Math Grade 6

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
981963	Op	M25151	Pearson Spring 2018	130,590	Overall	0,1,2,3,4	25,899	68.6%	96.5%	35.2%	23.2%	19.4%	13.4%	6.2%	2.6%
981961	Op	VH082639	Pearson 2015 FT	1,500	Part A	0,1,2	348	90.2%	100.0%	54.6%	26.9%	14.3%			4.2%
				1,500	Part B	0,1	348	90.8%	100.0%	53.9%	39.5%			6.6%	
981954	Op	VH139067	Pearson Spring 2017	111,824	Part A	0,1,2	21,162	93.0%	98.4%	78.8%	5.4%	11.5%			4.3%
				111,824	Part B	0,1,2,3,4	21,162	86.7%	98.3%	59.0%	15.5%	8.5%	3.8%	9.2%	4.0%
981956	Op	VH220482	Pearson Spring 2017	111,824	Part B	0,1,2	22,112	92.4%	99.3%	31.8%	15.7%	49.6%			2.8%
914231	Op	1740-M23030	Pearson Spring 2017	89,916	Overall	0,1,2,3	8,905	70.5%	96.2%	40.2%	18.0%	20.4%	19.0%		2.3%
914231	Op	1740-M23030	DRC Spring 2018, Op	58,067	Overall	0,1,2,3	11,448	74.0%	96.0%	43.0%	18.0%	19.0%	17.0%		2.0%
903511	Op	N/A	DRC Spring 2018, FT	1,652	Part B	0,1,2,3	310	85.0%	98.0%	76.0%	10.0%	10.0%	5.0%		0.0%
914281	Op	M25152	Pearson Spring 2017	112,484	Overall	0,1,2,3	11,247	89.0%	99.0%	53.8%	14.1%	12.3%	17.2%		2.6%
914281	Op	M25152	DRC Spring 2018, Op	57,609	Overall	0,1,2,3	11,534	91.0%	99.0%	63.0%	13.0%	8.0%	14.0%		2.0%

Math Grade 7

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
914362	Op	VH083535	Pearson Spring 2016	100,577	Part A	0,1,2,3	19,892	90.4%	98.5%	74.7%	5.1%	4.9%	12.5%		2.8%
				100,577	Part B	0,1,2,3	19,892	90.3%	98.8%	70.8%	5.6%	5.6%	13.8%		4.3%
914362	Op	VH083535	DRC Spring 2018, Op	56,482	Part A	0,1,2,3	10,560	96.0%	100.0%	86.0%	3.0%	3.0%	7.0%		0.0%
				56,482	Part B	0,1,2,3	10,560	96.0%	100.0%	84.0%	3.0%	3.0%	9.0%		0.0%
982922	Op	M25544	Pearson 2015 FT	1,800	Overall	0,1,2,3	404	87.6%	99.0%	50.4%	13.7%	22.2%	7.0%		6.6%
868848	Op	M25578	Pearson Spring 2017	13,001	Overall	0,1,2,3	2,576	93.6%	99.0%	74.6%	5.4%	8.6%	1.4%		10.1%
900539	Op	N/A	DRC Spring 2018, FT	1,646	Part A	0,1,2	316	91.0%	99.0%	46.0%	37.0%	17.0%			0.0%
				1,646	Part B	0,1	316	97.0%	100.0%	62.0%	38.0%				0.0%
982929	Op	M22009	Pearson Spring 2018	124,808	Overall	0,1,2,3	24,757	83.2%	98.8%	45.6%	20.8%	20.4%	11.3%		2.0%
900520	Op	N/A	DRC Spring 2018, FT	1,624	Overall	0,1,2,3	348	97.0%	100.0%	77.0%	6.0%	4.0%	9.0%		3.0%
914339	Op	VH151385	Pearson Spring 2017	88,725	Part A	0,1,2	8,838	95.4%	99.4%	66.9%	7.8%	20.9%			4.4%
				88,725	Part B	0,1,2	8,838	95.5%	99.7%	77.2%	5.7%	9.7%			7.4%
914339	Op	VH151385	DRC Spring 2018, Op	56,454	Part A	0,1,2	10,887	98.0%	100.0%	73.0%	7.0%	19.0%			2.0%
				56,454	Part B	0,1,2	10,887	98.0%	100.0%	83.0%	6.0%	10.0%			2.0%

Math Grade 8

IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
983010	Op	VH097312	Pearson Spring 2018	28,653	Part A	0,1,2	5,561	95.7%	99.8%	63.6%	19.8%	8.2%			8.5%
				28,653	Part B	0,1,2,3,4	5,561	90.6%	98.9%	72.2%	9.5%	6.1%	1.6%	0.3%	10.4%
982987	Op	M800114	Pearson 2017 FT	1,500	Part A	0,1,2	300	93.3%	98.0%	73.6%	7.9%	15.3%			3.2%
				1,500	Part B	0,1,2	300	89.3%	98.7%	69.6%	12.2%	13.7%			4.5%
982999	Op	M22203	Pearson Spring 2017	69,637	Overall	0,1,2,3	13,500	84.2%	96.8%	54.6%	23.7%	8.5%	8.8%		4.4%
870899	Op	1282-M21381	Pearson Spring 2015	48,511	Part A	0,1,2	9,762	89.2%	97.5%	72.2%	9.3%	8.9%			9.7%
				48,511	Part B	0,1	9,762	91.0%	99.2%	65.5%	22.2%				12.2%
899312	Op	N/A	DRC Spring 2018, FT	1,648	Part B	0,1,2	318	85.0%	98.0%	27.0%	30.0%	43.0%			0.0%
914381	Op	M25425	Pearson Spring 2017	69,637	Overall	0,1,2,3,4	6,943	90.8%	99.1%	52.2%	12.5%	26.2%	1.9%	1.0%	6.2%
914381	Op	M25425	DRC Spring 2018, Op	49,280	Overall	0,1,2,3,4	10,088	94.0%	100.0%	59.0%	16.0%	20.0%	2.0%	0.0%	2.0%
899329	Op	N/A	DRC Spring 2018, FT	1,653	Part B	0,1	314	90.0%	100.0%	51.0%	49.0%				0.0%
				1,653	Part C	0,1	314	94.0%	100.0%	57.0%	43.0%				0.0%

English I

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
LAT	902152	D	VH017536_2T	Pearson Spr 17	126,939	RCWE	0,1,2,3,4	8,443	14,816	114,377	23,741	73.6%	99.3%	36.7%	36.2%	16.6%	5.3%	1.8%	3.4%
LAT	902152	D	VH017536_2T	DRC Spr 18	126,939	Conv	0,1,2,3	8,443	14,816	114,377	23,741	73.4%	99.6%	31.5%	37.1%	20.4%	7.5%		3.4%
LAT	902152	D	VH017536_2T	DRC Fall 18	51,374	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,844	81.0%	100%	32.0%	44.0%	19.0%	3.0%	0%	2.0%
LAT	902152	D	VH017536_2T	DRC Fall 18	51,374	Conv	0,1,2,3	n/a	n/a	n/a	10,844	82.0%	100%	26.0%	45.0%	23.0%	3.0%		3.0%
RST	914552	D, E	GG431834057	Pearson Spr 18	7,444	RCWE	0,1,2,3,4	n/a	n/a	n/a	1,870	86.0%	100%	44.0%	32.0%	16.0%	3.0%	0.0%	3.0%
RST	914552	D, E	GG431834057	Pearson Spr 18	7,444	Conv	0,1,2,3	n/a	n/a	n/a	1,870	85.0%	100%	39.0%	34.0%	20.0%	4.0%		3.0%
NWT	983215	E	GG604245591	Pearson 17 FT	66,624	RCWE	0,1,2,3,4	2,058	7,456	62,441	13,132	75.6%	99.6%	26.3%	28.9%	27.4%	11.3%	2.2%	4.0%
NWT	983215	E	GG604245591	Pearson 17 FT	66,624	Conv	0,1,2,3	2,058	7,456	62,441	13,132	76.1%	99.5%	25.2%	30.3%	27.4%	12.1%		4.0%
RST	902161	A	VH017542_2T	Pearson Spr 17	1,696	Expr	0,1,2,3,4	1,430	155	0	299	74.9%	97.0%	24.7%	25.1%	26.3%	11.6%	5.0%	7.3%
RST	902161	A	VH017542_2T	Pearson Spr 17	1,696	Conv	0,1,2,3	1,430	155	0	299	72.6%	100.0%	22.8%	22.8%	27.5%	14.7%		7.3%
RST	902161	A	VH017542_2T	DRC Fall 17*	123,860	RCWE	0,1,2,3,4	2,656	13,063	116,406	23,334	76.1%	99.5%	22.2%	33.3%	24.3%	12.4%	3.7%	4.1%
RST	902161	A	VH017542_2T	DRC Fall 17*	123,860	Conv	0,1,2,3	2,656	13,063	116,407	23,334	76.1%	99.6%	23.1%	32.7%	23.5%	16.6%		4.1%
RST	902161	A	VH017542_2T	DRC Spr 18	4,674	RCWE	0,1,2,3,4	n/a	n/a	n/a	982	78.0%	99.0%	12.0%	34.0%	40.0%	13.0%	0.0%	0.0%
RST	902161	A	VH017542_2T	DRC Spr 18	4,674	Conv	0,1,2,3	n/a	n/a	n/a	982	78.0%	99.0%	14.0%	32.0%	38.0%	15.0%		0.0%
RST	902161	A	VH017542_2T	DRC Fall 18	50,817	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,136	81.0%	100%	17.0%	37.0%	32.0%	11.0%	1.0%	2.0%
RST	902161	A	VH017542_2T	DRC Fall 18	50,817	Conv	0,1,2,3	n/a	n/a	n/a	10,136	79.0%	100%	17.0%	36.0%	32.0%	13.0%		2.0%
RST	902161	A	VH017542_2T	DRC Fall 18	7,444	RCWE	0,1,2,3,4	n/a	n/a	n/a	1,870	84.0%	100%	30.0%	30.0%	24.0%	10.0%	1.0%	3.0%
RST	902161	A	VH017542_2T	DRC Fall 18	7,444	Conv	0,1,2,3	n/a	n/a	n/a	1,870	84.0%	100%	30.0%	29.0%	25.0%	12.0%		3.0%
NWT	906512	A	VH084830	Pearson Spr 17	61,936	Expr	0,1,2,3,4	3,125	7,776	53,955	10,498	73.3%	98.7%	30.3%	21.8%	27.3%	8.7%	4.2%	7.6%
NWT	906512	A	VH084830	Pearson Spr 17	61,936	Conv	0,1,2,3	3,125	7,776	53,955	10,498	74.4%	99.4%	28.1%	27.7%	25.4%	11.2%		7.6%
NWT	906512	A	VH084830	DRC Fall 17*	5,047	Expr	0,1,2,3,4	n/a	n/a	n/a	1,076	81.0%	99.0%	22.0%	34.0%	29.0%	10.0%	1.0%	2.0%
NWT	906512	A	VH084830	DRC Fall 17*	5,047	Conv	0,1,2,3	n/a	n/a	n/a	1,076	78.0%	99.0%	25.0%	36.0%	26.0%	10.0%		2.0%
RST	902194	C	VH017614_2T	Pearson Spr 17	3,179	RCWE	0,1,2,3,4	3,012	317	0	620	82.6%	99.4%	43.9%	33.3%	12.0%	2.5%	0.8%	7.6%
RST	902194	C	VH017614_2T	Pearson Spr 17	3,179	Conv	0,1,2,3	3,012	317	0	620	79.4%	99.7%	47.8%	30.2%	11.7%	2.8%		7.6%
RST	902194	C	VH017614_2T	DRC Sum 18	1,546	RCWE	0,1,2,3,4	n/a	n/a	n/a	338	86.0%	100%	56.0%	32.0%	7.0%	1.0%	0%	4.0%
RST	902194	C	VH017614_2T	DRC Sum 18	1,546	Conv	0,1,2,3	n/a	n/a	n/a	338	82.0%	100%	57.0%	32.0%	7.0%	1.0%		4.0%
NWT	902203	C	6139	Pearson Spr 17	126,941	Expr	0,1,2,3,4	7,555	14,727	112,973	24,056	76.8%	99.6%	22.5%	33.7%	25.4%	9.3%	2.5%	6.6%
NWT	902203	C	6139	Pearson Spr 17	126,941	Conv	0,1,2,3	7,555	14,727	112,973	24,056	76.4%	99.8%	26.8%	30.6%	25.9%	10.1%		6.6%
NWT	902203	C	6139	DRC Sum 18	1,510	WE	0,1,2,3,4	n/a	n/a	n/a	408	87.0%	100%	63.0%	27.0%	1.0%	0%	0%	9.0%
NWT	902203	C	6139	DRC Sum 18	1,510	Conv	0,1,2,3	n/a	n/a	n/a	408	93.0%	100%	73.0%	17.0%	1.0%	0%		9.0%

Form Key: Forms D and E = Operational, Form A = Seniors only, Form C = Administrative Error (AE)

\*Handscored by DRC in Fall of 2017



English II

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
LAT	906197	D	HH428127697	Pearson Spr 17	57,407	RCWE	0,1,2,3,4	28,405	6,463	27,217	14,673	75.4%	99.2%	27.0%	30.6%	27.3%	9.0%	1.3%	4.8%
LAT	906197	D	HH428127697	DRC Spr 18	48,997	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,486	77.0%	99.0%	22.0%	37.0%	32.0%	6.0%	0.0%	3.0%
LAT	906197	D	HH428127697	DRC Fall 18	10,724	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,872	80.0%	99.0%	38.0%	31.0%	21.0%	5.0%	1.0%	4.0%
RST	983688	D, E	HH607742252	Pearson 2017 FT	1,604	RCWE	0,1,2,3,4	1,487	162	0	312	78.2%	100.0%	28.1%	29.5%	20.2%	7.2%	2.2%	12.8%
NWT	983642	E	HH432845949	Pearson Spr 17	57,527	Expr	0,1,2,3,4	28,646	6,810	26,290	13,745	76.6%	99.6%	16.3%	23.2%	32.9%	15.6%	4.9%	7.1%
RST	902331	A	VH004490	Pearson Spr 17**	2,605	RCWE	0,1,2,3,4	1,915	263	646	827	81.9%	99.3%	51.5%	28.1%	7.2%	0.9%	0.1%	12.3%
RST	902331	A	VH004490	Pearson Spr 16**	126,270	RCWE	0,1,2,3,4	121,660	n/a	n/a	16,036	76.6%	99.7%	22.7%	34.8%	23.4%	8.3%	2.0%	8.8%
RST	902331	A	VH004490	DRC Fall 17*	9,305	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,020	79.0%	100.0%	37.0%	24.0%	25.0%	11.0%	2.0%	2.0%
RST	902331	A	VH004490	DRC Spr 18	48,949	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,460	79.0%	100.0%	15.0%	35.0%	34.0%	11.0%	2.0%	3.0%
RST	902331	A	VH004490	DRC Fall 18	10,714	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,826	84.0%	100.0%	30.0%	33.0%	22.0%	9.0%	2.0%	3.0%
NWT	902354	A	7064	Pearson Spr 17	4,409	Expr	0,1,2,3,4	4,189	435	0	844	84.5%	100.0%	42.5%	19.5%	13.5%	6.0%	2.0%	16.5%
NWT	902354	A	7064	DRC Fall 17*	9,721	Expr	0,1,2,3,4	n/a	n/a	n/a	2,098	81.0%	100.0%	46.0%	17.0%	19.0%	12.0%	2.0%	2.0%
LAT	906181	C	HH431436431	Pearson Spr 17	57,534	RCWE	0,1,2,3,4	28,697	6,808	27,606	14,813	75.9%	98.9%	33.6%	37.4%	17.6%	5.6%	1.6%	4.2%
LAT	906181	C	HH431436431	DRC Sum18	2,632	RCWE	0,1,2,3,4	n/a	n/a	n/a	864	90.0%	100%	75.0%	18.0%	1.0%	0%	0%	6.0%
RST	906190	C	HH433954866	Pearson Spr 17	57,526	RCWE	0,1,2,3,4	26,197	5,981	27,528	13,108	76.8%	99.8%	28.7%	30.7%	21.0%	8.4%	1.9%	9.2%
RST	906190	C	HH433954866	DRC Sum18	2,440	RCWE	0,1,2,3,4	n/a	n/a	n/a	636	93.0%	99.0%	70.0%	19.0%	1.0%	0%	0%	9.0%

Form Key: Forms D and E = Operational, Form A =Seniors only, Form C = Administrative Error (AE)

\* Handscored by DRC in Fall of 2017

\*\* Pearson – Statistics from 2017 and 2016 are included for 902331/VH004490. Volumes were significantly higher in 2016, but reports in 2016 did not split out human and AI scoring, so the 2016 and 2017 column headers are different.

EOC English III (All Report Data Scored by DRC and MI [AI])

IDEAS ID	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact%	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	NS%
851370	Form W, Op	Op Fall 2017	8,382	Content	1, 2, 3, 4	3,672	91%	9%	0%		18%	34%	28%	11%	10%
			8,382	Style	1, 2, 3, 4	3,672	91%	9%	0%		14%	34%	33%	9%	10%
			8,382	Formation (F)	0, 1	3,672	94%	6%		27%	63%				10%
			8,382	Usage (U)	0, 1	3,672	93%	7%		46%	44%				10%
			8,382	Mechanics (M)	0, 1	3,672	93%	7%		28%	62%				10%
8,382	Spelling (S)	0, 1	3,672	96%	4%		21%	69%					10%		
851368	Form X, AE	Op Fall 2016	7,783	Content	1, 2, 3, 4	2,204	83%	17%	0%		17%	39%	38%	5%	2%
			7,783	Style	1, 2, 3, 4	2,204	83%	17%	0%		12%	32%	44%	11%	2%
			7,783	Formation (F)	0, 1	2,204	83%	17%		37%	61%				2%
			7,783	Usage (U)	0, 1	2,204	78%	22%		21%	78%				2%
			7,783	Mechanics (M)	0, 1	2,204	89%	11%		22%	77%				2%
7,783	Spelling (S)	0, 1	2,204	83%	17%		25%	73%					2%		
851368	Form X, AE	Op Summer 2017	686	Content	1, 2, 3, 4	486	94%	6%	0%		56%	26%	2%	0%	15%
			686	Style	1, 2, 3, 4	486	94%	6%	0%		48%	30%	6%	1%	15%
			686	Formation (F)	0, 1	486	95%	5%		57%	27%				15%
			686	Usage (U)	0, 1	486	96%	4%		71%	14%				15%
			686	Mechanics (M)	0, 1	486	93%	7%		57%	28%				15%
686	Spelling (S)	0, 1	486	93%	7%		35%	50%					15%		
851368	Form X, AE	Op Spring 2018	59,941	Content	1, 2, 3, 4	51,890	91%	9%	0%		14%	40%	38%	6%	2%
			59,941	Style	1, 2, 3, 4	51,890	93%	7%	0%		9%	32%	45%	13%	2%
			59,941	Formation (F)	0, 1	51,890	97%	3%		15%	83%				2%
			59,941	Usage (U)	0, 1	51,890	95%	5%		26%	73%				2%
			59,941	Mechanics (M)	0, 1	51,890	96%	4%		20%	78%				2%
59,941	Spelling (S)	0, 1	51,890	97%	3%		13%	86%					2%		
Form W (851370) will be AI scored by MI with human backreads by DRC. Form X (851368), the AE form, will be handscored by DRC supervisors.															

### ELA Grade 3

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
RST	915227	Op	A1598	Pearson 2016 FT	1,582	RCWE	0,1,2,3	n/a	n/a	n/a	339	69.3%	99.4%	52.7%	39.1%	7.1%	0.0%		1.1%
					1,582	Conventions	0,1,2,3	n/a	n/a	n/a	339	69.3%	98.2%	56.6%	32.7%	8.1%	1.5%		1.1%
NWT	913497	Op	AA431426588	Pearson Spring 17	118,416	Expression	0,1,2,3	34,298	13,546	84,911	27,299	71.2%	98.6%	30.0%	56.0%	10.6%	1.6%		1.8%
					118,416	Conventions	0,1,2,3	34,298	13,546	84,910	27,299	68.6%	98.6%	33.3%	46.8%	16.0%	2.1%		1.8%
NWT	913497	Op	AA431426588	DRC Spring 18	62,260	Expression	0,1,2,3	n/a	n/a	n/a	13,242	80%	99%	31%	50%	13%	2%		4%
					62,260	Conventions	0,1,2,3	n/a	n/a	n/a	13,242	77%	99%	16%	58%	20%	2%		4%

### ELA Grade 4

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
LAT	913567	Op	VH170170	Pearson Spring 2017	121,461	RCWE	0,1,2,3,4	35,658	13,901	87,168	28,425	67.3%	98.5%	33.5%	40.0%	21.2%	3.5%	0.5%	1.3%
					121,461	Conventions	0,1,2,3	35,659	13,893	87,168	28,418	69.3%	99.1%	28.1%	45.7%	21.0%	3.9%		1.3%
LAT	913567	Op	VH170170	DRC Spring 2018	62,127	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,196	83%	100%	26%	36%	34%	3%	0%	1%
					62,127	Conventions	0,1,2,3	n/a	n/a	n/a	12,196	81%	100%	25%	36%	34%	4%		1%
RST	982233	Op	VH060330	Pearson 2017 FT	1,500	RCWE	0,1,2,3,4	1,468	150	0	300	77.7%	100.0%	26.0%	51.8%	17.5%	3.1%	0.0%	1.6%
					1,500	Conventions	0,1,2,3	1,468	150	0	300	78.0%	100.0%	19.7%	56.2%	19.4%	3.2%		1.6%

ELA Grade 5

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
LAT	801310	Op	VF821667	DRC Spring 2016	60,357	RCWE	0,1,2,3,4	n/a	n/a	n/a	14,914	77%	99%	45%	42%	11%	1%	0%	1%
					60,357	Conventions	0,1,2,3	n/a	n/a	n/a	14,914	75%	98%	24%	50%	22%	3%		1%
LAT	801310	Op	VF821667	Pearson Spring 2017	11,258	RCWE	0,1,2,3,4	11,045	1,127	0	2,231	86.8%	99.6%	79.7%	13.3%	1.2%	0.1%	0.0%	5.7%
					11,258	Conventions	0,1,2,3	11,045	1,127	0	2,231	81.6%	99.2%	64.7%	25.3%	3.9%	0.3%		5.7%
RST	915510	Op	VH198972	Pearson 2016 FT	1,561	RCWE	0,1,2,3,4	n/a	n/a	n/a	332	69.3%	100.0%	38.5%	40.9%	15.8%	3.7%	0.3%	0.8%
					1,561	Conventions	0,1,2,3	n/a	n/a	n/a	332	69.9%	98.8%	28.4%	42.9%	22.8%	5.1%		0.8%

ELA Grade 6

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
RST	913715	Op	DD502035970	Pearson Spring 2017	128,716	RCWE	0,1,2,3,4	36,320	13,240	93,042	29,065	72.5%	99.0%	32.1%	35.1%	24.5%	5.8%	1.0%	1.4%
					128,716	Conventions	0,1,2,3	36,320	13,240	93,042	29,065	71.3%	98.8%	32.2%	32.5%	25.6%	8.3%		1.4%
NWT	913694	Op	D1466	Pearson Spring 2017	127,628	Expression	0,1,2,3,4	34,718	14,034	93,800	29,433	75.9%	99.4%	40.3%	20.6%	22.9%	10.0%	4.3%	1.8%
					127,628	Conventions	0,1,2,3	34,718	14,034	93,800	29,433	75.0%	99.6%	33.4%	30.2%	23.3%	11.3%		1.8%
NWT	913694	Op	D1466	DRC Spring 2018	58,773	Expression	0,1,2,3,4	n/a	n/a	n/a	11,768	74%	99%	41%	24%	25%	6%	2%	0%
					58,773	Conventions	0,1,2,3	n/a	n/a	n/a	11,768	71%	99%	31%	38%	23%	6%		0%

ELA Grade 7

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
RST	915582	Op	E1567	Pearson Spring 2017	1,630	RCWE	0,1,2,3,4	n/a	n/a	n/a	345	75.7%	99.4%	30.8%	32.5%	22.9%	8.0%	2.8%	3.0%
					1,630	Conventions	0,1,2,3	n/a	n/a	n/a	345	76.2%	100.0%	30.5%	32.6%	23.2%	10.8%		3.0%
NWT	913842	Op	EE43013306	Pearson Spring 2017	128,845	Expression	0,1,2,3,4	37,606	14,582	91,555	30,289	72.7%	98.6%	34.2%	12.7%	20.4%	17.9%	12.6%	2.2%
					128,845	Conventions	0,1,2,3	37,605	14,582	91,555	30,289	71.5%	99.0%	28.5%	20.9%	23.8%	24.7%		2.2%
NWT	913842	Op	EE43013306	DRC Spring 2018	57,320	Expression	0,1,2,3,4	n/a	n/a	n/a	11,538	73%	99%	35%	13%	25%	18%	8%	0%
					57,320	Conventions	0,1,2,3	n/a	n/a	n/a	11,538	70%	99%	27%	23%	29%	20%		0%

ELA Grade 8

Task	IDEAS ID	Spring 2019 Form	PARCC UIN	Source of IRR and SPD Data	Responses Available	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	AI 1st & 2nd Score Count	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Cond Code %
LAT	913958	Op	F1460	Pearson Spring 2017	128,084	RCWE	0,1,2,3,4	36,606	4,234	89,633	19,154	69.6%	99.8%	26.4%	31.2%	25.9%	10.9%	3.1%	2.4%
					128,084	Conventions	0,1,2,3	36,606	4,234	89,634	19,154	71.9%	99.7%	22.9%	30.7%	29.2%	14.7%		2.4%
LAT	913958	Op	F1460	DRC Spring 2018	57,038	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,090	73%	99%	18%	32%	35%	125%	1%	0%
					57,038	Conventions	0,1,2,3	n/a	n/a	n/a	12,090	76%	100%	14%	31%	39%	15%		0%
RST	982327	Op	FF506834510	Pearson 2017 FT	1,625	RCWE	0,1,2,3,4	1,496	165	0	317	74.8%	99.4%	42.8%	22.6%	16.7%	6.1%	2.5%	9.2%
					1,625	Conventions	0,1,2,3	1,496	165	0	317	74.1%	98.1%	35.4%	23.1%	23.4%	8.9%		9.2%

Biology (EOC)

IDEAS ID	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	NS%
845227	Form H, Op	DRC Spring 2017, OP	48,129	0-4	11,158	90%	8%	1%	42%	25%	18%	10%	3%	2%
845227	Form H, Op	DRC Fall 2017, OP	12,420	0-4	2,790	94%	6%	0%	47%	27%	13%	6%	3%	4%
845227	Form H, Op	DRC Summer 2018, OP	3,156	0-4	990	99%	1%	0%	60%	23%	4%	0%	0%	12%
845226	Form K, AE	DRC Fall 2016, OP	12,560	0-4	3,680	91%	9%	0%	45%	24%	11%	6%	3%	12%
845226	Form K, AE	DRC Summer 2017, OP	2,952	0-4	1,128	98%	2%	0%	61%	16%	2%	0%	0%	20%

Biology ERs and CRs (LEAP 2025)

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
965124	ER	B	Spring 2018 FT	4842	Part A (0-3)	4842	69%	29%	2%	9%	34%	36%	20%			
				4842	Part B (0-6)	4842	61%	26%	14%	36%	19%	21%	8%	10%	2%	3%
965129	CR	B, C	Spring 2018 FT	332	0-2	1566	81%	19%	1%	58%	29%	10%				
965237	CR	B, C	Spring 2018 FT	360	0-2	1607	96%	4%	0%	82%	14%	3%				
965295	CR	B, C	Spring 2018 FT	318	0-2	1622	76%	23%	1%	57%	33%	10%				
965286*	ER	A, C	Spring 2018 FT (re-scored Oct. 2018)	5,140	Part A (0-6)	5,140	82%	15%	3%	47%	13%	13%	15%	2%	1%	2%
				5,140	Part B (0-3)	5,140	84%	13%	3%	36%	30%	12%	16%			
965286	ER	A, C	Fall 2018 Op	7,446	Part A (0-6)	1,588	87%	10%	3%	55%	13%	13%	14%	2%	1%	1%
				7,446	Part B (0-3)	1,588	85%	14%	1%	41%	35%	11%	11%			
965190	CR	A	Spring 2018 FT	324	0-2	1626	84%	15%	1%	65%	20%	14%				
965190	CR	A	Fall 2018 Op	7,357	0-2	1,448	93%	7%	0%	78%	13%	7%				
965222	CR	A	Spring 2018 FT	350	0-2	1592	93%	7%	0%	64%	28%	4%				
965222	CR	A	Fall 2018 Op	7,279	0-2	1,516	92%	8%	0%	71%	25%	2%				
965279	CR	A	Spring 2018 FT	316	0-2	1625	75%	25%	1%	46%	31%	22%				
965279	CR	A	Fall 2018 Op	7,540	0-2	1,484	86%	14%	0%	57%	31%	11%				

Form Key: Form A = Administrative Error (AE), Forms B and C = Operational

All data from DRC 2018 field test handscoring. Nonscores excluded.  
 \*ER 965286 FT item was re-scored in October 2018 using updated rubric.

*Grade 3 Science*

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj %	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
957382	ER	Op	Spring 2018 FT	2,768	0-6	536	78%	18%	4%	63%	13%	15%	4%	4%	0%	0%
957435	CR	Op	Spring 2018 FT	1,660	0-2	320	87%	13%	0%	58%	33%	7%				
957418	CR	Op	Spring 2018 FT	1,661	0-2	322	88%	12%	0%	36%	61%	2%				
957409	CR	Op	Spring 2018 FT	1,675	0-2	350	87%	13%	0%	40%	40%	19%				

*Grade 4 Science*

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj %	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
957054	ER	Op	Spring 2018 FT	2,778	0-6	556	74%	23%	3%	6%	13%	40%	37%	3%	0%	0%
957144	CR	Op	Spring 2018 FT	1,668	0-2	326	88%	12%	0%	83%	14%	1%				
957090	CR	Op	Spring 2018 FT	1,665	0-2	330	79%	21%	0%	45%	49%	6%				
957099	CR	Op	Spring 2018 FT	1,657	0-2	314	96%	4%	0%	71%	25%	3%				

*Grade 5 Science*

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj %	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %	SP 7 %	SP 8 %	SP 9 %
959503	ER	Op	Spring 2018 FT	4,992	0-9	4,992	67%	23%	10%	42%	12%	11%	9%	9%	6%	5%	3%	2%	1%
959557	CR	Op	Spring 2018 FT	1,667	0-2	346	89%	7%	4%	29%	51%	19%							
959548	CR	Op	Spring 2018 FT	1,658	0-2	324	96%	4%	1%	69%	12%	19%							
959530	CR	Op	Spring 2018 FT	1,690	0-2	382	98%	2%	0%	56%	7%	37%							



### Grade 6 Science

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
958421	ER	Op	Spring 2018 FT	4,988	Part A (0-3)	4,988	86%	8%	6%	68%	19%	0%	13%			
				4,988	Part B (0-3)	4,988	80%	19%	2%	58%	29%	11%	2%			
				4,988	Part C (0-3)	4,988	85%	12%	3%	62%	17%	19%	2%			
958378	CR	Op	Spring 2018 FT	1,652	0-2	314	86%	14%	0%	81%	14%	55				
958308	CR	Op	Spring 2018 FT	1,653	0-2	316	88%	11%	1%	68%	29%	3%				
958359	CR	Op	Spring 2018 FT	1,648	0-2	320	91%	95	0%	74%	20%	6%				

### Grade 7 Science

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
959635	ER	Op	Spring 2018 FT	4,952	Part A (0-3)	4,952	78%	16%	6%	71%	17%	10%	2%			
				4,952	Part B (0-4)	4,952	81%	15%	4%	71%	19%	8%	1%	0%		
				4,952	Part C (0-2)	4,952	96%	4%	0%	88%	10%	1%				
959748	CR	Op	Spring 2018 FT	1,646	0-2	312	82%	18%	0%	30%	50%	20%				
959657	CR	Op	Spring 2018 FT	1,651	0-2	332	92%	8%	0%	39%	42%	19%				
959715	CR	Op	Spring 2018 FT	1,647	0-2	336	92%	8%	0%	92%	6%	1%				

### Grade 8 Science

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	SP 5 %	SP 6 %
959334	ER	Op	Spring 2018 FT	4,950	Part A (0-3)	4,950	62%	30%	8%	28%	28%	28%	15%			
				4,950	Part B (0-6)	4,950	49%	32%	20%	12%	13%	20%	21%	17%	10%	7%
959309	CR	Op	Spring 2018 FT	1,656	0-2	324	90%	10%	0%	87%	11%	1%				
959291	CR	Op	Spring 2018 FT	1,639	0-2	320	86%	13%	1%	42%	51%	6%				
959221	CR	Op	Spring 2018 FT	1,648	0-2	326	88%	12%	0%	76%	20%	3%				

U.S. History ERs and CRs (LEAP 2025)

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
894256	ER	B	Spring 2017 FT	5,000	Content	0-4	5,000	66%	32%	2%	28%	37%	24%	8%	3%
				5,000	Claims	0-4	5,000	64%	33%	3%	34%	33%	21%	8%	3%
892956	ER	B	Spring 2018 OP	16,722	Content	0-4	10142	93%	7%	0%	27%	30%	26%	11%	4%
				16,722	Claims	0-4	10142	93%	7%	0%	32%	27%	25%	10%	4%
892956	ER	B	Fall 2018 OP	12,519	Content	0-4	8,054	91%	9%	0%	39%	32%	19%	5%	2%
				12,519	Claims	0-4	8,054	92%	8%	0%	45%	27%	18%	5%	1%
894104	ER	A	Spring 2017 FT	5,000	Content	0-4	5,000	62%	33%	5%	31%	34%	22%	9%	4%
				5,000	Claims	0-4	5,000	61%	32%	7%	39%	28%	21%	9%	4%
894104	ER	A	Fall 2017 OP	7,649	Content	0-4	4028	90%	9%	0%	36%	34%	20%	6%	2%
				7,649	Claims	0-4	4028	89%	11%	0%	45%	30%	17%	6%	1%
894104	ER	A	Spring 2018 OP	14,069	Content	0-4	9990	96%	4%	0%	21%	34%	26%	11%	6%
				14,069	Claims	0-4	9990	95%	5%	0%	31%	33%	21%	10%	3%
894104	ER	A	Sum 2018 OP	215	Content	0-4	152	96%	4%	0%	75%	17%	6%	1%	0%
				215	Claims	0-4	152	99%	1%	0%	83%	12%	3%	1%	0%
892955	ER	F	Spring 2017 FT	5,000	Content	0-4	5,000	65%	32%	3%	34%	29%	25%	9%	3%
				5,000	Claims	0-4	5,000	64%	32%	4%	37%	26%	25%	10%	3%
892955	ER	F	Spring 2018 OP	10,506	Content	0-4	5,426	94%	6%	0%	16%	32%	31%	15%	3%
				10,506	Claims	0-4	5,426	93%	7%	0%	21%	28%	30%	15%	3%
894271	CR	F	Spring 2017 FT	1,658		0-2	316	66%	34%	1%	54%	37%	8%		
894271	CR	F	Spring 2018 FT	1,331		0-2	254	82%	18%	0%	29%	48%	23%		
957768	CR	F	Spring 2018 FT	1,557		0-2	294	86%	14%	0%	48%	27%	25%		
894225	CR	B	Spring 2017 FT	1,660		0-2	320	71%	29%	0%	44%	34%	22%		
894225	CR	B	Spring 2018 OP	39,705		0-2	7600	80%	19%	0%	55%	24%	21%		
894225	CR	B	Fall 2018 OP	9,205		0-2	1,694	88%	12%	0%	75%	15%	10%		
892994	CR	B	Spring 2017 FT	1,659		0-2	318	68%	31%	1%	13%	43%	44%		
892994	CR	B	Spring 2018 OP	39,867		0-2	7282	78%	22%	0%	22%	55%	23%		
892994	CR	B	Fall 2018 OP	9,375		0-2	1,728	80%	20%	0%	43%	39%	18%		
894188	CR	A	Fall 2017 OP	6,150		0-2	1,190	85%	15%	0%	53%	33%	14%		
894188	CR	A	Spring 2017 FT	1,655		0-2	310	74%	25%	1%	33%	38%	28%		
894188	CR	A	Spring 2018 FT	1,382		0-2	248	87%	13%	0%	55%	31%	13%		
894188	CR	A	Sum 2018 OP	154		0-2	30	73%	27%	0%	69%	28%	3%		
894149	CR	A	Spring 2017 FT	1,653		0-2	306	76%	21%	3%	44%	26%	30%		
894149	CR	A	Fall 2017 OP	6,056		0-2	1,132	87%	13%	0%	68%	18%	14%		
894149	CR	A	Spring 2018 FT	1,339		0-2	252	84%	16%	0%	64%	19%	17%		
894149	CR	A	Sum 2018 OP	145		0-2	26	92%	8%	0%	88%	10%	3%		
Form Key: Form A = Seniors only, Form B = Administrative Error (AE), Form F = Operational															
All data from DRC handscoring. Nonscores excluded.															
AI scoring models for ER items 892955 and 894104 built by MI using 2500 FT responses, scored 2x and resolved. Op ER responses will be AI scored by MI with DRC human backreads. (See U.S. History 5-level model building data earlier in Appendix.)															
All other Spring 2019 LEAP 2025 USH items (Op 2-point CRs and all AE form items) will be handscored.															

### Social Studies Grade 3

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
801184	CR	Op	Spring 2016 FT	1,281		0-2	248	78%	20%	2%	76%	15%	9%		
801184	CR	Op	Spring 2017 Op	62,961		0-2	11,436	89%	10%	1%	53%	18%	22%		
890683	CR	Op	Spring 2017 FT	1,654		0-2	308	81%	18%	2%	42%	38%	16%		

### Social Studies Grade 4

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
801539	CR	Op	Spring 2016 FT	1,654		0-2	308	71%	25%	3%	29%	37%	30%		
801539	CR	Op	Spring 2017 Op	62,340		0-2	11,406	82%	17%	1%	40%	36%	20%		
890820	CR	Op	Spring 2017 FT	1,654		0-2	308	85%	15%	0%	80%	17%	2%		

### Social Studies Grade 5

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
807773	ER	Op	Spring 2016 FT	5,668	Content	0-4	5,668	78%	21%	1%	62%	25%	12%	2%	0%
					Claims	0-4		79%	20%	1%	67%	23%	9%	1%	0%
890885	CR	Op	Spring 2017 FT	1,650		0-2	300	76%	23%	1%	54%	39%	6%		
890920	CR	Op	Spring 2017 FT	1,647		0-2	294	71%	29%	0%	63%	28%	9%		

### Social Studies Grade 6

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
804889	ER	Op	Spring 2016 FT	5,108	Content	0-4	5,108	67%	32%	1%	42%	44%	12%	1%	0%
					Claims	0-4		68%	31%	1%	52%	38%	9%	1%	0%
804889	ER	Op	Spring 2017 Op	71,724	Content	0-4	39,110	93%	6%	0%	56%	32%	10%	2%	0%
					Claims	0-4		93%	7%	0%	66%	24%	8%	1%	0%
804851	CR	Op	Spring 2016 FT	1,632		0-2	320	73%	28%	0%	46%	50%	5%		
804851	CR	Op	Spring 2017 Op	56,842		0-2	10,362	80%	20%	0%	41%	53%	6%		
949224	CR	Op	Spring 2018 FT	1,629		0-2	300	87%	13%	0%	45%	53%	2%		

### Social Studies Grade 7

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
805627	ER	Op	Spring 2016 FT	5,066	Content	0-4	5,066	73%	25%	2%	45%	41%	12%	2%	0%
					Claims	0-4		73%	25%	2%	57%	31%	11%	2%	0%
805627	ER	Op	Spring 2017 Op	68,833	Content	0-4	34,732	91%	9%	0%	48%	34%	13%	4%	1%
					Claims	0-4		91%	8%	0%	56%	28%	11%	4%	1%
891266	CR	Op	Spring 2017 FT	1,648		0-2	296	75%	25%	0%	43%	43%	14%		
805632	CR	Op	Spring 2016 FT	1,626		0-2	314	83%	17%	0%	42%	34%	24%		
805632	CR	Op	Spring 2017 Op	56,280		0-2	10,274	80%	19%	1%	47%	28%	25%		

### Social Studies Grade 8

IDEAS ID	Item Type	Spring 2019 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %
808905	ER	Op	Spring 2016 FT	5,068	Content	0-4	5,068	65%	33%	2%	30%	36%	25%	7%	2%
					Claims	0-4		64%	34%	2%	30%	37%	25%	7%	2%
808905	ER	Op	Spring 2017 Op	65,286	Content	0-4	30,674	89%	11%	1%	32%	30%	25%	9%	3%
					Claims	0-4		88%	11%	1%	32%	29%	25%	9%	4%
808955	CR	Op	Spring 2016 FT	1,623		0-2	320	79%	21%	0%	39%	40%	21%		
808955	CR	Op	Spring 2017 Op	54,395		0-2	10,174	77%	22%	0%	32%	51%	17%		
892278	CR	Op	Spring 2017 FT	1,656		0-2	312	79%	20%	1%	43%	41%	15%		
892278	CR	Op	Spring 2018 Op	55,340		0-2	10,110	78%	21%	0%	43%	44%	13%		

## References

---

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012, January). Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments. Austin, TX: Pearson. Retrieved from [http://researchnetwork.pearson.com/wpcontent/uploads/Bulletin21\\_Evidence\\_Based\\_Standard\\_Setting.pdf](http://researchnetwork.pearson.com/wpcontent/uploads/Bulletin21_Evidence_Based_Standard_Setting.pdf)
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publication.
- Center for Assessment. (2017, June). LEAP 2017: English language arts -grade 6 summary – comparability with PARCC performance standards (Memorandum). Dove, NH.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Data Recognition Corporation. (2016). *Interpretive guide: Grades 3–8 ELA and math* Maple Grove, MN.
- Dorans, N. J., & Schmitt, M. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47) Princeton, NJ: Educational Testing Service.
- Educational Testing Service, Pearson, & Measured Progress. (2016). *Final technical report for 2015 administration. PARCC*. Retrieved from <https://eric.ed.gov/?q=source%3a%22Partnership+for+Assessment+of+Readiness+for+College+and+Careers%22&id=ED599097>
- Green, D. R. (1975). Procedures for assessing bias in achievement tests. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

- Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15, 1-5.
- Kim, S., & Kolen, M. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models (Version 1.0) [Computer software]. Iowa City, IA: University of Iowa.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer-Verlag.
- 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.
- Lu, Y., & Sireci, S. G., (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(40), 29-37.
- 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.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176.
- Muraki, E., & Bock, R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago, IL: Scientific Software.
- Pearson. (2015). Performance level setting technical report. PARCC. Retrieved from <https://eric.ed.gov/?q=source%3a%22Partnership+for+Assessment+of+Readiness+for+College+and+Careers%22&id=ED599097/>.
- Pearson. (2017). PARCC: Final technical report for 2016 administration. PARCC. Retrieved from <https://eric.ed.gov/?q=source%3a%22Partnership+for+Assessment+of+Readiness+for+College+and+Careers%22&id=ED599197>.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Schumacker, R. E. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions*, 10(1), 479.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210.
- Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from <http://www.cehd.umn.edu/NCEO/OnlinePUBs/Policy14.htm>
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30(3), 233–251.