2021 LEAP 2025 Grades 3-8
Operational Technical Report
English Language Arts and Mathematics

Submitted to the
Louisiana Department of Education

January 2022

DATA RECOGNITION
DRC
CORPORATION

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 6
E.1 Overview of This Report 6
E.2 Administration 7
E.3 Student Performance 8
E.4 Validity and Test Scores 8
Chapter 1: Introduction 10
1.1 Background 10
1.2 Purpose of the LEAP 2025 10
1.3 Design of the LEAP 2025 11
Chapter 2: The Uses of Test Scores 14
2.1 Uses of Test Scores 14
2.2 Test-Level Scores 14
2.3 Scale Scores 15
2.4 Levels of Achievement 15
2.5 Use of Test-Level Scores 15
2.6 Category- and Subcategory-Level Subscores 15
2.7 Use of the Reporting Category- and Subcategory-Level Ratings 16
Chapter 3: Test Content Development 17
3.1 Defining the Specific Test Blueprint 19
3.2 English Language Arts Test Blueprints and Test Designs 19
3.3 Mathematics Test Blueprints and Test Designs 29
3.4 Item Development and Selection 42
3.5 Considerations of Test Fairness in Item Development 42
3.6 New Meridian Item Reviews 42
3.7 Operational Test Selection 43
3.8 Universal Design 43
3.9 Accommodations and Designated Supports 45
3.10 Item and Task Specifications 46
3.11 Summary 47
Chapter 4: Test Administration 49
4.1 Return Material Forms and Guidelines 55

4.2 Security Checklists 55
4.3 Interpretive Guides 58
4.4 Test Security Measures 58
4.5 Data Forensic Analyses 58
4.5.1 Response Change Analysis 58
4.5.2 Score Fluctuation Analysis 58
4.5.3 Item Exposure Monitoring 59
4.5.4 Web Monitoring 59
4.5.5 Plagiarism Detection 59
4.6 Test Administration 59
4.6.1 Time 59
4.6.2 Accommodations 60
4.7 Summary 65
Chapter 5: Scoring of Constructed-Response and Technology-Enhanced Items 67
5.1 Constructed-Response Item Scoring Process 67
5.1.1 Selection of Scoring Evaluators 68
5.1.2 Security 68
5.1.3 Handscoring Training Process 69
5.1.4 Monitoring the Scoring Process 73
5.2 Inter-Rater Reliability 75
5.3 Multiple-Choice and Multiple-Select Item Scoring Process 81
5.4 Summary 81
Chapter 6: Operational Data Analyses 83
6.1 Test-Level Statistics 83
6.2 Item-Level Statistics 85
6.3 Item Response Theory 109
6.4 Calibration and Linking 109
6.5 Summary 128
Chapter 7: Test Results 129
7.1 Current Administration Data 136
7.1.1 Description of Each Type of Report 139
Chapter 8: Performance-Level Setting 141
8.1 PARCC Performance-Level Setting Process for English Language Arts and Mathematics 141
8.2 Cut Scores 141

8.2.1 Reporting Category Cut Scores 142
8.3 Summary 143
Chapter 9: Evidence of Validity 144
9.1 Construct-Irrelevant Variance and Construct Underrepresentation 145
9.2 Reliability 145
9.2.1 Test Reliability 146
9.2.2 Standard Error of Measurement 147
9.2.3 Conditional Standard Error of Measurement 147
9.2.4 Classification Accuracy and Consistency 152
9.2.5 Convergent Validity 155
9.3 Principal Components Analysis 156
9.4 Analyses by Reporting Categories and Subcategories 158
9.4.1 Correlations among Reporting Categories and Subcategories 158
9.4.2 Reliability of Reporting Categories and Subcategories 163
9.4.3 Standard Error of Measurement of Reporting Categories and Subcategories 164
9.5 Divergent (Discriminant) Validity 168
9.6 Regression of LEAP 2025 from 2019 to 2021 168
9.7 Summary 171
Chapter 10: Fairness 173
10.1 Minimizing Bias through Careful Test Development 174
10.2 Evaluating Bias through Differential Item Functioning (DIF) Statistics 174
10.2.1 DIF Statistics for Demographic Groups 176
10.2.2 DIF Statistics for Test Language 179
10.3 Evaluating Bias through Impact Analysis 179
10.3.1 Reliability 180
10.3.2 Effect Size 188
10.4 Mode Effect Study 203
10.5 Summary 204
Appendix A—Accommodated Print Form Creation 205
Appendix B—Transadaptation Process for Spanish Mathematics Forms 207
Appendix C—LEAP 2025 Spring 2021 Handscoring/AI Documentation 209
References 365

Executive Summary

This report is a technical summary of the 2021 administration of the Louisiana Educational Assessment Program (LEAP 2025) in English language arts (ELA) and mathematics for grades 3 through 8. The LEAP 2025 summative assessments in ELA and mathematics are 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 New Meridian 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 2021 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 2021 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 2021 LEAP 2025 assessments, which include the following major parts: the classical item analysis; calibration, scaling, and linking using IRT models; and student scoring.

Test Results (Chapter 7)

Chapter 7 of the technical report contains information on the results of the spring 2021 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 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 2021, Louisiana administered the LEAP 2025 summative assessments in ELA and mathematics to students in grades 3–8. A paper-based test (PBT) option was administered in grades 3 and 4, and the computer-based test (CBT) was administered in grades 3–8. The CBTs were administered from April 26 to May 26, 2021. The PBTs were administered from April 28 to 30, 2021. Test administration is discussed in Chapter 4 of this report.

A total of 103 school systems and 32 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 2021 LEAP 2025 Grades 3-8

Grade	Accountable in ELA	Percentage Reportable in ELA	Accountable in Mathematics	Percentage Reportable in Mathematics*
3	≥50,130	98.75%	≥50,540	98.79%
4	≥50,290	98.67%	≥50,590	98.69%
5	≥50,270	98.95%	≥50,270	98.97%
6	≥52,240	98.50%	≥52,240	98.51%
7	≥53,190	98.24%	≥53,210	98.29%
8	≥52,780	98.31%	≥52,820	98.38%

**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 in 2021 who were classified in each of the achievement levels for ELA and mathematics.

Table E.2 Percentage of Students Classified in Achievement Levels Using 2021 Census Data: English Language Arts

Grade	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
3	19.3	19.0	23.1	33.4	5.2
4	13.7	19.1	25.7	32.3	9.3
5	10.7	24.0	28.1	32.7	4.4
6	12.1	26.1	28.3	28.7	4.9
7	13.4	18.3	26.2	29.1	13.0
8	14.3	16.4	25.2	34.9	9.2

Table E.3 Percentage of Students Classified in Achievement Levels Using 2021 Census Data: Mathematics

Grade	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
3	18.2	22.9	25.3	28.3	5.3
4	20.0	23.1	25.2	29.7	2.1
5	18.5	28.6	26.7	23.2	3.1
6	18.8	27.9	28.9	21.9	2.5
7	12.0	33.0	32.6	20.5	1.9
8	27.3	25.8	25.2	20.2	1.5

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 2021 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 Partnership for Assessment of Readiness for College and Careers (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 2021 and presents evidence for the validity of the 2021 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, 2018–2019, and 2020–2021 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 in subsequent years), were administered forms for ELA and mathematics that consisted of New Meridian (formerly PARCC) assessment items while developing some Louisiana-owned items to enhance the New Meridian item bank. This allowed for the continued comparability to forms administered in the 2014–2015 and 2015–2016 school years. Louisiana received approval from the federal and state governments to waive the requirement to adminsiter the spring 2020 assessment due to school facilities closing in March 2020 due to COVID-19.

The information that follows describes the technical aspects of the 2021 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 Believe to Achieve (www.louisianabelieves.com/resources/about-us/believe-to-achieve). The LEAP 2025 supports this vision by measuring the full range of student performance 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 2021 LEAP 2025 test blueprints and test design for ELA and mathematics are based on the ELA https://resources.newmeridiancorp.org/ela-test-design/ and mathematics https://resources.newmeridiancorp.org/math-test-design/ blueprints of New Meridian's full forms. The 2021 LEAP 2025 test blueprints and test design for ELA and mathematics differ from the New Meridian blueprints and design in order to reduce testing time while maintaining full coverage and including a variety of standards.

The 2021 LEAP 2025 ELA blueprints kept a similar design as the design of New Meridian's full 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 and Literary Analysis Task **or** Narrative Writing Task—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 2021 LEAP 2025 Spring ELA blueprints are also different 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 2021 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 New Meridian's 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 assessment designs have three sessions and a total testing time of 240 minutes. However, New Meridian 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. New Meridian 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 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 New Meridian'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.

2.3 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.

2.4 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 2025 *Interpretive Guide* (see Chapter 7).

2.5 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.

2.6 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 reporting categories: Major Content, Additional & Supporting Content, Expressing Mathematical Reasoning, or Modeling & Application., as described in Table 3.8, *Overview of LEAP 2025 Mathematics Task Types and Reporting Categories*. Reporting categories are further broken down into subcategories, which vary by grade level. 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.

2.7 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 2021 LEAP 2025 test specifications consisted of a test blueprint and a test design for each grade and content area. The 2021 blueprints and test designs were closely aligned to blueprints of New Meridian's full forms. The specific content area and grade-level test blueprints for the 2021 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 2021 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 2021 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, 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 2021 LEAP 2025 assessments for grades 3–8 use the same reporting categories that were used in spring 2019. 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 2021 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 New Meridian and Louisiana-

owned item banks. DRC contracted with New Meridian and was provided access to the entire bank of items and passage sets that could potentially be used on operational forms. The acquired items and passages and the Louisiana-owned items and passage sets make up the available item pool for the 2021 LEAP 2025 forms construction. 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 [English Learners];
- 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[1] 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 New Meridian and Louisiana-owned items aligned to the Louisiana Student Standards. Louisiana had access to the complete New Meridian item bank for forms administered in spring 2021. For grades 3 and 4, the contract with New Meridian provided for the use of enough tems 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 and passage sets developed specifically for Louisiana. These items and passage sets became the available item pool used to construct the 2021 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 New

[1] 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.

Meridian's 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 2020-2021 *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 26-May 26, 2021) and during a PBT testing window (April 28-April 30, 2021). The 2021 ELA assessment was the same as the 2019 assessment with one exception. An item in grade 7 was edited from its previous use. 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
Reading	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.
Writing	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 study 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 New Meridian 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	1	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	1	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	Reading: Reading Literary Text/Reading Informational Text/Reading Vocab*	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	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	2	18	24	
			Writing: Knowledge and Use of Language Conventions	0	0		6		6		
			Total	22	44	2	27	24	71	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	1	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 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	
	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	1	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	0	6	12		
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	2	24	30	
			Writing: Knowledge and Use of Language Conventions	0	0		6		6		
			Total	24	48	2	38	26	86	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	1	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 Literary / Informational Texts	1	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	1	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	2	24	30	
			Writing: Knowledge and Use of Language Conventions	0	0		6		6		
			Total	24	48	2	38	26	86	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	1	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	1	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	Reading (Reading Literary Text/Reading Informational Text/Reading Vocabulary)	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		Reading: Reading Informational Text/Reading Vocab*			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	2	24	30	
			Writing: Knowledge and Use of Language Conventions	0	0		6		6		
			Total	28	56	2	34	30	90	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	1	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 Literary / Informational Texts	1	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	1	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	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		Reading: Reading Informational Text/Reading Vocab*			0	0			RI standards 1-3, 5, 7-10; vocabulary standards RI.4, L.4, L.5	
	Totals	2		10	20	0	0	10	20		
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 Conventions	0	0		6		6		
			Total	28	56	2	38	30	94	94	

**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, retell the story in another narrative form or from a different point of view).
    - 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. 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 26-May 26, 2021) or during a PBT testing window (April 28-April 30, 2021). The 2021 mathematics assessment was the same as the 2019 assessment, with the following exceptions: grade 3 had one item replaced and grades 3, 4, 5, 6, and 8 had some items changed from operational status to field test/placeholder stature. Each test session included the four mathematics categories, 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	*Major Content:* solve problems involving the major content for the grade level. *Additional & Supporting Content:* solve problems involving the additional and supporting content for the grade level.	Can involve any or all practices
Type II	Written arguments/ justifications, critique of reasoning, or precision in mathematical statements	*Expressing Mathematical Reasoning:* express mathematical reasoning 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	*Modeling & Application:* solve real-world problems engaging particularly in the modeling 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 study 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 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
Total	**62**	**62**	**62**	**66**	**66**	**66**

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	• Products and Quotients/Solve Multiplication and Division Problems • Solve Problems with Any Operation • Fractions as Numbers and Equivalence • Solve Time, Area, Measurement, and Estimation Problems
4	• Compare and Solve Problems with Fractions • Solve Multi-step Problems • Multiplicative Comparison and Place Value
5	• Operations with Decimals/Read, Write, and Compare Decimals • Solve Fraction Problems • Interpret Fractions, Place Value, and Scaling • Recognize, Represent, and Determine Volume/Multiply and Divide Whole Numbers
6	• Rational Numbers/Multiply and Divide Fractions • Ratio and Rate • Expressions, Inequalities, and Equations
7	• Analyze Proportional Relationships and Solve Problems • Operations with Rational Numbers • Expressions, Inequalities, and Equations
8	• Radicals, Integer Exponents, and Scientific Notation • Proportional Relationships, Linear Equations, and Functions • Solving Linear Equations/Systems of Linear Equations • Congruence and Similarity/Pythagorean Theorem

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
TOTAL	**37**	**40**	**3**	**10**	**3**	**12**	
	TOTAL TASKS		**43**	**TOTAL POINTS**		**62**	

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
TOTAL	**37**	**40**	**3**	**10**	**3**	**12**	
	TOTAL TASKS		**43**	**TOTAL POINTS**		**62**	

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
TOTAL	**37**	**40**	**3**	**10**	**3**	**12**	
	TOTAL TASKS		**43**	**TOTAL POINTS**		**62**	

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
TOTAL	**36**	**40**	**4**	**14**	**3**	**12**	
	TOTAL TASKS		**43**	**TOTAL POINTS**		**66**	

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
TOTAL	**36**	**40**	**4**	**14**	**3**	**12**	
	TOTAL TASKS		**43**	**TOTAL POINTS**		**66**	

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	
	TOTAL TASKS		**42**	**TOTAL POINTS**		**66**	

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 documented calculator accommodation. For session one of the mathematics test in grades 6–8, students are prohibited from using calculators, except those students with a documented 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
TOTAL (Operational Only)	**15**	**21**	**14**	**20**	**14**	**21**	**43**	**62**
Test Duration (minutes)*	**75**		**85**		**75**		**235**	

**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
TOTAL (Operational Only)	**15**	**21**	**14**	**20**	**14**	**21**	**43**	**62**
Test Duration (minutes)*	**75**		**85**		**75**		**235**	

**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
TOTAL (Operational Only)	**15**	**21**	**14**	**20**	**14**	**21**	**43**	**62**
Test Duration (minutes)*	**75**		**85**		**75**		**235**	

**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
TOTAL (Operational Only)	**16–20**	**20**	**12-13**	**26**	**11–13**	**20**	**43**	**66**
Test Duration (minutes)*	**60**		**90**		**90**		**240**	

**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
TOTAL (Operational Only)	**16–20**	**20**	**12-13**	**26**	**11–13**	**20**	**43**	**66**
Test Duration (minutes)*	**60**		**90**		**90**		**240**	

**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
TOTAL (Operational Only)	**15–20**	**22**	**10–13**	**25**	**10–12**	**19**	**42**	**66**
Test Duration (minutes)*	**60**		**90**		**90**		**240**	

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

The following item types were used in the 2021 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.

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	2	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 2021 LEAP 2025 ELA and mathematics assessments came from New Meridian's and Louisiana-owned item banks.

The items selected for use on the 2021 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 New Meridian 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 New Meridian 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 New Meridian Item Reviews

As part of New Meridian's ongoing item development practices, several educator committees had already been convened to conduct rigorous reviews of every passage and item developed for the New Meridian 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, accessibility guidelines, selected metadata fields, and a 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 New Meridian's item review processes and procedures can be found at the New Meridian Resource Center. 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. 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 New Meridian 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 Operational Test Selection

The operational test administered in the 2021 spring administration were the same forms used in the 2019 spring administration, with the following exceptions. In grade 3 math, one item was replaced with an operational item from 2018 that aligned to the same sublcaim and had the same score point. In grade 7 ELA, one item was edited from its previous use. For information regarding item and form seletion, please refer to the *2019 LEAP 2025 Grades 3-8 Operational Technical Report: English Language Arts and Mathematics.*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.

3.8 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 test forms administered in 2021 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. 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 the requirements followed by PARCC and New Meridian.

Table 3.24 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.

3.9 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 IEPs, 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 A, *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 B, "Forms Development Process for Spanish Translations Forms."

3.10 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.

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, 2021 LEAP 2025 items were selected according to the elements of universal design.

3.11 Summary

In summary, the overall purpose of this chapter is to explicate the procedures used in the development of the forms administered during the spring 2021 LEAP 2025 administation. . 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 2021 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 2021 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 Test Coordinator Manuals: *LEAP 2025 Computer-Based Test Coordinator Manual* and *LEAP 2025 Paper-Based Test Coordinator 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 Coordinator 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. Test Schedule
   7.4. Extended Time for Testing
   7.5. Extended Breaks
   7.6. Makeup Testing
   7.7. Test Administration Resources
8. Testing Times
9. District Test Coordinator
   9.1. Conduct Training Session
   9.2. Receive Test Materials
   9.3. Large-print and Braille Test Materials and Communication Assistance Scripts (CAS)
   9.4. Accommodated Materials
   9.5. Verify and Distribute Test Materials to School Test Coordinators
   9.6. Request Additional Test Materials and Bar-code Labels
   9.7. Collect Materials from Schools After Testing
   9.8. Used and Unused Consumable Test Booklets (Defined)
   9.9. Unscorable Documents and Unscorable Document Labels
10. Directions for Returning Test Materials to DRC in May
   10.1. Pickup 1
   10.2. Pickup 2
   10.3. Pickup 3
   10.4. Final Checklist for Returning Test Materials to DRC
11. School Test Coordinator
   11.1. Receive and Verify Test Materials
   11.2. Conduct Test Administration and Security Training Session
   11.3. Supervise Application of Bar-code Labels and Coding of Consumable Test Booklets
   11.4. Soiled, Damaged, and Other Unscorable Consumable Test Booklets
   11.5. Verify and Distribute Materials to Test Administrators
   11.6. Supervise Test Administration
   11.7. Collect Test Materials
   11.8. Used and Unused Consumable Test Booklets (Defined)
   11.9. Coding Responsibilities of Principals—Before Testing
   11.10. Coding Responsibilities of Principals—Before and After Testing
   11.11. Coding Responsibilities of Principals—After Testing
12. Directions for Returning Test Materials to the DTC
   12.1. Pickup 1
   12.2. Pickup 2
   12.3. Pickup 3
   12.4. Final Checklist for Returning Materials to the DTC
13. Void Notification
14. Index

Computer-Based Administration *Test Coordinator Manual* Table of Contents

1. Key Dates
2. Resources Available in DRC INSIGHT Portal (eDIRECT) Spring 2021
3. Alerts
4. Oath of Security and Confidentiality Statements
5. General Information
   5.1. DRC INSIGHT Portal (eDIRECT) and INSIGHT
6. LEAP 2025
7. Test Security
   7.1. Key Definitions
   7.2. Violations of Test Security
8. Testing Guidelines
   8.1. Testing Eligibility
   8.2. Testing Conditions
   8.3. Test Schedule
   8.4. Extended Time for Testing
   8.5. Extended Breaks
   8.6. Accommodations
   8.7. Makeup Testing
   8.8. Test Administration Resources
9. Testing Times for Grades 3 through 8
10. Roles and Responsibilities
    10.1. District Test Coordinator
    10.2. School Test Coordinator
    10.3. Technology Coordinator
11. Managing Test Tickets
    11.1. Student Transfers
    11.2. Locked Test Tickets
    11.3. Technical Issues
    11.4. Invalidating Test Tickets
12. Resources for Online Testing
    12.1. Test Administration Manuals
    12.2. DRC INSIGHT Portal (eDIRECT) User Guides
    12.3. LEAP 2025 Accommodations and Accessibility Features User Guide
    12.4. INSIGHT Technology User Guide
    12.5. Online Tools Training (OTT)
    12.6. Student Tutorials
13. Void Notification

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 and 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 Learners (ELs)
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 through 8
    9.1. Makeup Testing
    9.2. Testing Conditions
10. Online Tools Training
11. Student Tutorials
12. Special Populations and Accommodations
    12.1. IDEA Special Education Students
    12.2. Students with One or More Disabilities According to Section 504
    12.3. Gifted and Talented Special Education Students
    12.4. Test Accommodations for Special Education and Section 504 Students
    12.5. Special Considerations for Deaf and Hard of Hearing Students
    12.6. English Learners (ELs)
13. Test Materials
    13.1. Receipt Directions to Students
14. General Instructions
    14.1.Reading Directions to Students
15. LEAP 2025: Grades 3-8 English Language Arts (All Sessions)
16. LEAP 2025: Grades 3-8 Mathematics (All Sessions)
17. LEAP 2025: Grades 3-8 Science (Sessions 1-3)
18. LEAP 2025: Grades 3-8 Social Studies (Grades 3-4 Sessions 1-2, Grades 5-8 Sessions 1-3)

19. Post-test Procedures
    19.1. Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    19.2. Returning Test Materials to the School Test Coordinator
20. 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 Coordinator 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 place 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

Summary for District		
Science and ELA/Math Test Materials		Exact Number of Boxes Shipped to DRC
Pickup 1: UPS Ground Service (automatic pickup date)	**SCORABLE MATERIALS:**	5
	Used Science answer documents	
	Used ELA and Math consumable test booklets	
Pickup 2: UPS Ground Service (automatic pickup date)	**SCORABLE MATERIALS:**	
	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	
	NONSCORABLE MATERIALS:	
	All unused Science answer documents	
	All unused ELA/Math consumable test booklets	
Pickup 3: Assessment Distribution Services (ADS)	**NONSCORABLE MATERIALS:**	
	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		
Social Studies Test Materials		Exact Number of Boxes Shipped to DRC
Pickup 1: UPS Ground Service (automatic pickup date)	**SCORABLE AND NONSCORABLE MATERIALS:**	
	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	
District	Status
	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 Coordinator 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.

4.5 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 Fluctuation Analysis
- Item Exposure Monitoring
- Web Monitoring
- Plagiarism Detection

4.5.1 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.

4.5.2 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 administators' change in

performance during the same time frame. Schools and test administrators were identified when the level of change was statistically unexpected.

4.5.3 Item Exposure Monitoring

Due to re-use of the 2019 operational forms for the spring of 2021 administration, item performance was examined to ensure that item content had not been exposed. Freuquently during the testing window, every item's moving *p*-value and point-biserial averages were produced. If an item's moving average *p*-value was larger than expected compared to the previous administration's the item was flagged. Additionally, plots were produced for a visual inspection of the day-to-day patterns of item performance.

4.5.4 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.

4.5.5 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.

4.6 Test Administration

The 2021 assessments were administered to students within the state testing window of April 26 through May 26, 2021. The paper testing window was April 28 through 30, 2021. Each session of the assessment within each content area of the LEAP 2025 assessments was required to be administered in one block of 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

4.6.2 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 2021 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	≥1,590	4.24%	≥1,570	4.19%
3	Braille	<50	NR	<50	NR
3	Large Print	<50	NR	<50	NR
3	Answers Recorded	≥520	1.40%	≥520	1.39%
3	Extended Time	≥3,380	9.01%	≥3,370	8.99%
3	Transferred Answers	≥110	0.31%	≥110	0.31%
3	Individual/Small Group Administration	≥3,200	8.55%	≥3,180	8.50%
3	Tests Read Aloud	≥2,400	6.39%	≥2,650	7.07%
4	No Accommodation	≥1,420	4.31%	≥1,440	4.36%
4	Braille	<50	NR	<50	NR
4	Large Print	<50	NR	<50	NR
4	Answers Recorded	≥400	1.22%	≥400	1.22%
4	Extended Time	≥3,120	9.45%	≥3,130	9.49%
4	Transferred Answers	≥130	0.42%	≥130	0.42%
4	Individual/Small Group Administration	≥2,930	8.86%	≥2,940	8.91%
4	Tests Read Aloud	≥2,340	7.09%	≥2,500	7.57%

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

English Learner Accommodation Type					
		English Language Arts		Mathematics	
Grade	Accommodation	Number	Percentage	Number	Percentage
3	No Accommodation	≥180	0.50%	≥150	0.40%
3	Extended Time	≥1,080	2.89%	≥1,100	2.94%
3	Individual/Small Group Administration	≥740	1.98%	≥760	2.04%
3	English/Native Language Word-to-Word Dictionary	≥160	0.43%	≥140	0.38%
3	Test Administered by ESL Teacher	≥50	0.15%	<50	NR
3	Directions Read Aloud/Clarified in Native Language	≥50	0.14%	<50	NR
4	No Accommodation	≥150	0.47%	≥130	0.39%
4	Extended Time	≥790	2.40%	≥830	2.52%
4	Individual/Small Group Administration	≥540	1.66%	≥560	1.70%
4	English/Native Language Word-to-Word Dictionary	≥160	0.51%	≥150	0.48%
4	Test Administered by ESL Teacher	<50	NR	<50	NR
4	Directions Read Aloud/Clarified in Native Language	<50	NR	<50	NR

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	≥230	0.62%	≥230	0.62%
3	Large Print	<50	NR	<50	NR
3	Answers Recorded	≥80	0.22%	≥80	0.23%
3	Extended Time	≥2,370	6.32%	≥2,370	6.33%
3	Transferred Answers	<50	NR	<50	NR
3	Individual/Small Group Administration	≥1,870	4.98%	≥1,870	4.99%
3	Tests Read Aloud	≥790	2.11%	≥1,020	2.73%
4	No Accommodation	≥250	0.77%	≥240	0.73%
4	Large Print	<50	NR	<50	NR
4	Answers Recorded	≥70	0.22%	≥70	0.23%
4	Extended Time	≥2,670	8.08%	≥2,660	8.06%
4	Transferred Answers	<50	NR	<50	NR
4	Individual/Small Group Administration	≥2,020	6.13%	≥2,030	6.15%
4	Tests Read Aloud	≥850	2.57%	≥1,080	3.27%

Table 4.5 Number and Percentage of Students Receiving Online Accommodations by Accommodation Type, as valued in DRC INSIGHT (eDIRECT)

Online Accommodation Type					
		English Language Arts		Mathematics	
Grade	Accommodation	Number	Percentage	Number	Percentage
3	Text-to-Speech	≥810	6.74%	≥2,450	20.35%
3	Human Read Aloud	≥460	3.80%	≥740	6.20%
3	Native Language Word-to-Word Dictionary	≥210	1.76%	≥200	1.67%
3	Directions in Native Language	≥110	0.95%	≥90	0.80%
3	Transferred Answers	≥60	0.53%	≥60	0.52%
3	Answers Recorded	≥180	1.51%	≥180	1.50%
3	Extended Time	≥2,680	22.21%	≥2,670	22.18%
3	Individual/Small Group Administration	≥1,990	16.51%	≥1,990	16.55%
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	≥1,420	8.66%	≥3,240	19.73%
4	Human Read Aloud	≥730	4.48%	≥1,160	7.08%
4	Native Language Word-to-Word Dictionary	≥250	1.55%	≥240	1.49%
4	Directions in Native Language	≥100	0.64%	≥90	0.58%
4	Transferred Answers	≥90	0.60%	≥90	0.60%
4	Answers Recorded	≥270	1.64%	≥270	1.64%
4	Extended Time	≥3,980	24.17%	≥3,980	24.22%
4	Individual/Small Group Administration	≥3,320	20.19%	≥3,330	20.26%
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	≥4,860	9.77%	≥7,830	15.77%
5	Human Read Aloud	≥2,510	5.05%	≥3,310	6.68%
5	Native Language Word-to-Word Dictionary	≥450	0.91%	≥400	0.80%
5	Directions in Native Language	≥170	0.36%	≥140	0.28%
5	Transferred Answers	≥210	0.44%	≥220	0.44%
5	Answers Recorded	≥630	1.27%	≥630	1.27%
5	Extended Time	≥11,520	23.15%	≥11,500	23.15%
5	Individual/Small Group Administration	≥9,130	18.35%	≥9,150	18.42%
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	≥5,140	10.00%	≥7,520	14.65%
6	Human Read Aloud	≥2,120	4.12%	≥2,720	5.31%
6	Native Language Word-to-Word Dictionary	≥620	1.22%	≥560	1.11%
6	Directions in Native Language	≥120	0.25%	≥80	0.16%
6	Transferred Answers	≥140	0.28%	≥140	0.28%
6	Answers Recorded	≥370	0.73%	≥370	0.73%
6	Extended Time	≥11,690	22.73%	≥11,660	22.72%
6	Individual/Small Group Administration	≥8,230	16.01%	≥8,230	16.04%
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	≥5,000	9.59%	≥7,160	13.76%
7	Human Read Aloud	≥1,890	3.62%	≥2,330	4.48%
7	Native Language Word-to-Word Dictionary	≥660	1.28%	≥590	1.14%
7	Directions in Native Language	≥110	0.22%	≥60	0.13%
7	Transferred Answers	≥120	0.24%	≥120	0.23%
7	Answers Recorded	≥190	0.37%	≥190	0.37%
7	Extended Time	≥11,390	21.83%	≥11,340	21.78%
7	Individual/Small Group Administration	≥7,470	14.32%	≥7,440	14.30%
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	≥4,730	9.17%	≥6,790	14.82%
8	Human Read Aloud	≥1,720	3.33%	≥2,150	4.69%
8	Native Language Word-to-Word Dictionary	≥750	1.47%	≥680	1.50%
8	Directions in Native Language	≥130	0.26%	≥90	0.20%
8	Transferred Answers	≥80	0.17%	≥80	0.19%
8	Answers Recorded	≥160	0.31%	≥150	0.34%
8	Extended Time	≥10, 960	21.21%	≥10,640	23.22%
8	Individual/Small Group Administration	≥6,880	13.33%	≥6,700	14.62%
8	Accommodated Paper	<50	NR	<50	NR
8	Braille	<50	NR	<50	NR
8	Communication Assistance Scripts	<50	NR	<50	NR

4.7 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 2021 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 C. Seven 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	Q20	Q7	Pearson
ELA grade 6	N/A	Q9, Q14	Pearson
ELA grade 7	N/A	Q9, Q14	Pearson
ELA grade 8	N/A	Q7, Q20	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 2021 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

Whether training and scoring are conducted within a DRC facility or done remotely, security is essential to our handscoring process. When users log into DRC's secure, web-based scoring application, ScoreBoard, they are required to read and accept our security policy before they are allowed to access any project. For each project, scorers are also required to read and sign non-disclosure agreements, and during training emphasis is always given to what security means, the importance of maintaining security, and how this is accomplished.

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.

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 New Meridian for ELA and math. These materials 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 2021 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.[2] 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, readers received additional item-specific training on the abbreviated item they were going to score. This consisted of an item-specific anchor set and two item-specific practice sets. After completing the abbreviated item training, the readers 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

Set Type	Prototype Item Training	Abbreviated Item Training	Annotated
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.			

[2] 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 2021 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 in 2019 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). The sets updated in 2019 were used during the 2021 scoring process.

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

	Perfect Agreement	Perfect Plus Adjacent Agreement
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)*

Composite (multipart) 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 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, they 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. 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

Pearson 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 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 with the readers, 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 C.

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, and Table 5.10 for Spanish mathematics items.

As shown in Table 5.8, raters demonstrated at least 99% perfect and adjacent agreement for all ELA handscored items. As shown in Table 5.9 raters demonstrated at least 98% 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	≥58,860	≥12,330	83	17	100
			Knowledge and Use of Language Conventions	≥58,860	≥12,330	81	19	100
	Narrative Writing	12	Written Expression	≥58,900	≥12,280	77	22	99
			Knowledge and Use of Language Conventions	≥58,900	≥12,280	87	13	100
4	Literary Analysis	7	Reading Comprehension and Written Expression	≥58,100	≥11,370	75	24	99
			Knowledge and Use of Language Conventions	≥58,100	≥11,370	71	28	99
	Research Simulation	20	Reading Comprehension and Written Expression	≥56,850	≥8,870	88	12	100
			Knowledge and Use of Language Conventions	≥56,850	≥8,870	84	16	100
5	Literary Analysis	7	Reading Comprehension and Written Expression	≥56,840	≥14,310	86	13	99
			Knowledge and Use of Language Conventions	≥56,840	≥14,310	85	15	100
	Research Simulation	20	Reading Comprehension and Written Expression	≥54,180	≥8,550	77	23	100
			Knowledge and Use of Language Conventions	≥54,180	≥8,550	73	27	100
6	Research Simulation (AI)	9	Reading Comprehension and Written Expression	≥57,610	≥12,490	74	25	99
			Knowledge and Use of Language Conventions	≥57,610	≥12,490	72	27	99
	Narrative Writing (AI)	14	Written Expression	≥57,500	≥12,320	80	19	99
			Knowledge and Use of Language Conventions	≥57,500	≥12,320	77	23	100
7	Research Simulation	9	Reading Comprehension and Written Expression	≥58,280	≥12,140	83	17	100
			Knowledge and Use of Language Conventions	≥58,280	≥12,140	82	18	100
	Narrative Writing (AI)	14	Written Expression	≥58,590	≥12,930	85	15	100
			Knowledge and Use of Language Conventions	≥58,590	≥12,930	83	17	100
8	Literary Analysis (AI)	7	Reading Comprehension and Written Expression	≥58,470	≥13,500	81	19	100
			Knowledge and Use of Language Conventions	≥58,470	≥13,500	81	19	100
	Research Simulation	20	Reading Comprehension and Written Expression	≥58,030	≥12,630	81	19	100
			Knowledge and Use of LanguageConventions	≥58,030	≥12,630	81	19	100

**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	≥58,240	≥11,160	87	13	100
		Part B	≥58,240	≥11,160	93	6	99*
	18	N/A	≥58,190	≥11,400	96	4	100
	32	Part A	≥58,410	≥11,500	97	3	100
		Part B	≥58,410	≥11,500	99	1	100
	33	Part B (CBT)	≥13,290	≥2,430	98	2	100
		Part B (PBT)	≥44,750	≥8,190	97	3	100
	48	N/A	≥58,300	≥11,260	95	5	100
	49	Part B (CBT)	≥13,350	≥2,410	97	2	99*
		Part C (CBT)	≥13,350	≥2,410	96	4	100
		Part B (PBT)	≥44,800	≥8,280	96	4	100
		Part C (PBT)	≥44,800	≥8,280	96	4	100
4	17	Part C (CBT)	≥18,250	≥3,360	96	4	100
		Part C (PBT)	≥41,240	≥10,750	96	4	100
	18	N/A	≥57,710	≥11,300	94	6	100
	32	N/A	≥58,070	≥11,570	90	10	100
	33	N/A	≥58,270	≥11,950	91	8	99*
	48	Part A	≥57,790	≥11,650	96	4	100
		Part B	≥57,790	≥11,650	98	2	100
	49	Part A	≥57,650	≥11,210	94	6	100
		Part B	≥57,650	≥11,210	98	2	100
		Part C	≥57,650	≥11,210	96	3	99*
5	17	N/A	≥55,000	≥10,970	85	15	100
	18	N/A	≥54,680	≥10,640	88	11	99
	32	Part B	≥54,960	≥10,020	87	12	99*
	33	N/A	≥54,730	≥11,030	93	7	100
	48	Part B	≥54,860	≥9,960	91	9	100
	49	Part B	≥54,930	≥10,040	95	5	100
		Part C	≥54,930	≥10,040	91	8	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	≥56,330	≥11,070	80	20	100
	34	Part A	≥56,910	≥10,720	91	9	100
		Part B	≥56,910	≥10,720	96	4	100
	35	Part A	≥55,950	≥11,110	97	3	100
		Part B	≥55,950	≥11,110	95	5	100
	36	Part B	≥55,850	≥10,150	92	7	99
	47	N/A	≥56,460	≥11,740	88	11	99
	48	Part B	≥56,700	≥10,320	91	9	100
	49	N/A	≥56,310	≥11,400	93	6	99
7	31	Part A	≥57,440	≥11,140	96	4	100
		Part B	≥57,440	≥11,140	96	3	99
	34	N/A	≥56,880	≥11,920	92	7	99
	36	N/A	≥56,670	≥12,560	97	2	99*
	37	Part A	≥56,530	≥10,300	88	12	100
		Part B	≥56,530	≥10,300	96	4	100
	47	N/A	≥56,750	≥12,290	96	4	100
	48	Part A	≥57,240	≥11,430	97	3	100
		Part B	≥57,240	≥11,430	97	3	100
	49	N/A	≥56,780	≥11,590	93	7	100
8	31	Part A	≥50,850	≥10,040	94	6	100
		Part B	≥50,850	≥10,040	84	14	98*
	34	Part A	≥50,400	≥10,470	91	8	99
		Part B	≥50,400	≥10,470	89	10	99
	35	N/A	≥50,080	≥10,700	90	8	98
	36	Part A	≥49,730	≥10,630	94	5	99
		Part B	≥49,730	≥10,630	97	3	100
	42	Part B	≥50,600	≥9,140	93	7	100
	46	N/A	≥50,590	≥11,430	94	6	100
	48	Part B	≥50,650	≥9,230	90	10	100
		Part C	≥50,650	≥9,230	91	9	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	≥30	<10	NR	NR	NR
		Part B	≥30	<10	NR	NR	NR
	18	N/A	≥40	≥20	100	0	100
	32	Part A	≥40	≥10	100	0	100
		Part B	≥40	≥10	100	0	100
	33	Part B (CBT)	≥10	<10	NR	NR	NR
		Part B (PBT)	≥20	<10	NR	NR	NR
	48	N/A	≥40	≥10	100	0	100
	49	Part B	≥10	<10	NR	NR	NR
		Part C	≥10	<10	NR	NR	NR
	49	Part B	≥20	<10	NR	NR	NR
		Part C	≥20	<10	NR	NR	NR
4	17	Part C (CBT)	≥10	<10	N/A	N/A	N/A
		Part C (PBT)	≥10	<10	N/A	N/A	N/A
	18	N/A	≥20	<10	N/A	N/A	N/A
	32	N/A	≥20	<10	N/A	N/A	N/A
	33	N/A	≥20	<10	N/A	N/A	N/A
	48	Part A	≥20	<10	N/A	N/A	N/A
		Part B	≥20	<10	N/A	N/A	N/A
	49	Part A	≥20	<10	N/A	N/A	N/A
		Part B	≥20	<10	N/A	N/A	N/A
		Part C	≥20	<10	N/A	N/A	N/A
5	17	N/A	≥70	≥10	100	0	100
	18	N/A	≥70	≥10	100	0	100
	32	Part B	≥70	≥10	100	0	100
	33	N/A	≥70	≥20	100	0	100
	48	Part B	≥70	≥10	100	0	100
	49	Part B	≥70	≥10	100	0	100
		Part C	≥70	≥10	86	14	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	≥90	≥20	92	8	100
	34	Part A	≥80	≥10	100	0	100
		Part B	≥80	≥10	100	0	100
	35	Part A	≥80	≥10	100	0	100
		Part B	≥80	≥10	100	0	100
	36	Part B	≥80	≥10	100	0	100
	47	N/A	≥80	≥10	100	0	100
	48	Part B	≥90	≥10	100	0	100
	49	N/A	≥90	≥20	100	0	100
7	31	Part A	≥100	≥30	100	0	100
		Part B	≥100	≥30	100	0	100
	34	N/A	≥100	≥30	88	13	101*
	36	N/A	≥100	≥40	100	0	100
	37	Part A	≥90	≥10	100	0	100
		Part B	≥90	≥10	100	0	100
	47	N/A	≥100	≥30	100	0	100
	48	Part A	≥100	≥30	100	0	100
		Part B	≥100	≥30	100	0	100
	49	N/A	≥100	≥30	100	0	100
8	31	Part A	≥90	≥20	100	0	100
		Part B	≥90	≥20	100	0	100
	34	Part A	≥90	≥20	92	8	100
		Part B	≥90	≥20	100	0	100
	35	N/A	≥90	≥20	100	0	100
	36	Part A	≥90	≥30	100	0	100
		Part B	≥90	≥30	100	0	100
	42	Part B	≥90	≥20	100	0	100
	46	N/A	≥90	≥20	100	0	100
	48	Part B	≥90	≥10	100	0	100
		Part C	≥90	≥10	89	11	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*

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.

5.3 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.

5.4 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 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 2021 LEAP 2025 analyses, the Louisiana student population was used. 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 0.30 to 0.43. Average form difficulty for mathematics ranged from 0.28 to 0.47. In general, the 2021 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 0.30 average form difficulty, and the grade 7 was the easiest, with 0.43 average form difficulty. For mathematics, the grade 8 test was the most difficult, with 0.28 average form difficulty, and the grade 3 paper-based test (PBT) test was the easiest, with 0.47 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	21.12 (12.05)	0.30 (0.12)
	3	PBT	27	71	25.42 (12.72)	0.37 (0.12)
	4	CBT	30	86	27.42 (16.08)	0.32 (0.13)
	4	PBT	30	86	32.31 (16.60)	0.38 (0.11)
	5	CBT	30	86	28.09 (15.81)	0.33 (0.16)
	6	CBT	33	90	33.63 (17.55)	0.38 (0.13)
	7	CBT	33	90	38.39 (19.52)	0.43 (0.12)
	8	CBT	34	94	37.23 (18.39)	0.40 (0.10)
Mathematics	3	CBT	43	62	24.27 (13.47)	0.39 (0.18)
	3	PBT	43	62	28.70 (14.23)	0.47 (0.17)
	4	CBT	42	61	23.40 (13.52)	0.39 (0.20)
	4	PBT	42	61	25.59 (13.90)	0.43 (0.19)
	5	CBT	38	56	21.36 (12.25)	0.38 (0.16)
	6	CBT	40	63	21.61 (13.58)	0.35 (0.17)
	7	CBT	43	66	21.31 (13.22)	0.33 (0.18)
	8	CBT	37	60	16.58 (11.01)	0.28 (0.16)

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})

				Item *p*-Value				Item-Total Correlation			
Content	Grade	Mode	N of Items	Mean	Std. Dev.	Min.	Max	Mean	Std. Dev.	Min.	Max
ELA	3	CBT	27	0.34	0.12	0.15	0.59	0.45	0.13	0.22	0.66
	3	PBT	27	0.40	0.13	0.24	0.65	0.44	0.12	0.23	0.65
	4	CBT	30	0.36	0.13	0.18	0.67	0.50	0.16	0.25	0.80
	4	PBT	30	0.42	0.11	0.22	0.64	0.48	0.15	0.28	0.78
	5	CBT	30	0.39	0.16	0.13	0.74	0.50	0.17	0.19	0.79
	6	CBT	33	0.41	0.13	0.21	0.71	0.48	0.15	0.25	0.79
	7	CBT	33	0.46	0.12	0.27	0.70	0.50	0.15	0.26	0.81
	8	CBT	34	0.43	0.11	0.27	0.65	0.47	0.18	0.14	0.83
Mathematics	3	CBT	43	0.45	0.17	0.10	0.82	0.49	0.12	0.21	0.76
	3	PBT	43	0.53	0.17	0.17	0.89	0.50	0.11	0.26	0.78
	4	CBT	42	0.45	0.19	0.13	0.80	0.52	0.09	0.33	0.71
	4	PBT	42	0.48	0.18	0.19	0.84	0.51	0.09	0.31	0.71
	5	CBT	38	0.44	0.15	0.14	0.73	0.49	0.12	0.29	0.71
	6	CBT	40	0.40	0.16	0.10	0.69	0.51	0.10	0.26	0.68
	7	CBT	43	0.39	0.17	0.06	0.82	0.45	0.14	0.07	0.68
	8	CBT	37	0.31	0.17	0.08	0.75	0.47	0.13	0.08	0.68

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. The *p*-value and item-total correlations calculations used the adjusted N to determine the values. The rest of the statistics in the table are based on the total N.

***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 0.34 in the grade 3 CBT form to 0.46 in grade 7. The average *p*-values associated with the mathematics forms range from 0.31 in grade 8 CBT to 0.53 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 students of all levels of achievement, 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 0.13 to 0.74 (see Tables 6.3–6.10) across all grade levels. The *p*-values on the mathematics forms range from 0.06 to 0.89 (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 students across all levels of achievement. 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	≥12,160	≥12,150	0.39	0.47	0.08	46.29	28.53	25.10	
915224	ESR	≥12,160	≥12,110	0.37	0.25	0.39	54.31	17.65	27.66	
915228	TE	≥12,160	≥12,120	0.33	0.39	0.32	49.67	35.13	14.88	
915230	ESR	≥12,160	≥12,120	0.42	0.37	0.29	44.73	26.88	28.10	
915220	TE	≥12,160	≥11,870	0.35	0.48	2.35	38.73	49.42	9.50	
915219	ESR	≥12,160	≥12,110	0.31	0.22	0.44	58.29	19.86	21.41	
915227O2	CR	≥12,160	≥11,600	0.16	0.65	1.75	56.22	32.83	5.88	0.47
915227O3	CR	≥12,160	≥11,600	0.15	0.61	1.75	61.71	25.12	7.69	0.88
936916	MS	≥12,160	≥12,150	0.23	0.39	0.08	65.47	23.70	10.74	
913494	ESR	≥12,160	≥12,130	0.41	0.44	0.21	53.17	12.13	34.49	
913495	TE	≥12,160	≥12,010	0.59	0.48	1.21	19.14	43.24	36.41	
913493	ESR	≥12,160	≥12,140	0.35	0.40	0.17	57.70	14.38	27.75	
913497O2	CR	≥12,160	≥11,690	0.20	0.65	1.13	45.33	45.48	4.90	0.47
913497O3	CR	≥12,160	≥11,690	0.22	0.66	1.13	43.79	42.74	8.68	0.96
913318	TE	≥12,160	≥12,110	0.38	0.42	0.38	30.51	62.61	6.50	
913308	ESR	≥12,160	≥12,100	0.42	0.54	0.50	49.13	17.20	33.17	
913314	ESR	≥12,160	≥12,090	0.41	0.55	0.56	48.50	19.51	31.43	
913310	ESR	≥12,160	≥12,090	0.24	0.25	0.55	63.84	24.05	11.56	
934821	ESR	≥12,160	≥12,150	0.30	0.25	0.09	59.42	21.65	18.83	
934823	ESR	≥12,160	≥12,140	0.49	0.47	0.13	30.08	42.26	27.53	
934822	TE	≥12,160	≥12,040	0.52	0.57	0.98	36.09	23.54	39.39	
934802	ESR	≥12,160	≥12,140	0.48	0.52	0.14	40.93	21.71	37.22	
915910	ESR	≥12,160	≥12,070	0.29	0.36	0.76	60.70	18.84	19.70	
915902	TE	≥12,160	≥12,060	0.48	0.39	0.84	42.59	18.14	38.43	
915908	MS	≥12,160	≥12,050	0.25	0.48	0.94	61.08	27.19	10.79	
915905	ESR	≥12,160	≥12,030	0.32	0.40	1.04	57.16	21.10	20.70	

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	≥37,530	≥37,330	0.48	0.47	0.55	37.83	26.92	34.70	
915224	ESR	≥37,530	≥37,260	0.42	0.27	0.73	48.64	17.80	32.82	
915225	ESR	≥37,530	≥37,240	0.26	0.31	0.78	66.14	14.14	18.94	
915230	ESR	≥37,530	≥37,250	0.49	0.37	0.75	37.46	25.73	36.05	
915229	ESR	≥37,530	≥37,150	0.24	0.33	1.02	70.20	10.27	18.51	
915219	ESR	≥37,530	≥37,130	0.37	0.28	1.07	53.22	18.76	26.95	
915227P2	CR	≥37,530	≥36,430	0.26	0.65	1.71	39.79	39.38	16.81	1.09
915227P3	CR	≥37,530	≥36,430	0.27	0.59	1.71	36.75	43.86	14.96	1.51
936916	MS	≥37,530	≥37,350	0.29	0.38	0.49	58.76	23.39	17.37	
913494	ESR	≥37,530	≥37,250	0.46	0.41	0.76	48.59	10.34	40.31	
913496	ESR	≥37,530	≥37,240	0.65	0.56	0.77	28.52	12.74	57.97	
913493	ESR	≥37,530	≥37,170	0.39	0.37	0.96	54.81	10.67	33.56	
913497P2	CR	≥37,530	≥36,710	0.26	0.58	1.10	33.50	53.60	9.41	1.28
913497P3	CR	≥37,530	≥36,710	0.27	0.56	1.10	35.04	46.63	14.77	1.36
913315	MS	≥37,530	≥36,130	0.43	0.43	3.75	22.48	63.87	9.90	
913308	ESR	≥37,530	≥35,840	0.54	0.54	4.50	35.90	15.39	44.22	
913314	ESR	≥37,530	≥36,050	0.54	0.51	3.96	32.66	23.70	39.69	
913310	ESR	≥37,530	≥35,690	0.27	0.23	4.92	58.75	22.11	14.22	
934821	ESR	≥37,530	≥37,240	0.33	0.24	0.78	57.19	18.82	23.21	
934823	ESR	≥37,530	≥37,070	0.58	0.41	1.24	20.65	42.38	35.73	
934806	ESR	≥37,530	≥37,140	0.44	0.47	1.05	52.67	4.90	41.38	
934802	ESR	≥37,530	≥37,030	0.57	0.53	1.34	32.52	20.09	46.05	
915910	ESR	≥37,530	≥36,880	0.36	0.42	1.74	55.31	15.22	27.72	
915909	ESR	≥37,530	≥36,600	0.62	0.33	2.49	31.36	10.63	55.52	
915908	MS	≥37,530	≥36,770	0.30	0.47	2.03	53.02	31.11	13.83	
915905	ESR	≥37,530	≥36,470	0.39	0.44	2.84	50.93	16.91	29.32	

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	≥16,600	≥16,590	0.52	0.48	0.06	39.55	17.69	42.70		
913562	TE	≥16,600	≥16,510	0.67	0.58	0.52	19.46	27.28	52.75		
913563	ESR	≥16,600	≥16,570	0.45	0.46	0.14	43.48	22.64	33.73		
946024	TE	≥16,600	≥16,510	0.23	0.42	0.49	65.25	22.81	11.45		
913564	ESR	≥16,600	≥16,570	0.54	0.49	0.16	41.27	10.16	48.42		
913566	MS	≥16,600	≥16,570	0.43	0.45	0.16	45.34	23.96	30.54		
913567O2	CR	≥16,600	≥16,300	0.22	0.77	0.76	36.97	40.81	17.50	2.58	0.33
913567O3	CR	≥16,600	≥16,300	0.30	0.73	0.76	34.61	40.90	19.33	3.36	
913592	ESR	≥16,600	≥16,530	0.41	0.39	0.39	49.52	18.27	31.83		
913594	TE	≥16,600	≥16,430	0.35	0.36	0.98	42.98	42.65	13.39		
998347	ESR	≥16,600	≥16,500	0.18	0.25	0.59	73.43	15.36	10.62		
913595	ESR	≥16,600	≥16,500	0.34	0.29	0.60	59.09	13.89	26.42		
982220	ESR	≥16,600	≥16,590	0.52	0.46	0.01	24.80	47.38	27.81		
982222	ESR	≥16,600	≥16,580	0.36	0.35	0.08	57.18	12.69	30.04		
982223	TE	≥16,600	≥16,570	0.38	0.51	0.18	41.66	40.90	17.27		
982225	ESR	≥16,600	≥16,580	0.43	0.52	0.12	47.01	20.66	32.21		
982227	TE	≥16,600	≥16,550	0.19	0.39	0.26	75.56	10.81	13.37		
982230	MS	≥16,600	≥16,570	0.30	0.47	0.14	54.26	31.08	14.52		
982228	ESR	≥16,600	≥16,560	0.40	0.48	0.20	55.80	7.45	36.55		
982229	ESR	≥16,600	≥16,570	0.60	0.47	0.16	34.13	11.03	54.68		
982233O2	CR	≥16,600	≥16,250	0.21	0.80	0.68	38.97	38.90	15.51	4.19	0.34
982233O3	CR	≥16,600	≥16,250	0.26	0.77	0.68	45.69	32.93	15.37	3.91	
915315	ESR	≥16,600	≥16,570	0.58	0.52	0.17	28.42	27.51	43.90		
915319	ESR	≥16,600	≥16,560	0.22	0.25	0.24	64.25	28.01	7.50		
915322	ESR	≥16,600	≥16,550	0.37	0.37	0.27	57.23	10.76	31.73		
915325	TE	≥16,600	≥16,530	0.34	0.50	0.40	45.02	41.87	12.71		
915316	ESR	≥16,600	≥16,530	0.36	0.48	0.42	55.83	15.22	28.53		
915317	ESR	≥16,600	≥16,520	0.37	0.50	0.44	54.02	17.29	28.25		

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	≥33,140	≥33,050	0.60	0.46	0.27	32.44	15.75	51.55		
946021	ESR	≥33,140	≥32,980	0.49	0.33	0.49	44.97	10.97	43.57		
913563	ESR	≥33,140	≥33,010	0.47	0.43	0.39	42.41	20.07	37.14		
946023	ESR	≥33,140	≥32,980	0.26	0.37	0.48	64.75	18.30	16.47		
913564	ESR	≥33,140	≥32,950	0.51	0.48	0.56	44.67	8.29	46.48		
913566	MS	≥33,140	≥32,990	0.46	0.44	0.45	41.73	23.86	33.96		
913567P2	CR	≥33,140	≥32,650	0.30	0.74	1.08	24.94	37.66	28.74	6.35	0.84
913567P3	CR	≥33,140	≥32,650	0.42	0.67	1.08	21.88	37.52	29.68	9.45	
913592	ESR	≥33,140	≥32,410	0.46	0.35	2.21	44.26	17.30	36.23		
913593	ESR	≥33,140	≥32,270	0.40	0.53	2.63	49.67	18.36	29.34		
998347	ESR	≥33,140	≥32,200	0.23	0.31	2.82	66.79	16.53	13.85		
913595	ESR	≥33,140	≥31,970	0.37	0.28	3.54	55.05	11.50	29.91		
982220	ESR	≥33,140	≥33,010	0.60	0.45	0.40	18.85	42.89	37.86		
982222	ESR	≥33,140	≥32,910	0.45	0.39	0.68	49.42	10.88	39.01		
982221	ESR	≥33,140	≥32,930	0.45	0.35	0.63	30.40	48.47	20.50		
982225	ESR	≥33,140	≥32,930	0.50	0.52	0.63	39.47	19.93	39.97		
982226	ESR	≥33,140	≥32,910	0.43	0.40	0.70	52.68	8.82	37.80		
982230	MS	≥33,140	≥32,890	0.33	0.42	0.75	51.58	30.35	17.32		
982228	ESR	≥33,140	≥32,930	0.48	0.47	0.64	48.42	6.16	44.78		
982229	ESR	≥33,140	≥32,850	0.64	0.40	0.86	30.95	9.19	58.99		
982233P2	CR	≥33,140	≥32,680	0.28	0.78	1.00	26.96	39.88	24.79	6.50	0.50
982233P3	CR	≥33,140	≥32,680	0.36	0.76	1.00	29.70	37.77	24.56	6.60	
915315	ESR	≥33,140	≥32,920	0.61	0.49	0.65	26.17	25.91	47.28		
915319	ESR	≥33,140	≥32,800	0.22	0.28	1.03	63.44	27.32	8.21		
915322	ESR	≥33,140	≥32,760	0.42	0.42	1.13	52.87	8.63	37.37		
915321	ESR	≥33,140	≥32,690	0.37	0.40	1.35	58.65	7.39	32.61		
915316	ESR	≥33,140	≥32,810	0.40	0.47	0.98	51.78	14.91	32.34		
915317	ESR	≥33,140	≥32,650	0.40	0.46	1.47	52.76	12.99	32.78		

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	≥50,010	≥49,980	0.40	0.29	0.06	53.73	11.60	34.60		
799889	MS	≥50,010	≥49,890	0.25	0.28	0.24	61.24	27.93	10.60		
799890	ESR	≥50,010	≥49,910	0.54	0.43	0.20	39.84	11.38	48.58		
799891	ESR	≥50,010	≥49,910	0.30	0.26	0.20	57.59	23.80	18.41		
799892	ESR	≥50,010	≥49,890	0.30	0.43	0.25	64.28	11.84	23.63		
995980	TE	≥50,010	≥49,830	0.62	0.40	0.36	18.36	38.40	42.88		
801310O2	CR	≥50,010	≥49,160	0.13	0.75	0.88	58.26	30.28	9.28	0.48	0.01
801310O3	CR	≥50,010	≥49,160	0.20	0.74	0.88	52.01	33.33	11.84	1.13	
932836	ESR	≥50,010	≥49,230	0.50	0.37	1.57	29.59	39.50	29.34		
932839	ESR	≥50,010	≥49,150	0.53	0.56	1.72	40.00	12.33	45.95		
932840	MS	≥50,010	≥48,960	0.43	0.59	2.10	43.16	24.73	30.00		
932837	TE	≥50,010	≥48,760	0.43	0.60	2.51	43.69	23.22	30.58		
915501	ESR	≥50,010	≥50,000	0.48	0.40	0.03	36.94	29.61	33.41		
915500	ESR	≥50,010	≥49,970	0.60	0.46	0.09	36.37	6.68	56.86		
915507	ESR	≥50,010	≥49,960	0.42	0.44	0.10	52.42	11.81	35.67		
915497	ESR	≥50,010	≥49,970	0.74	0.47	0.09	20.60	9.81	69.49		
915499	ESR	≥50,010	≥49,990	0.47	0.50	0.05	46.08	14.01	39.86		
915511	TE	≥50,010	≥49,950	0.26	0.19	0.12	74.24	0.01	25.64		
915512	TE	≥50,010	≥49,930	0.45	0.56	0.16	36.70	37.38	25.76		
915508	MS	≥50,010	≥49,960	0.28	0.38	0.11	59.23	25.46	15.19		
915510O2	CR	≥50,010	≥49,510	0.24	0.79	0.52	37.59	34.44	21.96	4.68	0.31
915510O3	CR	≥50,010	≥49,510	0.33	0.76	0.52	38.09	31.67	22.82	6.40	
913665	ESR	≥50,010	≥49,920	0.38	0.49	0.19	50.84	21.79	27.19		
913664	ESR	≥50,010	≥49,950	0.35	0.38	0.12	54.23	21.00	24.64		
913666	TE	≥50,010	≥49,890	0.67	0.44	0.25	10.59	44.02	45.14		
913668	ESR	≥50,010	≥49,880	0.48	0.51	0.26	48.75	5.87	45.12		
913667	MS	≥50,010	≥49,890	0.27	0.39	0.24	59.00	27.46	13.30		
913669	TE	≥50,010	≥49,820	0.26	0.52	0.38	61.57	25.28	12.78		

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	≥51,580	≥51,540	0.48	0.42	0.08	40.39	23.86	35.67		
913708	ESR	≥51,580	≥51,490	0.59	0.47	0.17	33.97	13.89	51.97		
913710	ESR	≥51,580	≥51,500	0.42	0.41	0.16	47.25	21.52	31.07		
913711	TE	≥51,580	≥51,320	0.40	0.31	0.49	33.29	52.18	14.04		
980309	ESR	≥51,580	≥51,500	0.50	0.41	0.14	43.94	11.06	44.86		
913712	TE	≥51,580	≥50,980	0.59	0.63	1.17	32.91	15.27	50.66		
913713	MS	≥51,580	≥51,390	0.34	0.41	0.37	53.41	25.57	20.65		
913714	ESR	≥51,580	≥51,370	0.45	0.51	0.39	45.00	18.70	35.90		
913715O2	CR	≥51,580	≥50,870	0.32	0.79	0.76	24.97	31.89	33.47	7.40	0.91
913715O3	CR	≥51,580	≥50,870	0.44	0.77	0.76	25.35	28.87	32.35	12.06	
913690	MS	≥51,580	≥51,560	0.40	0.46	0.03	39.75	40.29	19.93		
913691	TE	≥51,580	≥51,500	0.39	0.45	0.15	48.09	26.51	25.25		
913692	MS	≥51,580	≥51,510	0.27	0.34	0.12	57.32	30.40	12.15		
913693	TE	≥51,580	≥51,360	0.23	0.40	0.43	67.63	18.63	13.31		
913694O2	CR	≥51,580	≥50,720	0.21	0.75	0.89	49.70	21.84	20.29	5.25	1.26
913694O3	CR	≥51,580	≥50,720	0.29	0.76	0.89	40.22	34.97	18.55	4.59	
917785	ESR	≥51,580	≥51,470	0.43	0.51	0.21	41.62	29.96	28.21		
917781	ESR	≥51,580	≥51,480	0.38	0.35	0.19	49.03	25.54	25.25		
917755	MS	≥51,580	≥51,500	0.34	0.39	0.15	56.30	19.84	23.70		
917763	TE	≥51,580	≥51,470	0.35	0.39	0.22	47.81	34.70	17.27		
917778	TE	≥51,580	≥51,450	0.57	0.54	0.25	15.37	54.50	29.88		
917721	ESR	≥51,580	≥51,440	0.43	0.25	0.26	45.45	21.99	32.30		
913752	ESR	≥51,580	≥51,560	0.38	0.47	0.03	53.40	16.47	30.09		
913753	TE	≥51,580	≥51,460	0.70	0.54	0.22	20.34	20.15	59.29		
913754	TE	≥51,580	≥51,480	0.71	0.40	0.19	25.81	5.81	68.20		
913755	ESR	≥51,580	≥51,450	0.37	0.38	0.24	56.09	14.13	29.53		
913757	MS	≥51,580	≥51,470	0.31	0.52	0.22	58.92	19.14	21.73		
913756	MS	≥51,580	≥51,470	0.21	0.32	0.22	70.80	15.26	13.73		
980274	TE	≥51,580	≥51,330	0.30	0.26	0.48	53.03	33.97	12.53		
980271	ESR	≥51,580	≥51,350	0.43	0.47	0.44	41.29	31.20	27.07		
980273	MS	≥51,580	≥51,350	0.49	0.40	0.43	18.84	64.41	16.32		
980276	ESR	≥51,580	≥51,340	0.56	0.58	0.46	37.78	11.38	50.38		

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	≥52,350	≥52,340	0.70	0.34	0.03	18.02	24.51	57.44		
915572	ESR	≥52,350	≥52,270	0.55	0.42	0.16	33.86	21.56	44.42		
915573	ESR	≥52,350	≥52,260	0.45	0.32	0.17	47.17	14.58	38.08		
915574	TE	≥52,350	≥52,280	0.42	0.40	0.14	51.10	12.88	35.88		
915578	ESR	≥52,350	≥52,290	0.59	0.44	0.12	27.59	26.65	45.64		
915576	TE	≥52,350	≥52,230	0.50	0.52	0.24	35.44	27.90	36.42		
915579	ESR	≥52,350	≥52,230	0.64	0.47	0.23	19.79	32.48	47.50		
915583	MS	≥52,350	≥52,250	0.58	0.55	0.20	22.01	39.16	38.63		
915582O2	CR	≥52,350	≥51,680	0.36	0.81	0.69	20.40	31.47	33.42	10.65	2.77
915582O3	CR	≥52,350	≥51,680	0.44	0.80	0.69	23.01	31.71	32.31	11.67	
913840	TE	≥52,350	≥52,300	0.31	0.49	0.10	55.00	28.07	16.83		
913839	ESR	≥52,350	≥52,260	0.61	0.44	0.18	28.60	20.00	51.22		
913841	MS	≥52,350	≥52,300	0.28	0.41	0.09	51.80	40.46	7.65		
913838	TE	≥52,350	≥52,320	0.63	0.57	0.06	21.78	29.89	48.27		
913842O2	CR	≥52,350	≥51,420	0.32	0.80	0.85	41.99	13.99	20.90	14.01	7.32
913842O3	CR	≥52,350	≥51,420	0.41	0.81	0.85	34.69	23.12	23.91	16.50	
913807	ESR	≥52,350	≥52,280	0.59	0.38	0.13	38.11	5.29	56.47		
913808	ESR	≥52,350	≥52,300	0.58	0.51	0.10	32.71	18.37	48.82		
913811	ESR	≥52,350	≥52,300	0.32	0.38	0.09	57.50	20.14	22.27		
913810	ESR	≥52,350	≥52,270	0.31	0.44	0.16	59.16	19.45	21.23		
913812	TE	≥52,350	≥52,280	0.42	0.51	0.13	40.83	34.98	24.06		
913809	TE	≥52,350	≥52,250	0.27	0.48	0.19	56.54	31.69	11.59		
932822	ESR	≥52,350	≥52,310	0.46	0.37	0.07	40.83	26.02	33.07		
932782	ESR	≥52,350	≥52,280	0.59	0.36	0.14	37.23	7.14	55.49		
932785	ESR	≥52,350	≥52,220	0.37	0.44	0.25	56.56	12.74	30.45		
932810	MS	≥52,350	≥52,240	0.47	0.51	0.22	38.19	29.72	31.87		
932791	ESR	≥52,350	≥52,230	0.41	0.26	0.23	52.47	12.36	34.94		
932789	ESR	≥52,350	≥52,220	0.32	0.38	0.25	53.31	28.96	17.47		
932827	ESR	≥52,350	≥52,160	0.49	0.46	0.37	43.67	13.43	42.53		
953139	TE	≥52,350	≥52,160	0.59	0.56	0.36	32.16	16.77	50.70		
932821	MS	≥52,350	≥52,130	0.34	0.55	0.42	47.13	37.36	15.09		
933576	MS	≥52,350	≥52,110	0.40	0.42	0.46	49.24	21.57	28.72		

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	≥52,020	≥51,990	0.36	0.33	0.05	55.02	18.29	26.65		
913953	ESR	≥52,020	≥51,940	0.38	0.38	0.15	50.92	22.34	26.59		
913954	MS	≥52,020	≥51,960	0.46	0.43	0.11	47.28	14.15	38.47		
913955	TE	≥52,020	≥51,910	0.62	0.38	0.21	11.85	51.54	36.39		
913956	ESR	≥52,020	≥51,970	0.45	0.42	0.10	50.15	9.49	40.25		
913957	TE	≥52,020	≥51,910	0.39	0.38	0.21	33.80	54.78	11.21		
913958O2	CR	≥52,020	≥50,920	0.36	0.83	1.10	21.17	29.26	32.49	13.29	1.69
913958O3	CR	≥52,020	≥50,920	0.50	0.81	1.10	17.32	28.99	35.97	15.62	
982279	ESR	≥52,020	≥51,700	0.57	0.41	0.60	30.61	24.42	44.36		
982281	MS	≥52,020	≥51,620	0.35	0.43	0.77	47.00	35.96	16.28		
982276	ESR	≥52,020	≥51,580	0.57	0.38	0.83	34.11	16.75	48.31		
982278	TE	≥52,020	≥51,340	0.39	0.49	1.31	40.33	39.19	19.17		
982294	MS	≥52,020	≥52,010	0.33	0.38	0.02	43.52	47.93	8.53		
982297	TE	≥52,020	≥51,970	0.43	0.43	0.09	54.57	5.36	39.98		
982299	ESR	≥52,020	≥51,930	0.47	0.40	0.16	44.25	17.62	37.97		
982301	ESR	≥52,020	≥51,950	0.60	0.37	0.13	36.24	8.24	55.39		
982300	ESR	≥52,020	≥51,960	0.45	0.46	0.10	52.00	5.35	42.55		
982302	ESR	≥52,020	≥51,980	0.54	0.46	0.07	42.03	7.89	50.01		
982303	TE	≥52,020	≥51,940	0.59	0.42	0.15	27.66	27.51	44.68		
982304	ESR	≥52,020	≥51,960	0.38	0.43	0.11	59.27	6.17	34.45		
982327O2	CR	≥52,020	≥50,870	0.28	0.82	1.07	26.63	41.66	22.25	6.34	0.91
982327O3	CR	≥52,020	≥50,870	0.35	0.82	1.07	32.06	35.28	24.09	6.36	
982331	TE	≥52,020	≥51,960	0.65	0.49	0.12	14.06	40.95	44.88		
982330	ESR	≥52,020	≥51,940	0.43	0.50	0.15	54.29	5.08	40.48		
982333	ESR	≥52,020	≥51,970	0.38	0.27	0.10	53.72	17.20	28.98		
982332	TE	≥52,020	≥51,960	0.31	0.44	0.12	56.38	24.59	18.91		
913974	ESR	≥52,020	≥51,950	0.37	0.41	0.12	57.03	12.51	30.33		
913970	MS	≥52,020	≥51,960	0.50	0.29	0.12	22.59	55.29	22.00		
913971	ESR	≥52,020	≥51,910	0.27	0.14	0.21	65.70	14.68	19.40		
913972	MS	≥52,020	≥51,910	0.34	0.44	0.20	42.80	46.66	10.33		
913973	ESR	≥52,020	≥51,910	0.35	0.41	0.21	51.86	26.63	21.29		
913975	MS	≥52,020	≥51,910	0.49	0.45	0.20	33.51	34.52	31.77		

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	≥12,140	≥12,140	0.67	0.49	0.02							
913997	SA	≥12,140	≥12,040	0.41	0.62	0.81	58.89	40.30					
896772	MC	≥12,140	≥12,120	0.35	0.49	0.21							
914024	SA	≥12,140	≥12,050	0.43	0.30	0.72	56.15	43.14					
904404	MC	≥12,140	≥12,100	0.55	0.48	0.31							
914038	SA	≥12,140	≥12,080	0.39	0.46	0.52	60.44	39.04					
981774	MC	≥12,140	≥12,110	0.40	0.45	0.29							
981799	MC	≥12,140	≥12,100	0.56	0.40	0.30							
896859	SA	≥12,140	≥12,080	0.37	0.59	0.51	62.37	37.12					
981778	MC	≥12,140	≥12,110	0.57	0.35	0.29							
906209	MPSR	≥12,140	≥12,080	0.47	0.35	0.49	31.23	43.42	24.86				
981751	MC	≥12,140	≥12,110	0.64	0.41	0.26							
913987	MC	≥12,140	≥12,070	0.60	0.46	0.58							
981736	CR	≥12,140	≥11,750	0.24	0.60	1.44	46.36	23.23	16.26	7.84	3.12		
868619	CR	≥12,140	≥11,510	0.10	0.52	2.79	79.05	7.73	2.84	5.20			
981762	SA	≥12,140	≥12,110	0.69	0.21	0.29	30.79	68.92					
906210	MC	≥12,140	≥12,110	0.80	0.41	0.28							
896684	SA	≥12,140	≥12,080	0.26	0.43	0.47	73.77	25.76					
916044	SA	≥12,140	≥12,090	0.36	0.57	0.44	44.39	39.08	16.10				
935017	MS	≥12,140	≥12,120	0.19	0.39	0.21	81.33	18.47					
896862	MC	≥12,140	≥12,120	0.64	0.27	0.21							
981795	MC	≥12,140	≥12,080	0.61	0.48	0.54							
981767	MC	≥12,140	≥12,110	0.61	0.47	0.27							
914023	SA	≥12,140	≥12,100	0.60	0.56	0.34	40.00	59.66					
896902	SA	≥12,140	≥12,090	0.36	0.65	0.38	42.61	42.47	14.54				
914007	SA	≥12,140	≥12,070	0.23	0.62	0.56	76.17	23.27					
896860	SA	≥12,140	≥12,070	0.28	0.57	0.61	71.54	27.85					
898001	CR	≥12,140	≥11,760	0.20	0.60	1.43	61.55	14.80	18.75	1.75			
981742	CR	≥12,140	≥12,040	0.20	0.71	0.79	61.35	24.31	5.88	7.67			
981784	MC	≥12,140	≥12,130	0.58	0.55	0.12							
896770	SA	≥12,140	≥12,100	0.40	0.49	0.35	59.63	40.02					
981791	MC	≥12,140	≥12,120	0.53	0.53	0.14							
896868	MC	≥12,140	≥12,100	0.37	0.27	0.35							
896867	SA	≥12,140	≥12,080	0.57	0.52	0.49	43.22	56.29					
896863	MC	≥12,140	≥12,120	0.82	0.38	0.17							
896679	MC	≥12,140	≥12,110	0.60	0.51	0.21							
913991	MC	≥12,140	≥12,120	0.49	0.49	0.17							
914001	MS	≥12,140	≥12,110	0.29	0.48	0.26	70.48	29.25					
878608	MC	≥12,140	≥12,130	0.63	0.48	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	≥12,140	≥12,080	0.46	0.66	0.54	54.13	45.34					
914036	MS	≥12,140	≥12,120	0.37	0.54	0.19	63.24	36.57					
914039	CR	≥12,140	≥11,720	0.31	0.65	1.32	37.32	31.03	26.40	1.75			
981747	CR	≥12,140	≥12,110	0.29	0.76	0.21	23.09	34.42	16.27	11.39	6.08	4.41	4.13

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	≥37,530	≥37,120	0.76	0.49	1.03							
913997	SA	≥37,530	≥35,590	0.48	0.62	5.17	48.89	45.94					
896772	MC	≥37,530	≥36,660	0.44	0.51	2.26							
914024	SA	≥37,530	≥36,150	0.46	0.31	3.66	52.39	43.95					
904404	MC	≥37,530	≥36,030	0.63	0.50	3.07							
914038	SA	≥37,530	≥36,460	0.49	0.47	2.85	49.78	47.37					
981774	MC	≥37,530	≥36,760	0.47	0.48	1.45							
981799	MC	≥37,530	≥35,660	0.61	0.42	1.73							
896859	SA	≥37,530	≥36,280	0.44	0.57	3.31	53.97	42.71					
981778	MC	≥37,530	≥37,070	0.68	0.34	1.12							
906209	MPSR	≥37,530	≥37,130	0.52	0.41	1.07	27.90	39.29	31.74				
981751	MC	≥37,530	≥36,650	0.71	0.41	2.30							
913987	MC	≥37,530	≥36,190	0.68	0.48	2.99							
981736	CR	≥37,530	≥36,890	0.36	0.58	1.47	32.47	19.01	25.70	15.21	5.90		
868619	CR	≥37,530	≥33,790	0.17	0.55	9.58	67.15	8.12	5.31	9.45			
981762	SA	≥37,530	≥36,860	0.68	0.30	1.77	31.27	66.96					
906210	MC	≥37,530	≥37,070	0.85	0.39	0.96							
896684	SA	≥37,530	≥36,400	0.33	0.45	3.01	64.84	32.15					
916044	SA	≥37,530	≥36,980	0.42	0.60	1.45	37.97	38.91	21.67				
935017	MS	≥37,530	≥37,030	0.26	0.44	1.33	72.63	26.03					
896862	MC	≥37,530	≥36,230	0.66	0.26	3.31							
981795	MC	≥37,530	≥35,510	0.68	0.50	3.90							
981767	MC	≥37,530	≥36,940	0.70	0.45	1.51							
914023	SA	≥37,530	≥36,630	0.65	0.55	2.38	34.30	63.32					
896902	SA	≥37,530	≥37,170	0.44	0.68	0.95	34.25	42.03	22.77				
914007	SA	≥37,530	≥36,220	0.31	0.62	3.48	66.28	30.23					
896860	SA	≥37,530	≥36,300	0.35	0.56	3.28	62.62	34.11					
898001	CR	≥37,530	≥36,460	0.25	0.58	2.46	53.49	18.15	23.30	2.23			
981742	CR	≥37,530	≥37,210	0.28	0.70	0.84	49.69	28.32	7.35	13.80			
981784	MC	≥37,530	≥36,960	0.67	0.56	1.49							
896770	SA	≥37,530	≥36,510	0.53	0.45	2.72	46.04	51.24					
981791	MC	≥37,530	≥36,690	0.63	0.55	2.03							
896868	MC	≥37,530	≥36,710	0.42	0.26	1.87							
896867	SA	≥37,530	≥36,260	0.67	0.47	3.37	32.04	64.59					
896863	MC	≥37,530	≥36,980	0.89	0.35	1.41							
896679	MC	≥37,530	≥36,170	0.66	0.52	3.39							
913991	MC	≥37,530	≥36,800	0.60	0.52	1.91							
914001	MS	≥37,530	≥36,970	0.38	0.49	1.49	60.88	37.63					
878608	MC	≥37,530	≥36,770	0.73	0.51	1.87							

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	≥37,530	≥36,190	0.54	0.63	3.55	44.47	51.98					
914036	MS	≥37,530	≥36,510	0.43	0.55	2.72	55.09	42.19					
914039	CR	≥37,530	≥36,480	0.40	0.61	2.55	24.61	29.18	41.38	2.05			
981747	CR	≥37,530	≥37,370	0.41	0.78	0.42	17.33	22.49	14.99	16.56	10.68	6.21	11.32

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	≥16,550	≥16,520	0.63	0.52	0.16							
870319	SA	≥16,550	≥16,500	0.41	0.51	0.29	58.92	40.79					
981843	MS	≥16,550	≥16,510	0.30	0.56	0.21	70.35	29.45					
981835	SA	≥16,550	≥16,470	0.17	0.54	0.46	82.30	17.24					
897478	MC	≥16,550	≥16,530	0.52	0.35	0.12							
981874	MPSR	≥16,550	≥16,530	0.47	0.63	0.10	34.82	36.13	28.95				
981867	SA	≥16,550	≥16,490	0.38	0.59	0.33	61.46	38.21					
897446	SA	≥16,550	≥16,440	0.56	0.56	0.65	43.83	55.52					
914137	MC	≥16,550	≥16,490	0.53	0.44	0.36							
944080	MC	≥16,550	≥16,510	0.57	0.44	0.21							
981844	SA	≥16,550	≥16,350	0.20	0.55	1.17	79.26	19.57					
914080	MS	≥16,550	≥16,510	0.65	0.54	0.21	34.61	65.18					
914084	CR	≥16,550	≥16,510	0.29	0.71	0.24	31.16	34.38	23.33	9.78	1.11		
914086	CR	≥16,550	≥15,770	0.14	0.58	2.51	71.65	14.56	3.23	5.90			
914101	MC	≥16,550	≥16,540	0.73	0.41	0.06							
897470	MC	≥16,550	≥16,530	0.47	0.62	0.13							
897468	MC	≥16,550	≥16,530	0.36	0.44	0.09							
914082	SA	≥16,550	≥16,510	0.26	0.45	0.22	73.51	26.26					
897302	MC	≥16,550	≥16,530	0.50	0.50	0.12							
914121	SA	≥16,550	≥16,510	0.57	0.47	0.22	42.50	57.27					
914088	MC	≥16,550	≥16,530	0.50	0.52	0.13							
897444	SA	≥16,550	≥16,530	0.65	0.54	0.13	17.96	33.42	48.50				
878669	SA	≥16,550	≥16,520	0.52	0.52	0.19	22.87	49.92	27.03				
897475	SA	≥16,550	≥16,510	0.61	0.48	0.24	38.91	60.85					
897291	MS	≥16,550	≥16,530	0.63	0.57	0.13	37.03	62.84					
981838	MC	≥16,550	≥16,530	0.32	0.44	0.13							
981831	CR	≥16,550	≥16,000	0.20	0.69	1.15	63.22	15.41	12.14	5.90			
899959	CR	≥16,550	≥16,210	0.32	0.68	1.05	46.71	24.53	11.35	15.34			
897434	MC	≥16,550	≥16,530	0.76	0.43	0.10							
981850	MC	≥16,550	≥16,520	0.47	0.41	0.16							
898008	SA	≥16,550	≥16,500	0.56	0.51	0.31	44.33	55.36					
981890	MS	≥16,550	≥16,530	0.72	0.48	0.09	28.26	71.65					
914135	MC	≥16,550	≥16,520	0.80	0.40	0.14							
897305	MC	≥16,550	≥16,510	0.27	0.33	0.24							
897438	MC	≥16,550	≥16,520	0.71	0.43	0.19							
914099	SA	≥16,550	≥16,410	0.23	0.54	0.80	75.95	23.24					
897471	SA	≥16,550	≥16,320	0.37	0.50	1.39	62.22	36.39					
981866	SA	≥16,550	≥16,440	0.39	0.58	0.62	60.74	38.64					
981853	SA	≥16,550	≥16,430	0.40	0.55	0.68	59.55	39.77					

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**
914108	MS	≥16,550	≥16,510	0.31	0.48	0.22	68.44	31.33					
899955	CR	≥16,550	≥15,760	0.13	0.62	3.05	66.74	20.98	6.51	0.98			
981827	CR	≥16,550	≥15,840	0.15	0.65	2.45	61.24	10.03	11.79	3.50	5.33	1.84	1.96

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	≥33,140	≥32,410	0.68	0.53	2.16							
870319	SA	≥33,140	≥32,150	0.39	0.50	2.98	58.91	38.11					
981843	MS	≥33,140	≥32,680	0.37	0.55	1.37	62.37	36.26					
981835	SA	≥33,140	≥31,830	0.19	0.53	3.95	77.73	18.32					
897478	MC	≥33,140	≥32,590	0.55	0.33	1.46							
981874	MPSR	≥33,140	≥32,910	0.51	0.65	0.67	30.81	35.65	32.87				
981867	SA	≥33,140	≥31,670	0.39	0.57	4.44	58.61	36.95					
897446	SA	≥33,140	≥31,670	0.56	0.53	4.43	42.04	53.53					
914137	MC	≥33,140	≥32,330	0.59	0.46	2.39							
944080	MC	≥33,140	≥32,680	0.65	0.45	1.36							
981844	SA	≥33,140	≥31,190	0.24	0.55	5.87	71.22	22.91					
914080	MS	≥33,140	≥32,650	0.74	0.50	1.46	25.48	73.06					
914084	CR	≥33,140	≥33,000	0.33	0.71	0.40	27.05	31.90	25.21	13.03	2.41		
914086	CR	≥33,140	≥31,360	0.21	0.60	5.05	57.27	22.99	5.40	8.99			
914101	MC	≥33,140	≥32,840	0.78	0.40	0.79							
897470	MC	≥33,140	≥32,800	0.51	0.63	0.96							
897468	MC	≥33,140	≥32,730	0.38	0.44	1.20							
914082	SA	≥33,140	≥32,220	0.24	0.43	2.76	73.67	23.57					
897302	MC	≥33,140	≥32,600	0.49	0.51	1.25							
914121	SA	≥33,140	≥32,030	0.61	0.47	3.35	37.98	58.67					
914088	MC	≥33,140	≥32,260	0.51	0.49	1.78							
897444	SA	≥33,140	≥32,940	0.66	0.58	0.59	18.80	30.75	49.87				
878669	SA	≥33,140	≥32,910	0.56	0.48	0.69	19.04	49.05	31.22				
897475	SA	≥33,140	≥32,440	0.56	0.45	2.09	43.16	54.74					
897291	MS	≥33,140	≥32,830	0.64	0.55	0.91	35.40	63.69					
981838	MC	≥33,140	≥32,300	0.36	0.47	2.24							
981831	CR	≥33,140	≥32,290	0.24	0.69	2.13	54.97	18.74	18.95	4.78			
899959	CR	≥33,140	≥32,000	0.37	0.63	3.09	34.22	31.28	17.47	13.60			
897434	MC	≥33,140	≥32,760	0.81	0.41	1.13							
981850	MC	≥33,140	≥32,680	0.47	0.42	1.28							
898008	SA	≥33,140	≥32,100	0.59	0.50	3.11	39.58	57.31					
981890	MS	≥33,140	≥32,770	0.75	0.45	1.10	24.61	74.29					
914135	MC	≥33,140	≥32,400	0.84	0.39	2.18							
897305	MC	≥33,140	≥31,860	0.28	0.31	3.70							
897438	MC	≥33,140	≥31,860	0.74	0.44	3.83							
914099	SA	≥33,140	≥31,300	0.26	0.53	5.53	70.04	24.44					
897471	SA	≥33,140	≥30,800	0.43	0.52	7.06	52.78	40.16					
981866	SA	≥33,140	≥31,040	0.43	0.51	6.31	53.65	40.04					
981853	SA	≥33,140	≥31,710	0.45	0.53	4.30	52.29	43.41					

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**
914108	MS	≥33,140	≥32,560	0.37	0.47	1.74	61.46	36.81					
899955	CR	≥33,140	≥31,980	0.30	0.68	3.10	51.59	7.53	32.91	4.49			
981827	CR	≥33,140	≥32,330	0.20	0.65	2.17	55.43	9.26	14.14	5.84	7.31	3.08	2.50

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	≥49,930	≥49,900	0.57	0.45	0.05							
903245	MC	≥49,930	≥49,880	0.56	0.32	0.10							
914214	TE	≥49,930	≥49,900	0.73	0.36	0.05	26.87	73.08					
898173	SA	≥49,930	≥49,800	0.35	0.63	0.26	64.65	35.09					
800136	MC	≥49,930	≥49,860	0.59	0.39	0.13							
898145	MS	≥49,930	≥49,880	0.63	0.29	0.09	37.12	62.79					
898144	MC	≥49,930	≥49,840	0.56	0.55	0.18							
982506	SA	≥49,930	≥49,690	0.36	0.55	0.48	63.36	36.16					
898141	SA	≥49,930	≥49,890	0.40	0.66	0.07	45.51	29.46	24.96				
914209	MC	≥49,930	≥49,870	0.45	0.42	0.11							
898159	SA	≥49,930	≥49,780	0.39	0.63	0.30	60.34	39.35					
898151	MC	≥49,930	≥49,820	0.60	0.30	0.21							
914152	CR	≥49,930	≥49,010	0.31	0.71	1.07	36.93	25.23	17.98	12.65	5.38		
914148	CR	≥49,930	≥48,820	0.25	0.70	1.39	49.77	28.18	13.56	6.26			
870762	SA	≥49,930	≥49,690	0.20	0.59	0.48	68.20	22.08	9.24				
982499	SA	≥49,930	≥49,850	0.45	0.50	0.16	54.83	45.01					
914190	SA	≥49,930	≥49,520	0.44	0.49	0.82	55.87	43.31					
898152	MS	≥49,930	≥49,880	0.30	0.50	0.09	70.14	29.77					
898011	MC	≥49,930	≥49,840	0.50	0.42	0.17							
914215	MC	≥49,930	≥49,860	0.56	0.52	0.13							
897984	MC	≥49,930	≥49,800	0.38	0.52	0.25							
903244	MC	≥49,930	≥49,860	0.39	0.40	0.14							
982518	MS	≥49,930	≥49,850	0.71	0.47	0.15	29.14	70.71					
902410	CR	≥49,930	≥49,850	0.34	0.56	0.15	34.59	36.25	20.09	8.92			
902414	CR	≥49,930	≥48,470	0.14	0.55	1.64	74.62	8.71	10.24	3.53			
914140	SA	≥49,930	≥49,830	0.36	0.43	0.19	63.82	35.99					
914171	SA	≥49,930	≥49,800	0.57	0.55	0.26	43.16	56.57					
982538	MC	≥49,930	≥49,870	0.56	0.40	0.11							
914580	TE	≥49,930	≥49,870	0.64	0.42	0.12	36.32	63.56					
898162	MC	≥49,930	≥49,880	0.44	0.44	0.09							
914164	SA	≥49,930	≥49,710	0.26	0.31	0.43	73.86	25.71					
914155	TE	≥49,930	≥49,890	0.44	0.33	0.07	55.92	44.00					
914184	SA	≥49,930	≥49,770	0.63	0.38	0.32	36.54	63.14					
914198	SA	≥49,930	≥49,810	0.34	0.67	0.24	66.29	33.47					
982534	MS	≥49,930	≥49,860	0.25	0.51	0.13	74.83	25.04					
914203	MC	≥49,930	≥49,890	0.63	0.34	0.07							
914195	CR	≥49,930	≥49,810	0.26	0.69	0.24	54.86	22.39	11.43	11.07			
934015	CR	≥49,930	≥49,830	0.23	0.61	0.19	21.46	50.32	13.52	6.52	3.22	3.38	1.40

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	≥51,490	≥51,480	0.66	0.54	0.02	34.16	65.82					
901541	SA	≥51,490	≥50,910	0.64	0.44	1.12	35.18	63.70					
914223	TE	≥51,490	≥51,410	0.53	0.48	0.16	47.19	52.65					
914260	SA	≥51,490	≥50,940	0.37	0.45	1.06	61.92	37.02					
981981	TE	≥51,490	≥51,340	0.44	0.55	0.30	55.44	44.26					
916476	SA	≥51,490	≥51,160	0.29	0.36	0.64	70.86	28.50					
900521	SA	≥51,490	≥51,150	0.30	0.52	0.67	69.14	30.19					
901543	MS	≥51,490	≥51,400	0.45	0.46	0.18	55.17	44.65					
901534	MPSR	≥51,490	≥51,390	0.67	0.47	0.20	15.49	35.03	49.27				
878302	MC	≥51,490	≥51,390	0.60	0.48	0.20							
914237	TE	≥51,490	≥51,300	0.62	0.42	0.36	37.52	62.11					
914268	TE	≥51,490	≥51,220	0.37	0.46	0.53	62.30	37.17					
914230	SA	≥51,490	≥51,010	0.51	0.56	0.94	49.00	50.07					
878299	MC	≥51,490	≥51,150	0.46	0.36	0.67							
903077	SA	≥51,490	≥50,860	0.32	0.37	1.23	66.80	31.97					
914257	TE	≥51,490	≥51,010	0.67	0.54	0.94	32.20	66.86					
903099	MS	≥51,490	≥51,470	0.59	0.47	0.03	40.95	59.02					
982013	MC	≥51,490	≥51,450	0.35	0.43	0.09							
982025	TE	≥51,490	≥50,940	0.30	0.58	1.07	68.98	29.96					
901547	SA	≥51,490	≥51,310	0.48	0.63	0.34	42.26	19.66	37.73				
982019	SA	≥51,490	≥51,180	0.30	0.62	0.61	69.92	29.47					
903092	MC	≥51,490	≥51,420	0.45	0.26	0.14							
981963	CR	≥51,490	≥50,180	0.23	0.62	1.69	45.57	25.46	16.36	7.76	2.30		
982011	SA	≥51,490	≥51,240	0.24	0.52	0.48	75.34	24.18					
945486	SA	≥51,490	≥51,170	0.21	0.62	0.62	72.60	12.32	14.46				
981961	CR	≥51,490	≥50,070	0.23	0.63	2.13	52.01	28.02	12.82	4.39			
981954	CR	≥51,490	≥49,680	0.10	0.53	2.43	65.84	18.63	5.24	2.17	2.24	0.97	1.39
981956	CR	≥51,490	≥50,640	0.36	0.65	1.66	39.44	22.37	25.47	11.06			
914249	MC	≥51,490	≥51,420	0.52	0.31	0.14							
914271	SA	≥51,490	≥51,240	0.27	0.60	0.48	73.12	26.40					
901536	SA	≥51,490	≥51,380	0.44	0.60	0.21	55.47	44.31					
914273	SA	≥51,490	≥51,250	0.48	0.62	0.48	28.75	46.01	24.77				
914233	MS	≥51,490	≥51,420	0.19	0.53	0.14	80.60	19.26					
902741	TE	≥51,490	≥51,340	0.69	0.43	0.29	31.15	68.56					
903102	SA	≥51,490	≥51,150	0.23	0.57	0.67	76.84	22.49					
902748	MC	≥51,490	≥51,420	0.46	0.38	0.14							
914280	SA	≥51,490	≥51,340	0.20	0.58	0.30	79.44	20.26					
914231	CR	≥51,490	≥49,510	0.25	0.68	1.97	56.94	15.11	14.15	9.96			
903511	CR	≥51,490	≥51,430	0.23	0.56	0.12	36.81	45.51	9.20	6.25	2.10		

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**
914281	CR	≥51,490	≥49,790	0.23	0.66	1.95	63.00	13.42	8.88	11.39			

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	≥52,250	≥52,200	0.73	0.45	0.10							
914294	MS	≥52,250	≥52,070	0.20	0.51	0.35	79.92	19.73					
870847	SA	≥52,250	≥52,140	0.51	0.34	0.22	48.74	51.04					
982970	TE	≥52,250	≥52,120	0.20	0.49	0.25	80.08	19.67					
914299	MC	≥52,250	≥52,170	0.37	0.38	0.16							
914324	SA	≥52,250	≥51,740	0.50	0.47	0.98	49.31	49.71					
899318	MS	≥52,250	≥52,150	0.30	0.21	0.19	70.02	29.78					
983000	TE	≥52,250	≥52,130	0.43	0.33	0.22	56.65	43.12					
914359	SA	≥52,250	≥52,000	0.50	0.54	0.48	50.09	49.43					
914293	MS	≥52,250	≥52,110	0.29	0.57	0.28	70.32	29.40					
899322	SA	≥52,250	≥52,130	0.41	0.67	0.23	41.13	34.92	23.71				
983019	MC	≥52,250	≥52,050	0.63	0.42	0.39							
983004	SA	≥52,250	≥51,670	0.54	0.35	1.10	45.51	53.39					
914340	MC	≥52,250	≥51,850	0.44	0.44	0.77							
982988	MS	≥52,250	≥51,880	0.21	0.53	0.72	78.23	21.05					
983009	MC	≥52,250	≥51,790	0.25	0.28	0.89							
897990	SA	≥52,250	≥51,280	0.37	0.57	1.85	61.96	36.19					
983024	MC	≥52,250	≥51,660	0.43	0.29	1.12							
798344	MC	≥52,250	≥51,590	0.49	0.50	1.26							
899859	MC	≥52,250	≥52,220	0.33	0.32	0.06							
914330	MS	≥52,250	≥52,190	0.70	0.34	0.12	29.88	70.00					
982964	TE	≥52,250	≥52,210	0.38	0.42	0.08	62.40	37.52					
914633	MS	≥52,250	≥52,200	0.67	0.43	0.10	33.31	66.59					
899323	SA	≥52,250	≥52,070	0.54	0.50	0.34	45.91	53.75					
982941	MC	≥52,250	≥52,150	0.38	0.07	0.19							
982954	TE	≥52,250	≥52,140	0.29	0.59	0.21	61.75	17.44	20.61				
914362	CR	≥52,250	≥51,310	0.13	0.65	1.08	79.51	1.94	2.07	3.41	2.18	3.56	5.52
914316	TE	≥52,250	≥52,050	0.33	0.48	0.38	67.09	32.54					
902446	MC	≥52,250	≥52,180	0.37	0.29	0.14							
982922	CR	≥52,250	≥49,850	0.23	0.66	2.89	61.86	8.14	19.55	5.87			
868848	CR	≥52,250	≥48,890	0.06	0.53	3.89	83.38	3.88	5.09	1.21			
900539	CR	≥52,250	≥51,230	0.30	0.68	1.96	42.47	20.59	14.12	13.06	7.80		
914342	MC	≥52,250	≥52,200	0.48	0.35	0.10							
914319	SA	≥52,250	≥52,210	0.33	0.57	0.07	47.33	39.90	12.70				
982947	MC	≥52,250	≥52,200	0.38	0.31	0.10							
898444	SA	≥52,250	≥52,090	0.61	0.56	0.30	38.96	60.74					
900174	MC	≥52,250	≥52,210	0.82	0.37	0.08							
982935	MC	≥52,250	≥52,190	0.45	0.22	0.11							
914335	MPSR	≥52,250	≥52,210	0.34	0.43	0.09	45.88	40.37	13.67				

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	≥52,250	≥52,180	0.51	0.24	0.13							
900520	CR	≥52,250	≥49,340	0.10	0.60	3.44	81.69	3.17	2.82	6.74			
914339	CR	≥52,250	≥50,800	0.16	0.58	1.65	66.02	8.80	17.45	1.69	3.27		
982929	CR	≥52,250	≥50,080	0.21	0.63	2.72	60.71	15.93	13.51	5.69			

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
983049	MS	≥46,150	≥45,930	0.30	0.52	0.47	69.33	30.20					
914366	SA	≥46,150	≥45,060	0.29	0.46	2.36	69.28	28.37					
983076	MC	≥46,150	≥45,940	0.46	0.18	0.46							
897458	SA	≥46,150	≥44,630	0.16	0.50	3.30	80.77	15.93					
903089	MS	≥46,150	≥45,950	0.41	0.51	0.43	58.39	41.18					
914367	MC	≥46,150	≥45,890	0.40	0.31	0.56							
983117	MC	≥46,150	≥46,060	0.68	0.29	0.19							
983063	TE	≥46,150	≥45,960	0.42	0.54	0.42	37.98	38.66	22.94				
897074	MS	≥46,150	≥46,050	0.25	0.43	0.22	74.63	25.15					
914427	MS	≥46,150	≥45,800	0.34	0.57	0.76	65.65	33.59					
868884	MS	≥46,150	≥45,840	0.25	0.56	0.67	74.47	24.86					
896995	MS	≥46,150	≥45,870	0.27	0.38	0.60	73.01	26.39					
983032	SA	≥46,150	≥44,680	0.14	0.56	3.18	83.25	13.57					
891485	MS	≥46,150	≥45,610	0.16	0.37	1.17	82.70	16.13					
914431	SA	≥46,150	≥44,660	0.22	0.59	3.23	75.96	20.81					
896996	MC	≥46,150	≥45,470	0.52	0.30	1.48							
944912	MPSR	≥46,150	≥45,380	0.34	0.46	1.66	46.01	38.70	13.63				
914370	MS	≥46,150	≥45,060	0.23	0.56	2.36	75.26	22.38					
983074	MC	≥46,150	≥46,060	0.42	0.40	0.20							
903088	MPSR	≥46,150	≥46,130	0.75	0.48	0.04	13.37	22.44	64.14				
914433	MC	≥46,150	≥46,050	0.40	0.08	0.22							
983034	TE	≥46,150	≥46,040	0.19	0.59	0.23	80.90	18.87					
983010	CR	≥46,150	≥45,240	0.19	0.59	1.09	40.93	25.11	16.55	10.00	3.68	1.36	0.40
897072	SA	≥46,150	≥45,780	0.18	0.58	0.79	81.30	17.91					
982987	CR	≥46,150	≥44,220	0.15	0.49	2.54	66.36	13.34	10.21	1.61	4.30		
982999	CR	≥46,150	≥43,580	0.12	0.51	3.56	71.00	15.43	4.63	3.38			
870899	CR	≥46,150	≥43,300	0.10	0.52	4.21	75.63	10.51	5.09	2.59			
983109	TE	≥46,150	≥46,080	0.66	0.44	0.15	34.00	65.85					
914436	SA	≥46,150	≥45,750	0.23	0.68	0.88	65.62	20.82	12.69				
914396	MC	≥46,150	≥46,100	0.41	0.49	0.10							
914397	MC	≥46,150	≥46,090	0.27	0.50	0.12							
899312	CR	≥46,150	≥45,900	0.36	0.62	0.54	38.38	26.42	24.32	10.33			
914430	MS	≥46,150	≥46,070	0.21	0.54	0.16	78.51	21.33					
914426	MC	≥46,150	≥46,100	0.58	0.40	0.10							
914381	CR	≥46,150	≥43,420	0.14	0.64	3.15	61.18	14.35	16.76	1.23	0.57		
982967	TE	≥46,150	≥46,060	0.08	0.32	0.18	91.64	8.18					
899329	CR	≥46,150	≥45,910	0.25	0.51	0.52	49.89	31.48	12.10	6.01			

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 move 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 θ 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 θ will respond in category *x* to partial-credit item *j* is

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

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

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.

6.4 Calibration and Linking

Item calibration and linking for the LEAP 2025 forms were not performed in the spring of 2021. ELA forms used in the 2020-2021 administration were intact forms previously used in the 2018-2019 administration. New Meridian released some of the mathematics items used on the 2019 operational forms in all grades, but

grade 7. Most items were multiple-choice items. In mathematics grade 3, one released item was replaced with a spring 2018 operational item with the same subclaim and score point. For the other grades, the released items were administered, but not used for scoring. Table 6.19 summarizes the number of released items and the change to the score points when the released items were not counted. The number of released items ranged from one to four across grades. Adjusted scoring tables were created for math using these previously-calibrated and scaled items. For information regarding calibration and linking of these forms, please see the *2019 LEAP 2025 Grades 3-8 Operational Technical Report: English Language Arts and Mathematics*.

Table 6.19 Item Statistics—Number of Released Mathematics Items and Final Score Points

Grade	Spring 2019 Original Form		Number of Released Items	Spring 2021 Without Released Items	
	Total Items	Score Points		Total Items	Score Points
4	43	62	1	42	61
5	41	60	3	38	56
6	42	65	2	40	63
8	41	65	4	37	60

6.4.1.1. 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.

6.4.1.2. 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 subcategories that differ by grade. Subcategory performance is reported in one of three ratings of achievement: *Strong*, *Moderate*, or *Weak*. The 2021 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 Θ_{L3} and Θ_{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 Θ_{L3} and Θ_{L4} are cut scores for the category (mathematics) and subcategory (ELA).

This is also illustrated in Table 6.20.

Table 6.20 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	Θ_{L3}
4	750	50	35	Θ_{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 ELA forms used in the spring 2021 were intact forms with pre-existing raw-to-scale-score tables. The math forms that had released items on them had adjusted raw-to-scale-score tables. The grade 3 math form had a scoring table created using previously-used operational items. Tables 6.21 through 6.32. provide scale

scores at selected percentiles that can be used to compare the distributional characteristics of the spring 2021 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 6.21 Comparisons of Scale Scores at Selected Percentiles—Grade 3 ELA

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Form D
99	822	839	842	845	845
95	796	810	810	816	812
90	783	793	797	802	795
85	774	784	788	792	785
80	768	775	779	782	776
75	762	770	773	776	767
70	757	762	768	770	761
65	751	757	762	764	755
60	746	752	757	758	749
55	741	748	752	752	743
50	738	743	746	746	737
45	732	739	741	740	731
40	727	734	736	734	725
35	721	727	730	728	719
30	715	723	724	722	712
25	712	718	715	715	708
20	706	710	708	708	700
15	695	701	701	700	690
10	687	695	692	690	679
5	676	679	676	679	664
1	654	655	650	650	650

Table 6.22 Comparisons of Scale Scores at Selected Percentiles—Grade 4 ELA

	2016	2017	2018	2019	2021
Percentile	**Form A**	**Form B**	**Form C**	**Form D**	**Form D**
99	816	818	821	824	828
95	794	796	800	801	802
90	785	785	789	789	789
85	777	777	778	780	780
80	769	771	774	774	772
75	765	765	767	768	766
70	760	761	763	762	761
65	755	756	757	758	755
60	751	752	753	753	751
55	746	748	749	750	746
50	744	744	744	744	742
45	740	741	740	741	737
40	735	737	736	736	732
35	731	733	731	731	727
30	727	728	727	726	721
25	722	724	721	721	716
20	715	717	714	714	709
15	709	711	707	706	703
10	701	702	698	699	693
5	691	691	687	688	684
1	666	670	668	665	664

Table 6.23 Comparisons of Scale Scores at Selected Percentiles—Grade 5 ELA

	2016	2017	2018	2019	2021
Percentile	**Form A**	**Form B**	**Form C**	**Form D**	**Form D**
99	816	813	817	821	821
95	792	793	795	798	798
90	782	782	782	784	784
85	774	775	777	776	776
80	767	769	769	770	768
75	763	763	765	765	763
70	758	758	760	759	758
65	754	754	756	754	752
60	749	750	753	751	747
55	745	747	749	745	742
50	740	743	746	742	738
45	738	739	740	737	733
40	733	735	736	733	729
35	728	731	732	729	725
30	723	727	728	725	718
25	720	721	724	718	713
20	714	716	716	713	710
15	708	709	711	707	704
10	701	701	702	701	697
5	692	691	691	693	688
1	675	673	676	676	676

Table 6.24 Comparisons of Scale Scores at Selected Percentiles—Grade 6 ELA

	2016	2017	2018	2019	2021
Percentile	**Form A**	**Form B**	**Form C**	**Form D**	**Form D**
99	813	814	808	812	812
95	792	790	789	791	788
90	780	779	777	778	776
85	772	770	770	771	769
80	765	763	763	766	762
75	760	759	758	761	758
70	756	754	753	756	753
65	752	748	749	751	748
60	748	745	746	747	744
55	745	741	742	743	740
50	741	736	737	740	735
45	737	733	735	735	731
40	734	729	730	731	726
35	730	724	726	728	723
30	727	721	721	723	718
25	723	716	718	718	714
20	718	711	713	714	708
15	713	705	707	708	703
10	706	698	700	701	698
5	696	689	691	692	688
1	676	671	675	675	675

Table 6.25 Comparisons of Scale Scores at Selected Percentiles—Grade 7 ELA

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Form D
99	825	826	831	826	834
95	800	800	801	804	804
90	787	786	789	789	789
85	777	778	780	782	780
80	771	770	774	775	773
75	766	765	767	769	767
70	761	759	762	764	761
65	756	756	757	759	756
60	751	751	752	756	751
55	747	745	749	750	747
50	742	742	744	747	742
45	740	737	740	741	738
40	735	733	735	736	733
35	730	728	730	731	728
30	726	723	726	727	722
25	721	717	719	720	716
20	714	711	713	714	710
15	706	702	707	705	703
10	697	692	697	695	692
5	683	675	685	681	681
1	655	654	662	659	659

Table 6.26 Comparisons of Scale Scores at Selected Percentiles—Grade 8 ELA

	2016	2017	2018	2019	2021
Percentile	**Form A**	**Form B**	**Form C**	**Form D**	**Form D**
99	825	834	824	831	831
95	804	806	801	804	806
90	790	791	789	793	793
85	781	782	781	785	783
80	775	776	774	777	775
75	770	770	768	771	769
70	764	764	764	766	764
65	759	758	758	760	758
60	754	754	754	755	753
55	752	749	751	750	748
50	747	745	745	746	743
45	743	740	741	741	738
40	739	734	737	736	734
35	735	731	732	732	728
30	731	725	726	727	723
25	727	719	722	721	717
20	721	714	716	714	710
15	714	707	708	707	702
10	706	696	699	696	693
5	693	681	683	686	682
1	670	651	657	667	660

Table 6.27 Comparisons of Scale Scores at Selected Percentiles—Grade 3 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Revised Form D
99	824	822	817	815	816
95	802	796	793	796	790
90	789	786	783	784	778
85	781	776	775	776	768
80	775	772	771	771	764
75	770	765	764	764	758
70	765	761	759	760	752
65	760	756	755	756	748
60	756	752	750	752	742
55	751	747	746	748	738
50	746	743	742	744	734
45	741	738	740	738	727
40	738	733	735	735	723
35	733	728	731	731	719
30	728	725	726	724	711
25	722	720	719	720	706
20	716	715	713	713	700
15	710	706	708	705	694
10	703	699	698	700	686
5	692	689	686	686	677
1	672	667	664	672	658

Table 6.28 Comparisons of Scale Scores at Selected Percentiles—Grade 4 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Revised Form D
99	819	812	812	813	803
95	797	792	790	792	785
90	786	779	780	781	775
85	777	774	772	774	768
80	771	767	768	769	762
75	766	762	762	763	757
70	761	756	757	759	751
65	756	752	753	755	746
60	752	748	749	750	741
55	747	744	744	746	737
50	743	740	740	742	732
45	738	736	735	737	726
40	732	732	733	732	722
35	728	727	728	728	717
30	723	722	723	724	711
25	718	717	718	719	706
20	713	712	715	712	699
15	708	706	710	706	693
10	703	700	700	699	688
5	693	693	689	688	679
1	677	674	670	673	658

Table 6.29 Comparisons of Scale Scores at Selected Percentiles—Grade 5 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Revised Form D
99	819	808	810	809	803
95	792	784	784	788	782
90	779	774	774	778	772
85	771	767	765	769	765
80	766	760	759	763	757
75	759	755	755	757	751
70	754	751	749	753	747
65	749	747	745	748	741
60	745	742	743	744	737
55	740	740	738	740	733
50	735	735	734	737	729
45	731	730	729	733	724
40	728	728	727	728	719
35	722	723	722	724	716
30	720	720	720	719	710
25	714	715	714	714	707
20	711	709	711	711	703
15	705	706	705	705	699
10	699	699	698	699	690
5	691	691	689	690	685
1	678	675	672	674	671

Table 6.30 Comparisons of Scale Scores at Selected Percentiles—Grade 6 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Revised Form D
99	803	808	800	804	798
95	783	781	780	783	777
90	771	771	770	773	768
85	765	762	762	765	760
80	758	757	757	758	754
75	753	752	752	754	749
70	747	746	748	750	743
65	744	742	743	745	740
60	740	738	739	742	735
55	735	734	736	739	731
50	731	732	732	733	727
45	729	727	728	729	723
40	724	724	723	725	718
35	722	719	721	721	713
30	717	717	716	717	710
25	714	711	713	714	704
20	709	708	707	709	701
15	706	701	704	703	693
10	699	697	696	696	689
5	692	688	686	687	683
1	679	671	672	667	656

Table 6.31 Comparisons of Scale Scores at Selected Percentiles—Grade 7 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Form D
99	797	796	797	796	793
95	779	777	777	776	773
90	768	766	766	766	764
85	760	760	759	761	757
80	754	754	755	756	752
75	750	749	750	752	748
70	746	746	745	748	743
65	742	741	742	743	740
60	738	737	739	740	736
55	734	734	735	736	732
50	730	731	731	732	728
45	728	727	729	730	724
40	723	723	725	726	722
35	721	721	721	722	719
30	719	717	718	719	714
25	714	712	713	714	711
20	712	709	710	711	708
15	706	706	706	705	701
10	703	699	702	701	697
5	695	694	693	692	687
1	678	673	679	680	671

Table 6.32 Comparisons of Scale Scores at Selected Percentiles—Grade 8 Mathematics

	2016	2017	2018	2019	2021
Percentile	Form A	Form B	Form C	Form D	Revised Form D
99	808	809	807	812	806
95	787	784	784	788	781
90	775	771	773	775	768
85	766	763	764	766	759
80	761	757	757	758	751
75	753	751	752	752	747
70	749	746	746	746	740
65	744	741	742	742	735
60	737	736	737	737	732
55	734	730	732	732	726
50	731	727	727	730	723
45	727	724	721	724	716
40	724	718	718	721	712
35	720	714	715	715	708
30	712	710	707	711	703
25	708	706	702	707	698
20	704	698	697	699	693
15	699	693	691	694	686
10	695	687	684	689	679
5	684	674	676	677	671
1	663	656	654	659	650

Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) for the LEAP 2025 across administrations, see figures 6.1 and 6.2. For ELA forms and grade 7 mathematics, the 2021 form is the intact 2019 form, and since they would share a curve, they are labeled LEAP2019_2021. For most content areas and grades, the TCCs for the three years were similar across ability ranges. For ELA grade 5 and grade 6, the 2018 forms were slightly easier than the 2017 and 2019/2021 forms for high-performing students. For grade 7, the 2018 and 2019/2021 forms were slightly easier than the 2017 forms. Grade 3 forms have been gradually becoming more difficult from 2017 to 2019/2021. For grade 8, the 2019/2021 form was more difficult than the 2017 and 2018 forms across all ability levels.

For mathematics grades 7 and 8, the 2019 and 2021 and 2017 forms were slightly easier than the 2018 form for low-performing students. For grades 3 and 4, the 2019 and 2021 forms were slightly more difficult than the 2018 forms for low-performing students. For grade 5, the 2019 and 2021 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 6.3 and 6.4 show SEMs for the 2017- 2021 LEAP 2025 assessments. For most content areas and grades, the SEMs were similar across ability ranges, especially in the middle ability ranges.

Figure 6.1 TCCs Across Years: ELA

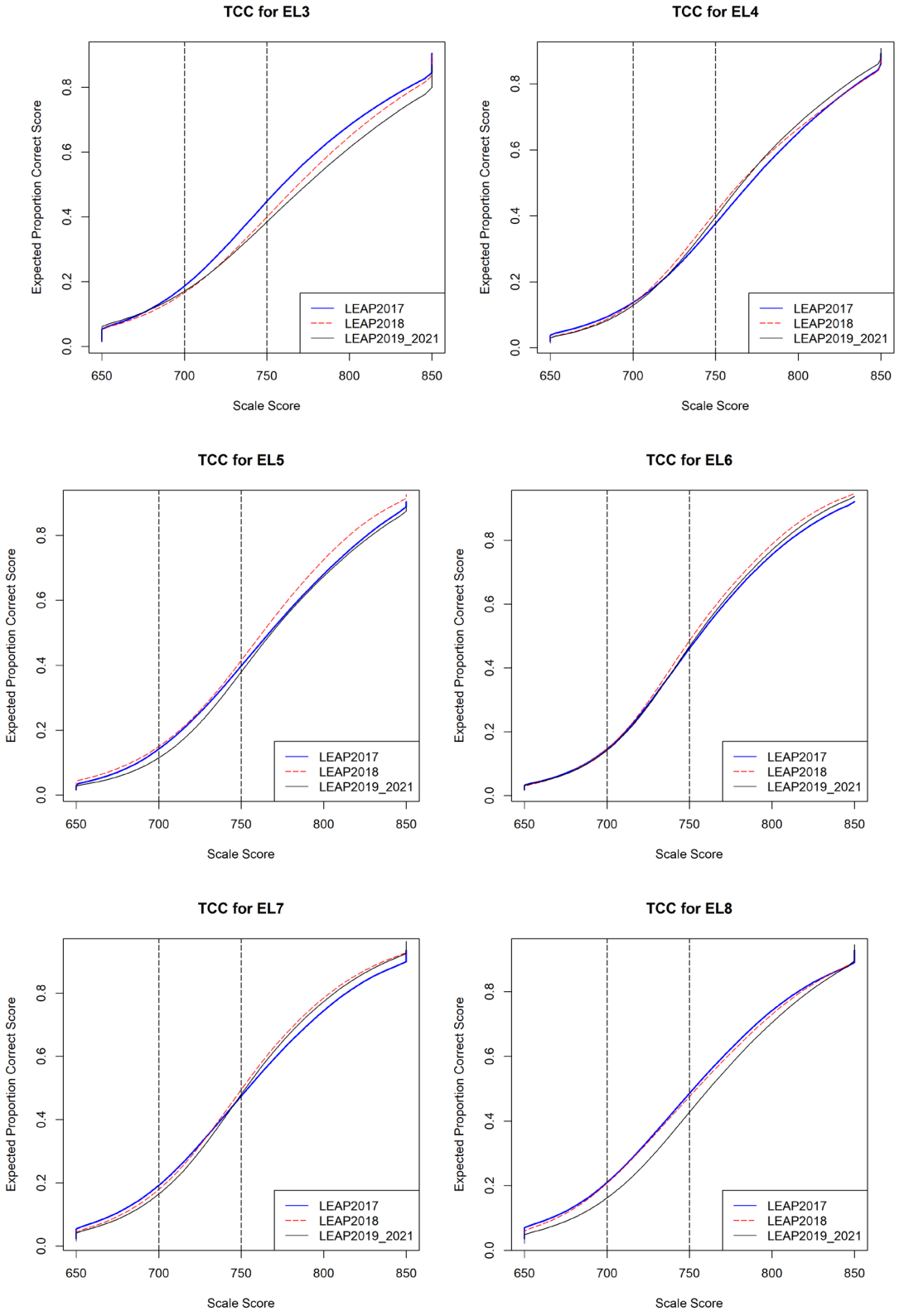


Figure 6.2 TCCs Across Years: Mathematics

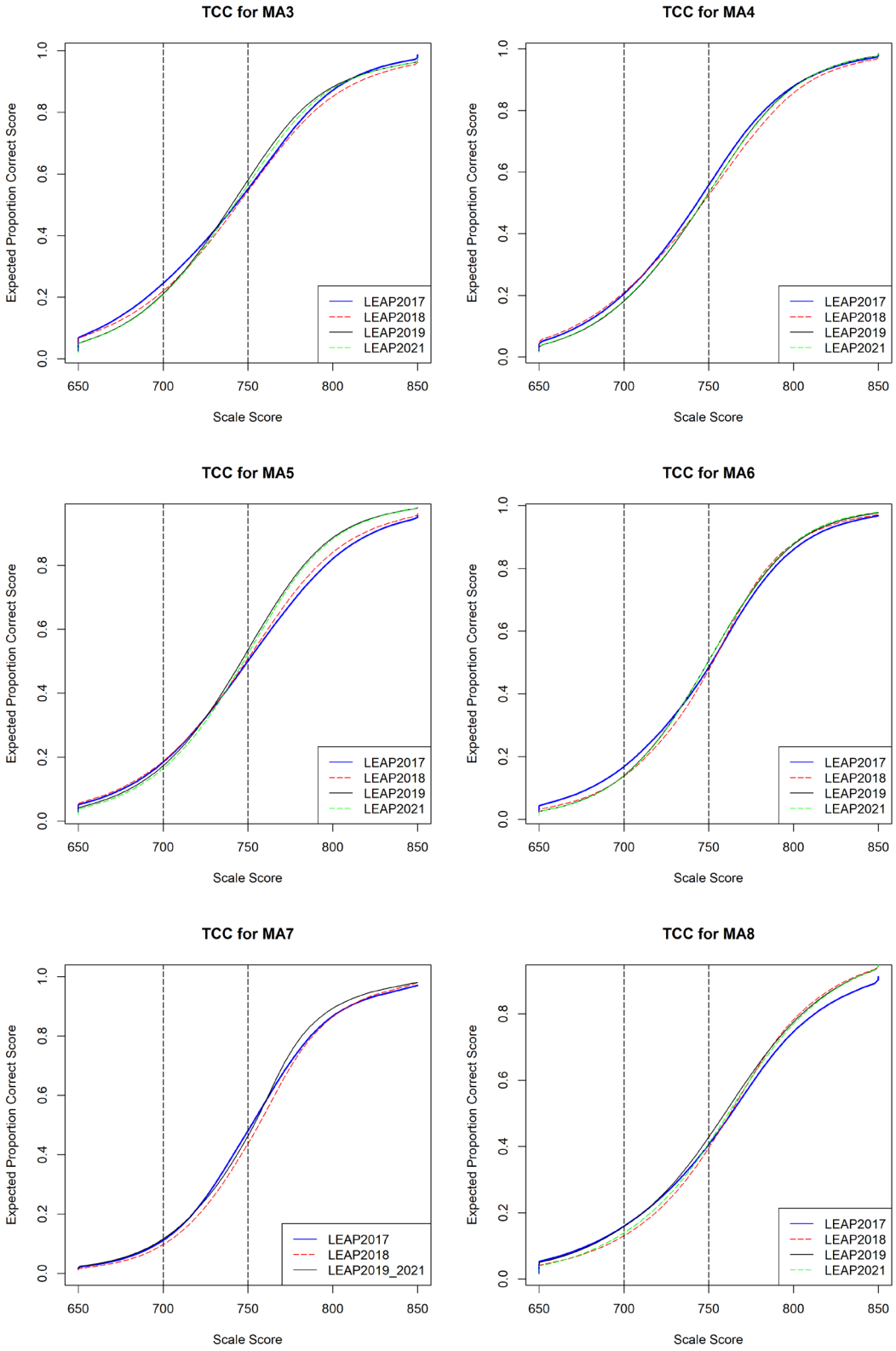


Figure 6.3 SEM Across Years: ELA

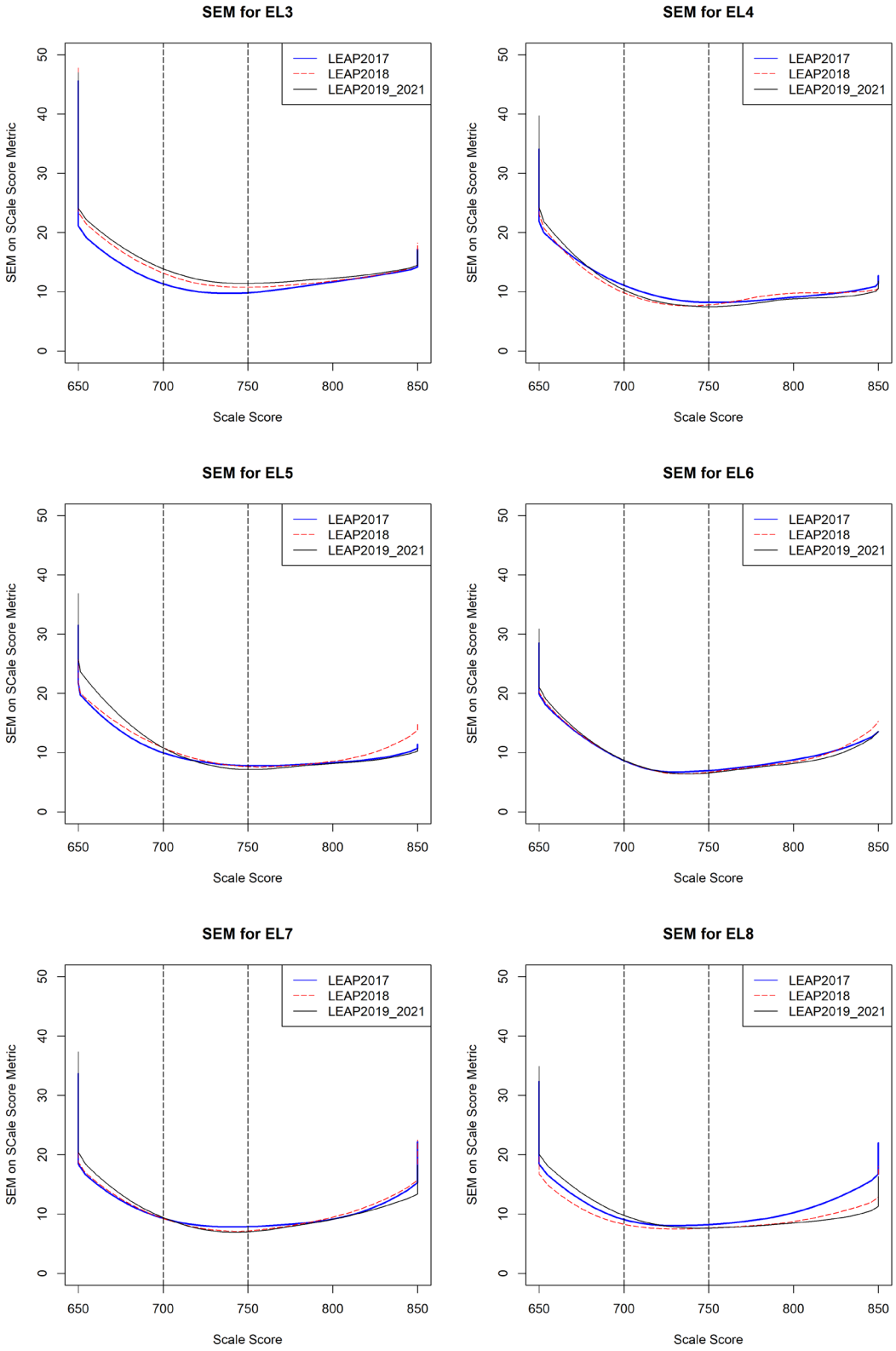
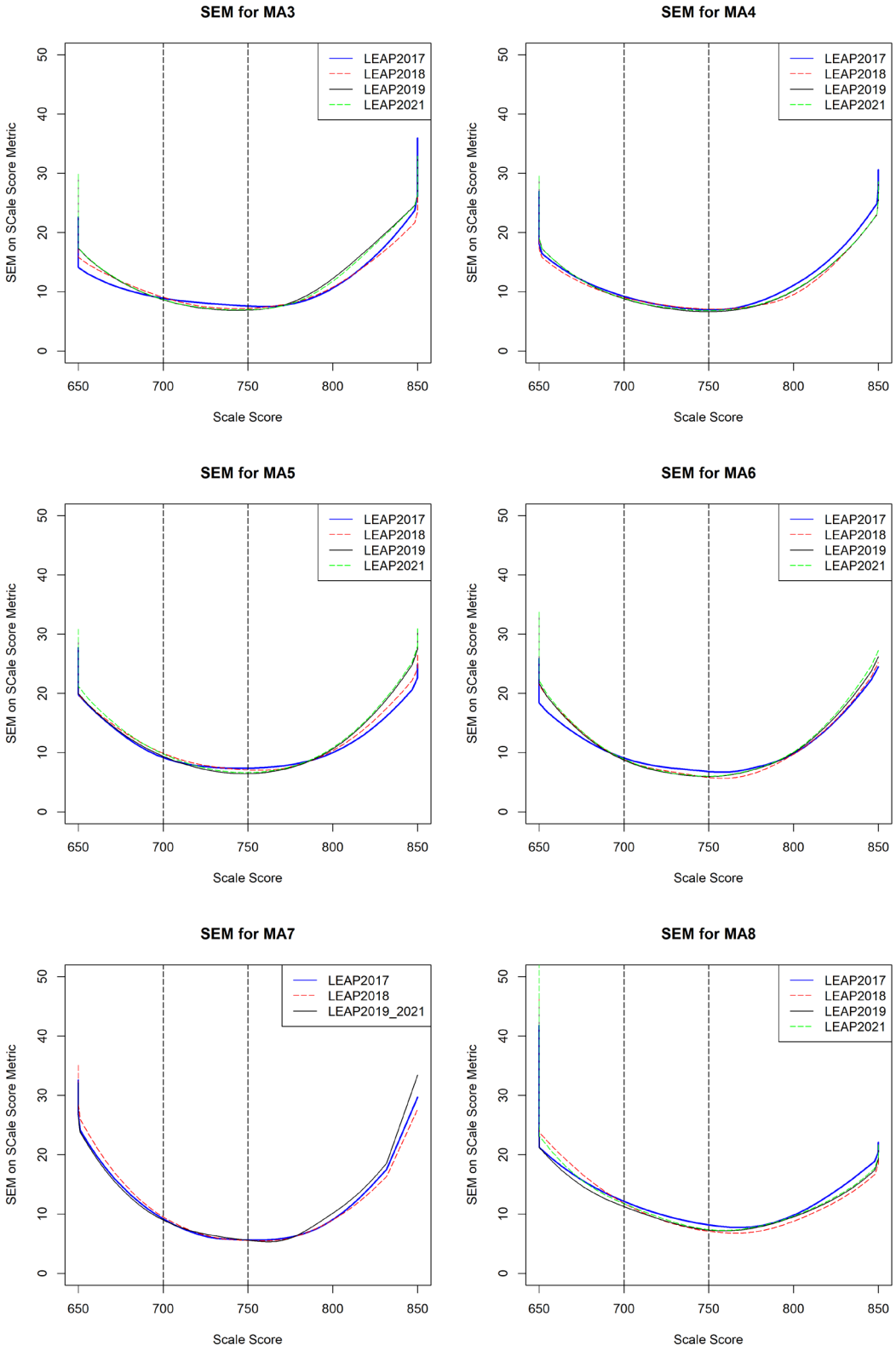


Figure 6.4 SEM Across Years: Mathematics



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 2021 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	≥50,130	98.75%	≥50,540	98.79%
	Gender				
	Female	≥24,530	98.79%	≥24,690	98.82%
	Male	≥25,560	98.74%	≥25,760	98.79%
	Ethnicity				
	Hispanic/Latino	≥4,620	98.90%	≥4,670	98.95%
	American Indian or Alaska Native	≥280	98.96%	≥290	99.32%
	Asian	≥820	99.15%	≥820	99.15%
	Black or African American	≥21,370	98.54%	≥21,530	98.60%
	Native Hawaiian or Other Pacific	≥40	97.78%	≥40	97.87%
	White	≥21,240	98.97%	≥21,360	98.98%
	Two or More Races	≥1,680	98.39%	≥1,690	98.46%
	Education Classification				
	Regular	≥43,820	98.83%	≥44,190	98.86%
	Special	≥6,300	98.19%	≥6,350	98.27%
	Economic Status				
	Economically Disadvantaged	≥37,140	98.68%	≥37,380	98.73%
	Not Economically Disadvantaged	≥12,980	98.98%	≥13,160	98.94%
	English Learner Status				
	Not English Learner	≥47,920	98.74%	≥48,310	98.76%
	English Learner	≥2,200	99.09%	≥2,230	99.24%
	Migrant Status				
	Not Migrant	≥46,770	98.73%	≥47,170	98.75%
	Migrant	≥3,350	99.14%	≥3,370	99.23%
	Section 504 Status				
	Not Section 504	≥50,030	98.75%	≥50,440	98.79%
	Section 504	≥100	98.02%	≥100	98.02%
	Homeless Status				
	Not Homeless	≥49,070	98.76%	≥49,460	98.79%
	Homeless	≥1,050	98.49%	≥1,080	98.43%
	Foster Care Status				
	Not in Foster Care	≥49,890	98.75%	≥50,290	98.78%
	In Foster Care	≥240	100.00%	≥240	100.00%
	Military Affiliation				
	Not Military Affiliated	≥49,150	98.74%	≥49,550	98.77%
	Military Affiliated	≥980	99.39%	≥990	99.39%

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	≥50,290	98.67%	≥50,590	98.69%
	Gender				
	Female	≥24,470	98.70%	≥24,590	98.74%
	Male	≥25,780	98.70%	≥25,940	98.73%
	Ethnicity				
	Hispanic/Latino	≥4,450	98.97%	≥4,490	99.00%
	American Indian or Alaska Native	≥260	99.26%	≥270	99.26%
	Asian	≥770	98.84%	≥770	98.84%
	Black or African American	≥21,600	98.54%	≥21,750	98.62%
	Native Hawaiian or Other Pacific	≥30	100.00%	≥30	100.00%
	White	≥21,430	98.79%	≥21,510	98.79%
	Two or More Races	≥1,660	98.67%	≥1,660	98.68%
	Education Classification				
	Regular	≥44,040	98.73%	≥44,300	98.75%
	Special	≥6,240	98.24%	≥6,280	98.25%
	Economic Status				
	Economically Disadvantaged	≥36,870	98.62%	≥37,070	98.67%
	Not Economically Disadvantaged	≥13,420	98.79%	≥13,520	98.73%
	English Learner Status				
	Not English Learner	≥48,370	98.64%	≥48,650	98.66%
	English Learner	≥1,910	99.32%	≥1,930	99.33%
	Migrant Status				
	Not Migrant	≥45,810	98.62%	≥46,090	98.64%
	Migrant	≥4,480	99.11%	≥4,500	99.16%
	Section 504 Status				
	Not Section 504	≥50,200	98.67%	≥50,510	98.68%
	Section 504	≥80	100.00%	≥80	100.00%
	Homeless Status				
	Not Homeless	≥49,320	98.68%	≥49,610	98.70%
	Homeless	≥960	97.93%	≥970	97.96%
	Foster Care Status				
	Not in Foster Care	≥50,060	98.67%	≥50,360	98.69%
	In Foster Care	≥220	97.81%	≥220	97.82%
	Military Affiliation				
	Not Military Affiliated	≥49,370	98.66%	≥49,670	98.68%
	Military Affiliated	≥910	98.90%	≥910	99.02%

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	≥50,270	98.95%	≥50,270	98.97%
	Gender				
	Female	≥24,390	99.02%	≥24,390	99.04%
	Male	≥25,870	98.89%	≥25,880	98.91%
	Ethnicity				
	Hispanic/Latino	≥4,570	99.39%	≥4,570	99.41%
	American Indian or Alaska Native	≥290	99.32%	≥290	99.32%
	Asian	≥810	99.01%	≥810	99.01%
	Black or African American	≥21,470	98.67%	≥21,480	98.70%
	Native Hawaiian or Other Pacific	≥20	100.00%	≥20	100.00%
	White	≥21,420	99.18%	≥21,420	99.19%
	Two or More Races	≥1,630	98.35%	≥1,630	98.35%
	Education Classification				
	Regular	≥44,190	99.03%	≥44,200	99.05%
	Special	≥6,070	98.39%	≥6,070	98.39%
	Economic Status				
	Economically Disadvantaged	≥37,060	98.82%	≥37,070	98.84%
	Not Economically Disadvantaged	≥13,200	99.33%	≥13,200	99.33%
	English Learner Status				
	Not English Learner	≥48,300	98.93%	≥48,310	98.95%
	English Learner	≥1,960	99.44%	≥1,960	99.49%
	Migrant Status				
	Not Migrant	≥45,130	98.93%	≥45,130	98.95%
	Migrant	≥5,140	99.12%	≥5,140	99.18%
	Section 504 Status				
	Not Section 504	≥50,220	98.95%	≥50,230	98.97%
	Section 504	≥40	100.00%	≥40	100.00%
	Homeless Status				
	Not Homeless	≥49,160	98.97%	≥49,160	98.99%
	Homeless	≥1,110	98.29%	≥1,110	98.29%
	Foster Care Status				
	Not in Foster Care	≥50,070	98.95%	≥50,080	98.97%
	In Foster Care	≥190	98.47%	≥190	98.47%
	Military Affiliation				
	Not Military Affiliated	≥49,390	98.94%	≥49,400	98.96%
	Military Affiliated	≥870	99.54%	≥870	99.54%

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	≥52,240	98.50%	≥52,240	98.51%
	Gender				
	Female	≥25,660	98.59%	≥25,660	98.61%
	Male	≥26,570	98.41%	≥26,580	98.42%
	Ethnicity				
	Hispanic/Latino	≥4,470	98.79%	≥4,470	98.79%
	American Indian or Alaska Native	≥300	99.35%	≥300	99.35%
	Asian	≥760	99.48%	≥760	99.48%
	Black or African American	≥22,790	98.08%	≥22,790	98.10%
	Native Hawaiian or Other Pacific	≥40	97.92%	≥40	97.92%
	White	≥22,130	98.87%	≥22,130	98.87%
	Two or More Races	≥1,710	98.07%	≥1,710	98.07%
	Education Classification				
	Regular	≥46,250	98.58%	≥46,260	98.59%
	Special	≥5,980	97.91%	≥5,980	97.93%
	Economic Status				
	Economically Disadvantaged	≥38,560	98.26%	≥38,570	98.28%
	Not Economically Disadvantaged	≥13,670	99.17%	≥13,670	99.18%
	English Learner Status				
	Not English Learner	≥50,440	98.48%	≥50,450	98.50%
	English Learner	≥1,790	99.00%	≥1,790	99.00%
	Migrant Status				
	Not Migrant	≥46,680	98.52%	≥46,680	98.53%
	Migrant	≥5,560	98.35%	≥5,560	98.35%
	Section 504 Status				
	Not Section 504	≥52,180	98.50%	≥52,190	98.51%
	Section 504	≥50	100.00%	≥50	100.00%
	Homeless Status				
	Not Homeless	≥51,110	98.55%	≥51,110	98.56%
	Homeless	≥1,130	96.28%	≥1,130	96.46%
	Foster Care Status				
	Not in Foster Care	≥52,040	98.50%	≥52,050	98.51%
	In Foster Care	≥190	98.97%	≥190	98.97%
	Military Affiliation				
	Not Military Affiliated	≥51,370	98.48%	≥51,370	98.50%
	Military Affiliated	≥870	99.54%	≥870	99.54%

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	≥53,190	98.24%	≥53,210	98.29%
	Gender				
	Female	≥26,090	98.27%	≥26,100	98.31%
	Male	≥27,100	98.21%	≥27,100	98.27%
	Ethnicity				
	Hispanic/Latino	≥4,530	98.83%	≥4,530	98.92%
	American Indian or Alaska Native	≥300	98.06%	≥300	98.06%
	Asian	≥830	98.57%	≥830	98.57%
	Black or African American	≥23,030	97.86%	≥23,040	97.95%
	Native Hawaiian or Other Pacific	≥40	100.00%	≥40	100.00%
	White	≥22,790	98.54%	≥22,800	98.56%
	Two or More Races	≥1,630	97.61%	≥1,630	97.61%
	Education Classification				
	Regular	≥47,430	98.35%	≥47,440	98.40%
	Special	≥5,760	97.36%	≥5,760	97.40%
	Economic Status				
	Economically Disadvantaged	≥38,610	97.86%	≥38,630	97.93%
	Not Economically Disadvantaged	≥14,580	99.26%	≥14,580	99.27%
	English Learner Status				
	Not English Learner	≥51,430	98.24%	≥51,450	98.29%
	English Learner	≥1,750	98.18%	≥1,750	98.29%
	Migrant Status				
	Not Migrant	≥47,570	98.20%	≥47,590	98.25%
	Migrant	≥5,610	98.58%	≥5,610	98.65%
	Section 504 Status				
	Not Section 504	≥53,130	98.24%	≥53,140	98.29%
	Section 504	≥60	98.44%	≥60	98.44%
	Homeless Status				
	Not Homeless	≥52,070	98.30%	≥52,080	98.35%
	Homeless	≥1,120	95.64%	≥1,120	95.73%
	Foster Care Status				
	Not in Foster Care	≥52,980	98.25%	≥53,000	98.30%
	In Foster Care	≥210	96.67%	≥210	96.68%
	Military Affiliation				
	Not Military Affiliated	≥52,290	98.22%	≥52,300	98.27%
	Military Affiliated	≥900	99.45%	≥900	99.45%

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	≥52,780	98.31%	≥52,820	98.38%
	Gender				
	Female	≥25,980	98.43%	≥25,990	98.51%
	Male	≥26,790	98.19%	≥26,820	98.25%
	Ethnicity				
	Hispanic/Latino	≥4,100	98.71%	≥4,100	98.78%
	American Indian or Alaska Native	≥310	99.37%	≥310	99.37%
	Asian	≥800	99.13%	≥800	99.13%
	Black or African American	≥22,690	98.05%	≥22,720	98.17%
	Native Hawaiian or Other Pacific	≥40	97.73%	≥40	97.73%
	White	≥23,240	98.48%	≥23,250	98.51%
	Two or More Races	≥1,570	97.77%	≥1,570	97.77%
	Education Classification				
	Regular	≥47,430	98.41%	≥47,470	98.49%
	Special	≥5,340	97.38%	≥5,350	97.37%
	Economic Status				
	Economically Disadvantaged	≥37,640	97.94%	≥37,680	98.04%
	Not Economically Disadvantaged	≥15,130	99.21%	≥15,130	99.23%
	English Learner Status				
	Not English Learner	≥51,050	98.34%	≥51,090	98.38%
	English Learner	≥1,720	97.16%	≥1,730	98.27%
	Migrant Status				
	Not Migrant	≥47,370	98.36%	≥47,410	98.39%
	Migrant	≥5,400	97.84%	≥5,410	98.23%
	Section 504 Status				
	Not Section 504	≥52,720	98.31%	≥52,760	98.38%
	Section 504	≥50	98.21%	≥50	98.21%
	Homeless Status				
	Not Homeless	≥51,800	98.38%	≥51,850	98.42%
	Homeless	≥970	94.13%	≥970	96.19%
	Foster Care Status				
	Not in Foster Care	≥52,560	98.32%	≥52,600	98.39%
	In Foster Care	≥220	95.45%	≥220	95.93%
	Military Affiliation				
	Not Military Affiliated	≥51,990	98.29%	≥52,030	98.36%
	Military Affiliated	≥780	99.36%	≥780	99.36%

**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 2021 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
2021	≥49,630	19.3	19.0	23.1	33.4	5.2

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
2021	≥49,550	13.7	19.1	25.7	32.3	9.3

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
2021	≥49,780	10.7	24.0	28.1	32.7	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
2021	≥51,430	12.1	26.1	28.3	28.7	4.9

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
2021	≥52,180	13.4	18.3	26.2	29.1	13.0

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
2021	≥51,680	14.3	16.4	25.2	34.9	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
2021	≥49,590	18.2	22.9	25.3	28.3	5.3

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
2021	≥49,490	20.0	23.1	25.2	29.7	2.1

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
2021	≥49,700	18.5	28.6	26.7	23.2	3.1

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
2021	≥51,340	18.8	27.9	28.9	21.9	2.5

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
2021	≥52,080	12.0	33.0	32.6	20.5	1.9

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
2021	≥45,840	27.3	25.8	25.2	20.2	1.5

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* (2021).

In addition to providing interpretationof 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 2021 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 their peers, are found at the bottom of the page. When a student does not receive a scale score, their 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			
	Approaching Basic	***Basic***	***Mastery***	***Advanced***
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			
	Approaching Basic	***Basic***	***Mastery***	***Advanced***
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 study 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 study 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 study 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 study 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 study 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**
Advanced	810–850	790–850	799–850	790–850	785–850	794–850
Mastery	750–809	750–789	750–798	750–789	750–784	750–793
Basic	725–749					
Approaching Basic	700–724					
Unsatisfactory	650–699					
MATHEMATICS						
Achievement Level	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
Advanced	790–850	796–850	790–850	788–850	786–850	801–850
Mastery	750–789	750–795	750–789	750–787	750–785	750–800
Basic	725–749					
Approaching Basic	700–724					
Unsatisfactory	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 2020–2021 LEAP Grades 3-8 ELA and Mathematics Assessment Frameworks
	Test Blueprint and Item Alignment to Curriculum and Standards	Chapter 3 2020–2021 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	2020–2021 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 2019 to 2021	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

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

where n is the number of items on the test, σ_i^2 is the variance of item i, and σ_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 0.87 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 0.91 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	4.21	0.88	≥12,090
ELA	3	PBT	26	71	4.58	0.87	≥37,540
ELA	4	CBT	28	86	4.97	0.90	≥16,480
ELA	4	PBT	28	86	5.39	0.89	≥33,070
ELA	5	CBT	28	86	4.97	0.90	≥49,780
ELA	6	CBT	32	90	5.20	0.91	≥51,430
ELA	7	CBT	32	90	5.60	0.92	≥52,180
ELA	8	CBT	32	94	5.71	0.90	≥51,680
Mathematics	3	CBT	43	62	3.47	0.93	≥12,070
Mathematics	3	PBT	43	62	3.72	0.93	≥37,520
Mathematics	4	CBT	42	61	3.35	0.94	≥16,430
Mathematics	4	PBT	42	61	3.53	0.94	≥33,050
Mathematics	5	CBT	38	56	3.33	0.93	≥49,700
Mathematics	6	CBT	40	63	3.46	0.94	≥51,340
Mathematics	7	CBT	43	66	3.80	0.92	≥52,080
Mathematics	8	CBT	37	60	3.23	0.91	≥45,840

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:

$$\text{SEM} = SD\sqrt{1 - R_{xx'}}\ , \tag{9.2}$$

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:

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

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

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

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. Since LEAP 2025 was first administered, every effort has been made to make the TCC and CSEM values at the cut scores between the PARCC assessments and the LEAP 2025 assessments similar. Both TCC and CSEM values have been similar across the LEAP 2025 alternate forms given the same content because similar or the same statistical properties are important for alternate forms. To provide context regarding the magnitude of the CSEMs, it is important to also refer to sections 9.2.1 Test Reliability and 9.2.4 Classification Accuracy and Consistency where evidence is provided of high measures of form reliability and levels of accurate student classification at the cutpoints to support the use of the LEAP 2025 assessments.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

			Approaching Basic		*Basic*		*Mastery*		*Advanced*	
Content Area	**Grade**	**Mode**	**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	10
Mathematics	3	PBT	700	9	725	7	750	7	790	10
Mathematics	4	CBT	700	9	725	7	750	7	796	10
Mathematics	4	PBT	700	9	725	7	750	7	796	10
Mathematics	5	CBT	700	9	725	7	750	7	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

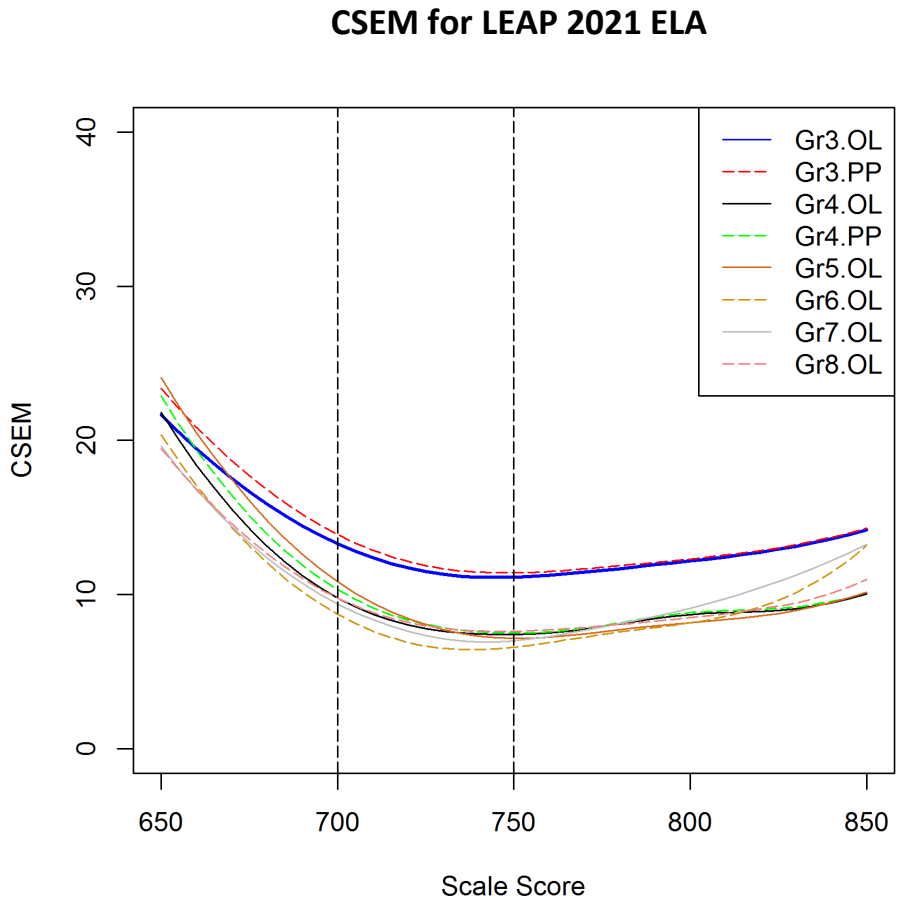
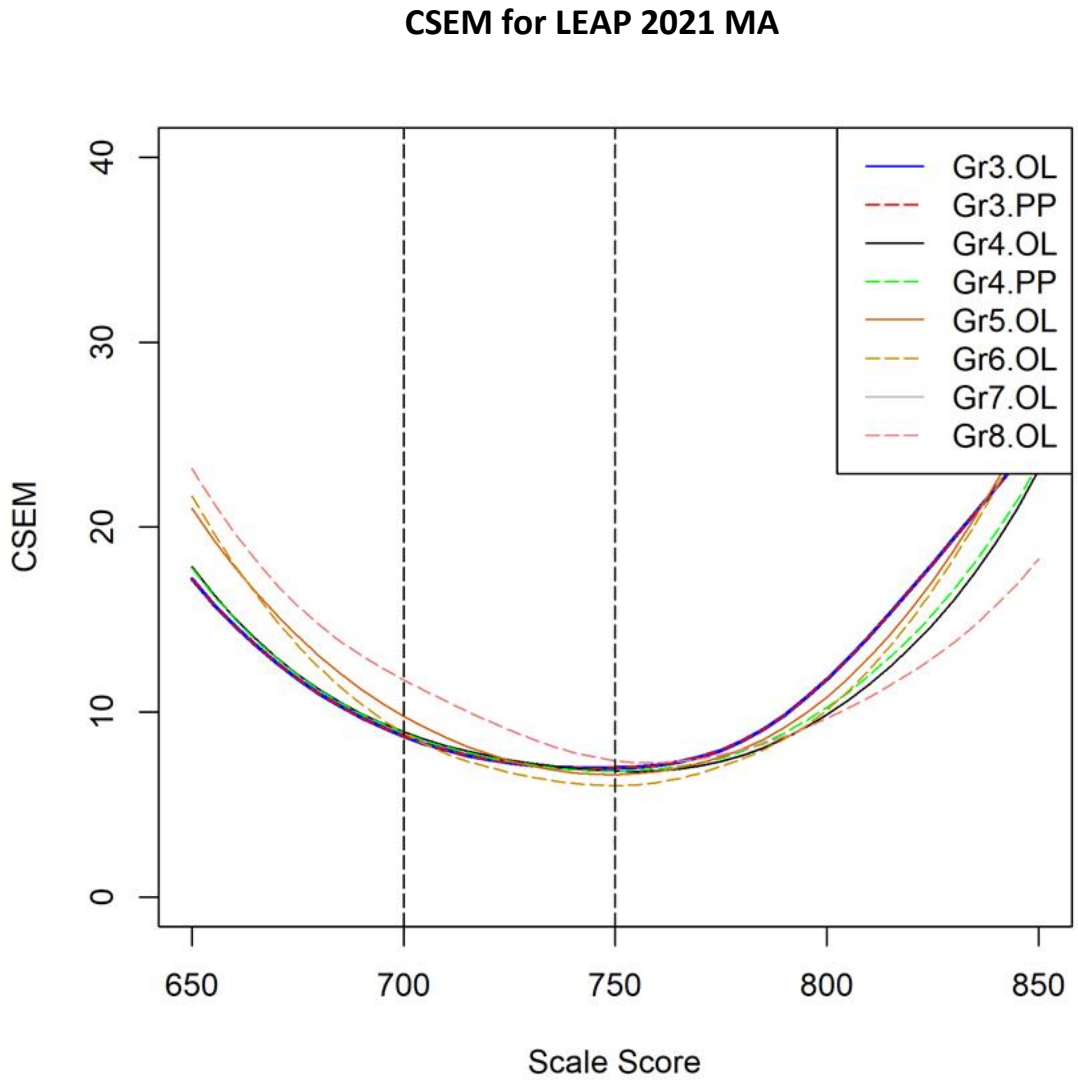


Figure 9.2 CSEM Curves for Mathematics Grades 3 through 8



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 2021 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 2021 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 0.53 to 0.84 for ELA and 0.22 to 0.88 for mathematics. Classification consistency (see Table 9.5) conditioned on achievement level ranged from 0.43 to 0.75 for ELA and 0.35 to 0.82 for mathematics. Table 9.6 shows that classification accuracy at achievement cut points ranged from 0.89 to 0.97 for ELA and 0.88 to 0.99 for mathematics. Classification consistency (see Table 9.7) conditioned at achievement cut points ranged from 0.85 to 0.97 for ELA and 0.84 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

	Classification Accuracy						
Content Area	Grade	Mode	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
ELA	3	CBT	0.80	0.58	0.63	0.77	0.53
ELA	3	PBT	0.71	0.57	0.59	0.82	0.57
ELA	4	CBT	0.77	0.68	0.69	0.78	0.66
ELA	4	PBT	0.68	0.62	0.71	0.79	0.69
ELA	5	CBT	0.61	0.64	0.71	0.84	0.58
ELA	6	CBT	0.65	0.75	0.74	0.80	0.63
ELA	7	CBT	0.76	0.66	0.73	0.77	0.76
ELA	8	CBT	0.72	0.63	0.68	0.80	0.70
Mathematics	3	CBT	0.83	0.72	0.74	0.84	0.62
Mathematics	3	PBT	0.80	0.71	0.73	0.85	0.59
Mathematics	4	CBT	0.82	0.72	0.74	0.88	0.43
Mathematics	4	PBT	0.84	0.67	0.76	0.88	0.22
Mathematics	5	CBT	0.76	0.66	0.77	0.83	0.56
Mathematics	6	CBT	0.77	0.74	0.78	0.84	0.60
Mathematics	7	CBT	0.43	0.77	0.75	0.84	0.68
Mathematics	8	CBT	0.81	0.59	0.72	0.82	0.66

Table 9.5 Classification Consistency Conditioned on Level of Achievement

	Classification Consistency						
Content Area	Grade	Mode	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
ELA	3	CBT	0.75	0.43	0.48	0.69	0.52
ELA	3	PBT	0.66	0.43	0.47	0.71	0.54
ELA	4	CBT	0.73	0.53	0.56	0.68	0.63
ELA	4	PBT	0.64	0.52	0.55	0.70	0.65
ELA	5	CBT	0.52	0.53	0.60	0.76	0.52
ELA	6	CBT	0.63	0.61	0.62	0.73	0.58
ELA	7	CBT	0.70	0.54	0.59	0.67	0.72
ELA	8	CBT	0.68	0.47	0.55	0.72	0.66
Mathematics	3	CBT	0.77	0.61	0.64	0.76	0.61
Mathematics	3	PBT	0.75	0.59	0.62	0.76	0.56
Mathematics	4	CBT	0.76	0.60	0.65	0.82	0.46
Mathematics	4	PBT	0.78	0.58	0.64	0.82	0.35
Mathematics	5	CBT	0.64	0.55	0.65	0.77	0.54
Mathematics	6	CBT	0.71	0.60	0.69	0.78	0.61
Mathematics	7	CBT	0.45	0.61	0.66	0.78	0.64
Mathematics	8	CBT	0.72	0.47	0.59	0.78	0.62

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 0.84. 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

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

Table 9.7 Classification Consistency at Achievement Cut Points

			Classification Consistency			
Content Area	Grade	Mode	Unsatisfactory/ Approaching Basic	Approaching Basic/ Basic	Basic/ Mastery	Mastery/ Advanced
ELA	3	CBT	0.85	0.86	0.89	0.97
ELA	3	PBT	0.89	0.86	0.85	0.95
ELA	4	CBT	0.90	0.88	0.88	0.95
ELA	4	PBT	0.93	0.88	0.87	0.92
ELA	5	CBT	0.89	0.87	0.89	0.96
ELA	6	CBT	0.91	0.88	0.89	0.96
ELA	7	CBT	0.92	0.89	0.88	0.93
ELA	8	CBT	0.91	0.88	0.87	0.94
Mathematics	3	CBT	0.88	0.90	0.92	0.97
Mathematics	3	PBT	0.92	0.90	0.90	0.95
Mathematics	4	CBT	0.90	0.90	0.92	0.98
Mathematics	4	PBT	0.91	0.90	0.91	0.97
Mathematics	5	CBT	0.86	0.89	0.92	0.97
Mathematics	6	CBT	0.89	0.89	0.92	0.98
Mathematics	7	CBT	0.86	0.87	0.92	0.99
Mathematics	8	CBT	0.84	0.87	0.92	0.99

9.2.5 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 0.87 to 0.92 for ELA forms and from 0.91 to 0.94 for mathematics forms, indicating that items on the 2021 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

Grade	Mode	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
3	CBT	First Component	6.93	26.65	26.65
3	CBT	Second Component	1.09	4.19	30.83
3	CBT	Ratio (First/Second)	6.36		
3	PBT	First Component	6.47	24.88	24.88
3	PBT	Second Component	1.15	4.42	29.29
3	PBT	Ratio (First/Second)	5.63		
4	CBT	First Component	8.20	29.28	29.28
4	CBT	Second Component	1.20	4.27	33.55
4	CBT	Ratio (First/Second)	6.85		
4	PBT	First Component	7.54	26.94	26.94
4	PBT	Second Component	1.30	4.64	31.59
4	PBT	Ratio (First/Second)	5.80		
5	CBT	First Component	8.12	29.00	29.00
5	CBT	Second Component	1.31	4.69	33.69
5	CBT	Ratio (First/Second)	6.19		
6	CBT	First Component	8.90	27.82	27.82
6	CBT	Second Component	1.38	4.31	32.12
6	CBT	Ratio (First/Second)	6.46		
7	CBT	First Component	9.37	29.27	29.27
7	CBT	Second Component	1.23	3.86	33.12
7	CBT	Ratio (First/Second)	7.59		
8	CBT	First Component	8.45	26.42	26.42
8	CBT	Second Component	1.37	4.28	30.70
8	CBT	Ratio (First/Second)	6.18		

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	12.50	29.07	29.07
3	CBT	Second Component	1.52	3.53	32.60
3	CBT	Ratio (First/Second)	8.23		
3	PBT	First Component	12.67	29.46	29.46
3	PBT	Second Component	1.53	3.56	33.03
3	PBT	Ratio (First/Second)	8.27		
4	CBT	First Component	12.93	30.78	30.78
4	CBT	Second Component	1.65	3.92	34.70
4	CBT	Ratio (First/Second)	7.86		
4	PBT	First Component	12.70	30.24	30.24
4	PBT	Second Component	1.58	3.76	34.00
4	PBT	Ratio (First/Second)	8.05		
5	CBT	First Component	11.02	28.99	28.99
5	CBT	Second Component	1.37	3.61	32.60
5	CBT	Ratio (First/Second)	8.03		
6	CBT	First Component	12.40	31.00	31.00
6	CBT	Second Component	1.41	3.51	34.52
6	CBT	Ratio (First/Second)	8.82		
7	CBT	First Component	10.83	25.18	25.18
7	CBT	Second Component	1.79	4.15	29.34
7	CBT	Ratio (First/Second)	6.06		
8	CBT	First Component	10.32	27.89	27.89
8	CBT	Second Component	1.38	3.73	31.63
8	CBT	Ratio (First/Second)	7.47		

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 (CAPPM), 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:

$$CAPPM = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}, \tag{9.5}$$

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		0.87
	CBT	2	Writing	4	0.72	
	PBT	1	Reading	22		0.84
	PBT	2	Writing	4	0.68	
4	CBT	1	Reading	24		0.89
	CBT	2	Writing	4	0.79	
	PBT	1	Reading	24		0.87
	PBT	2	Writing	4	0.76	
5	CBT	1	Reading	24		0.85
	CBT	2	Writing	4	0.75	
6	CBT	1	Reading	28		0.81
	CBT	2	Writing	4	0.72	
7	CBT	1	Reading	28		0.85
	CBT	2	Writing	4	0.77	
8	CBT	1	Reading	28		0.85
	CBT	2	Writing	4	0.77	

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	.	1.00	0.87	0.96	0.86
	CBT	2	Reading Information Text	7	0.67	.	1.00	1.17	0.95
	CBT	3	Reading Vocabulary	4	0.57	0.57	.	0.91	0.81
	CBT	4	Written Expression	4	0.60	0.65	0.49	.	1.26
	CBT	5	Knowledge & Use of Language	2	0.61	0.60	0.49	0.75	.
	PBT	1	Reading Literary Text	11	.	1.00	0.86	0.94	0.80
	PBT	2	Reading Information Text	7	0.65	.	0.94	1.20	0.91
	PBT	3	Reading Vocabulary	4	0.55	0.51	.	0.87	0.73
	PBT	4	Written Expression	4	0.57	0.62	0.44	.	1.30
	PBT	5	Knowledge & Use of Language	2	0.55	0.54	0.43	0.70	.
4	CBT	1	Reading Literary Text	7	.	1.06	1.04	1.09	1.00
	CBT	2	Reading Information Text	9	0.70	.	1.02	1.00	0.94
	CBT	3	Reading Vocabulary	8	0.71	0.74	.	0.88	0.83
	CBT	4	Written Expression	4	0.74	0.73	0.66	.	1.17
	CBT	5	Knowledge & Use of Language	2	0.69	0.70	0.63	0.89	.
	PBT	1	Reading Literary Text	7	.	1.06	1.05	1.07	0.97
	PBT	2	Reading Information Text	9	0.67	.	1.03	0.98	0.93
	PBT	3	Reading Vocabulary	8	0.69	0.72	.	0.87	0.82
	PBT	4	Written Expression	4	0.70	0.68	0.63	.	1.22
	PBT	5	Knowledge & Use of Language	2	0.65	0.66	0.60	0.88	.
5	CBT	1	Reading Literary Text	8	.	1.01	1.00	0.93	0.88
	CBT	2	Reading Information Text	10	0.73	.	0.98	0.95	0.89
	CBT	3	Reading Vocabulary	6	0.66	0.67	.	0.83	0.79
	CBT	4	Written Expression	4	0.67	0.72	0.57	.	1.19
	CBT	5	Knowledge & Use of Language	2	0.65	0.68	0.55	0.92	.
6	CBT	1	Reading Literary Text	9	.	0.99	1.00	0.81	0.78
	CBT	2	Reading Information Text	13	0.73	.	1.02	0.92	0.88
	CBT	3	Reading Vocabulary	6	0.66	0.72	.	0.82	0.79
	CBT	4	Written Expression	4	0.59	0.72	0.57	.	1.16
	CBT	5	Knowledge & Use of Language	2	0.59	0.71	0.57	0.92	.
7	CBT	1	Reading Literary Text	10	.	0.99	0.95	0.86	0.82
	CBT	2	Reading Information Text	13	0.77	.	0.97	0.96	0.92
	CBT	3	Reading Vocabulary	5	0.64	0.65	.	0.82	0.79
	CBT	4	Written Expression	4	0.67	0.75	0.56	.	1.14
	CBT	5	Knowledge & Use of Language	2	0.66	0.74	0.55	0.93	.
8	CBT	1	Reading Literary Text	7	.	1.06	1.05	1.00	0.99
	CBT	2	Reading Information Text	13	0.73	.	1.01	0.89	0.90
	CBT	3	Reading Vocabulary	8	0.67	0.71	.	0.76	0.77
	CBT	4	Written Expression	4	0.74	0.72	0.57	.	1.10
	CBT	5	Knowledge & Use of Language	2	0.73	0.73	0.58	0.95	.

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	.	1.01	0.98	0.96
	CBT	2	Additional & Supporting Con	10	0.79	.	1.00	1.01
	CBT	3	Expressing Mathematical Rea	3	0.76	0.68	.	1.04
	CBT	4	Modeling & Application	3	0.77	0.71	0.72	.
	PBT	1	Major Content	27	.	1.00	0.99	0.98
	PBT	2	Additional & Supporting Con	10	0.79	.	1.00	1.03
	PBT	3	Expressing Mathematical Rea	3	0.75	0.66	.	1.03
	PBT	4	Modeling & Application	3	0.79	0.72	0.69	.
4	CBT	1	Major Content	28	.	0.97	0.95	0.90
	CBT	2	Additional & Supporting Con	8	0.78	.	0.96	0.93
	CBT	3	Expressing Mathematical Rea	3	0.78	0.70	.	0.98
	CBT	4	Modeling & Application	3	0.73	0.66	0.71	.
	PBT	1	Major Content	28	.	0.98	0.95	0.93
	PBT	2	Additional & Supporting Con	8	0.78	.	0.95	0.94
	PBT	3	Expressing Mathematical Rea	3	0.80	0.71	.	0.94
	PBT	4	Modeling & Application	3	0.74	0.66	0.69	.
5	CBT	1	Major Content	24	.	0.98	1.00	0.91
	CBT	2	Additional & Supporting Con	8	0.76	.	0.98	0.91
	CBT	3	Expressing Mathematical Rea	3	0.79	0.69	.	0.98
	CBT	4	Modeling & Application	3	0.72	0.63	0.70	.
6	CBT	1	Major Content	27	.	0.98	0.95	0.95
	CBT	2	Additional & Supporting Con	6	0.76	.	0.94	0.92
	CBT	3	Expressing Mathematical Rea	4	0.78	0.66	.	1.01
	CBT	4	Modeling & Application	3	0.74	0.62	0.72	.
7	CBT	1	Major Content	27	.	1.00	0.98	0.94
	CBT	2	Additional & Supporting Con	9	0.69	.	0.99	0.95
	CBT	3	Expressing Mathematical Rea	4	0.77	0.61	.	1.03
	CBT	4	Modeling & Application	3	0.72	0.57	0.71	.
8	CBT	1	Major Content	24	.	1.03	0.98	0.93
	CBT	2	Additional & Supporting Con	6	0.78	.	1.00	0.94
	CBT	3	Expressing Mathematical Rea	4	0.73	0.67	.	0.92
	CBT	4	Modeling & Application	3	0.70	0.63	0.60	.

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	.	0.94	0.89	0.99
	CBT	2	A2	3	0.65	.	0.89	0.98
	CBT	3	A3	7	0.66	0.57	.	0.92
	CBT	4	A4	8	0.75	0.64	0.64	.
	PBT	1	A1	9	.	0.94	0.90	0.99
	PBT	2	A2	3	0.66	.	0.92	0.97
	PBT	3	A3	7	0.67	0.60	.	0.92
	PBT	4	A4	8	0.76	0.65	0.66	.
4	CBT	1	A1	7	.	0.87	0.93	.
	CBT	2	A2	7	0.65	.	0.79	.
	CBT	3	A3	7	0.68	0.58	.	.
	PBT	1	A1	7	.	0.90	0.95	.
	PBT	2	A2	7	0.67	.	0.83	.
	PBT	3	A3	7	0.68	0.60	.	.
5	CBT	1	A1	5	.	0.97	1.08	0.98
	CBT	2	A2	6	0.59	.	1.02	0.97
	CBT	3	A3	6	0.63	0.65	.	0.97
	CBT	4	A4	6	0.60	0.66	0.64	.
6	CBT	1	A1	8	.	0.95	0.92	.
	CBT	2	A2	7	0.70	.	0.97	.
	CBT	3	A3	12	0.71	0.75	.	.
7	CBT	1	A1	8	.	1.01	1.06	.
	CBT	2	A2	15	0.75	.	1.06	.
	CBT	3	A3	4	0.70	0.73	.	.
8	CBT	1	A1	4	.	1.03	1.04	0.93
	CBT	2	A2	8	0.48	.	1.01	0.96
	CBT	3	A3	4	0.48	0.59	.	0.97
	CBT	4	A4	8	0.50	0.64	0.64	.

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 itemson 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	16.97	8.85	3.49	0.84
	CBT	Writing	4	24	4.14	4.01	1.80	0.80
	PBT	Reading	23	47	19.27	9.28	3.73	0.84
	PBT	Writing	4	24	6.16	4.46	2.16	0.77
4	CBT	Reading	26	56	20.69	11.00	3.95	0.87
	CBT	Writing	4	30	6.73	5.91	1.88	0.90
	PBT	Reading	26	56	23.13	11.29	4.23	0.86
	PBT	Writing	4	30	9.18	6.32	2.17	0.88
5	CBT	Reading	26	56	22.19	11.05	3.99	0.87
	CBT	Writing	4	30	5.90	5.74	1.77	0.90
6	CBT	Reading	29	60	25.23	11.86	4.03	0.88
	CBT	Writing	4	30	8.40	6.94	2.02	0.91
7	CBT	Reading	29	60	27.81	12.40	4.15	0.89
	CBT	Writing	4	30	10.58	8.31	2.35	0.92
8	CBT	Reading	30	64	27.21	12.26	4.47	0.87
	CBT	Writing	4	30	10.02	7.21	1.70	0.94

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	7.71	4.66	2.29	0.76
	CBT	Reading Information Text	8	17	5.73	3.35	2.14	0.59
	CBT	Reading Vocabulary	4	8	3.54	2.18	1.45	0.56
	CBT	Written Expression	2	18	3.08	3.06	2.11	0.52
	CBT	Knowledge & Use of Language	2	6	1.06	1.18	0.67	0.67
	PBT	Reading Literary Text	11	22	9.01	5.03	2.43	0.77
	PBT	Reading Information Text	8	17	5.96	3.54	2.35	0.56
	PBT	Reading Vocabulary	4	8	4.29	2.20	1.49	0.54
	PBT	Written Expression	2	18	4.57	3.48	2.53	0.47
	PBT	Knowledge & Use of Language	2	6	1.58	1.26	0.77	0.62
4	CBT	Reading Literary Text	8	18	6.12	3.53	2.19	0.61
	CBT	Reading Information Text	10	22	7.09	4.48	2.42	0.71
	CBT	Reading Vocabulary	8	16	7.49	4.20	2.12	0.74
	CBT	Written Expression	2	24	5.07	4.51	2.27	0.75
	CBT	Knowledge & Use of Language	2	6	1.65	1.52	0.71	0.78
	PBT	Reading Literary Text	8	18	6.45	3.79	2.39	0.60
	PBT	Reading Information Text	10	22	8.44	4.66	2.64	0.68
	PBT	Reading Vocabulary	8	16	8.24	4.18	2.18	0.73
	PBT	Written Expression	2	24	6.86	4.85	2.58	0.72
	PBT	Knowledge & Use of Language	2	6	2.32	1.62	0.84	0.73
5	CBT	Reading Literary Text	9	20	7.40	4.15	2.31	0.69
	CBT	Reading Information Text	11	24	8.78	5.16	2.56	0.75
	CBT	Reading Vocabulary	6	12	6.01	3.05	1.86	0.63
	CBT	Written Expression	2	24	4.33	4.31	2.14	0.75
	CBT	Knowledge & Use of Language	2	6	1.57	1.53	0.71	0.78
6	CBT	Reading Literary Text	9	18	6.87	3.77	2.09	0.69
	CBT	Reading Information Text	14	30	12.78	6.25	2.86	0.79
	CBT	Reading Vocabulary	6	12	5.58	3.10	1.88	0.63
	CBT	Written Expression	2	24	6.24	5.32	2.53	0.77
	CBT	Knowledge & Use of Language	2	6	2.16	1.72	0.74	0.82
7	CBT	Reading Literary Text	10	20	8.71	4.69	2.24	0.77
	CBT	Reading Information Text	14	30	12.89	6.38	3.01	0.78
	CBT	Reading Vocabulary	5	10	6.21	2.67	1.73	0.58
	CBT	Written Expression	2	24	8.06	6.50	3.02	0.78
	CBT	Knowledge & Use of Language	2	6	2.52	1.93	0.77	0.84
8	CBT	Reading Literary Text	8	18	7.43	3.87	2.37	0.62
	CBT	Reading Information Text	14	30	12.16	6.18	3.04	0.76
	CBT	Reading Vocabulary	8	16	7.61	3.57	2.11	0.65
	CBT	Written Expression	2	24	7.51	5.50	2.05	0.86
	CBT	Knowledge & Use of Language	2	6	2.51	1.78	0.65	0.86

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	14.33	7.42	2.32	0.90
	CBT	Additional & Supporting Content	10	10	4.94	2.31	1.30	0.68
	CBT	Expressing Mathematical Reasoning	3	10	2.39	2.21	1.27	0.67
	CBT	Modeling & Application	3	12	2.60	2.76	1.49	0.71
	PBT	Major Content	27	30	16.21	7.58	2.32	0.91
	PBT	Additional & Supporting Content	10	10	5.45	2.35	1.31	0.69
	PBT	Expressing Mathematical Reasoning	3	10	3.29	2.32	1.41	0.63
	PBT	Modeling & Application	3	12	3.75	3.30	1.77	0.71
4	CBT	Major Content	28	29	14.79	7.42	2.21	0.91
	CBT	Additional & Supporting Content	8	10	4.32	2.50	1.33	0.72
	CBT	Expressing Mathematical Reasoning	3	10	2.10	2.13	1.08	0.74
	CBT	Modeling & Application	3	12	2.19	2.78	1.48	0.72
	PBT	Major Content	28	29	15.35	7.27	2.21	0.91
	PBT	Additional & Supporting Content	8	10	4.53	2.51	1.35	0.71
	PBT	Expressing Mathematical Reasoning	3	10	2.89	2.53	1.19	0.78
	PBT	Modeling & Application	3	12	2.83	2.96	1.63	0.70
5	CBT	Major Content	24	26	11.81	6.33	2.17	0.88
	CBT	Additional & Supporting Content	8	8	4.03	2.20	1.22	0.69
	CBT	Expressing Mathematical Reasoning	3	10	3.03	2.58	1.38	0.71
	CBT	Modeling & Application	3	12	2.49	2.44	1.33	0.70
6	CBT	Major Content	27	30	13.54	7.45	2.31	0.90
	CBT	Additional & Supporting Content	6	7	2.54	1.94	1.12	0.67
	CBT	Expressing Mathematical Reasoning	4	14	3.22	2.98	1.48	0.75
	CBT	Modeling & Application	3	12	2.30	2.55	1.44	0.68
7	CBT	Major Content	27	30	12.92	6.84	2.36	0.88
	CBT	Additional & Supporting Content	9	10	4.10	2.11	1.44	0.53
	CBT	Expressing Mathematical Reasoning	4	14	2.59	2.93	1.58	0.71
	CBT	Modeling & Application	3	12	1.70	2.92	1.67	0.67
8	CBT	Major Content	24	27	9.45	5.66	2.14	0.86
	CBT	Additional & Supporting Content	6	7	2.48	1.84	1.04	0.68
	CBT	Expressing Mathematical Reasoning	4	14	2.18	2.47	1.45	0.65
	CBT	Modeling & Application	3	12	2.47	2.37	1.38	0.66

Table 9.17 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Mathematics Reporting Subcategories

Mean, Standard Deviation, and SEM: Mathematics								
Grade	Mode	Major Content Subcategory	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
3	CBT	A1	9	9	4.99	2.74	1.24	0.80
	CBT	A2	3	4	1.36	1.27	0.80	0.60
	CBT	A3	7	8	3.52	2.14	1.20	0.68
	CBT	A4	8	9	4.46	2.39	1.28	0.71
	PBT	A1	9	9	5.56	2.67	1.21	0.79
	PBT	A2	3	4	1.59	1.33	0.83	0.61
	PBT	A3	7	8	4.05	2.21	1.21	0.70
	PBT	A4	8	9	5.01	2.47	1.28	0.73
4	CBT	A1	7	8	4.46	2.30	1.18	0.74
	CBT	A2	7	7	2.83	2.14	1.02	0.77
	CBT	A3	7	7	3.73	1.99	1.07	0.71
	PBT	A1	7	8	4.65	2.26	1.18	0.73
	PBT	A2	7	7	2.98	2.12	1.04	0.76
	PBT	A3	7	7	3.87	1.92	1.06	0.69
5	CBT	A1	5	5	2.66	1.44	0.97	0.55
	CBT	A2	6	6	2.87	1.80	1.04	0.66
	CBT	A3	6	7	2.80	1.89	1.17	0.62
	CBT	A4	6	7	3.13	1.95	1.08	0.69
6	CBT	A1	8	9	5.10	2.42	1.26	0.73
	CBT	A2	7	8	3.20	2.39	1.23	0.73
	CBT	A3	12	13	5.25	3.43	1.47	0.82
7	CBT	A1	8	9	3.90	2.45	1.34	0.70
	CBT	A2	15	16	7.27	3.57	1.71	0.77
	CBT	A3	4	5	1.76	1.48	0.92	0.62
8	CBT	A1	4	4	1.37	1.03	0.81	0.38
	CBT	A2	8	8	2.70	1.88	1.20	0.59
	CBT	A3	4	5	2.13	1.33	0.87	0.57
	CBT	A4	8	10	3.24	2.56	1.27	0.75

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 2021. These correlations are based on the census data, and the results are shown in Table 9.18. The correlation coefficients ranged from 0.71 (between mathematics and social studies in grades 3 and 5) to 0.84 (between ELA and social studies in grade). 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	0.75	0.76	0.78	0.71	0.76	0.77
4	0.76	0.79	0.78	0.75	0.77	0.80
5	0.76	0.79	0.81	0.71	0.78	0.77
6	0.80	0.83	0.79	0.79	0.78	0.81
7	0.79	0.82	0.79	0.76	0.80	0.80
8	0.74	0.84	0.78	0.74	0.74	0.83

9.6 Regression of LEAP 2025 from 2019 to 2021

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.

In prior years, this study required the collection of data from adjacent grades for each content area. However, since LEAP 2025 was not administered in 2020, "adjacent grades" for this administration's study had to be defined differently. For this purpose, matched longitudinal LEAP 2025 test data from spring 2019 and spring 2021 were used. For example, grade 3 students were matched with grade 5 students, and only matched students were used to estimate correlation and perform linear regression from 2019 to 2021.

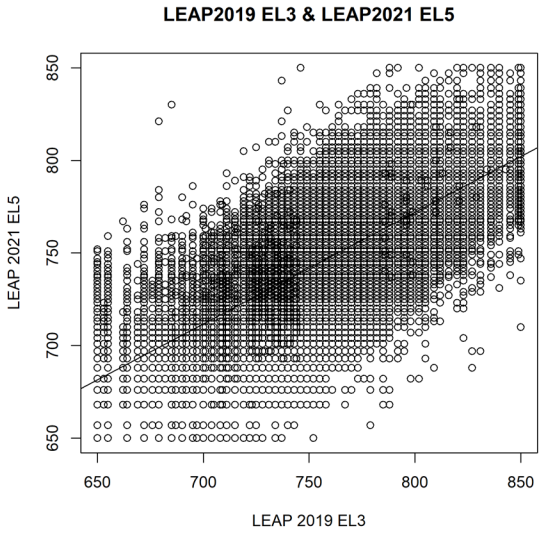
Table 9.19 summarizes the correlation and regression results for 2019 and 2021 LEAP 2025. For ELA, the correlation ranged from 0.75 to 0.81, and for mathematics, the correlation ranged from 0.75 to 0.80. Correlations for both content areas can be considered moderate, which can often be found in state assessments. R^2 indicates how much of the 2019 performance can explain the 2021 performance. For example, 0.56 for ELA 2019 grade 3 and 2021 grade 5 means that 2019's grade 3 performance can explain (predict) about 56% of 2021's grade 5 performance. This R^2 value is generally the power of 2 for the matching correlation. The R^2 values for ELA range from 0.56 to 0.66, and those for mathematics range from 0.57 to 0.65.

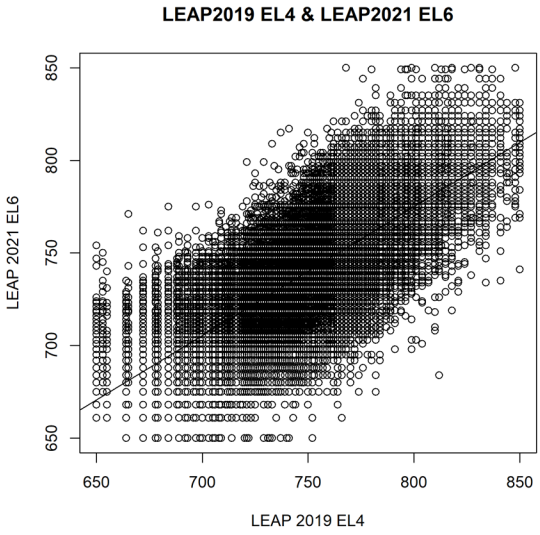
Table 9.19 Correlation and Regression Summary for 2019 and 2021 LEAP 2025

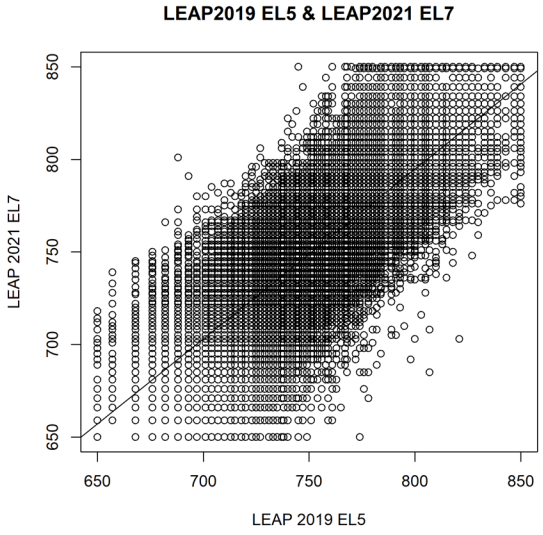
Content	2019 Grade	2021 Grade	N	Correlation	R^2
ELA	3	5	≥45,590	0.75	0.56
	4	6	≥47,280	0.76	0.58
	5	7	≥47,720	0.79	0.62
	6	8	≥47,480	0.81	0.66
Mathematics	3	5	≥45,470	0.75	0.57
	4	6	≥47,170	0.79	0.63
	5	7	≥47,540	0.8	0.65
	6	8	≥41,860	0.78	0.61

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 2019 to 2021. 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 2019 and 2021 adjacent grades for both ELA and mathematics.

Figure 9.3 Regression Line and Scatter Plots:







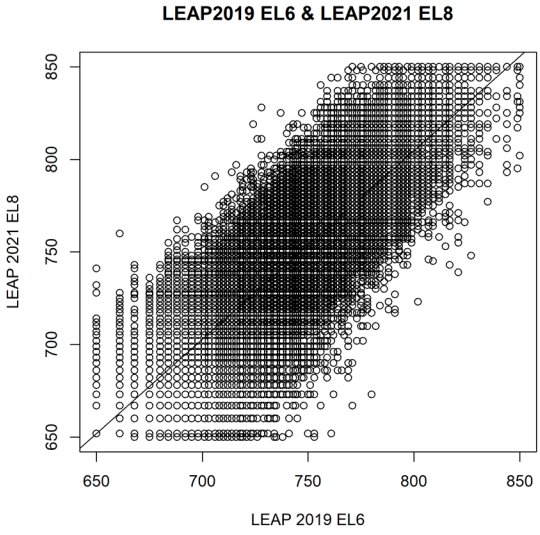
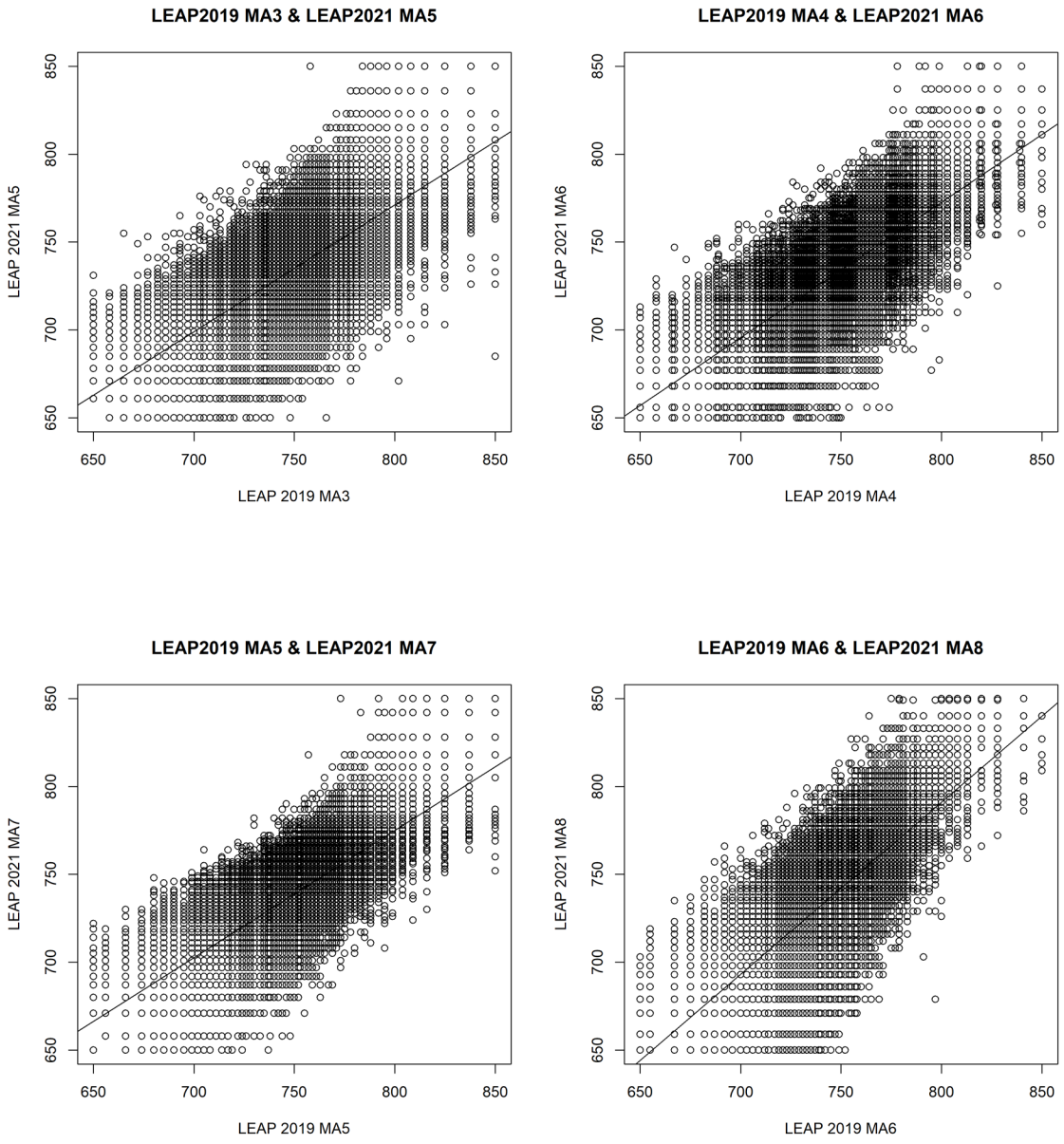


Figure 9.4 Regression Line and Scatter Plots: Mathematics



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 kth 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 (χ^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 α is assumed to be constant across the *K* matched levels. The χ^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 (Δ_{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):

$$SMD = p_{Fk}(\sum_k m_{Fk} - \sum_k m_{Rk}),$$

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 summarizes 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.

DIF statistics are produced and examined for all newly field-tested items and for all items being administered for the first time operationally in Louisiana. In the spring 2021 administration, items were field tested in grades 3 and 6 ELA.

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	6	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	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	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
3	PBT	6	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	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	0	0
			Economic Status	Economically Disadvantaged	0	0	0	0
			Section 504 Status	Section 504	0	0	0	0
6	CBT	6	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	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	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 differencesand 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 C 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 if analyses was performed in 2019, when most of the 2021 items were originally administered. 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 10.2 summarizes how many items overall exhibited moderate or large DIF in mathematics.

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	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 10.3–10.10 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 0.75 to 0.93. For mathematics the reliability coefficients ranged from 0.82 to 0.94. 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. 3 Grade 3 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥12,090	0.88	4.21	≥12,070	0.93	3.47
Gender						
Female	≥6,160	0.88	4.16	≥6,150	0.94	3.46
Male	≥5,920	0.87	4.26	≥5,910	0.93	3.49
Ethnicity						
Hispanic/Latino	≥1,880	0.88	4.13	≥1,860	0.93	3.45
American Indian or Alaska Native	≥70	0.88	4.48	≥70	0.93	3.54
Asian	≥250	0.89	4.48	≥250	0.93	3.75
Black or African American	≥5,320	0.85	4.04	≥5,320	0.92	3.25
Native Hawaiian or Other Pacific	<10	NR	NR	<10	NR	NR
White	≥4,170	0.87	4.38	≥4,180	0.93	3.63
Two or More Races	≥360	0.88	4.29	≥360	0.94	3.55
Education Classification						
Regular	≥10,670	0.88	4.24	≥10,650	0.93	3.50
Special	≥1,410	0.84	3.92	≥1,410	0.92	3.25
English Learner Status						
Not English Learner	≥10,990	0.88	4.24	≥10,990	0.93	3.49
English Learner	≥1,090	0.79	3.86	≥1,080	0.91	3.24
Economic Status						
Economically Disadvantaged	≥9,650	0.86	4.13	≥9,640	0.93	3.38
Not Economically Disadvantaged	≥2,430	0.88	4.52	≥2,430	0.93	3.76
Section 504 Status						
Not Section 504	≥11,570	0.88	4.22	≥11,550	0.93	3.48
Section 504	≥510	0.85	4.03	≥510	0.92	3.39

Table 10.4 Grade 3 Paper-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥37,540	0.87	4.58	≥37,520	0.93	3.72
Gender						
Female	≥19,150	0.87	4.52	≥19,140	0.93	3.72
Male	≥18,360	0.87	4.64	≥18,330	0.93	3.73
Ethnicity						
Hispanic/Latino	≥2,950	0.88	4.51	≥2,940	0.93	3.68
American Indian or Alaska Native	≥210	0.85	4.74	≥210	0.92	3.74
Asian	≥560	0.88	4.72	≥560	0.93	3.70
Black or African American	≥15,660	0.84	4.41	≥15,640	0.92	3.49
Native Hawaiian or Other Pacific	≥30	0.83	4.62	≥30	0.93	3.76
White	≥16,780	0.85	4.71	≥16,770	0.92	3.80
Two or More Races	≥1,280	0.85	4.62	≥1,280	0.92	3.71
Education Classification						
Regular	≥32,770	0.87	4.61	≥32,750	0.93	3.74
Special	≥4,760	0.85	4.35	≥4,760	0.93	3.53
English Learner Status						
Not English Learner	≥36,160	0.87	4.60	≥36,150	0.93	3.73
English Learner	≥1,380	0.78	4.15	≥1,360	0.91	3.45
Economic Status						
Economically Disadvantaged	≥27,130	0.85	4.48	≥27,080	0.93	3.61
Not Economically Disadvantaged	≥10,400	0.85	4.81	≥10,430	0.91	3.82
Section 504 Status						
Not Section 504	≥34,720	0.87	4.59	≥34,700	0.93	3.73
Section 504	≥2,810	0.85	4.43	≥2,810	0.92	3.62

Table 10.5 Grade 4 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥16,480	0.90	4.97	≥16,430	0.94	3.35
Gender						
Female	≥8,370	0.91	4.83	≥8,350	0.94	3.35
Male	≥8,100	0.90	5.08	≥8,080	0.94	3.36
Ethnicity						
Hispanic/Latino	≥2,230	0.90	4.83	≥2,210	0.93	3.31
American Indian or Alaska Native	≥70	0.88	5.06	≥70	0.93	3.31
Asian	≥290	0.91	5.29	≥290	0.93	3.59
Black or African American	≥6,680	0.88	4.83	≥6,670	0.92	3.08
Native Hawaiian or Other Pacific	<10	NR	NR	<10	NR	NR
White	≥6,650	0.89	5.11	≥6,640	0.93	3.53
Two or More Races	≥520	0.90	4.98	≥520	0.94	3.35
Education Classification						
Regular	≥14,360	0.90	5.02	≥14,310	0.94	3.38
Special	≥2,120	0.89	4.36	≥2,120	0.93	3.04
English Learner Status						
Not English Learner	≥15,400	0.90	5.00	≥15,370	0.94	3.37
English Learner	≥1,080	0.80	4.41	≥1,060	0.91	2.97
Economic Status						
Economically Disadvantaged	≥12,350	0.89	4.88	≥12,320	0.93	3.23
Not Economically Disadvantaged	≥4,120	0.89	5.24	≥4,110	0.93	3.65
Section 504 Status						
Not Section 504	≥15,210	0.91	4.98	≥15,160	0.94	3.37
Section 504	≥1,260	0.88	4.72	≥1,260	0.93	3.16

Table 10.6 Grade 4 Paper-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥33,070	0.89	5.39	≥33,050	0.94	3.53
Gender						
Female	≥17,050	0.90	5.26	≥17,040	0.94	3.52
Male	≥16,010	0.89	5.48	≥16,000	0.93	3.54
Ethnicity						
Hispanic/Latino	≥2,340	0.90	5.32	≥2,330	0.93	3.50
American Indian or Alaska Native	≥190	0.87	5.53	≥190	0.93	3.59
Asian	≥470	0.91	5.37	≥470	0.93	3.73
Black or African American	≥14,430	0.87	5.28	≥14,430	0.92	3.24
Native Hawaiian or Other Pacific	≥30	0.86	5.21	≥30	0.92	3.73
White	≥14,450	0.88	5.47	≥14,450	0.92	3.67
Two or More Races	≥1,110	0.88	5.46	≥1,110	0.93	3.61
Education Classification						
Regular	≥29,060	0.89	5.40	≥29,040	0.93	3.55
Special	≥4,010	0.88	4.98	≥4,010	0.93	3.21
English Learner Status						
Not English Learner	≥32,020	0.89	5.40	≥32,020	0.94	3.54
English Learner	≥1,040	0.82	4.99	≥1,030	0.92	3.15
Economic Status						
Economically Disadvantaged	≥23,950	0.88	5.33	≥23,940	0.93	3.40
Not Economically Disadvantaged	≥9,110	0.88	5.52	≥9,110	0.91	3.71
Section 504 Status						
Not Section 504	≥29,900	0.90	5.40	≥29,890	0.94	3.54
Section 504	≥3,170	0.87	5.22	≥3,160	0.93	3.39

Table 10.7 Grade 5 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥49,780	0.90	4.97	≥49,700	0.93	3.33
Gender						
Female	≥25,610	0.90	4.83	≥25,570	0.93	3.28
Male	≥24,170	0.90	5.08	≥24,130	0.92	3.37
Ethnicity						
Hispanic/Latino	≥4,760	0.90	4.88	≥4,700	0.92	3.29
American Indian or Alaska Native	≥290	0.86	5.04	≥290	0.91	3.37
Asian	≥800	0.92	5.17	≥800	0.93	3.56
Black or African American	≥21,040	0.87	4.79	≥21,020	0.90	3.09
Native Hawaiian or Other Pacific	≥20	0.92	5.27	≥20	0.93	3.48
White	≥21,220	0.89	5.12	≥21,220	0.92	3.47
Two or More Races	≥1,600	0.90	5.02	≥1,600	0.92	3.37
Education Classification						
Regular	≥43,820	0.90	5.03	≥43,740	0.93	3.36
Special	≥5,960	0.87	4.30	≥5,950	0.89	2.90
English Learner Status						
Not English Learner	≥47,580	0.90	5.00	≥47,550	0.93	3.34
English Learner	≥2,200	0.83	4.26	≥2,140	0.89	2.93
Economic Status						
Economically Disadvantaged	≥36,690	0.89	4.86	≥36,600	0.91	3.21
Not Economically Disadvantaged	≥13,090	0.89	5.23	≥13,090	0.91	3.58
Section 504 Status						
Not Section 504	≥44,680	0.90	4.99	≥44,600	0.93	3.35
Section 504	≥5,100	0.87	4.68	≥5,100	0.91	3.10

Table 10.8 Grade 6 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥51,430	0.91	5.20	≥51,340	0.94	3.46
Gender						
Female	≥26,130	0.91	5.08	≥26,080	0.94	3.41
Male	≥25,300	0.91	5.26	≥25,250	0.93	3.50
Ethnicity						
Hispanic/Latino	≥4,600	0.92	5.12	≥4,520	0.93	3.39
American Indian or Alaska Native	≥300	0.90	5.26	≥300	0.92	3.52
Asian	≥760	0.92	5.33	≥760	0.93	3.93
Black or African American	≥22,200	0.89	5.03	≥22,190	0.91	3.12
Native Hawaiian or Other Pacific	≥40	0.90	5.45	≥40	0.93	3.72
White	≥21,830	0.91	5.31	≥21,830	0.93	3.69
Two or More Races	≥1,670	0.90	5.32	≥1,670	0.93	3.57
Education Classification						
Regular	≥45,600	0.91	5.24	≥45,520	0.93	3.51
Special	≥5,820	0.88	4.53	≥5,820	0.91	2.80
English Learner Status						
Not English Learner	≥49,460	0.91	5.22	≥49,450	0.93	3.48
English Learner	≥1,970	0.86	4.68	≥1,890	0.91	2.91
Economic Status						
Economically Disadvantaged	≥37,890	0.90	5.12	≥37,820	0.93	3.30
Not Economically Disadvantaged	≥13,530	0.90	5.37	≥13,520	0.92	3.82
Section 504 Status						
Not Section 504	≥45,980	0.91	5.22	≥45,890	0.94	3.49
Section 504	≥5,450	0.89	4.97	≥5,450	0.92	3.15

Table 10.9 Grade 7 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥52,180	0.92	5.60	≥52,080	0.92	3.80
Gender						
Female	≥26,590	0.92	5.42	≥26,530	0.92	3.75
Male	≥25,590	0.91	5.69	≥25,540	0.91	3.83
Ethnicity						
Hispanic/Latino	≥4,640	0.93	5.48	≥4,550	0.91	3.73
American Indian or Alaska Native	≥300	0.91	5.55	≥300	0.91	3.85
Asian	≥830	0.92	5.66	≥830	0.92	4.54
Black or African American	≥22,350	0.90	5.42	≥22,340	0.89	3.26
Native Hawaiian or Other Pacific	≥40	0.92	5.67	≥40	0.93	4.21
White	≥22,400	0.91	5.71	≥22,400	0.91	4.12
Two or More Races	≥1,590	0.91	5.68	≥1,590	0.91	3.86
Education Classification						
Regular	≥46,600	0.91	5.65	≥46,500	0.91	3.87
Special	≥5,580	0.89	4.79	≥5,580	0.89	2.90
English Learner Status						
Not English Learner	≥50,270	0.92	5.62	≥50,260	0.92	3.82
English Learner	≥1,910	0.89	4.90	≥1,820	0.88	3.03
Economic Status						
Economically Disadvantaged	≥37,760	0.91	5.50	≥37,660	0.90	3.51
Not Economically Disadvantaged	≥14,420	0.90	5.81	≥14,410	0.91	4.33
Section 504 Status						
Not Section 504	≥46,660	0.92	5.63	≥46,570	0.92	3.84
Section 504	≥5,520	0.90	5.33	≥5,510	0.90	3.36

Table 10.10 Grade 8 Computer-Based Test Administration Reliability and SEM by Subgroup

	ELA			Mathematics		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥51,680	0.90	5.71	≥45,840	0.91	3.23
Gender						
Female	≥26,210	0.90	5.58	≥23,490	0.92	3.16
Male	≥25,470	0.90	5.73	≥22,350	0.91	3.30
Ethnicity						
Hispanic/Latino	≥4,200	0.91	5.65	≥3,640	0.91	3.13
American Indian or Alaska Native	≥310	0.90	5.86	≥290	0.91	3.32
Asian	≥800	0.91	5.81	≥540	0.94	3.71
Black or African American	≥22,030	0.88	5.60	≥20,730	0.88	2.96
Native Hawaiian or Other Pacific	≥40	0.89	5.82	≥30	0.93	3.47
White	≥22,740	0.90	5.79	≥19,220	0.91	3.44
Two or More Races	≥1,520	0.90	5.75	≥1,350	0.92	3.32
Education Classification						
Regular	≥46,510	0.90	5.74	≥40,790	0.91	3.28
Special	≥5,170	0.85	4.97	≥5,040	0.86	2.58
English Learner Status						
Not English Learner	≥49,820	0.90	5.73	≥44,110	0.91	3.25
English Learner	≥1,850	0.84	5.08	≥1,720	0.87	2.74
Economic Status						
Economically Disadvantaged	≥36,790	0.89	5.65	≥33,940	0.90	3.09
Not Economically Disadvantaged	≥14,890	0.89	5.86	≥11,900	0.91	3.56
Section 504 Status						
Not Section 504	≥46,410	0.90	5.73	≥40,790	0.91	3.26
Section 504	≥5,270	0.88	5.48	≥5,040	0.90	2.99

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:

$$d = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}},$$

where $\overline{x_a}$ is the mean score of group A, $\overline{x_b}$ 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 10.11–10.18. Results are NR for subgroups with fewer than 10 students. If the effect size is negative, that means the group performs at a higher level than the group to which it's being compared. A positive effect size indicats the group performs at a lower level than the group to which it is being compared. For example, in Table 10.11 in regards to the ELA test, the effect size for the group female is -0.10 indicating that althouth there is less than a small difference in performance, females are scoring higher than males. On the ELA test in most grades, there are small differences in mean test scores at grades, 6, 7, and 8 between females and males where females outperform males. 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, and non-EL and EL students in EL status.

Table 10.11 Impact Analysis, Grade 3 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥12,090	724.61	42.09		≥12,440	725.01	33.17	
Gender								
Male	≥6,160	722.64	42.52		≥6,160	725.36	33.95	
Female	≥5,920	726.66	41.55	-0.10	≥5,920	724.65	32.34	0.02
Ethnicity								
White	≥4,170	739.56	41.28		≥4,180	738.30	32.32	
Hispanic/Latino	≥1,880	718.66	43.00	0.50	≥1,880	723.75	32.39	0.45
American Indian or Alaska Native	≥70	734.03	44.06	0.13	≥70	727.96	32.85	0.32
Asian	≥250	752.30	44.47	-0.31	≥250	754.25	34.88	-0.49
Black or African American	≥5,320	712.77	37.54	0.68	≥5,320	713.08	28.68	0.83
Native Hawaiian or Other Pacific	<10	NR	NR	NR	<10	NR	NR	NR
Two or More Races	≥360	734.79	43.02	0.12	≥360	731.92	35.00	0.20
Education Classification								
Regular	≥10,670	727.30	42.02		≥10,670	727.00	33.14	
Special	≥1,410	704.30	36.81	0.55	≥1,410	710.06	29.41	0.52
Economic Status								
Not Economically Disadvantaged	≥2,430	750.14	43.01		≥2,430	746.85	33.85	
Economically Disadvantaged	≥9,650	718.17	39.33	0.80	≥9,650	719.51	30.64	0.87
English Learner Status								
Not English Learner	≥10,990	727.22	42.04		≥10,990	726.22	33.40	
English Learner	≥1,090	698.45	32.62	0.70	≥1,090	712.92	28.07	0.40
Migrant Status								
Not Migrant	≥12,070	724.63	42.10		≥12,070	725.02	33.17	
Migrant	≥10	710.12	35.43	0.34	≥10	723.47	35.31	0.05
Section 504 Status								
Not Section 504	≥11,570	725.02	42.23		≥11,570	725.28	33.25	
Section 504	≥510	715.43	37.79	0.23	≥510	718.96	30.72	0.19

Table 10.12 Impact Analysis, Grade 3 Paper-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥37,540	742.07	43.02		≥37,190	735.66	34.57	
Gender								
Male	≥19,150	740.04	43.26		≥18,940	735.92	35.11	
Female	≥18,360	744.20	42.67	-0.10	≥18,180	735.40	33.99	0.02
Ethnicity								
White	≥16,780	757.16	40.41		≥16,660	749.11	31.74	
Hispanic/Latino	≥2,950	732.66	44.70	0.60	≥2,900	733.66	32.67	0.48
American Indian or Alaska Native	≥210	748.20	40.51	0.22	≥210	741.30	33.08	0.25
Asian	≥560	768.24	44.67	-0.27	≥550	763.16	34.39	-0.44
Black or African American	≥15,660	726.15	39.16	0.78	≥15,460	720.20	31.22	0.92
Native Hawaiian or Other Pacific	≥30	746.46	36.17	0.26	≥30	746.08	34.79	0.10
Two or More Races	≥1,280	748.24	40.18	0.22	≥1,270	738.92	32.06	0.32
Education Classification								
Regular	≥32,770	745.02	42.69		≥32,410	737.91	34.30	
Special	≥4,760	721.81	39.70	0.55	≥4,710	720.26	32.45	0.52
Economic Status								
Not Economically Disadvantaged	≥10,400	766.23	40.22		≥10,270	756.90	30.88	
Economically Disadvantaged	≥27,130	732.80	40.38	0.83	≥26,850	727.53	32.39	0.92
English Learner Status								
Not English Learner	≥36,160	743.48	42.69		≥35,770	736.27	34.61	
English Learner	≥1,380	705.26	34.42	0.90	≥1,350	719.55	29.40	0.49
Migrant Status								
Not Migrant	≥37,460	742.11	43.00		≥37,040	735.68	34.57	
Migrant	≥80	723.87	49.23	0.42	≥80	728.02	34.06	0.22
Section 504 Status								
Not Section 504	≥34,720	743.10	43.15		≥34,330	736.43	34.71	
Section 504	≥2,810	729.36	39.24	0.32	≥2,790	726.23	31.32	0.30

Table 10.13 Impact Analysis, Grade 4 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥16,480	733.71	36.24		≥16,720	728.41	33.51	
Gender								
Male	≥8,370	731.09	36.16		≥8,360	729.30	33.99	
Female	≥8,100	736.41	36.13	-0.15	≥8,090	727.50	32.99	0.05
Ethnicity								
White	≥6,650	746.46	34.21		≥6,640	742.14	31.78	
Hispanic/Latino	≥2,230	726.87	36.01	0.57	≥2,220	726.40	31.82	0.49
American Indian or Alaska Native	≥70	738.92	31.53	0.22	≥70	736.11	28.35	0.19
Asian	≥290	760.56	40.08	-0.41	≥290	758.49	31.79	-0.51
Black or African American	≥6,680	721.63	33.16	0.74	≥6,670	713.78	29.01	0.93
Native Hawaiian or Other Pacific	<10	NR	NR	NR	<10	NR	NR	NR
Two or More Races	≥520	739.15	35.63	0.21	≥520	731.33	32.86	0.34
Education Classification								
Regular	≥14,360	737.41	35.27		≥14,330	731.02	33.30	
Special	≥2,120	708.66	32.65	0.82	≥2,120	710.77	29.35	0.62
Economic Status								
Not Economically Disadvantaged	≥4,120	754.58	34.55		≥4,110	749.01	31.67	
Economically Disadvantaged	≥12,350	726.75	34.06	0.81	≥12,340	721.55	31.22	0.88
English Learner Status								
Not English Learner	≥15,400	735.85	35.88		≥15,370	729.59	33.57	
English Learner	≥1,080	703.21	26.38	0.92	≥1,080	711.57	27.67	0.54
Migrant Status								
Not Migrant	≥16,450	733.72	36.25		≥16,430	728.43	33.51	
Migrant	≥20	726.83	30.00	0.19	≥20	717.26	32.91	0.33
Section 504 Status								
Not Section 504	≥15,210	734.62	36.47		≥15,190	729.18	33.67	
Section 504	≥1,260	722.75	31.44	0.33	≥1,260	719.23	30.07	0.30

Table 10.14 Impact Analysis, Grade 4 Paper-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥33,070	745.24	36.09		≥32,830	733.06	33.77	
Gender								
Male	≥17,050	742.44	36.09		≥16,890	733.38	34.45	
Female	≥16,010	748.24	35.87	-0.16	≥15,880	732.72	33.02	0.02
Ethnicity								
White	≥14,450	759.01	33.22		≥14,380	747.48	30.52	
Hispanic/Latino	≥2,340	738.22	37.96	0.61	≥2,310	730.87	33.42	0.54
American Indian or Alaska Native	≥190	749.01	32.05	0.30	≥190	735.94	31.30	0.38
Asian	≥470	767.88	39.09	-0.27	≥460	759.68	33.68	-0.40
Black or African American	≥14,430	731.17	32.76	0.84	≥14,280	717.48	29.73	1.00
Native Hawaiian or Other Pacific	≥30	762.03	28.66	-0.09	≥30	745.00	32.79	0.08
Two or More Races	≥1,110	753.38	33.34	0.17	≥1,100	739.11	32.54	0.27
Education Classification								
Regular	≥29,060	748.41	35.33		≥28,800	735.67	33.35	
Special	≥4,010	722.20	33.07	0.75	≥3,970	714.08	30.56	0.65
Economic Status								
Not Economically Disadvantaged	≥9,110	766.29	33.59		≥9,030	754.48	29.47	
Economically Disadvantaged	≥23,950	737.22	33.71	0.86	≥23,740	724.90	31.66	0.95
English Learner Status								
Not English Learner	≥32,020	746.33	35.77		≥31,750	733.70	33.72	
English Learner	≥1,040	711.93	29.48	0.97	≥1,020	713.25	28.91	0.61
Migrant Status								
Not Migrant	≥33,010	745.28	36.09		≥32,720	733.09	33.77	
Migrant	≥60	723.25	28.31	0.61	≥50	716.41	27.19	0.49
Section 504 Status								
Not Section 504	≥29,900	746.36	36.34		≥29,620	733.97	33.93	
Section 504	≥3,170	734.61	31.80	0.33	≥3,150	724.49	30.86	0.28

Table 10.15 Impact Analysis, Grade 5 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥49,780	739.53	33.27		≥49,780	730.00	30.67	
Gender								
Male	≥25,610	736.86	32.74		≥25,600	729.35	30.89	
Female	≥24,170	742.35	33.61	-0.17	≥24,160	730.68	30.43	-0.04
Ethnicity								
White	≥21,220	751.70	31.96		≥21,210	742.09	29.40	
Hispanic/Latino	≥4,760	733.77	33.02	0.56	≥4,760	725.86	29.82	0.55
American Indian or Alaska Native	≥290	739.95	27.75	0.37	≥290	731.76	27.83	0.35
Asian	≥800	764.41	37.70	-0.39	≥800	758.18	34.32	-0.54
Black or African American	≥21,040	727.22	29.32	0.80	≥21,020	717.42	26.24	0.89
Native Hawaiian or Other Pacific	≥20	753.50	37.60	-0.06	≥20	746.93	30.59	-0.16
Two or More Races	≥1,600	744.27	32.67	0.23	≥1,600	732.71	29.28	0.32
Education Classification								
Regular	≥43,820	742.72	32.73		≥43,800	732.48	30.58	
Special	≥5,960	716.06	27.29	0.83	≥5,950	711.70	24.57	0.69
Economic Status								
Not Economically Disadvantaged	≥13,090	759.23	31.91		≥13,090	749.34	29.21	
Economically Disadvantaged	≥36,690	732.49	30.84	0.86	≥36,660	723.09	28.12	0.92
English Learner Status								
Not English Learner	≥47,580	740.77	33.09		≥47,550	730.88	30.61	
English Learner	≥2,200	712.65	24.73	0.86	≥2,200	711.01	25.39	0.65
Migrant Status								
Not Migrant	≥49,730	739.54	33.27		≥49,710	730.00	30.67	
Migrant	≥40	729.94	38.33	0.29	≥40	726.94	35.47	0.10
Section 504 Status								
Not Section 504	≥44,680	740.93	33.50		≥44,660	731.11	30.90	
Section 504	≥5,100	727.27	28.42	0.41	≥5,100	720.28	26.77	0.36

Table 10.16 Impact Analysis, Grade 6 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥51,430	736.21	31.03		≥51,430	727.35	30.46	
Gender								
Male	≥26,130	731.68	30.55		≥26,120	726.17	30.81	
Female	≥25,300	740.89	30.82	-0.30	≥25,290	728.57	30.03	-0.08
Ethnicity								
White	≥21,830	747.37	29.99		≥21,820	740.15	28.57	
Hispanic/Latino	≥4,600	731.58	31.93	0.52	≥4,600	723.16	29.68	0.59
American Indian or Alaska Native	≥300	739.16	28.70	0.27	≥300	728.28	27.84	0.42
Asian	≥760	761.30	33.37	-0.46	≥760	755.10	31.41	-0.52
Black or African American	≥22,200	724.86	27.13	0.79	≥22,190	714.31	26.28	0.94
Native Hawaiian or Other Pacific	≥40	741.35	28.50	0.20	≥40	734.58	32.08	0.19
Two or More Races	≥1,670	741.90	29.66	0.18	≥1,670	732.02	28.94	0.28
Education Classification								
Regular	≥45,600	739.52	30.12		≥45,590	730.15	29.99	
Special	≥5,820	710.32	25.25	0.99	≥5,820	705.39	24.56	0.84
Economic Status								
Not Economically Disadvantaged	≥13,530	753.90	29.76		≥13,530	746.25	28.68	
Economically Disadvantaged	≥37,890	729.89	28.95	0.82	≥37,880	720.60	28.14	0.91
English Learner Status								
Not English Learner	≥49,460	737.21	30.84		≥49,440	728.19	30.33	
English Learner	≥1,970	711.20	24.46	0.85	≥1,970	706.12	25.61	0.73
Migrant Status								
Not Migrant	≥51,380	736.22	31.02		≥51,360	727.35	30.46	
Migrant	≥50	731.17	33.13	0.16	≥50	727.41	28.91	0.00
Section 504 Status								
Not Section 504	≥45,980	737.68	31.14		≥45,960	728.66	30.60	
Section 504	≥5,450	723.81	27.03	0.45	≥5,450	716.26	26.79	0.41

Table 10.17 Impact Analysis, Grade 7 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥52,180	741.81	37.27		≥52,180	729.68	26.48	
Gender								
Male	≥26,590	735.62	36.47		≥26,580	729.20	27.07	
Female	≥25,590	748.23	37.00	-0.34	≥25,570	730.18	25.84	-0.04
Ethnicity								
White	≥22,400	754.35	35.17		≥22,390	740.04	25.15	
Hispanic/Latino	≥4,640	735.33	39.56	0.53	≥4,640	726.75	26.30	0.52
American Indian or Alaska Native	≥300	745.73	35.30	0.24	≥300	733.61	26.28	0.26
Asian	≥830	772.46	40.84	-0.51	≥830	757.70	31.57	-0.69
Black or African American	≥22,350	728.96	33.73	0.74	≥22,330	718.53	22.40	0.90
Native Hawaiian or Other Pacific	≥40	756.55	39.35	-0.06	≥40	740.82	30.97	-0.03
Two or More Races	≥1,590	747.36	35.16	0.20	≥1,590	733.08	24.97	0.28
Education Classification								
Regular	≥46,600	745.92	35.80		≥46,580	732.33	25.68	
Special	≥5,580	707.46	30.99	1.09	≥5,570	707.55	22.37	0.98
Economic Status								
Not Economically Disadvantaged	≥14,420	762.44	34.93		≥14,420	745.59	25.52	
Economically Disadvantaged	≥37,760	733.93	35.05	0.81	≥37,730	723.60	24.22	0.89
English Learner Status								
Not English Learner	≥50,270	743.09	36.85		≥50,240	730.38	26.39	
English Learner	≥1,910	708.02	31.81	0.96	≥1,910	711.34	21.77	0.73
Migrant Status								
Not Migrant	≥52,120	741.82	37.26		≥52,090	729.68	26.48	
Migrant	≥60	732.33	37.79	0.25	≥60	724.98	25.90	0.18
Section 504 Status								
Not Section 504	≥46,660	743.67	37.32		≥46,640	730.88	26.54	
Section 504	≥5,520	726.10	32.83	0.48	≥5,510	719.49	23.60	0.43

Table 10.18 Impact Analysis, Grade 8 Computer-Based Test Administration

	ELA				Mathematics			
Group	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**	**N**	**Scale Score Mean**	**Scale Score Std. Dev.**	**Effect Size**
All Students	≥51,680	743.34	37.69		≥51,680	722.94	34.21	
Gender								
Male	≥26,210	735.89	37.08		≥23,510	720.82	34.92	
Female	≥25,470	751.01	36.77	-0.41	≥22,360	725.18	33.30	-0.13
Ethnicity								
White	≥22,740	754.97	36.12		≥19,220	736.37	32.79	
Hispanic/Latino	≥4,200	735.78	39.46	0.52	≥3,720	717.93	32.94	0.56
American Indian or Alaska Native	≥310	746.80	37.02	0.23	≥290	727.82	34.54	0.26
Asian	≥800	776.13	40.13	-0.58	≥540	756.06	40.96	-0.60
Black or African American	≥22,030	731.13	34.19	0.68	≥20,680	710.10	29.84	0.84
Native Hawaiian or Other Pacific	≥40	750.50	35.95	0.12	≥30	732.97	40.66	0.10
Two or More Races	≥1,520	748.91	37.46	0.17	≥1,350	727.42	34.49	0.27
Education Classification								
Regular	≥46,510	747.31	36.38		≥40,840	726.20	33.54	
Special	≥5,170	707.63	29.67	1.11	≥5,040	696.56	27.45	0.90
Economic Status								
Not Economically Disadvantaged	≥14,890	762.98	35.62		≥11,900	742.15	33.36	
Economically Disadvantaged	≥36,790	735.39	35.54	0.78	≥33,970	716.21	31.87	0.80
English Learner Status								
Not English Learner	≥49,820	744.64	37.33		≥44,090	723.80	34.17	
English Learner	≥1,850	708.41	29.93	0.98	≥1,790	701.97	27.96	0.64
Migrant Status								
Not Migrant	≥51,620	743.35	37.69		≥45,820	722.95	34.21	
Migrant	≥50	732.83	35.20	0.28	≥50	720.00	30.94	0.09
Section 504 Status								
Not Section 504	≥46,410	745.16	37.74		≥40,850	724.26	34.31	
Section 504	≥5,270	727.34	33.28	0.48	≥5,020	712.23	31.40	0.35

Additional data for mean scale scores are provided in Tables 10.19 and 10.20. 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 10.19 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	≥780	809.58	29.13	≥48,840	736.66	42.66
	Talented	≥510	778.26	38.16	≥49,120	737.39	43.30
	Autism	≥390	706.32	36.14	≥49,240	738.07	43.41
	Deaf-Blindness	<50	NR	NR	≥49,630	737.82	43.45
	Developmental Delay	≥700	709.03	33.46	≥48,930	738.23	43.44
	Emotional Disturbance	≥50	712.31	44.41	≥49,580	737.84	43.44
	HI—Deaf	<50	NR	NR	≥49,610	737.83	43.44
	HI—Hard-of-Hearing	≥50	716.06	38.61	≥49,580	737.84	43.45
	Mild Mental Disability	≥360	689.98	25.83	≥49,270	738.17	43.36
	Moderate Mental Disability	<50	NR	NR	≥49,610	737.84	43.44
	Orthopedic Impairment	≥50	738.78	45.12	≥49,580	737.81	43.45
	Other Health Impairment	≥690	709.99	36.46	≥48,940	738.21	43.41
	Specific Learning Disability	≥2,040	711.87	32.32	≥47,590	738.93	43.52
	Speech or Language Impairment	≥1,740	739.30	44.14	≥47,890	737.76	43.42
	Traumatic Brain Injury	<50	NR	NR	≥49,620	737.82	43.45
	Visual Impairment	<50	NR	NR	≥49,590	737.82	43.45
	Other	<50	NR	NR	≥49,630	737.82	43.45
	HI—Hearing Impairment	<50	NR	NR	≥49,630	737.82	43.45
	Unknown	<50	NR	NR	≥49,630	737.82	43.45
4	Gifted	≥1,030	798.36	26.45	≥48,520	740.19	35.76
	Talented	≥850	772.36	30.42	≥48,700	740.86	36.41
	Autism	≥390	707.94	34.43	≥49,160	741.67	36.44
	Deaf-Blindness	<50	NR	NR	≥49,550	741.40	36.55
	Developmental Delay	<50	NR	NR	≥49,520	741.43	36.54
	Emotional Disturbance	≥90	717.73	34.10	≥49,460	741.45	36.54
	HI—Deaf	<50	NR	NR	≥49,530	741.42	36.54
	HI—Hard-of-Hearing	≥50	726.57	38.30	≥49,500	741.42	36.54
	Mild Mental Disability	≥440	691.76	19.20	≥49,110	741.85	36.36
	Moderate Mental Disability	<50	NR	NR	≥49,540	741.41	36.54
	Orthopedic Impairment	<50	NR	NR	≥49,510	741.41	36.55
	Other Health Impairment	≥970	714.19	31.06	≥48,580	741.95	36.44
	Specific Learning Disability	≥2,640	711.83	26.96	≥46,910	743.07	36.31
	Speech or Language Impairment	≥1,350	741.66	36.58	≥48,200	741.40	36.55
	Traumatic Brain Injury	<50	NR	NR	≥49,540	741.41	36.55
	Visual Impairment	<50	NR	NR	≥49,520	741.41	36.55
	Other	<50	NR	NR	≥49,550	741.41	36.55
	HI—Hearing Impairment	<50	NR	NR	≥49,550	741.40	36.55
	Unknown	<50	NR	NR	≥49,550	741.40	36.55

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	≥1,150	792.25	25.19	≥48,630	738.27	32.41
	Talented	≥1,330	761.39	32.16	≥48,450	738.92	33.10
	Autism	≥350	715.77	30.05	≥49,430	739.70	33.24
	Deaf-Blindness	<50	NR	NR	≥49,780	739.53	33.27
	Developmental Delay	<50	NR	NR	≥49,740	739.55	33.27
	Emotional Disturbance	≥110	716.99	27.05	≥49,660	739.58	33.27
	HI—Deaf	<50	NR	NR	≥49,760	739.54	33.27
	HI—Hard-of-Hearing	<50	NR	NR	≥49,740	739.55	33.27
	Mild Mental Disability	≥440	696.95	16.18	≥49,340	739.91	33.14
	Moderate Mental Disability	<50	NR	NR	≥49,770	739.54	33.27
	Orthopedic Impairment	≥50	731.82	35.65	≥49,730	739.53	33.27
	Other Health Impairment	≥1,020	715.92	25.87	≥48,760	740.02	33.23
	Specific Learning Disability	≥2,810	711.38	21.70	≥46,970	741.21	33.09
	Speech or Language Impairment	≥970	737.77	32.11	≥48,810	739.56	33.30
	Traumatic Brain Injury	<50	NR	NR	≥49,770	739.54	33.27
	Visual Impairment	<50	NR	NR	≥49,750	739.53	33.28
	Other	<50	NR	NR	≥49,770	739.53	33.27
	HI—Hearing Impairment	<50	NR	NR	≥49,780	739.53	33.27
	Unknown	<50	NR	NR	≥49,780	739.53	33.27
6	Gifted	≥1,180	785.88	24.70	≥50,250	735.04	30.19
	Talented	≥1,680	757.54	29.49	≥49,740	735.49	30.82
	Autism	≥290	714.02	29.71	≥51,140	736.34	30.99
	Deaf-Blindness	<50	NR	NR	≥51,430	736.21	31.03
	Developmental Delay	<50	NR	NR	≥51,400	736.23	31.02
	Emotional Disturbance	≥150	709.78	25.90	≥51,270	736.29	31.01
	HI—Deaf	<50	NR	NR	≥51,410	736.22	31.02
	HI—Hard-of-Hearing	≥50	723.10	28.59	≥51,370	736.23	31.02
	Mild Mental Disability	≥300	691.64	15.45	≥51,130	736.48	30.90
	Moderate Mental Disability	<50	NR	NR	≥51,430	736.21	31.02
	Orthopedic Impairment	≥50	722.46	28.76	≥51,380	736.23	31.02
	Other Health Impairment	≥1,160	709.20	24.40	≥50,270	736.84	30.88
	Specific Learning Disability	≥3,020	706.98	21.05	≥48,400	738.04	30.63
	Speech or Language Impairment	≥670	732.07	29.66	≥50,760	736.27	31.04
	Traumatic Brain Injury	<50	NR	NR	≥51,420	736.22	31.03
	Visual Impairment	<50	NR	NR	≥51,390	736.22	31.02
	Other	<50	NR	NR	≥51,420	736.22	31.02
	HI—Hearing Impairment	<50	NR	NR	≥51,430	736.21	31.03
	Unknown	<50	NR	NR	≥51,430	736.21	31.03

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	≥1,350	797.45	27.44	≥50,820	740.32	36.34
	Talented	≥1,860	767.99	34.32	≥50,320	740.84	37.02
	Autism	≥290	715.91	37.77	≥51,890	741.96	37.21
	Deaf-Blindness	<50	NR	NR	≥52,180	741.81	37.26
	Developmental Delay	<50	NR	NR	≥52,180	741.81	37.27
	Emotional Disturbance	≥170	705.57	29.65	≥52,010	741.93	37.23
	HI—Deaf	<50	NR	NR	≥52,160	741.83	37.26
	HI—Hard-of-Hearing	≥50	728.05	42.87	≥52,130	741.82	37.26
	Mild Mental Disability	≥230	682.92	18.58	≥51,950	742.07	37.12
	Moderate Mental Disability	<50	NR	NR	≥52,180	741.81	37.27
	Orthopedic Impairment	≥50	727.56	35.94	≥52,130	741.82	37.26
	Other Health Impairment	≥1,230	709.24	29.74	≥50,950	742.59	37.08
	Specific Learning Disability	≥3,000	702.73	26.43	≥49,180	744.20	36.50
	Speech or Language Impairment	≥460	736.01	35.82	≥51,720	741.86	37.28
	Traumatic Brain Injury	<50	NR	NR	≥52,170	741.81	37.27
	Visual Impairment	<50	NR	NR	≥52,160	741.82	37.26
	Other	<50	NR	NR	≥52,170	741.81	37.26
	HI—Hearing Impairment	<50	NR	NR	≥52,180	741.81	37.27
	Unknown	<50	NR	NR	≥52,180	741.81	37.27
8	Gifted	≥1,470	798.35	27.19	≥50,200	741.72	36.73
	Talented	≥1,900	769.90	34.01	≥49,770	742.32	37.45
	Autism	≥250	716.73	33.33	≥51,430	743.47	37.67
	Deaf-Blindness	<50	NR	NR	≥51,680	743.34	37.69
	Developmental Delay	<50	NR	NR	≥51,680	743.34	37.69
	Emotional Disturbance	≥190	707.63	31.27	≥51,480	743.48	37.65
	HI—Deaf	<50	NR	NR	≥51,670	743.35	37.69
	HI—Hard-of-Hearing	≥60	725.77	33.49	≥51,610	743.36	37.69
	Mild Mental Disability	≥180	688.39	20.85	≥51,490	743.54	37.59
	Moderate Mental Disability	<50	NR	NR	≥51,670	743.35	37.69
	Orthopedic Impairment	<50	NR	NR	≥51,630	743.35	37.69
	Other Health Impairment	≥1,120	708.73	30.29	≥50,550	744.11	37.48
	Specific Learning Disability	≥2,920	703.86	25.83	≥48,760	745.71	36.97
	Speech or Language Impairment	≥300	733.94	36.52	≥51,370	743.40	37.69
	Traumatic Brain Injury	<50	NR	NR	≥51,670	743.35	37.69
	Visual Impairment	<50	NR	NR	≥51,650	743.35	37.69
	Other	<50	NR	NR	≥51,660	743.35	37.69
	HI—Hearing Impairment	<50	NR	NR	≥51,680	743.34	37.69
	Unknown	<50	NR	NR	≥51,680	743.34	37.69

Table 10.20 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	≥780	792.88	24.13	≥48,810	732.00	33.84
	Talented	≥510	762.01	28.62	≥49,070	732.66	34.48
	Autism	≥380	713.52	33.20	≥49,200	733.12	34.52
	Deaf-Blindness	<50	NR	NR	≥49,590	732.96	34.55
	Developmental Delay	≥700	710.34	27.94	≥48,890	733.29	34.53
	Emotional Disturbance	≥50	716.89	37.99	≥49,540	732.98	34.54
	HI—Deaf	<50	NR	NR	≥49,570	732.97	34.55
	HI—Hard-of-Hearing	≥50	721.74	30.64	≥49,540	732.97	34.55
	Mild Mental Disability	≥350	693.87	19.66	≥49,230	733.25	34.47
	Moderate Mental Disability	<50	NR	NR	≥49,570	732.98	34.54
	Orthopedic Impairment	≥50	724.90	33.82	≥49,540	732.97	34.55
	Other Health Impairment	≥690	709.99	29.00	≥48,900	733.29	34.51
	Specific Learning Disability	≥2,040	711.90	25.38	≥47,540	733.87	34.60
	Speech or Language Impairment	≥1,740	736.52	34.51	≥47,840	732.83	34.54
	Traumatic Brain Injury	<50	NR	NR	≥49,580	732.97	34.55
	Visual Impairment	<50	NR	NR	≥49,550	732.97	34.55
	Other	<50	NR	NR	≥49,590	732.97	34.55
	HI—Hearing Impairment	<50	NR	NR	≥49,590	732.96	34.55
	Unknown	<50	NR	NR	≥49,590	732.96	34.55
4	Gifted	≥1,030	783.70	22.37	≥48,450	730.32	33.07
	Talented	≥850	757.21	28.27	≥48,630	730.99	33.67
	Autism	≥390	710.53	30.79	≥49,090	731.61	33.73
	Deaf-Blindness	<50	NR	NR	≥49,490	731.44	33.76
	Developmental Delay	<50	NR	NR	≥49,450	731.46	33.75
	Emotional Disturbance	≥90	710.45	30.07	≥49,390	731.48	33.76
	HI—Deaf	<50	NR	NR	≥49,460	731.45	33.76
	HI—Hard-of-Hearing	≥50	726.59	32.39	≥49,430	731.44	33.76
	Mild Mental Disability	≥440	690.43	16.69	≥49,050	731.81	33.65
	Moderate Mental Disability	<50	NR	NR	≥49,480	731.45	33.76
	Orthopedic Impairment	<50	NR	NR	≥49,450	731.44	33.76
	Other Health Impairment	≥970	709.67	27.82	≥48,510	731.88	33.73
	Specific Learning Disability	≥2,640	707.18	23.29	≥46,840	732.81	33.74
	Speech or Language Impairment	≥1,340	734.41	34.81	≥48,140	731.36	33.73
	Traumatic Brain Injury	<50	NR	NR	≥49,480	731.45	33.76
	Visual Impairment	<50	NR	NR	≥49,450	731.44	33.76
	Other	<50	NR	NR	≥49,480	731.44	33.76
	HI—Hearing Impairment	<50	NR	NR	≥49,490	731.44	33.76
	Unknown	<50	NR	NR	≥49,490	731.44	33.76

Special Education Classification Scale-Score Means and Standard Deviations: Mathematics							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
5	Gifted	≥1,150	779.21	23.40	≥48,540	728.88	29.82
	Talented	≥1,330	747.87	29.93	≥48,360	729.56	30.51
	Autism	≥350	713.73	29.10	≥49,340	730.17	30.62
	Deaf-Blindness	<50	NR	NR	≥49,700	730.06	30.64
	Developmental Delay	<50	NR	NR	≥49,650	730.08	30.64
	Emotional Disturbance	≥110	709.04	22.88	≥49,580	730.11	30.64
	HI—Deaf	<50	NR	NR	≥49,680	730.06	30.64
	HI—Hard-of-Hearing	<50	NR	NR	≥49,650	730.07	30.64
	Mild Mental Disability	≥440	694.83	15.46	≥49,260	730.37	30.56
	Moderate Mental Disability	<50	NR	NR	≥49,690	730.07	30.63
	Orthopedic Impairment	≥50	721.46	24.96	≥49,650	730.06	30.64
	Other Health Impairment	≥1,020	711.44	23.40	≥48,670	730.45	30.65
	Specific Learning Disability	≥2,810	707.64	18.92	≥46,890	731.40	30.69
	Speech or Language Impairment	≥970	730.53	29.99	≥48,730	730.05	30.65
	Traumatic Brain Injury	<50	NR	NR	≥49,690	730.06	30.64
	Visual Impairment	<50	NR	NR	≥49,670	730.05	30.64
	Other	<50	NR	NR	≥49,690	730.06	30.64
	HI—Hearing Impairment	<50	NR	NR	≥49,700	730.06	30.64
	Unknown	<50	NR	NR	≥49,700	730.06	30.64
6	Gifted	≥1,180	778.39	24.55	≥50,160	726.21	29.50
	Talented	≥1,680	745.11	28.82	≥49,660	726.81	30.29
	Autism	≥280	711.73	29.53	≥51,050	727.50	30.40
	Deaf-Blindness	<50	NR	NR	≥51,340	727.41	30.42
	Developmental Delay	<50	NR	NR	≥51,310	727.42	30.42
	Emotional Disturbance	≥150	703.23	26.86	≥51,190	727.48	30.40
	HI—Deaf	<50	NR	NR	≥51,320	727.42	30.42
	HI—Hard-of-Hearing	≥50	721.68	31.75	≥51,280	727.42	30.42
	Mild Mental Disability	≥300	684.43	15.82	≥51,040	727.67	30.31
	Moderate Mental Disability	<50	NR	NR	≥51,340	727.41	30.42
	Orthopedic Impairment	≥50	713.02	30.11	≥51,290	727.42	30.42
	Other Health Impairment	≥1,160	704.59	23.39	≥50,180	727.94	30.36
	Specific Learning Disability	≥3,020	702.63	19.92	≥48,320	728.96	30.29
	Speech or Language Impairment	≥670	724.44	30.09	≥50,670	727.45	30.42
	Traumatic Brain Injury	<50	NR	NR	≥51,330	727.42	30.42
	Visual Impairment	<50	NR	NR	≥51,310	727.42	30.42
	Other	<50	NR	NR	≥51,330	727.41	30.42
	HI—Hearing Impairment	<50	NR	NR	≥51,340	727.41	30.42
	Unknown	<50	NR	NR	≥51,340	727.41	30.42

Special Education Classification Scale-Score Means and Standard Deviations: Mathematics							
Grade	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
7	Gifted	≥1,350	775.23	22.41	≥50,720	728.52	25.46
	Talented	≥1,860	745.25	24.42	≥50,220	729.16	26.35
	Autism	≥290	716.43	28.25	≥51,790	729.81	26.42
	Deaf-Blindness	<50	NR	NR	≥52,080	729.73	26.45
	Developmental Delay	<50	NR	NR	≥52,080	729.73	26.45
	Emotional Disturbance	≥170	706.42	22.77	≥51,900	729.81	26.43
	HI—Deaf	<50	NR	NR	≥52,060	729.74	26.45
	HI—Hard-of-Hearing	≥50	719.61	28.81	≥52,030	729.74	26.45
	Mild Mental Disability	≥230	690.17	14.98	≥51,850	729.91	26.36
	Moderate Mental Disability	<50	NR	NR	≥52,080	729.73	26.45
	Orthopedic Impairment	≥50	718.82	26.22	≥52,020	729.74	26.45
	Other Health Impairment	≥1,220	707.75	22.12	≥50,850	730.26	26.32
	Specific Learning Disability	≥3,000	704.35	18.63	≥49,080	731.28	26.06
	Speech or Language Impairment	≥460	728.03	25.64	≥51,610	729.75	26.46
	Traumatic Brain Injury	<50	NR	NR	≥52,070	729.74	26.45
	Visual Impairment	<50	NR	NR	≥52,060	729.74	26.45
	Other	<50	NR	NR	≥52,070	729.74	26.45
	HI—Hearing Impairment	<50	NR	NR	≥52,080	729.73	26.45
	Unknown	<50	NR	NR	≥52,080	729.73	26.45
8	Gifted	≥690	781.11	32.16	≥45,140	722.13	33.42
	Talented	≥1,470	742.97	33.58	≥44,360	722.36	33.99
	Autism	≥240	706.19	33.54	≥45,600	723.11	34.16
	Deaf-Blindness	<50	NR	NR	≥45,840	723.02	34.18
	Developmental Delay	<50	NR	NR	≥45,840	723.02	34.18
	Emotional Disturbance	≥180	694.10	28.82	≥45,650	723.14	34.14
	HI—Deaf	<50	NR	NR	≥45,830	723.03	34.17
	HI—Hard-of-Hearing	≥50	714.45	27.10	≥45,780	723.03	34.18
	Mild Mental Disability	≥180	682.07	17.98	≥45,650	723.19	34.12
	Moderate Mental Disability	<50	NR	NR	≥45,840	723.03	34.17
	Orthopedic Impairment	<50	NR	NR	≥45,800	723.04	34.18
	Other Health Impairment	≥1,110	698.17	27.80	≥44,730	723.64	34.09
	Specific Learning Disability	≥2,880	693.78	24.66	≥42,950	724.99	33.83
	Speech or Language Impairment	≥270	714.34	34.31	≥45,570	723.08	34.17
	Traumatic Brain Injury	<50	NR	NR	≥45,830	723.03	34.17
	Visual Impairment	<50	NR	NR	≥45,820	723.03	34.17
	Other	<50	NR	NR	≥45,820	723.03	34.18
	HI—Hearing Impairment	<50	NR	NR	≥45,840	723.02	34.18
	Unknown	<50	NR	NR	≥45,840	723.02	34.18

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 2021 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. A mode comparability study was not conducted in 2021 as the forms were primarily the intact forms from 2019. Only one item on the grade 3 mathematics form was from 2018. In both 2018 and 2019, mode comparability studies were conducted. For details regarding the mode comparability study, see the *2019 LEAP 2025 Grades 3-8 Operational Technical Report: English Language Arts and Mathematics*. At a summary level, 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.

Although the PSM mode study was not conducted in 2021, DIF statistics were used to identify item performance differences across mode of administration. The Mantel-Haenszel (MH) statistic and the standardized mean difference (SMD) were used as DIF statitiscs on the grades 3 and 4 ELA and mathematics operational items to determine if the items performed differently across modes. In this analysis, the PBT administration was considered the focal group and the CBT administration the reference group. Table 10.21 summarizes the count of DIF flags for each subject. In ELA, only a few items displayed moderate DIF, one item in grade 3 and two items in grade 4. Large DIF was displayed on two items in grade 3 mathematics one item in grade 4. Of the flagged items, only one, in grade 4 mathematics, was also flagged in 2019.

Table 10.21 2021 LEAP 2025 DIF Statistics: Number of Flagged Items, Mode

DIF Statistics: Mode				Count of Items at DIF Magnitude			
				Moderate		Large	
Subject	Grade	Number of Items	Group	B-	B+	C-	C+
ELA	3	32	Paper	0	1	0	0
	4	34	Paper	1	1	0	0
Mathematics	3	43	Paper	0	0	2	0
	4	42	Paper	0	0	0	1

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—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 B—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 C—LEAP 2025 Spring 2021 Handscoring/AI Documentation

LEAP 2025 Spring 2021 Handscoring/AI Documentation

LEAP 2025 Grades 3-8

ELA, Math, Science, and Social Studies

LEAP 2025 High School

Algebra I, Geometry, English I, English II, Biology, and U.S. History

Contents

Staffing and Schedule 1

Training and Scoring Schedule 1

Scorer Degree Requirements 2

Security 2

Remote Scoring Overview 3

Background 3

Experience 3

System Tools – Scoring, Training, Chat 3

Content Training with Moodle 4

Quality Control 4

Content-Specific Training 4

Training Materials 5

LEAP 2025 Biology, LEAP 2025 U.S. History, and LEAP 2025 Grades 3-8 Science and Social Studies 5

LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math (Items and Materials Developed by DRC) 6

LEAP 2025 Algebra I, Geometry, English I, English II, and Grades 3-8 ELA and Math (Items and Materials Developed by PARCC) 7

Algebra I, Geometry, and Grades 3-8 Math Training Set Composition 8

English I, English II, and Grades 3-8 ELA Training Set Composition 9

English I, English II, and Grades 3-8 ELA Training Set Composition (continued) 10

Algebra I Items and Associated Training Materials 10

Geometry Items and Associated Training Materials 11

Grade 3 Math Items and Associated Training Materials 11

Grade 4 Math Items and Associated Training Materials 12

Grade 5 Math Items and Associated Training Materials 12

Grade 6 Math Items and Associated Training Materials 13

Grade 7 Math Items and Associated Training Materials 13

Grade 8 Math Items and Associated Training Materials 13

English I Items and Associated Training Materials 14

English II Items and Associated Training Materials 14
Grades 3-8 ELA Items and Associated Training Materials 14
Qualifying 15
LEAP 2025 Constructed-Response and Extended-Response Items 15
LEAP 2025 English I, English II, and Grades 3-8 ELA 15
LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math 16
LEAP 2025 U.S. History and Grades 3-8 Social Studies 17
LEAP 2025 Biology and Grades 3-8 Science 17
Spring 2021 Scoring Plan 18
LEAP 2025 High School 18
LEAP 2025 Grades 3-8 18
Handscoring Rules 19
AI Scoring 19
Handscoring 19
Calculating the Final Score: 19
Reader Monitoring Procedures 20
Team Leader Read-Behinds 20
Validity Sets and Inter-Rater Reliability 20
Calibration Sets 23
Handscoring Quality Control Reports 23
Scoring Summary Report Sample 23
Scoring Summary Report Sample with AI (Reader ID #3) 24
Reader Feedback Logs 24
Handling Unusual Responses 25
Nonscore Codes and Definitions 25
Nonscore Code Definitions 25
Nonscore Codes by Test 25
Alerts 26
Artificial Intelligence Scoring 27

AI Scoring – Measurement, Inc. 27
Model Building 27
Evaluation Metric 28
Scoring Responses with the AI Engine 30
Quality Control of the AI Engine 30
Identifying Responses for Human Review 31
Alert Detection System 31
Identification of Nonscorable Responses 32
Identifying Copied Text and Plagiarism with the AI Engine 33
AI Scoring – Pearson 36
The Intelligent Essay Assessor 36
How the Intelligent Essay Assessor was Trained 38
Quality Monitoring 39
Scoring (DRC) 40
Rescores 40
Appendix A 41
DRC-MI Streaming Scoring Documentation 41
SECTION 1 – General Information 42
SECTION 2 – SCHEMA SUPPLEMENT 43
SECTION 3 – STATUS CODE INFORMATION 46
Appendix B 47
DRC Distributed Scoring Process 47
Appendix C 56
AI Model Data – LEAP 2025 U.S. History ERs (Spring 2021) 56
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 56
AI Model Building – Social Studies Grades 5-8 ERs (Spring 2021) 57
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 57
AI Model CR Performance – ELA Grades 5-8, English I, and English II (Spring 2021) 58
Spring 2021 LEAP 2025 Items – IRR and SPD from Previous Administrations 59

Algebra I 59
Algebra I (continued) 60
Algebra I (continued) 61
Geometry 62
Math Grade 3 64
Math Grade 5 66
Math Grade 6 67
Math Grade 7 68
Math Grade 8 69
English II 71
ELA Grade 3 72
ELA Grade 4 72
ELA Grade 5 73
ELA Grade 6 73
ELA Grade 7 74
ELA Grade 8 74
Biology ERs and CRs 75
Biology ERs and CRs (continued) 76
Grade 3 Science 77
Grade 4 Science 77
Grade 5 Science 78
Grade 6 Science 78
Grade 7 Science 79
Grade 8 Science 79
U.S. History ERs and CRs 80
Social Studies Grade 3 82
Social Studies Grade 4 82
Social Studies Grade 5 83
Social Studies Grade 6 83

Social Studies Grade 7 84
Social Studies Grade 8 84

Staffing and Schedule

Training and Scoring Schedule

DRC's spring 2021 reader training and scoring schedule is based on the spring testing windows of April 15, 2021 – May 21, 2021 (LEAP 2025 high school) and April 26, 2021 – May 26, 2021 (LEAP 2025 grades 3-8). High school Administrative Error (AE) testing ends on May 25, 2021. Anticipated reader training and scoring dates are noted below.

Due to site capacity limitations in DRC scoring facilities necessitated by the COVID-19 pandemic, reader training and handscoring for the spring 2021 administration of LEAP 2025 high school and grades 3-8 assessments will be conducted using both site-based and remote approaches (as indicated in the Training and Scoring Schedule below). Site-based projects will take place in the secure DRC scoring facilities in the locations noted. Remote project training and scoring will be conducted from within DRC's secure, remote online training/scoring environment.

Grade/Content Area or Test	DRC Scoring Location	Anticipated Staffing	2021 Reader Training and Scoring Window
3 ELA	Remote	2 Scoring Directors, 6 Team Leaders, 42 Readers	June 1 – June 18
4 ELA	Remote	2 Scoring Directors, 6 Team Leaders, 42 Readers	June 1 – June 18
5 ELA	Remote	1 Scoring Director, 2 Team Leaders, 22 Readers	April 26 – May 28
6 ELA	Remote	1 Scoring Director, 1 Team Leader, 6 Readers	April 26 – May 28
7 ELA	Remote	1 Scoring Director, 1 Team Leader, 6 Readers	April 28 – June 1
8 ELA	Remote	1 Scoring Director, 1 Team Leader, 6 Readers	April 28 – June 1
3 Math	Remote	2 Scoring Directors, 5 Team Leaders, 50 Readers	June 1– June 18
4 Math	Plymouth, MN	2 Scoring Directors, 4 Team Leaders, 30 Readers	June 1 – June 18
5 Math	Remote	2 Scoring Directors, 4 Team Leaders, 40 Readers	April 27 – June 1
6 Math	Remote	2 Scoring Directors, 4 Team Leaders, 26 Readers	April 28 – June 1
7 Math	Remote	2 Scoring Directors, 5 Team Leaders, 50 Readers	April 26 – May 28
8 Math	Remote	2 Scoring Directors, 4 Team Leaders, 40 Readers	April 27 – June 1
3 Science (CRs)	Remote	1 Scoring Director, 3 Team Leaders, 16 Readers	June 1 – June 18
4 Science (CRs)	Remote	1 Scoring Director, 3 Team Leaders, 16 Readers	June 1 – June 18
3 & 4 Science (ERs)	Remote	1 Scoring Director, 1 Assistant Scoring Director, 7 Team Leaders, 48 Readers	June 1 – June 18
5 Science	Remote	1 Scoring Director, 6 Team Leaders, 40 Readers	April 27 – June 1
6 Science (ER)	Remote	1 Scoring Director, 3 Team Leaders, 16 Readers	April 27 – May 28
6 & 7 Science (CRs)	Remote	2 Scoring Directors, 4 Team Leaders, 36 Readers	April 27 – June 1
7 Science (ER)	Remote	1 Scoring Director, 3 Team Leaders, 16 Readers	April 27 – May 28
8 Science	Remote	1 Scoring Director, 1 Assistant Scoring Director, 7 Team Leaders, 48 Readers	April 27 – May 28
3 & 4 SS	Remote	1 Scoring Director, 1 Assistant Scoring Director, 6 Team Leaders, 18 Readers	June 1 – June 18
5, 6, 7, & 8 SS	Remote	1 Scoring Director, 1 Assistant Scoring Director, 6 Team Leaders, 48 Readers	April 26 – May 28

Grade/Content Area or Test	DRC Scoring Location	Anticipated Staffing	2021 Reader Training and Scoring Window
LEAP 2025 Algebra I	Remote	2 Scoring Directors, 4 Team Leaders, 37 Readers	April 14 – May 26
LEAP 2025 Geometry	Remote	2 Scoring Directors, 3 Team Leaders, 23 Readers	April 13 – May 26
LEAP 2025 English I	Remote	1 Scoring Director, 1 Team Leader, 6 Readers	April 13 – June 1
LEAP 2025 English II	Remote	1 Scoring Director, 1 Team Leader, 6 Readers	April 13 – June 1
LEAP 2025 Biology	Remote	1 Scoring Director, 1 Assistant Scoring Director, 7 Team Leaders, 48 Readers	April 14 – May 26
LEAP 2025 U.S. History	Remote	1 Scoring Director, 1 Assistant Scoring Director, 6 Team Leaders, 48 Readers	April 14 – May 26

Scorers will be divided by test as detailed in the 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.

Security

Whether training and scoring are conducted within a DRC facility or done remotely, security is essential to our handscoring process. When users log into DRC's secure, web-based scoring application, ScoreBoard, they are required to read and accept our security policy before they are allowed to access any project. For each project, scorers are also required to read and sign non-disclosure agreements, and during training emphasis is always given to what security means, the importance of maintaining security, and how this is accomplished.

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.

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. 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.

In a remote environment, security reminders are given on a daily basis. Similar to the work that occurs within DRC scoring sites, in a remote environment, education about security expectations is the best way to maintain security of any project materials. DRC requires scorers working remotely to work in a private environment away from other people (including family members). Restrictions are in place that define the hours during the day scorers are able to log into the system. If any type of security breach were to occur, immediate action would be taken to secure materials, and the employee would be terminated. DRC has the same policy within our scoring sites.

Remote Scoring Overview

Background

DRC's remote scoring is designed to very closely emulate the work that is done in our physical scoring locations. The platform, content, and expectations for quality remain the same, and interactive technology and content training and discussions are conducted live (virtually). The differences come with the method through which training is delivered (online), and in the modes of communication that are used (web screen sharing, webcast, video chat, and chat). Our scoring leaders are equipped with a variety of tools to ensure every scorer is successful in understanding and applying scoring criteria to student responses. For a detailed explanation of DRC's remote scoring process, refer to Appendix B.

Experience

Of the assessments from other clients' programs that continued to be administered and scored during spring 2020 and winter 2021, DRC successfully utilized over 900 DRC scoring professionals working remotely to meet both timelines and quality expectations. The successful transition from site-based to remote scoring was made possible by leveraging existing tools with modified and enhanced procedures to provide our teams the needed resources and support. Our team looks forward to collaborating with LDOE to refine and modify existing remote scoring processes to reflect their unique requirements.

System Tools – Scoring, Training, Chat

ScoreBoard is DRC's secure, web-based scoring application that is designed to be used in a distributed environment. Our platform is used within our scoring centers and in remote locations (e.g., in a scorer's home). Our integrated training resources provide the capability to securely maintain digital training materials within the scoring platform itself.

Live, interactive training is conducted via Moodle Learning Management System, which mirrors aspects of the scoring room and provides a versatile platform for training. It also serves as a place to share files of important documents such as daily scoring statistics, selected training materials, and platform user guides. Through embedded communication tools, Scoring Directors, Assistant Scoring Directors, and Team Leaders are able to facilitate group or one-on-one training sessions and discussions using audio and video.

Zulip is the chat tool used in conjunction with ScoreBoard and Moodle to facilitate instant communication between Scoring Directors, Assistant Scoring Directors, Team Leaders, and Scorers. Zulip provides a tool for scorers to be able to directly ask supervisors questions about responses and allows supervisors to direct individual or groups of scorers to join Moodle training rooms for important discussions and retraining.

Content Training with Moodle

Content training remains an interactive, comprehensive, hands-on experience. Scoring Directors train each scoring group by screensharing PDFs of training materials as they progress through training. Each training example is displayed individually, and supervisors are able to use text highlighting, etc., to draw scorers' attention to relevant parts of the responses. Throughout the training, supervisors continue to guide the discussion, and scorers continue to be able to pose questions to supervisors. All secure materials such as passages/sources, anchors, training sets, and/or qualifying sets are accessible for scorers and Team Leaders in ScoreBoard, which does not permit anything to be downloaded or printed. Scorers are not permitted to download, print, or take screenshots of any confidential materials, including test items and student responses. Supplemental documents that are not secure, such as the ELA writing task rubrics and nonscore definitions may be located in Moodle where users have the capability to download or print. The Scoring Director directs the Team Leaders and scorers to take their training and qualifying sets, following the same training flow as they would in the scoring facility. This is described in the Content-Specific Training section below (see Appendix B for a detailed description of tools and procedures used in remote training/scoring).

Quality Control

Our robust quality control processes and handscoring metrics (detailed later in this document) remain in place for all projects scored remotely. Scored responses are monitored with second reads exactly as they are at the scoring sites. Read-behinds are also conducted in the exact same manner; however, any conversations and/or retraining needed as a result of the monitoring are held in one-on-one video chat sessions. DRC scoring leadership has found this to be a very effective and efficient way to adjust any training and clarify scoring decisions for scorers. Handscoring quality reports continue to be available for all projects on a regular basis for both project leadership and LDOE.

Content-Specific Training

In preparation for the scoring of all LEAP 2025 items, DRC scoring supervisors will train readers using the same content-specific training materials that were used for prior administrations of the same items. These training materials originated from the sources noted below.

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 11-13)
- LEAP 2025 Biology and U.S. History, as well as select items for Algebra I and Geometry (noted as DRC Material Type on pages 10-11)

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 source/stimuli for applicable tests and item types;
- Rubrics;
- Anchor Sets;
- Training Sets (or Practice Sets); and
- Qualifying Sets.

DRC will start the training with a review of passages/sources, 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 the spring 2021 administration of the LEAP 2025 grades 3-8 and high school assessments.

Training Materials

LEAP 2025 Biology, U.S. History, and Grades 3-8 Science and Social Studies

Reader training for LEAP 2025 Biology, U.S. History, and grades 3-8 science and social studies is conducted using item-specific anchor sets, training sets, and qualifying sets developed by DRC.

Set Type	**Biology, U.S. History, and Grades 3-8 Science and Social Studies Training Materials**	**Annotated**
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 ● 10 responses per training set ● All numeric score points are represented*	No
Qualifying Sets	There are two qualifying sets for each item ● 10 responses per qualifying set ● All numeric score points are represented*	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 identify examples of these scores during live scoring to supplement later reader retraining/recalibration as needed.		

LEAP 2025 Algebra I, Geometry, and Grades 3-8 Math (Items and Materials Developed by DRC)

Training materials for math items developed and field tested by DRC are made up of item-specific anchor sets, training sets, and qualifying sets developed by DRC.

Set Type	Algebra I, Geometry, and Grades 3-8 Math Training Materials	Annotated
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 ● 10 responses per training set ● All numeric score points are represented	No
Qualifying Sets	There are three qualifying sets for each item ● 10 responses per qualifying set ● All numeric score points are represented	No

LEAP 2025 Algebra I, Geometry, English I, English II, and Grades 3-8 ELA and Math (Items and Materials Developed by PARCC)

DRC will use the PARCC-approved training and qualifying materials provided by New Meridian for all English I, English II, and grades 3-8 ELA items as well as for the Algebra I, Geometry, and grades 3-8 math items not developed by DRC. Training materials for each item can be grouped into one of two categories: "prototype" item materials or "abbreviated" item materials. [Note: 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 that DRC uses for these items is the same process as outlined below for PARCC-approved math "prototype" items.]

Prototype Item Materials

PARCC selected one item that was representative of each PARCC task type to serve as a prototype item. For each prototype item, full sets of training materials were developed which consist of Anchor Sets, Practice Sets, and Qualifying Sets. DRC will start the training with a review of prototype 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. If the prototype is not one of the items included on the current Louisiana form, qualified readers will complete their training using abbreviated item training materials for the item that they will score as described below.

Abbreviated Item Materials

Unlike prototype items, abbreviated 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 the training materials for an associated prototype item that is similar to the abbreviated one they will be scoring on the Louisiana form.[1] 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, the actual item, they are going to score. This consists of an item-specific Anchor Set and two item-specific Practice Sets. After completing the training for the abbreviated item, readers may begin scoring live responses for the item.

[1] 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 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

Set Type	Mathematics Prototype Item Training	Annotated
Anchor Set	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	Yes
Practice Set 2	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

Set Type	Mathematics Abbreviated Item Training	Annotated
Anchor Set	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	Yes
Practice Set 2	10 responses representing the range of responses	Yes

English I, English II, and Grades 3-8 ELA Training Set Composition

Set Type	English Prototype Item Training	Annotated
Anchor Set (for the RCWE and WE traits)	3 responses per score point ● Anchor Sets for prototype RST and LAT item training include scores for the combined trait Reading Comprehension and Written Expression (RCWE). ● Anchor sets for prototype NWT item training include scores for Written Expression (WE).	Yes
Anchor Set (for the Knowledge and Use of Language Conventions trait)	● There are 3 responses per score point in each set. ● 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. ● Subsequent Practice Sets for prototype and abbreviated items will require readers to practice scoring the Knowledge and Use of Language Conventions trait along with the RCWE trait (for LAT or RST) or with the WE trait (for NWT). ● In addition, readers will be required to qualify on the Knowledge and Use of Language Conventions trait during each prototype item qualifying session.	Yes
Practice Set 1	5 responses representing the range of responses for ● the RCWE trait (for LAT and RST items) ● the WE trait (for NWT items)	Yes
Practice Set 2	5 responses representing the range of responses for the Knowledge and Use of Language 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 (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 the 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.		

English I, English II, and Grades 3-8 ELA Training Set Composition (continued)

Set Type	English Abbreviated Item Training	Annotated
Anchor Set (for the RCWE and WE traits)	3 responses per score point ● Anchor Sets for abbreviated RST and LAT item training include scores for the combined trait Reading Comprehension and Written Expression (RCWE). ● Anchor Sets for abbreviated NWT item training include scores for Written Expression (WE).	Yes
Practice Set 1	10 responses representing the range of responses for both traits appropriate to the task type (the two traits appropriate to LAT and RST items are RCWE and Knowledge and Use of Language Conventions; the two traits appropriate to NWT items are WE and Knowledge and Use of Language Conventions)	Yes
Practice Set 2	10 responses representing the range of responses for both traits appropriate to the task type (the two traits appropriate to LAT and RST items are RCWE and Knowledge and Use of Language Conventions; the two traits appropriate to NWT items are WE and Knowledge and Use of Language Conventions)	Yes

Algebra I Items and Associated Training Materials

Question	Form	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
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	BR (AE)	901832	3031 M44083P	Abbreviated	3003_M43111
15	BR (AE)	901882	VH196970	Abbreviated	VH046614
28	BR (AE)	901687	2407_M41752_AT	Prototype	N/A
29	BR (AE)	938737	MA10139 (DRC ID)	DRC	N/A
43	BR (AE)	901851	M41726	Abbreviated	3003_M43111
44	BR (AE)	901705	VF883359_AT	Abbreviated	VH046614
45	BR (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	E	902012	M41169	Abbreviated	VF935309
15	E	980937	M43798	Abbreviated	2904-M43021
25	E	980929	M1000516	Abbreviated	2904-M43021
28	E	902042	3020-M44058	Abbreviated	3042-M44133
43	E	980930	M1000518	Abbreviated	2904-M43021
44	E	980938	M100106	Abbreviated	VF935309
45	E	980936	VH239429	Abbreviated	2904-M43021
13	BR (AE)	902012	M41169	Abbreviated	VF935309
15	BR (AE)	902046	M46668	Abbreviated	3042_M44133
27	BR (AE)	902027	M43233	Abbreviated	VH001716
28	BR (AE)	902042	3020-M44058	Abbreviated	3042-M44133
43	BR (AE)	902062	VH150384	Abbreviated	VF613786
44	BR (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.					

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	868619	M00848	Prototype	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	981827	0318-M01475	Abbreviated	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.				

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	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.				
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.				

English I Items and Associated Training Materials

Question	Form	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
9	E	RST	914552	GG431834057	Abbreviated	VH017542 2T
14	E	NWT	983215	GG604245591	Abbreviated	6139
9	A (SR)	RST	902161	VH017542_2T	Prototype	N/A
14	A (SR)	NWT	906152	VH084830	Abbreviated	6139
*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.						

English II Items and Associated Training Materials

Question	Form	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
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
*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 Items and Associated Training Materials

Grade	Question	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
3	7	RST	915227	A1598	Abbreviated	VF906000
	12	NWT	913497	AA431426588	Abbreviated	VF910093
4	7	LAT	913567	VH170170	Abbreviated	VF925727
	20	RST	982233	VH060330	Abbreviated	VF653524
5	7	LAT	801310	VF821667	Abbreviated	VF882724
	20	RST	915510	VH198972	Abbreviated	2208
6	9	RST	913715	DD502035970	Abbreviated	3538
	14	NWT	913694	D1466	Abbreviated	VH000592
7	9	RST	915582	E1567	Abbreviated	VH014400
	14	NWT	913842	EE430133306	Abbreviated	4284
8	7	LAT	913958	F1460	Abbreviated	5271
	20	RST	982327	FF506834510	Abbreviated	VH007336
*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.						

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 2021 administration of the LEAP 2025 items as were established when these items were scored previously.

LEAP 2025 Constructed-Response and Extended-Response Items

For all LEAP 2025 ELA and math CR items, DRC will follow the same qualification standards determined by PARCC. A description of these qualifying standards is below.

LEAP 2025 English I, English II, and Grades 3-8 ELA

Test	**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 a reader's scores 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

Test	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

Test	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

Test and Item Type	Qualifying Standard
U.S. History and Grades 3-8 Social Studies 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.
U.S. History and Grades 5-8 Social Studies 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

Test and Item Type	Qualifying Standard	
Biology and Grades 3-8 Science 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.
Biology and Grades 3-8 Science 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.
Grades 3 and 4 Science 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.
Biology and Grades 5-8 Science 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 composite (multi-part) 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 2021 Scoring Plan

The charts below provide an overview of the Spring 2021 LEAP 2025 scoring plan, detailing the types of scoring that will be done for each course/grade.

LEAP 2025 High School

Test	Handscoring Only	AI Scoring	AI Vendor
LEAP 2025 English I	RST_VH017542_2T (Form A – AE) NWT_VH084830 (Form A –AE)	NWT_GG604245591 (Form E) RST_GG431834057 (Form E)	Pearson
LEAP 2025 English II	NWT_7064 (Form A – AE) RST_VH004490 (Form A – AE)	NWT_HH432845949 (Form E) RST_HH607742252 (Form E)	Pearson
LEAP 2025 Algebra I	All CRs	N/A	
LEAP 2025 Geometry	All CRs	N/A	
LEAP 2025 Biology	All CRs and ERs	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.			
* DRC's handscoring teams will provide a second read for at least ten percent of all AI-scored responses.			

LEAP 2025 Grades 3-8

Test	Handscoring Only	AI Scoring*	AI Vendor
ELA grade 3	Both PCRs	N/A	
ELA grade 4	Both PCRs	N/A	
ELA grade 5	RST_ VH198972/915510	LAT_VF821667/801310	Pearson
ELA grade 6	N/A	Both PCRs	Pearson
ELA grade 7	N/A	Both PCRs	Pearson
ELA grade 8	N/A	Both PCRs	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 the LEAP 2025 U.S. History ER and grades 5-8 Social Studies 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 5-8 ELA, 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 25.) 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 2021 Scoring Plan) will be provided by DRC's handscoring team. The score associated with the first scorer 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 and Use of Language Conventions score is not weighted.

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 approximately every 10 readers. (When test numbers are low and smaller groups totaling 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 using PARCC-approved validity responses supplied by New Meridian (for ELA and math) as well as LDOE-approved validity responses identified by DRC scoring supervisors for DRC-developed math items and WestEd-developed Biology, U.S. History, and grades 3-8 Science and Social Studies items. The validity responses that will be used in spring of 2021 are the same ones that were used when these items were previously administered and scored by DRC.

The validity responses will be added to DRC's image handscoring system prior to the beginning of 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 group 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 samples on pages 23-24) 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 the LEAP 2025 U.S. History and grades 5-8 Social Studies operational ERs. For select CRs in LEAP 2025 English I, English II, and grades 5-8 ELA (see Spring 2021 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 27. 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 in the table, Agreement Rate Expectations for Validity and Inter-Rater Reliability, on page 22. 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 LEAP 2025 items are shown below.

Agreement Rate Expectations for Validity and Inter-Rater Reliability – LEAP 2025			
Content Area/Course	**Score Point Range**	**Perfect Agreement**	**Perfect Agreement + Adjacent**
English I, English II, Grades 3-8 ELA	0-3 or 0-4 Rubric, Multi-trait	65% (each trait)	96% (each trait)
Algebra I, Geometry, Grades 3-8 Math	0-1 Rubric	90%	95%
Algebra I, Geometry, Grades 3-8 Math	0-2 Rubric	80%	95%
Algebra I, Geometry, Grades 3-8 Math	0-3 Rubric	70%	95%
Algebra I, Geometry, Grades 3-8 Math	0-4 Rubric	65%	95%
Biology, Grades 3-8 Science CR items	0-2 Rubric	80%	95%
Biology, Grades 3-8 Science 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%
Grades 3 and 4 Science Comprehensive (single part) ER items	0-6 Rubric	60%	93%
Biology, Grades 5-8 Science Comprehensive (single part) ER items	0-9 Rubric	60%	90%
U.S. History, Grades 3-8 Social Studies CR items	0-2	80%	95%
U.S. History, Grades 5-8 Social Studies 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 table above.

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 were chosen by DRC scoring supervisors during live scoring or supplied by New Meridian (for ELA and math). After the readers score one of these calibration sets (usually 5-10 responses), the Scoring Director will review the set with the readers using rubric language and scoring concepts exemplified by 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 readers in need of additional clarification based on their performance.

Handscoring Quality Control Reports

The Scoring Summary reports show inter-rater reliability data and score point distribution information for each item (by part where appropriate).

Scoring Summary Report Sample

Grade 10 Biology Q000 — **Part B**

		Inter-rater Reliability				Score Point Distribution													
Totals		2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%4	%5	%6	%B	%F	%I	%N	%R	%U
	Current Handscore	6,260	79	17	4	24,197	31	20	19	10	8	2	3	0	0	4	1	0	1
	3426	53	81	16	3	241	32	22	22	9	10	2	1	0	0	2	0	0	0
	3468	101	77	19	4	500	37	23	19	7	7	2	1	0	0	2	0	0	1
	3556	70	82	16	2	363	35	20	19	8	11	2	2	0	0	2	0	0	1

Scoring Summary Report Sample with AI (Reader ID #3)

Grade 10 English II Q000 **Conventions**

Inter-rater Reliability **Score Point Distribution**

Totals

	2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%B	%F	%N	%R	%T	%U
Current Handscore	**636**	**87**	**13**	**0**	**2,440**	**62**	**26**	**3**	**0**	**0**	**0**	**3**	**0**	**0**	**3**
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	0	6	0	0
16132	76	83	17	0	76	64	33	0	0	0	0	3	0	0	0

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				●			●		

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.

The nonscore codes and the tests 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 Test

Test	B	F	I	N	R	T	U
LEAP 2025 Algebra I, Geometry, English I, English II, 3-8 ELA, and 3-8 Math	✓	✓	N/A	✓	✓	✓	✓
LEAP 2025 Biology, 3-8 Science, U.S. History, and 3-8 Social Studies ERs and CRs	✓	✓	✓	✓	✓	N/A	✓

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 processing of final alerts, see *Alerts* section on page 26).

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, test, 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 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 5-8 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 2021 administration. The AI scoring models were built by MI and followed the model-building process described below. (Model-building data for all items included on the spring 2021 test may be found in the Appendix.)

Test	Item Type	IDEAS ID	Model Built
LEAP 2025 U.S. History	ER	892955	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
*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. **The original 2016 model for grade 5 ER 807773 was similarly augmented prior to operational scoring in spring 2019 using a targeted sample of spring 2019 responses.			

Model Building

To build the model, PEG analyzed a set of inputs that were randomly pulled from the training set itself, which was 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 useful for determining the accuracy/suitability of the response 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[2] 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, therefore, 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.

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 κω(PEG, R1) and quadratic weighted kappa between Reader 1 and Reader 2 is referred to as κω(R1, R2).

---

[2] 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.

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 κω(R1, R2) will usually lead to a low κω(PEG, 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 κω(PEG, R1). In cases of four or more levels (e.g. a score ranging from 1-4, or broader) a κω(PEG, R1) of 0.7 has become a rule of thumb as a go-no-go metric. In these broader scales, a κω(PEG, 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 κω(PEG, R1) and κω(R1, R2) can be used as an additional evaluation metric. MI defines that value as follows:

$$\Delta\kappa = \kappa\omega(\text{PEG, R1}) - \kappa\omega(\text{R1, R2})$$

When Δκ is positive, PEG's scores are more in agreement with Reader 1 than Reader 1's scores are in agreement with Reader 2. When Δκ 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 Δκ maintains its weight as a relative value between the two kappa values. That is, a larger Δκ 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 ≥ 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.

Δκ 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, κω and Δκ 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 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 Appendix C.

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 page 28 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; 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, any U.S. History and Social Studies grades 5-8 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 receive a score of "I."

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 the LEAP 2025 U.S. History and Social Studies grades 5-8 ER item, this includes the prompt and associated source material (including MC/MS items) provided with the item. DRC also pre-identified websites that may be likely sources of external plagiarism. These include Wikipedia 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. These text references have been added to the scoring model.

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–3b below.

2. 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

3. 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. For 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 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 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 2021 administration of LEAP 2025. (Model-building data for all items included on the spring 2021 test may be found in the Appendix.)

Test	Task Type	IDEAS ID	PARCC UIN	Model Built
English I	NWT	983215	GG604245591	2021
English I	RST	914552	GG431834057	2018
English II	NWT	983642	HH432845949	2017
English II	RST	983688	HH607742252	2019
Grade 5 ELA	LAT	801310	VF821667	2021
Grade 6 ELA	RST	913715	DD502035970	2017
Grade 6 ELA	NWT	913694	D1466	2017
Grade 7 ELA	NWT	913842	EE430133306	2017
Grade 7 ELA	RST	915582	E1567	2021
Grade 8 ELA	LAT	913958	F1460	2017
Grade 8 ELA	RST	982327	FF506834510	2021

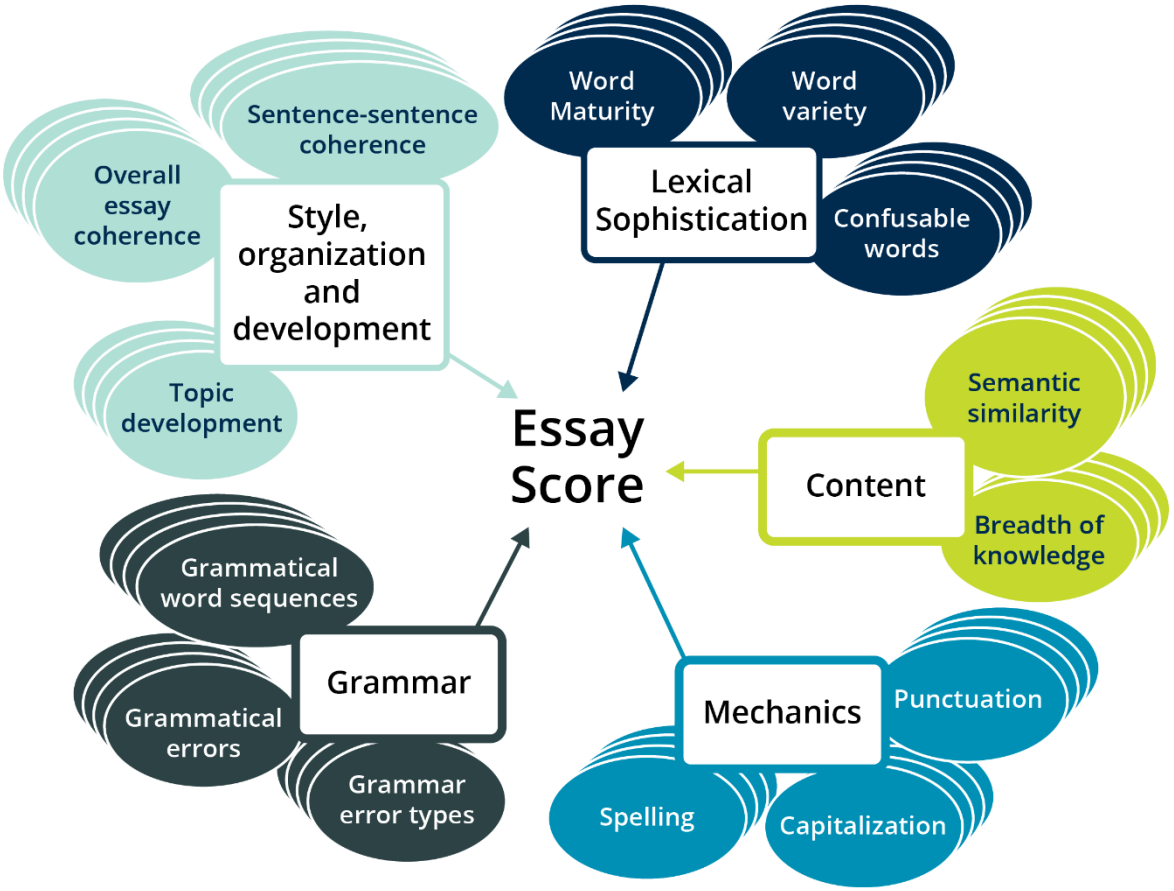
The Intelligent Essay Assessor

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 even though they have no words in common:

- 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 most of the ELA prompts that will be scored using AI, 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.

Evaluation of Automated Scoring Systems	
Criterion	**Threshold**
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 0.0525

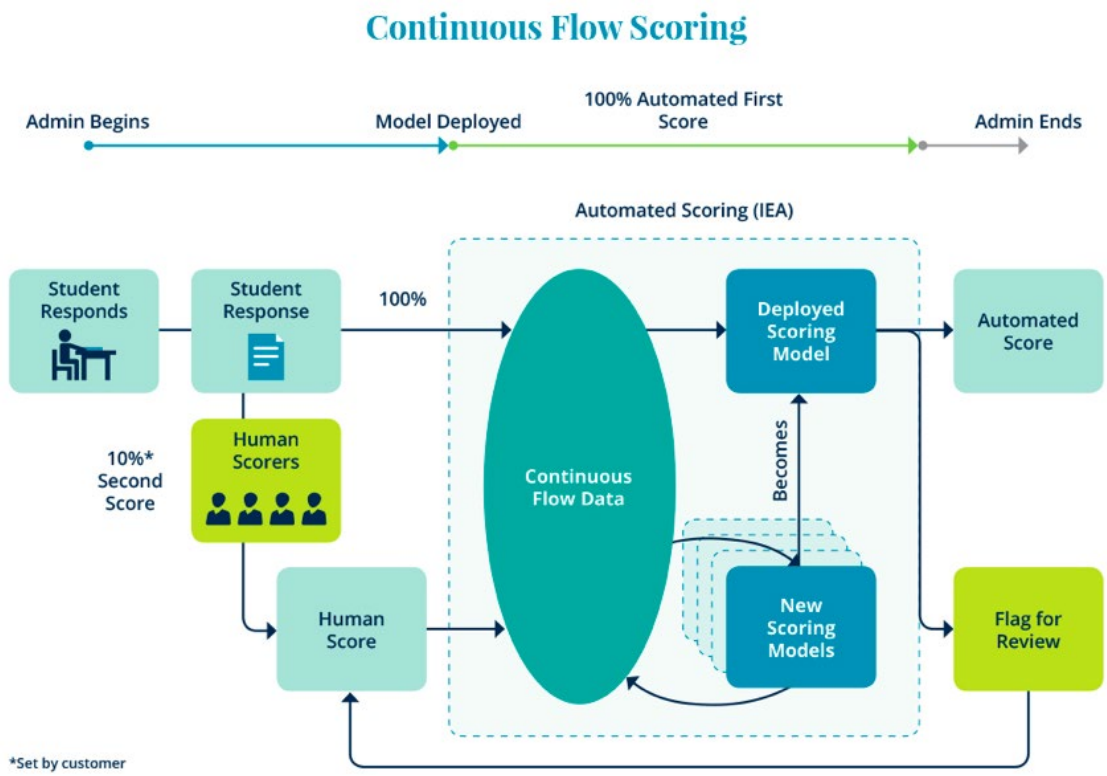
Evaluating Automated Scoring. *Statistical Criteria for the Evaluation of Automated Scoring Systems based on those used by Williamson et al, Smarter Balanced, and PARCC.*

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.

For the four prompts that were trained in 2021 (see prompt table on page 36), responses and DRC human scores from the spring 2019 administration of LEAP 2025 (rather than a prior PARCC administration) provided the inputs to training. A sample of approximately 6,000 responses representing the operational score distribution was selected for each prompt. Approximately two-thirds of those responses were used to train IEA and the remaining one-third were held out for evaluation. Performance on the evaluation set was measured using the same criteria as the PARCC-based prompts.

Responses for which IEA is less confident in its scores 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 below depicts the entire Continuous Flow process.



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. Disturbing content 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 2021 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 Processing Rules – Scoring.xlsx* on the LDOE Reporting SFTP site at /<2021> - 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 35

SECTION 2 – SCHEMA SUPPLEMENT 36-38

SECTION 3 – STATUS CODE INFORMATION 39

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 with 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	*Required.*

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	*Required.*

2.2.2 ConstructedResponseList

Name	Type	Min	Max	Description	Range
ConstructedResponse	ConstructedResponse	0	*unbounded*	List of individual CRs to be scored	*Required.* 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	*Required.*

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	*Required.* Will fall in the range 0-999. See **section 3** for more information
Description	string	0	1	Short description of status	*Required.* See **section 3** 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

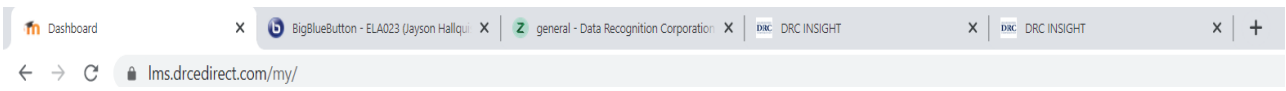
DRC Distributed Scoring Process

Web-based Platforms used	Description
ScoreBoard	Scoreboard is the same application used to handscore student responses.
Moodle	Moodle is a Learning Management Tool used by DRC as the interactive piece of remote training and scoring through the Big Blue Button application.
Zulip	Zulip is the chat tool used in conjunction with Scoreboard and Moodle to facilitate instant communication between Scoring Directors, Team Leaders, and Scorers. It is mainly used once training is complete and live scoring has begun.

LDOE will have login access to ScoreBoard and Moodle to be able to join reader training sessions.

Remote Technology Orientation

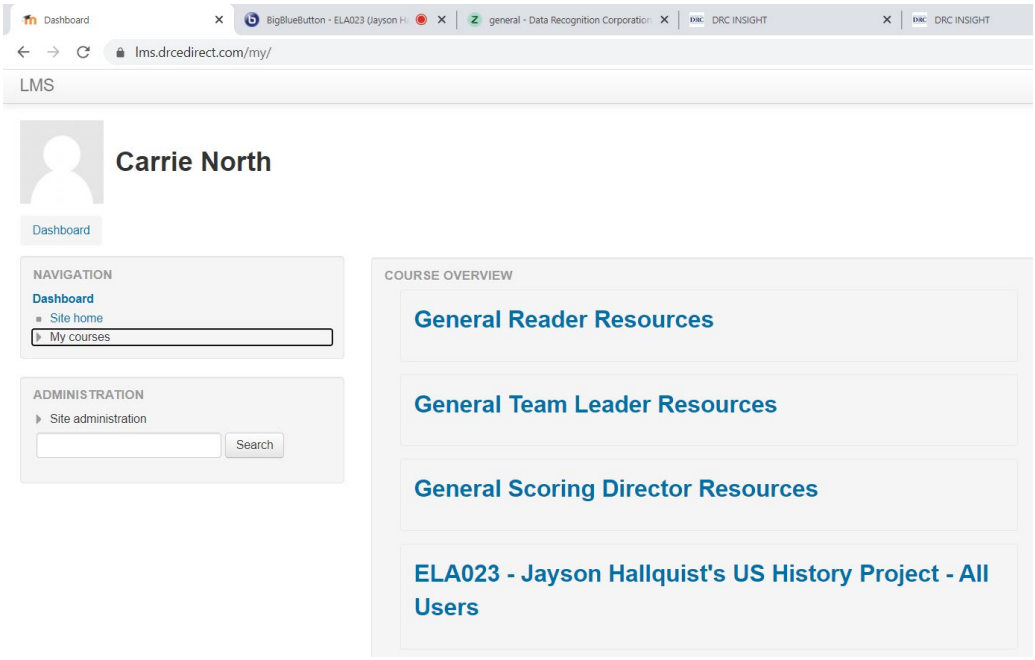
Supervisors and scorers go through a very thorough structured and layered remote technology orientation training of the DRC remote scoring process before beginning training on content. Scoring Directors (SDs) go through the training first, then Team Leaders (TLs), and finally scorers. This training is focused on security, the different platforms, and how to use them for remote scoring. Scoring from home is seen as an extension of the company, and a quiet area away from distractions and others is required. An entire day of the training is allotted to work out any technical issues and practice with the different applications. Users will work in the Chrome browser by navigating through the different tabs that they will keep open during a typical workday - Moodle Dashboard/Big Blue Button session, the Zulip chat tool, and two tabs of ScoreBoard (one for the Training/Qualifying/Recalibration application [TQR], with sets to take and notes to reference, and another one for scoring). The screenshot below shows these tabs (from left to right: Moodle, Big Blue Button, Zulip, and the two DRC INSIGHT tabs for TQR and scoring).



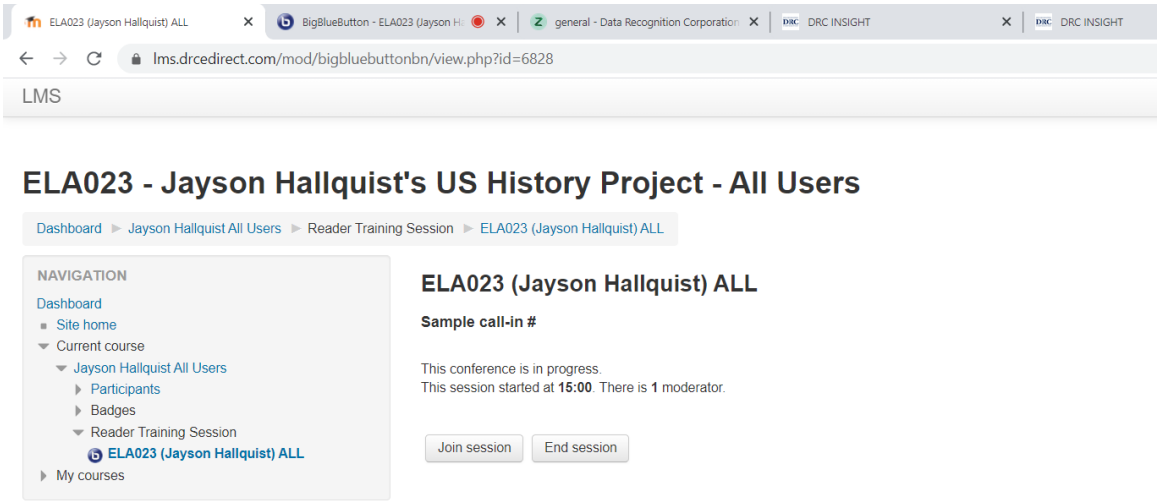
Moodle and Big Blue Button

Users are engaged in training inside the general training program known as Moodle; however, most interactive parts of the training process happen within the Big Blue Button (BBB) application. Big Blue Button is a "plug-in" to Moodle that allows everyone signed in to the session to hear the presenter's voice, ask questions that can be heard by the group, share screens, and "chat." Most interactive training will occur only after users click the "Join Session" button in Moodle to open a Big Blue Button interactive training session.

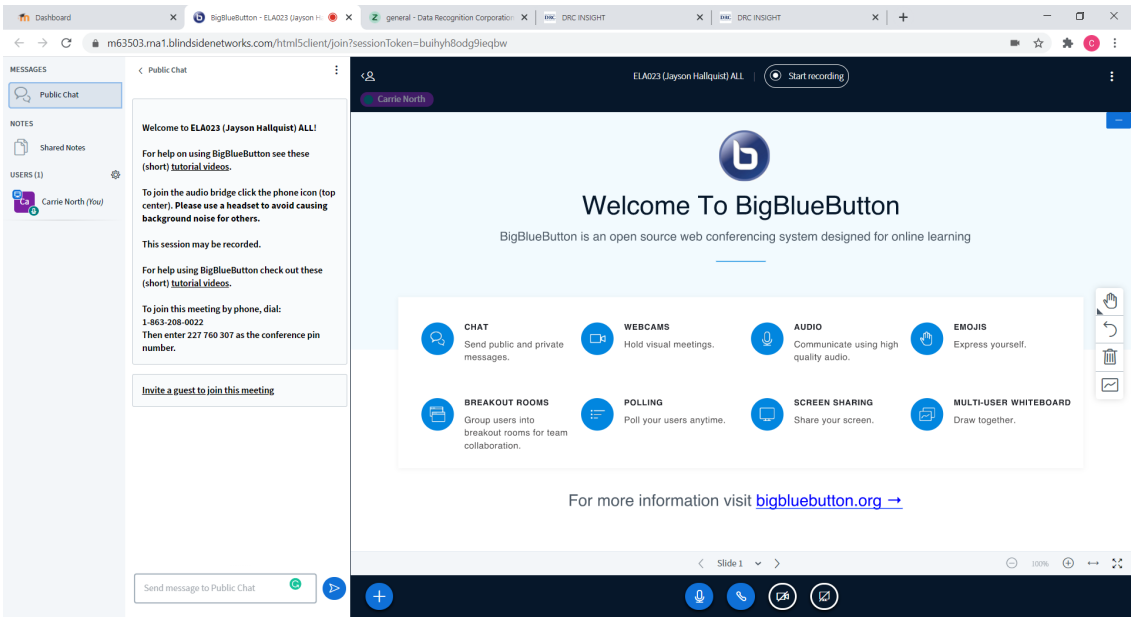
Below is a screenshot showing the Moodle dashboard of a user who is assigned to a U.S. History project.



After the user selects the course for the project (in this case, ELA023 – Jayson Hallquist's US History Project – All Users), the following screen with the "Join Session" button will become available.



After joining the session, the user will be brought to the Big Blue Button screen where the interactive portion of the session will begin once the session leader has determined that all parties have joined.



Moodle is frequently used in college classrooms in a manner similar to DRC's use; like a professor leading a class, our Scoring Directors lead our group training sessions and guide the ongoing learning process. Moodle mirrors aspects of the scoring room and provides a versatile platform for training. It serves as a place to share files of important documents such as scoring statistics, non-secure training

materials (e.g. nonscore definitions), and application help manuals. Moodle also provides a communication tool for SDs/TLs to host discussions with scorers. Through their Moodle Dashboard (the home screen they see upon logging into Moodle), users can navigate through different courses created to reflect a structure to the scoring room. Only users who are assigned to specific courses, including Big Blue Button sessions, may see what is visible in terms of course materials. The Scoring Director's - All Users course is similar to the entire scoring room, the Scoring Director's - Supervisor course is where Team Leader meetings are conducted, and the Team Leader's breakout room is where TLs go to have one-on-one discussions with their scorers or work in small groups with their team.

Content Training with Moodle and Big Blue Button

Scoring Directors will train the group within the Moodle Big Blue Button by screensharing PDFs of training materials as they progress through training. This ensures the audience has the clearest images of the training materials. Scorers are not permitted to download, print, or take screenshots of any confidential materials. All secure materials such as sources, rubrics, anchors, training sets, and/or qualifying sets will only be accessible to scorers and Team Leaders in ScoreBoard TQR (part of DRC's secure scoring platform), which does not allow anything to be downloaded or printed. (A copy of DRC's Scoring Security and Confidentiality Agreement, which all scorers must sign, can be found on the next page.) Supplemental documents that are not secure, such as nonscore definitions, will be located in Moodle where users may have the capability to download or print. When appropriate in the training, the Scoring Director will direct Team Leaders and scorers to take their TQR sets, following the same training flow as they would in a DRC scoring facility.

DRC Remote Test Scoring Security and Confidentiality Agreement

I understand that, as a Remote Test Scorer and employee for Data Recognition Corporation, the materials I work with (scoring rubrics, training materials, test questions, student responses) are secure and confidential. It is DRC's expectation that Remote Test Scorers score on devices using up-to-date operating systems.

I agree to the following terms:

All DRC technology, processes, records and information related to DRC and its customers are confidential and must be treated accordingly. DRC or DRC related information, including without limitation, documents, notes, files, records, oral information, computer files, or similar materials may not be saved, duplicated or removed from DRC premises or systems without permission from DRC. Additionally, the contents of DRC's records or information otherwise obtained regarding business may not be disclosed to anyone, except where required for a business purpose. Employees must not disclose any confidential information, purposefully or inadvertently, through casual conversation, with any unauthorized person inside or outside DRC at any time while employed, between projects or after termination of employment. Employees who are unsure about the confidential nature of specific information must ask their manager for clarification.

By signing below, I agree that:
- all training materials and student responses are the property of DRC.
- requests for information about particular projects are referred to DRC management.
- commenting on the content of items (test questions) or responses with non-project related personnel is prohibited.
- reproducing, in part or in whole, through means including but not limited to printing, taking pictures, downloading, or capturing screen shots of student responses, test questions, or training materials is expressly prohibited.
- the privacy of the students whose work I evaluate is to be respected, and all related data is to be protected from disclosure.
- I will work in a private environment, separate from others and free from distractions.
- I will be the only one to read and score student responses that have been assigned to me.
- I will adhere to the criteria defined by the rubric and training that I receive.
- during work hours, I will only use my cell phone to contact DRC support.
- I will not discuss test questions, student responses, and training materials with anyone except my Team Leader and Scoring Director.
- I will not share test questions, student responses or training materials on any media, including social media.
- I will score only on a
    - laptop or desktop; not on a cell phone or tablet.
    - device using a current and supported version of a Chrome browser.

Furthermore, I understand that violation of any of these security and confidentiality policies will be subject to appropriate disciplinary actions, up to and including termination of my employment with Data Recognition Corporation.

_____
Name Printed

_____
Signature

_____
Date

Remote Reader Training (Scorers)

- When scorers first log in to Moodle, they will enter the **Scoring Director's - All Users** course and join the **Reader Training** Session. They will also be assigned to a Team Leader, which they will see under their Moodle Dashboard. Each Team Leader will be assigned no more than 8 scorers for remote scoring. Scorers must be present each day of the training. The Big Blue Button set-up allows moderators of those sessions to track attendance via name/phone number sign-ins. All supervisors will be required to take attendance during the larger sessions as well as full team sessions.
- Scoring Directors will begin the training within Moodle and the Big Blue Button application with a thorough review of the scoring rubric, the prompt, sources, and annotated anchor papers.
- When scorers are ready to begin the first training set, the SD will explain the purpose of training papers, reinforcing the importance of using the Anchor set while assigning scores. The SD will instruct scorers to take each training set by navigating to their TQR tab within ScoreBoard which will record their scores electronically.
- Scorers will be given a suggested timeframe in which to finish each set.
- Scorers will electronically submit their scores upon completion and will be asked to use Zulip to inform their TL when sets are complete. After all the scores are submitted, reports will be generated. SDs will analyze group-wide trends and use that information to guide review within the Moodle Big Blue Button Session with all users.
- After the completion of each training set by scorers, the SD will announce the true score for each paper and discuss each response.

Zulip Chat Tool

Once content training is complete, scorers will use the ScoreBoard tab set up for standard scoring to score student responses, just as they do when scoring on-site in a DRC scoring facility. They will also continue to have access to the sources, rubrics, anchors, training sets, and/or qualifying sets through the use of a second ScoreBoard tab opened to TQR. They will be in contact with their Team Leader and Scoring Director through the Zulip chat tool throughout the course of their day.

Zulip is the chat tool used in conjunction with ScoreBoard and Moodle to facilitate instant communication between Scoring Directors, Team Leaders, and scorers. The first day of remote technology orientation is used to set up accounts for this instant messaging communication. All users will keep a separate Zulip tab open and navigate back to the Moodle and ScoreBoard tabs accordingly. While in the ScoreBoard tab or Moodle tab, they will be able to receive notifications of Zulip communication. Each user will be enrolled in Zulip "streams," which are group message forums with set, pre-designated enrollments. All stream messages are seen by all who are enrolled in a particular stream and anyone enrolled can post a message. Streams will be organized much like Moodle. There will be a

Stream for All Users, representing the Scoring Director's entire group, one for **Supervisors**, representing the Scoring Director and Team Leaders, and **Team** streams for the Team Leaders and their scorers.

- Zulip is used for short messages, unrelated to content, such as:
  - TL attendance notification. Scorers should send a private message to their TL every morning when they first log in, when they log out and come back from a lunch break, and when they log off at the end of a shift ("I'm starting my shift" or "I'm logging out for the day"). If the TL has not received a message within 5 minutes of the scheduled shift start time, the TL will send a private message asking if the scorer is logged in and ready to score.
  - Scorers may ask their TL for scoring help by sending a private message to TL ("I don't know how to score lithocode #######"). As mentioned earlier, it is acceptable to reference lithocodes in Zulip but not specific response content.
  - Scorers will receive team stream messages from their TL, like morning greetings or Moodle team meeting notifications.
  - Scorers may receive a private message from their TL to set up a scoring conference in Moodle ("Meet me in the Moodle team room at 11 am to go over a few responses").
  - Scorers may receive group stream messages from the SD to announce a group Moodle meeting place and time, scoring stats posting, break/lunch time, or log-out time.
  - All users may receive company announcements in the **PAS Announcements** stream.

- Moodle should be used for group meetings such as training, individual and team scoring conferences, or any other content-related communication.

Remote Work Scheduling

Remote work will adhere to the same work-day hours as scoring within a scoring facility, with scorers working core hours of 8:30 AM - 4:00 PM. The schedule's purpose is to provide the structure maintained from scoring on site to remote scoring, which also ensures frequent and regular communication while following a group-wide schedule. All users will join the large group Moodle session after returning from each break (10:30 AM, 2:30 PM). This is meant to be practiced for the first few days of scoring to keep everyone on the same page, practice with systems, and keep communication flowing. After the first few days (when the Project Manager and Scoring Director determine that things are moving along), these extra times can be cut out.

Example of Daily Remote Schedule for Scorers

8:30 AM – Start of shift

- Use Zulip chat tool to notify TL of start
- Log into Unanet and record start time
- Log into Moodle
- Log into DRC INSIGHT Portal
    - Check ScoreBoard dashboard for messages
    - Review rubric and sources
    - Begin scoring

8:45 AM – Morning check-in

- Join Moodle session for morning announcements from SD
    - Morning check-in and review
    - Begin/Continue scoring

10:15 AM – Morning break

- Log out of ScoreBoard for morning break (chat message from SD via Zulip)

10:30 AM – Return from morning break

- Join Moodle session
- Return to scoring (log back into ScoreBoard)

Noon – Lunch break

- Log out of ScoreBoard for lunch break (chat message from SD via Zulip)

12:30 PM – Return from lunch break

- Use Zulip chat tool to notify TL of return to work after lunch break
- Join Moodle session
- Return to scoring (log back into ScoreBoard)

2:15 PM – Afternoon break

- Log out of ScoreBoard for afternoon break (chat message from SD via Zulip)

2:30 PM – Return from afternoon break

- Join Moodle session
- Return to scoring (log back into ScoreBoard)

3:55 PM – Final check-in before end of core hours shift

- Watch for SD's end of the day chat message via Zulip
- Continue scoring until end of shift

4 PM – End of shift

- Use Zulip chat tool to notify TL of ending shift
- Log into Unanet and record end time
- Log out of ScoreBoard and the DRC INSIGHT Portal for the day

Attendance Policy

DRC's attendance policy has not changed. Scorers need to be present for all training. Team Leaders are in constant communication (via Zulip) with scorers throughout the day to ensure they are present. If anyone is going to be late or absent, they are instructed to call Human Resources.

Appendix C

AI Model Data – LEAP 2025 U.S. History ERs (Spring 2021)

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	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 2021)

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[1]	78	21	1	Human	62	25	12	2	0	0.88	H to H[1]	79	20	1	Human	67	23	9	1	0
		≈500		H to H[3]	92	7	1	Human	3	29	48	17	3		H to H[3]	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[1]	67	32	1	Human	42	44	12	1	0	0.76	H to H[1]	68	31	1	Human	52	38	9	1	0
		≈500		H to H[2]	98	2	0	Human	7	28	50	14	1		H to H[2]	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[1]	73	25	2	Human	45	41	12	2	0	0.83	H to H[1]	73	25	2	Human	57	31	11	2	0
		≈500		H to H[2]	98	1	0	Human	9	18	39	26	8		H to H[2]	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[2]	90	9	0	Human	1	6	34	35	24		H to H[2]	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[1] – Human scored 2016 Field Test sample of ≈ 2500 responses per item.

H to H[2], H to H[3] – Human scored targeted samples of ≈ 500 responses per item were used to augment and retrain the original AI models from 2016. These samples came from spring operational responses and 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. H to H[2] augmentation sample was scored in spring 2017. H to H[3] augmentation sample was 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 5-8, English I, and English II (Spring 2021)

				IEA-Human Agreement					
Prompt	**QWK***	**Grade**	**Trait**	**Exact**	**SP0**	**SP1**	**SP2**	**SP3**	**SP4**
E05_L_VF821667	0.77	5	1					To be human scored OP 2021**	
	0.79		2						
E06_N_D1466	–	6	1						
	–		2						
E06_R_DD502035970	–	6	1						
	–		2						
E07_N_EE430133306	–	7	1						
	–		2						
E07_R_E1567	0.88	7	1						
	0.86		2						
E08_L_F1460	–	8	1						
	–		2						
E08_R_ FF506834510	0.81	8	1						
	0.81		2						
E09_N_ GG604245591	0.89	9	1						
	0.86		2						
E09_R_GG431834057	–	9	1						
	–		2						
E10_N_HH432845949	–	10	1						
	–		2						
E10_R_HH607742252	–	10	1						
	–		2						

*QWK data is noted for item models built by Pearson for DRC LA scoring. The PARCC program does not require QWK data to be saved and stored each year, so Pearson does not have QWK data retained in their archives for item models that were initially built for PARCC scoring. Pearson is required to meet the PARCC quality criteria and they are confident these items met these criteria.
**E05_L_VF821667 – Very few students received high scores on this prompt in 2019, so there were insufficient examples of 3s and 4s available to train the model (refer to Grade 5 ELA SPD history on page 73 of Appendix C). As a result, responses identified by IEA as possible score point 3s and 4s for this item will be sent to human scorers for scoring.

- 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 2021 LEAP 2025 Items – IRR and SPD from Previous Administrations

Algebra I

IDEAS ID	Spring 2021 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 %
980924	E	M44463	Pearson Spring 2017	77,183	Overall	0,1,2,3	14,754	88	99	37	15	30	11		7
980924	E	M44463	DRC Spring 2019 (E)	23,688	Overall	0,1,2,3	4,860	88	100	39	18	28	12		2
980909	E	M43216	Pearson Spring 2018	98,152	Overall	0,1,2,3	18,677	88	99	62	14	11	4		10
980909	E	M43216	DRC Spring 2019 (E)	22,672	Overall	0,1,2,3	5,506	91	100	65	14	9	5		7
980927	E	VH251952	Pearson Spring 2018	124,433	Part A	0,1,2	23,748	97	100	70	15	5			11
				124,433	Part B	0,1,2	23,748	95	99	72	8	7			14
				124,433	Part C	0,1,2	23,748	91	99	68	12	7			14
980927	E	VH251952	DRC Spring 2019 (D, E)	52,828	Part A	0,1,2	11,128	98	100	79	14	4			3
				52,828	Part B	0,1,2	11,128	95	100	80	9	7			3
				52,828	Part C	0,1,2	11,128	93	100	73	15	8			3
980927	E	VH251952	DRC Fall 2019	6,338	Part A	0,1,2	1,538	99	100	83	8	2			7
				6,338	Part B	0,1,2	1,538	98	100	84	5	4			7
				6,338	Part C	0,1,2	1,538	97	100	80	9	4			7
980927	E	VH251952	DRC Summer 2020	489	Part A	0,1,2	142	100	100	84	2	1			13
				489	Part B	0,1,2	142	100	100	83	2	2			13
				489	Part C	0,1,2	142	99	100	80	3	3			13
980927	E	VH251952	DRC Fall 2020	5,456	Part A	0,1,2	1,254	99	100	86	7	2			5
				5,456	Part B	0,1,2	1,254	97	100	85	6	4			5
				5,456	Part C	0,1,2	1,254	95	100	79	11	4			5
980911	E	2679-M43312	Pearson 2015 FT	1,799	Part A	0,1,2	402	95	100	71	12	3			14
				1,799	Part B	0,1,2	402	95	100	19	63	3			15
980911	E	2679-M43312	DRC Spring 2019 (E)	22,976	Part A	0,1,2	5,159	98	100	75	15	5			5
				22,976	Part B	0,1,2	5,159	97	100	26	64	5			5
901851	BR, E	M41726	DRC Spring 2018	52,490	Overall	0,1,2,3	11,918	92	100	57	14	15	8		6
901851	BR, E	M41726	DRC Fall 2018	6,011	Overall	0,1,2,3	1,556	96	100	66	11	9	4		9
901851	BR, E	M41726	DRC Spring 2019 (E)	23,087	Overall	0,1,2,3	4,712	95	100	60	10	16	10		3
901851	BR, E	M41726	DRC Summer 2019	2,100	Overall	0,1,2,3	532	99	100	86	3	2	0		9
938737	BR, E	MA10139	DRC Spring 2018, FT	1,582	Overall	0,1,2,3,4	382	94	100	71	12	4	2	5	7
938737	BR, E	MA10139	DRC Spring 2019 (D)	28,926	Overall	0,1,2,3,4	8,624	97	100	67	10	3	2	4	13
938737	BR, E	MA10139	DRC Spring 2019 (E)	23,125	Overall	0,1,2,3,4	6,328	95	100	63	12	5	3	6	10
938737	BR, E	MA10139	DRC Summer 2019	2,086	Overall	0,1,2,3,4	624	100	100	83	2	0	0	0	14
938737	BR, E	MA10139	DRC Fall 2019	6,237	Overall	0,1,2,3,4	1,946	98	100	68	9	2	2	2	16
938737	BR, E	MA10139	DRC Summer 2020	498	Overall	0,1,2,3,4	194	99	100	77	3	1	1	1	17
938737	BR, E	MA10139	DRC Fall 2020	5,398	Overall	0,1,2,3,4	1,584	98	100	70	9	3	2	3	15
980923	E	M000312	Pearson 2017 FT	1,593	Overall	0,1,2,3	264	89	100	65	15	8	6		6
980923	E	M000312	DRC Spring 2019 (E)	22,990	Overall	0,1,2,3	5,366	97	100	68	15	6	5		6

Form Key: Form BR = Administrative Error (AE), Form E = Operational

Algebra I (continued)

IDEAS ID	Spring 2021 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	BR	3031-M44083P	Pearson Spring 2016	95,907	Part B	0,1,2	18,835	91	100	30	45	12			13
901832	BR	3031-M44083P	DRC Spring 2018	55,162	Part B	0,1,2	10,236	91	100	82	47	21			0
901832	BR	3031-M44083P	DRC Fall 2018	6,329	Part B	0,1,2	1,140	92	100	51	40	9			0
901832	BR	3031-M44083P	DRC Spring 2019 (D)	29,986	Part B	0,1,2	5,394	92	100	35	42	23			0
901832	BR	3031-M44083P	DRC Summer 2019	2,199	Part B	0,1,2	404	97	100	68	28	4			0
901832	BR	3031-M44083P	DRC Fall 2019	6,523	Part B	0,1,2	1,180	94	100	49	36	15			0
901832	BR	3031-M44083P	DRC Summer 2020	503	Part B	0,1,2	88	93	100	64	29	6			0
901832	BR	3031-M44083P	DRC Fall 2020	5,655	Part B	0,1,2	1,034	93	100	45	42	13			0
901882	BR	VH196970	Pearson Spring 2016	9,586	Part A	0,1	1,950	98	100	71	13				16
				9,586	Part B	0,1,2	1,950	90	97	66	7	4			23
901882	BR	VH196970	DRC Fall 2017	8,522	Part A	0,1	1,940	99	100	94	3				4
				8,522	Part B	0,1,2	1,940	99	100	94	2	1			4
901882	BR	VH196970	DRC Spring 2018	50,072	Part A	0,1	10,654	99	100	90	8				2
				50,072	Part B	0,1,2	10,654	97	100	93	3	2			2
901882	BR	VH196970	DRC Summer 2018	1,625	Part A	0,1	372	99	100	97	0				3
				1,625	Part B	0,1,2	372	99	100	96	1	0			3
901882	BR	VH196970	DRC Fall 2018	9,092	Part A	0,1	1,940	99	100	94	3				4
				9,092	Part B	0,1,2	1,940	99	100	94	2	1			4
901882	BR	VH196970	DRC Spring 2019 (SR)	265	Part A	0,1	18	100	100	92	3				4
				265	Part B	0,1,2	18	100	100	95	0	0			4
901882	BR	VH196970	DRC Summer 2019	2,122	Part A	0,1	462	100	100	96	0				4
				2,122	Part B	0,1,2	462	100	100	95	1	0			4
901687	BR	2407-M41752	DRC Spring 2018	53,117	Part A	0,1,2	11,413	98	100	74	3	19			4
				53,117	Part B	0,1,2	11,413	96	100	83	7	6			4
				53.117	Part C	0,1,2	11.413	98	100	89	4	3			4
901687	BR	2407-M41752	DRC Spring 2018	6,022	Part A	0,1,2	1,470	99	100	80	2	10			7
				6,022	Part B	0,1,2	1,470	99	100	87	3	2			7
				6,022	Part C	0,1,2	1,470	99	100	90	1	1			7
901687	BR	2407-M41752	DRC Summer 2019	2,114	Part A	0,1,2	530	100	100	90	0	2			8
				2,114	Part B	0,1,2	530	100	100	91	1	0			8
				2,114	Part C	0,1,2	530	100	100	91	0	0			8
901705	BR	VF883359	DRC Spring 2018	53,281	Part A	0,1,2,3	11,808	98	100	89	4	1	2		5
				53,281	Part B	0,1	11,808	93	100	84	11				5
901705	BR	VF883359	DRC Fall 2018	6,097	Part A	0,1,2,3	1,570	100	100	87	2	1	2		8
				6,097	Part B	0,1	1,570	98	100	94	7				8
901705	BR	VF883359	DRC Summer 2019	2,104	Part A	0,1,2,3	508	100	100	90	2	0	0		8
				2,104	Part B	0,1	508	100	100	89	3				8

Form Key: Form BR = Administrative Error (AE), Form E = Operational

Algebra I (continued)

IDEAS ID	Spring 2021 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 %
901857	BR	VH046479	Pearson Spring 2017	78,418	Part A	0,1,2	13,963	88	100	51	36	3			10
				78,418	Part B	0,1	13,963	92	100	69	19				12
901857	BR	VH046479	DRC Fall 2017	8,686	Part A	0,1,2	2,258	94	100	77	13	1			9
				8,686	Part B	0,1	2,258	97	100	86	5				9
901857	BR	VH046479	DRC Spring 2018	8,686	Part A	0,1,2	2,258	94	100	77	13	1			9
				8,686	Part B	0,1	2,258	97	100	86	5				9
901857	BR	VH046479	DRC Summer 2018	49,959	Part A	0,1,2	11,927	88	100	57	33	4			5
				49,959	Part B	0,1	11,927	94	100	80	14				5
901857	BR	VH046479	DRC Fall 2018	1,623	Part A	0,1,2	396	92	100	80	14	0			6
				1,623	Part B	0,1	396	99	100	93	1				6
901857	BR	VH046479	DRC Spring 2019 (SR)	227	Part A	0,1,2	8	100	100	86	7	0			7
				227	Part B	0,1	8	100	100	92	2				7
901857	BR	VH046479	DRC Summer 2019	2,084	Part A	0,1,2	570	97	100	77	12	0			11
				2,084	Part B	0,1	570	100	100	88	1				11

Form Key: Form BR = Administrative Error (AE), Form E = Operational

Geometry

IDEAS ID	Spring 2021 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	BR, E	M41169	Pearson Spring 2016	90,471	Overall	0,1,2,3	16,723	87	99	46	12	15	7		20
902012	BR, E	M41169	DRC Spring 2018	38,108	Overall	0,1,2,3	9,066	90	100	45	15	26	9		5
902012	BR, E	M41169	DRC Fall 2018	5,823	Overall	0,1,2,3	1,424	96	100	47	14	23	9		7
902012	BR, E	M41169	DRC Spring 2019 (D)	20,176	Overall	0,1,2,3	3,904	92	100	45	16	25	9		5
902012	BR, E	M41169	DRC Spring 2019 (E)	17,983	Overall	0,1,2,3	3,606	92	100	44	16	25	10		5
902012	BR, E	M41169	DRC Summer 2019	300	Overall	0,1,2,3	84	95	100	67	9	9	4		12
902012	BR, E	M41169	DRC Fall 2019	4,920	Overall	0,1,2,3	1,084	95	100	42	16	27	10		4
902012	BR, E	M41169	DRC Summer 2020	66	Overall	0,1,2,3	20	100	100	71	3	11	0		15
902012	BR, E	M41169	DRC Fall 2020	6,064	Overall	0,1,2,3	1,316	97	100	50	13	24	8		4
980937	E	M43798	Pearson Spring 2017	42,156	Overall	0,1,2,3	7,901	95	100	66	14	4	1		15
980937	E	M43798	DRC Spring 2019 (D)	19,879	Overall	0,1,2,3	4,942	99	100	80	10	2	0		8
980937	E	M43798	DRC Spring 2019 (E)	17,584	Overall	0,1,2,3	4,290	99	100	80	10	2	0		7
980937	E	M43798	DRC Fall 2019	4,878	Overall	0,1,2,3	1,174	98	100	78	10	4	1		6
980937	E	M43798	DRC Summer 2020	64	Overall	0,1,2,3	20	100	100	84	0	0	0		15
980937	E	M43798	DRC Fall 2020	5,952	Overall	0,1,2,3	1,326	99	100	82	9	3	1		5
980929	E	M1000516	Pearson 2017 FT	1,612	Overall	0,1,2,3,4	314	88	97	63	8	7	4	7	12
980929	E	M1000516	DRC Spring 2019 (E)	17,481	Overall	0,1,2,3,4	4,376	91	99	65	9	8	5	5	8
902042	BR, E	3020-M44058	Pearson Spring 2016	45,304	Part A	0,1,2,3	8,509	95	100	48	30	7	4		11
				45,304	Part B	0,1	8,509	96	100	61	22				17
				45,304	Part C	0,1,2	8,509	95	98	61	5	12			22
902042	BR, E	3020-M44058	DRC Spring 2018, Op	38,085	Part A	0,1,2,3	8,517	96	100	55	34	5	3		4
				38,085	Part B	0,1	8,517	97	100	78	19				4
				38,085	Part C	0,1,2	8,517	97	99	79	5	13			4
902042	BR, E	3020-M44058	DRC Fall 2018, Op	5,710	Part A	0,1,2,3	1,318	98	100	56	30	6	2		6
				5,710	Part B	0,1	1,318	98	100	77	17				6
				5,710	Part C	0,1,2	1,318	98	99	76	5	14			6
902042	BR, E	3020-M44058	DRC Spring 2019 (E)	17,677	Part A	0,1,2,3	2,866	97	100	50	35	7	4		3
				17,677	Part B	0,1	2,866	98	100	69	27				3
				17,677	Part C	0,1,2	2,866	98	99	75	5	18			3
902042	BR, E	3020-M44058	DRC Summer 2019	294	Part A	0,1,2,3	76	100	100	78	8	2	2		9
				294	Part B	0,1	76	100	100	83	7				9
				294	Part C	0,1,2	76	100	100	84	1	5			9
980930	E	M1000518	Pearson 2017 FT	1,500	Part B	0,1,2,3	298	95	100	60	11	12	1		15
980930	E	M1000518	DRC Spring 2019 (E)	18,605	Part B	0,1,2,3	3,396	97	100	76	9	14	1		0
980938	E	M100106	Pearson 2017 FT	1,635	Overall	0,1,2,3,4	314	93	99	74	5	6	4		11
980938	E	M100106	DRC Spring 2019 (D)	19,772	Overall	0,1,2,3,4	4,946	98	100	76	4	4	6		10
980938	E	M100106	DRC Spring 2019 (E)	17,503	Overall	0,1,2,3,4	4,374	99	100	75	4	4	7		10
980938	E	M100106	DRC Fall 2019	4,790	Overall	0,1,2,3,4	1,178	98	100	75	5	5	8		8
980938	E	M100106	DRC Summer 2020	62	Overall	0,1,2,3,4	22	100	100	84	0	0	0		16
980938	E	M100106	DRC Fall 2020	5,900	Overall	0,1,2,3,4	1,420	99	100	81	4	3	5		7

Form Key: Form BR = Administrative Error (AE), Form E = Operational

Geometry (continued)

IDEAS ID	Spring 2021 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 %
980936	E	VH239429	Pearson Spring 2017	42,154	Overall	0,1,2,3	8,173	84	99	72	16	4	2		6
980936	E	VH239429	DRC Spring 2019 (D)	20,142	Overall	0,1,2,3	4,570	95	100	75	10	7	2		5
980936	E	VH239429	DRC Spring 2019 (E)	17,729	Overall	0,1,2,3	3,930	94	100	74	12	8	2		5
980936	E	VH239429	DRC Fall 2019	4,768	Overall	0,1,2,3	1,034	97	100	76	6	10	2		6
980936	E	VH239429	DRC Summer 2020	66	Overall	0,1,2,3	22	100	100	82	0	0	0		18
980936	E	VH239429	DRC Fall 2020	5,930	Overall	0,1,2,3	1,286	95	100	80	6	8	1		5
902046	BR	M46668	Pearson Spring 2016	42,630	Overall	0,1,2,3	7,622	93	99	70	9	5	1		16
902046	BR	M46668	DRC Fall 2017	6,821	Overall	0,1,2,3	1,880	97	100	78	9	3	0		9
902046	BR	M46668	DRC Spring 2018	38,108	Overall	0,1,2,3	9,657	95	100	76	10	6	1		7
902046	BR	M46668	DRC Summer 2018	423	Overall	0,1,2,3	148	99	100	74	3	3	0		19
902046	BR	M46668	DRC Fall 2018	5,601	Overall	0,1,2,3	1,396	96	100	73	9	7	1		10
902046	BR	M46668	DRC Spring 2019 (SR)	403	Overall	0,1,2,3	116	98	100	78	3	0	0		18
902046	BR	M46668	DRC Summer 2019	291	Overall	0,1,2,3	78	100	100	80	2	3	1		12
902027	BR	M43233	Pearson Spring 2017	84,614	Overall	0,1,2,3,4	15,944	88	98	52	13	10	5	5	16
902027	BR	M43233	DRC Spring 2018	38,085	Overall	0,1,2,3,4	9,519	94	100	60	13	10	5	6	7
902027	BR	M43233	DRC Summer 2018	420	Overall	0,1,2,3,4	156	96	100	70	3	2	1	2	22
902027	BR	M43233	DRC Fall 2018	5,712	Overall	0,1,2,3,4	1,530	96	100	60	10	9	5	7	9
902027	BR	M43233	DRC Spring 2019 (SR)	398	Overall	0,1,2,3,4	102	100	100	79	2	2	0	1	17
902027	BR	M43233	DRC Summer 2019	294	Overall	0,1,2,3,4	96	96	100	72	5	1	0	3	17
902062	BR	VH150384	Pearson Spring 2016	2,581	Overall	0,1,2,3,4	542	89	97	57	6	4	2	1	31
902062	BR	VH150384	DRC Spring 2018	38,056	Overall	0,1,2,3,4	9,554	96	100	79	9	4	1	1	7
902062	BR	VH150384	DRC Fall 2018	5,747	Overall	0,1,2,3,4	1,452	97	100	76	9	4	2	1	9
902062	BR	VH150384	DRC Summer 2019	288	Overall	0,1,2,3,4	80	100	100	80	2	1	1	3	14
939101	BR	MGM0160	DRC Spring 2018, FT	1,665	Part C	0,1,2,3,4	336	80	97	73	15	8	2	1	1
939101	BR	MGM0160	DRC Spring 2019 (SR)	437	Part C	0,1,2,3,4	70	100	100	95	4	1	0	0	0
939101	BR	MGM0160	DRC Summer 2019	310	Part C	0,1,2,3,4	54	100	100	92	3	2	1	2	0

Form Key: Form BR = Administrative Error (AE), Form E = Operational

Math Grade 3

IDEAS ID	Spring 2021 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	99	47	33	17			3
				52,491	Part B	0,1,2	9,885	83	98	35	23	38			4
981736	Op	VH054794	DRC Spring 2019	58,729	Part A	0,1,2	11,036	86	99	55	30	13			1
				58,729	Part B	0,1,2	11,036	90	99	44	20	35			1
868619	Op	M00848	DRC Spring 2017	57,049	Overall	0, 1 2,3	11,716	93	99	64	9	5	13		8
868619	Op	M00848	DRC Spring 2018	61,311	Overall	0, 1 2,3	11,458	93	100	66	9	5	13		7
898001	Op	N/A	DRC Spring 2018, FT	1,659	Part A	0,1,2	318	94	100	41	21	37			1
				1,659	Part B	0,1	318	98	100	95	4				1
898001	Op	N/A	DRC Spring 2019	58,728	Part A	0,1,2	11,074	96	100	50	19	29			2
				58,728	Part B	0,1	11,074	99	100	94	3				2
981742	Op	M300388PD	Pearson 2017 FT	1,500	Part B	0,1,2	295	88	98	73	7	17			2
981742	Op	M300388PD	DRC Spring 2019 (Paper)	56,878	Part B	0,1,2	10,550	96	99	71	8	19			1
981742	Op	M300388PD	DRC Spring 2019 (Online)	1,768	Part B	0,1,2	348	98	100	86	8	6			0
914039	Op	M02527	Pearson Spring 2017	7,113	Overall	0,1,2,3	699	93	99	38	30	23	2		7
914039	Op	M02527	DRC Spring 2018	61,394	Overall	0,1,2,3	11,578	88	100	18	28	45	7		4
914039	Op	M02527	DRC Spring 2019	58,686	Overall	0,1,2,3	10,958	94	100	18	28	48	4		1
981747	Op	4127-M03599P	Pearson Spring 2018	102,233	Part B	0,1,2,3	20,403	91	99	48	26	9	13		4
				102,233	Part C	0,1,2	20,403	92	100	33	29	33			5
981747	Op	4127-M03599P	DRC Spring 2019 (Paper)	56,810	Part B	0,1,2,3	10,414	95	99	52	22	7	19		1
				56,810	Part C	0,1,2	10,414	96	100	33	21	45			1
981747	Op	4127-M03599P	DRC Spring 2019 (Online)	1,786	Part B	0,1,2,3	346	97	100	76	15	4	5		0
				1,786	Part C	0,1,2	346	97	100	64	21	15			0

Math Grade 4

IDEAS ID	Spring 2021 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	95	100	65	29	3			4
914084	Op	4112-M03491P	DRC Spring 2018 (Paper)	5,830	Part C	0,1,2	1,238	96	100	67	28	3			1
914084	Op	4112-M03491P	DRC Spring 2018 (Online)	56,155	Part C	0,1,2	10,776	95	100	63	28	5			4
914084	Op	4112-M03491P	DRC Spring 2019 (Paper)	52,276	Part C	0,1,2	9,874	92	100	59	32	5			4
914084	Op	4112-M03491P	DRC Spring 2019 (Online)	8,379	Part C	0,1,2	1,566	94	100	69	29	3			0
914086	Op	M04133	Pearson Spring 2017	107,359	Overall	0,1,2,3	10,670	91	99	53	24	7	15		1
914086	Op	M04133	DRC Spring 2018	61,742	Overall	0,1,2,3	11,702	95	100	54	24	7	9		5
914086	Op	M04133	DRC Spring 2019	60,533	Overall	0,1,2,3	11,438	93	99	54	24	5	12		4
981831	Op	M400526	Pearson 2017 FT	1,500	Overall	0,1,2,3	288	86	99	47	21	22	9		0
981831	Op	M400526	DRC Spring 2019	60,540	Overall	0,1,2,3	11,304	89	99	51	21	21	6		1
899959	Op	N/A	DRC Spring 2018, FT	1,622	Overall	0,1,2,3	302	82	99	34	24	11	30		0
899959	Op	N/A	DRC Spring 2019	60,611	Overall	0,1,2,3	11,420	89	100	31	32	19	16		1
899955	Op	N/A	DRC Spring 2018, FT	1,651	Part A	0,1,2	306	88	98	39	10	49			1
				1,651	Part B	0,1	306	96	100	88	11				1
899955	Op	N/A	DRC Spring 2019	60,626	Part A	0,1,2	11,651	94	99	47	12	39			1
				60,626	Part B	0,1	11,651	98	100	92	6				1
981827	Op	0318-M01475	Pearson 2017 FT	1,500	Part A	0,1,2	300	99	100	55	11	34			1
				1,500	Part B	0,1,2	300	99	100	80	3	15			2
				1,500	Part C	0,1,2	300	94	100	64	9	25			2
981827	Op	0318-M01475	DRC Spring 2019	60,421	Part A	0,1,2	11,306	93	99	59	11	28			1
				60,421	Part B	0,1,2	11,306	97	100	83	3	12			1
				60,421	Part C	0,1,2	11,306	95	100	74	6	18			1

Math Grade 5

IDEAS ID	Spring 2021 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	98	26	26	22	16	9	1
914152	Op	M03820	DRC Spring 2019	60,826	Overall	0,1,2,3,4	12,128	82	99	29	29	21	15	6	0
914148	Op	M03888	Pearson Spring 2017	72,736	Overall	0,1,2,3	7,272	87	99	40	28	13	19		1
914148	Op	M03888	DRC Spring 2018	59,662	Overall	0,1,2,3	11,464	93	99	57	22	8	12		1
914148	Op	M03888	DRC Spring 2019	60,403	Overall	0,1,2,3	11,584	90	99	49	27	16	7		1
902410	Op	N/A	DRC Spring 2018, FT	1,653	Part B	0,1,2	306	87	100	46	20	33			1
902410	Op	N/A	DRC Spring 2019	60,437	Part B	0,1,2	11,006	92	100	52	31	17			0
902414	Op	N/A	DRC Spring 2018, FT	1,651	Overall	0,1,2,3	318	87	99	63	11	20	7		0
902414	Op	N/A	DRC Spring 2019	60,212	Overall	0,1,2,3	11,750	92	99	73	10	12	4		1
914195	Op	0154-M00796	Pearson Spring 2017	92,904	Part B	0,1,2	9,282	96	100	80	8	6			5
914195	Op	0154-M00796	DRC Spring 2018	61,037	Part B	0,1,2	11,260	91	100	75	15	10			0
914195	Op	0154-M00796	DRC Spring 2019	60,444	Part B	0,1,2	11,022	91	100	70	16	14			0
934015	Op	N/A	DRC Spring 2018, FT	1,660	Part B	0,1	320	93	100	85	15				0
				1,660	Part C	0,1,2,3,4	320	89	98	58	19	11	4	7	0
934015	Op	N/A	DRC Spring 2019	60,399	Part B	0,1	11,016	94	100	85	15				0
				60,399	Part C	0,1,2,3,4	11,016	89	99	57	20	11	4	7	0

Math Grade 6

IDEAS ID	Spring 2021 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	67	97	35	23	19	13	6	3
981963	Op	M25151	DRC Spring 2019	60,801	Overall	0,1,2,3,4	12,446	75	98	40	26	19	11	4	1
981961	Op	VH082639	Pearson 2015 FT	1,500	Part A	0,1,2	348	90	100	55	27	14			4
				1,500	Part B	0,1	348	91	100	54	40				7
981961	Op	VH082639	DRC Spring 2019	60,295	Part A	0,1,2	11,792	92	99	69	23	7			2
				60,295	Part B	0,1	11,792	96	100	58	40				2
981954	Op	VH139067	Pearson Spring 2017	111,824	Part A	0,1,2	21,162	93	98	79	5	12			4
				111,824	Part B	0,1,2,3,4	21,162	87	98	59	16	9	4	9	4
981954	Op	VH139067	DRC Spring 2019	59,913	Part A	0,1,2	11,604	93	99	90	4	5			2
				59,913	Part B	0,1,2,3,4	11,604	87	99	74	15	5	2	3	2
981956	Op	VH220482	Pearson Spring 2017	111,824	Part B	0,1,2	22,112	92	99	32	16	50			3
981956	Op	VH220482	DRC Spring 2019	59,739	Part B	0,1,2	10,898	91	99	35	19	45			0
914231	Op	1740-M23030	Pearson Spring 2017	89,916	Overall	0,1,2,3	8,905	71	96	40	18	20	19		2
914231	Op	1740-M23030	DRC Spring 2018	58,067	Overall	0,1,2,3	11,448	74	96	43	18	19	17		2
914231	Op	1740-M23030	DRC Spring 2019	60,542	Overall	0,1,2,3	12,102	77	97	43	19	19	17		2
903511	Op	N/A	DRC Spring 2018, FT	1,652	Part B	0,1,2,3	310	85	98	76	10	10	5		0
903511	Op	N/A	DRC Spring 2019	60,453	Part B	0,1,2,3	10,992	91	100	79	10	9	3		0
914281	Op	M25152	Pearson Spring 2017	112,484	Overall	0,1,2,3	11,247	89	99	54	14	12	17		3
914281	Op	M25152	DRC Spring 2018	57,609	Overall	0,1,2,3	11,534	91	99	63	13	8	14		2
914281	Op	M25152	DRC Spring 2019	60,151	Overall	0,1,2,3	11,664	92	99	60	13	9	16		2

Math Grade 7

IDEAS ID	Spring 2021 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	99	75	5	5	13		3
				100,577	Part B	0,1,2,3	19,892	90	99	71	6	6	14		4
914362	Op	VH083535	DRC Spring 2018	56,482	Part A	0,1,2,3	10,560	96	100	86	3	3	7		0
				56,482	Part B	0,1,2,3	10,560	96	100	84	3	3	9		0
914362	Op	VH083535	DRC Spring 2019	57,526	Part A	0,1,2,3	10,992	95	99	82	4	4	9		0
				57,526	Part B	0,1,2,3	10,992	96	99	80	3	3	12		0
982922	Op	M25544	Pearson 2015 FT	1,800	Overall	0,1,2,3	404	88	99	50	14	22	7		7
982922	Op	M25544	DRC Spring 2019	56,961	Overall	0,1,2,3	11,714	94	99	59	10	22	7		2
868848	Op	M25578	Pearson Spring 2017	13,001	Overall	0,1,2,3	2,576	94	99	75	5	9	1		10
868848	Op	M25578	DRC Spring 2019	56,948	Overall	0,1,2,3	12,204	97	100	81	7	7	1		4
900539	Op	N/A	DRC Spring 2018, FT	1,646	Part A	0,1,2	316	91	99	46	37	17			0
				1,646	Part B	0,1	316	97	100	62	38				0
900539	Op	N/A	DRC Spring 2019	56,656	Part A	0,1,2	10,264	88	99	50	35	15			0
				56,656	Part B	0,1	10,264	96	100	66	34				0
982929	Op	M22009	Pearson Spring 2018	124,808	Overall	0,1,2,3	24,757	83	99	46	21	20	11		2
982929	Op	M22009	DRC Spring 2019	56,931	Overall	0,1,2,3	11,592	92	99	56	16	18	7		2
900520	Op	N/A	DRC Spring 2018, FT	1,624	Overall	0,1,2,3	348	97	100	77	6	4	9		3
900520	Op	N/A	DRC Spring 2019	56,781	Overall	0,1,2,3	11,972	96	99	81	4	3	8		3
914339	Op	VH151385	Pearson Spring 2017	88,725	Part A	0,1,2	8,838	95	99	67	8	21			4
				88,725	Part B	0,1,2	8,838	96	100	77	6	10			7
914339	Op	VH151385	DRC Spring 2018	56,454	Part A	0,1,2	10,887	98	100	73	7	19			2
				56,454	Part B	0,1,2	10,887	98	100	83	6	10			2
914339	Op	VH151385	DRC Spring 2019	57,375	Part A	0,1,2	11,238	98	100	71	7	20			2
				57,375	Part B	0,1,2	11,238	98	100	83	6	9			2

Math Grade 8

IDEAS ID	Spring 2021 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	96	100	64	20	8			9
				28,653	Part B	0,1,2,3,4	5,561	91	99	72	10	6	2	0	10
983010	Op	VH097312	DRC Spring 2019	49,262	Part A	0,1,2	9,484	92	100	46	42	11			0
				49,262	Part B	0,1,2,3,4	9,484	80	99	63	18	13	3	1	0
982987	Op	M800114	Pearson 2017 FT	1,500	Part A	0,1,2	300	93	98	74	8	15			3
				1,500	Part B	0,1,2	300	89	99	70	12	14			5
982987	Op	M800114	DRC Spring 2019	48,845	Part A	0,1,2	9,982	90	99	78	9	11			3
				48,845	Part B	0,1,2	9,982	89	98	76	11	10			3
982999	Op	M22203	Pearson Spring 2017	69,637	Overall	0,1,2,3	13,500	84	97	55	24	9	9		4
982999	Op	M22203	DRC Spring 2019	48,419	Overall	0,1,2,3	10,114	92	99	69	18	6	4		3
870899	Op	1282-M21381	Pearson Spring 2015	48,511	Part A	0,1,2	9,762	89	98	72	9	9			10
				48,511	Part B	0,1	9,762	91	99	66	22				12
870899	Op	1282-M21381	DRC Spring 2019	47,707	Part A	0,1,2	9,770	95	99	86	8	3			2
				47,707	Part B	0,1	9,770	96	100	82	15				2
899312	Op	N/A	DRC Spring 2018, FT	1,648	Part B	0,1,2	318	85	98	27	30	43			0
899312	Op	N/A	DRC Spring 2019	49,182	Part B	0,1,2	9,002	91	100	37	32	31			0
914381	Op	M25425	Pearson Spring 2017	69,637	Overall	0,1,2,3,4	6,943	91	99	52	13	26	2	1	6
914381	Op	M25425	DRC Spring 2018	49,280	Overall	0,1,2,3,4	10,088	94	100	59	16	20	2	0	2
914381	Op	M25425	DRC Spring 2019	49,073	Overall	0,1,2,3,4	10,614	93	99	57	15	21	2	1	4
899329	Op	N/A	DRC Spring 2018, FT	1,653	Part B	0,1	314	90	100	51	49				0
				1,653	Part C	0,1	314	94	100	57	43				0
899329	Op	N/A	DRC Spring 2019	49,151	Part B	0,1	8,962	94	100	75	25				0
				49,151	Part C	0,1	8,962	95	100	77	23				0

English I

Task	IDEAS ID	Spring 2021 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	914552	E	GG431834057	Pearson Spring 18	66,624	RCWE	0,1,2,3,4	2,058	7,456	62,441	13,132	76	100	26	29	27	11	2	4
					66,624	Conv	0,1,2,3	2,058	7,456	62,441	13,132	76	100	25	30	27	12		4
RST	914552	E	GG431834057	DRC Spring 19 (D, E)*	52,885	RCWE	0,1,2,3,4	n/a	n/a	n/a	11,496	76	100	21	37	33	6	0	2
					52,885	Conv	0,1,2,3	n/a	n/a	n/a	11,496	75	99	19	38	34	6		2
RST	914552	E	GG431834057	DRC Fall 19*	8,715	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,244	85	100	36	31	23	5	0	3
					8,715	Conv	0,1,2,3	n/a	n/a	n/a	2,244	82	100	33	33	24	5		3
RST	914552	E	GG431834057	DRC Summer 20*	754	RCWE	0,1,2,3,4	n/a	n/a	n/a	268	91	100	71	16	1	0	0	12
					754	Conv	0,1,2,3	n/a	n/a	n/a	268	90	100	68	20	1	0		12
NWT	983215	E	GG604245591	Pearson 17 FT	1,696	Expr	0,1,2,3,4	1,430	155	0	299	75	97	25	25	26	12	5	7
					1,696	Conv	0,1,2,3	1,430	155	0	299	73	100	23	23	28	15		7
NWT	983215	E	GG604245591	DRC Spring 19 (E)	23,695	Expr	0,1,2,3,4	n/a	n/a	n/a	4,870	79	99	26	32	30	9	2	3
					23,695	Conv	0,1,2,3	n/a	n/a	n/a	4,870	77	100	21	36	31	10		3
NWT	983215	E	GG604245591	DRC Fall 19	8,504	Expr	0,1,2,3,4	n/a	n/a	n/a	1,972	89	100	45	21	18	8	2	7
					8,504	Conv	0,1,2,3	n/a	n/a	n/a	1,972	87	100	43	23	19	9		7
RST	902161	A	VH017542_2T	Pearson Spring 17	123,860	RCWE	0,1,2,3,4	2,656	13,063	116,406	23,334	76	100	22	33	24	12	4	4
					123,860	Conv	0,1,2,3	2,656	13,063	116,407	23,334	76	100	23	33	24	17		4
RST	902161	A	VH017542_2T	DRC Fall 17	4,674	RCWE	0,1,2,3,4	n/a	n/a	n/a	982	78	99	12	34	40	13	0	0
					4,674	Conv	0,1,2,3	n/a	n/a	n/a	982	78	99	14	32	38	15		0
RST	902161	A	VH017542_2T	DRC Spring 18*	50,817	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,136	81	100	17	37	32	11	1	2
					50,817	Conv	0,1,2,3	n/a	n/a	n/a	10,136	79	100	17	36	32	13		2
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	100	30	30	24	10	1	3
					7,444	Conv	0,1,2,3	n/a	n/a	n/a	1,870	84	100	30	29	25	12		3
RST	902161	A	VH017542_2T	DRC Spring 19 (SR)	86	RCWE	0,1,2,3,4	n/a	n/a	n/a	12	100	100	44	37	15	2	0	1
					86	Conv	0,1,2,3	n/a	n/a	n/a	12	67	100	43	40	16	0		1
RST	902161	A	VH017542_2T	DRC Summer 19*	2,184	RCWE	0,1,2,3,4	n/a	n/a	n/a	554	87	100	51	38	5	1	0	4
					2,184	Conv	0,1,2,3	n/a	n/a	n/a	554	88	100	54	35	5	1		4
RST	902161	A	VH017542_2T	DRC Fall 20*	7,926	RCWE	0,1,2,3,4	n/a	n/a	n/a	1,778	84	100	23	35	27	12	2	2
					7,926	Conv	0,1,2,3	n/a	n/a	n/a	1,778	83	100	23	34	27	14		2
NWT	906152	A	VH084830	Pearson Spring 17	61,936	Expr	0,1,2,3,4	3,125	7,776	53,955	10,498	73	99	30	22	27	9	4	8
					61,936	Conv	0,1,2,3	3,125	7,776	53,955	10,498	74	99	28	28	25	11		8
NWT	906152	A	VH084830	DRC Fall 17	5,047	Expr	0,1,2,3,4	n/a	n/a	n/a	1,076	81	99	22	34	29	10	1	2
					5,047	Conv	0,1,2,3	n/a	n/a	n/a	1,076	78	99	25	36	26	10		2
NWT	906152	A	VH084830	DRC Spring 19 (SR)	78	Expr	0,1,2,3,4	n/a	n/a	n/a	10	100	100	47	36	10	0	0	7
					78	Conv	0,1,2,3	n/a	n/a	n/a	10	80	100	47	31	15	0		7
NWT	906152	A	VH084830	DRC Summer 19*	2,162	Expr	0,1,2,3,4	n/a	n/a	n/a	762	93	100	74	11	2	0	0	12
					2,162	Conv	0,1,2,3	n/a	n/a	n/a	762	92	100	71	15	2	0		12
NWT	906152	A	VH084830	DRC Fall 20*	7,823	Expr	0,1,2,3,4	n/a	n/a	n/a	2,080	85	100	37	23	24	7	2	6
					7,823	Conv	0,1,2,3	n/a	n/a	n/a	2,080	88	100	36	30	22	6		6

Form Key: Form E = Operational, Form A = Administrative Error (AE)

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021; an asterisk in the Source of IRR and SPD Data column (*) indicates previous AI scoring by Pearson for DRC

English II

Task	IDEAS ID	Spring 2021 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	983688	E	HH607742252	Pearson 2017 FT	1,604	RCWE	0,1,2,3,4	1,487	162	0	312	78	100	28	30	20	7	2	13
					1,604	Conv	0,1,2,3	1,487	162	0	312	78	100	26	31	21	9		13
RST	983688	E	HH607742252	Pearson 2019	81,553	RCWE	0,1,2,3,4	1,843	17,633	76,244	16,308	75	100	18	20	25	22	8	7
					81,553	Conv	0,1,2,3	1,843	17,633	76,244	16,308	79	100	16	21	27	29		7
RST	983688	E	HH607742252	DRC Spring 19 (D, E)	46,634	RCWE	0,1,2,3,4	n/a	n/a	n/a	9,264	78	100	20	28	41	8	0	2
					46,634	Conv	0,1,2,3	n/a	n/a	n/a	9,264	77	100	20	29	40	9		2
RST	983688	E	HH607742252	DRC Summer 20*	282	RCWE	0,1,2,3,4	n/a	n/a	n/a	104	94	100	68	16	5	0	0	12
					282	Conv	0,1,2,3	n/a	n/a	n/a	104	92	100	66	19	5	0		12
NWT	983642	E	HH432845949	Pearson Spring 17	57,527	Expr	0,1,2,3,4	28,646	6,810	26,290	13,745	77	100	16	23	33	16	5	7
					57,527	Conv	0,1,2,3	28,646	6,810	26,290	13,745	75	100	18	24	32	18		7
NWT	983642	E	HH432845949	DRC Spring 19 (E)*	21,673	Expr	0,1,2,3,4	n/a	n/a	n/a	4,650	84	100	11	26	43	15	3	2
					21,673	Conv	0,1,2,3	n/a	n/a	n/a	4,650	82	100	12	28	41	17		2
NWT	983642	E	HH432845949	DRC Fall 19*	8,878	Expr	0,1,2,3,4	n/a	n/a	n/a	2,264	92	100	26	25	29	13	3	4
					8,878	Conv	0,1,2,3	n/a	n/a	n/a	2,264	92	100	28	25	28	14		4
RST	902331	A	VH004490	Pearson Spring 17**	2,605	RCWE	0,1,2,3,4	1,915	263	646	827	82	99	52	28	7	1	0	12
					2,605	Conv	0,1,2,3	1,915	263	646	827	85	100	53	26	7	1		12
RST	902331	A	VH004490	Pearson Spring 16**	126,270	RCWE	0,1,2,3,4	121,660	n/a	n/a	16,036	77	100	23	35	23	8	2	9
					126,270	Conv	0,1,2,3	121,507	n/a	n/a	16,003	77	100	25	33	24	10		9
RST	902331	A	VH004490	DRC Fall 17	9,305	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,020	79	100	37	24	25	11	2	2
					9,305	Conv	0,1,2,3	n/a	n/a	n/a	2,020	77	99	35	23	27	14		2
RST	902331	A	VH004490	DRC Spring 18*	48,949	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,460	79	100	15	35	34	11	2	3
					48,949	Conv	0,1,2,3	n/a	n/a	n/a	10,460	78	99	17	35	34	11		3
RST	902331	A	VH004490	DRC Fall 18*	10,714	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,826	84	100	30	33	22	9	2	3
					10,714	Conv	0,1,2,3	n/a	n/a	n/a	2,826	81	100	33	32	21	9		3
RST	902331	A	VH004490	DRC Spring 19 (SR)	948	RCWE	0,1,2,3,4	n/a	n/a	n/a	164	94	100	68	24	3	0	0	5
					948	Conv	0,1,2,3	n/a	n/a	n/a	164	95	100	68	24	3	0		5
RST	902331	A	VH004490	DRC Summer 19*	1,870	RCWE	0,1,2,3,4	n/a	n/a	n/a	562	91	100	56	32	4	0	0	7
					1,870	Conv	0,1,2,3	n/a	n/a	n/a	562	90	100	60	29	3	0		7
RST	902331	A	VH004490	DRC Fall 20*	9,965	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,272	89	100	15	30	33	17	3	2
					9,965	Conv	0,1,2,3	n/a	n/a	n/a	2,272	89	100	17	30	34	17		2
NWT	902354	A	7064	Pearson Spring 17	4,409	Expr	0,1,2,3,4	4,189	435	0	844	85	100	43	20	14	6	2	17
					4,409	Conv	0,1,2,3	4,189	435	0	844	85	100	40	22	15	7		17
NWT	902354	A	7064	DRC Fall 17	9,721	Expr	0,1,2,3,4	n/a	n/a	n/a	2,098	81	100	46	17	19	12	2	2
					9,721	Conv	0,1,2,3	n/a	n/a	n/a	2,098	83	100	43	18	22	14		2
NWT	902354	A	7064	DRC Spring 19 (SR)	956	Expr	0,1,2,3,4	n/a	n/a	n/a	228	100	100	85	4	1	0	0	9
					956	Conv	0,1,2,3	n/a	n/a	n/a	228	97	100	81	8	2	0		9
NWT	902354	A	7064	DRC Fall 20	9,661	Expr	0,1,2,3,4	n/a	n/a	n/a	2,022	84	100	32	27	27	9	1	3
					9,661	Conv	0,1,2,3	n/a	n/a	n/a	2,022	82	100	33	27	26	10		3

Form Key: Form E = Operational, Form A = Administrative Error (AE)

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021; an asterisk in the Source of IRR and SPD Data column (*) indicates previous AI scoring by Pearson for DRC

** Pearson – Statistics from 2017 and 2016 are included for 902331/VH004490. Volumes were significantly higher in 2016, but Pearson reports in 2016 did not split out human and AI scoring.

ELA Grade 3

Task	IDEAS ID	Spring 2021 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	99	53	39	7	0		1
					1,582	Conventions	0,1,2,3	n/a	n/a	n/a	339	69	98	57	33	8	2		1
RST	915227	Op	A1598	DRC Spring 19	59,506	RCWE	0,1,2,3	n/a	n/a	n/a	12,420	80	99	36	47	13	0		3
					59,506	Conventions	0,1,2,3	n/a	n/a	n/a	12,420	80	100	37	46	12	3		3
NWT	913497	Op	AA431426588	Pearson Spring 17	118,416	Expression	0,1,2,3	34,298	13,546	84,911	27,299	71	99	30	56	11	2		2
					118,416	Conventions	0,1,2,3	34,298	13,546	84,910	27,299	69	99	33	47	16	2		2
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
NWT	913497	Op	AA431426588	DRC Spring 19	59,352	Expression	0,1,2,3	n/a	n/a	n/a	12,110	86	100	31	53	11	2		3
					59,352	Conventions	0,1,2,3	n/a	n/a	n/a	12,110	77	99	25	53	18	2		3

ELA Grade 4

Task	IDEAS ID	Spring 2021 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 17	121,461	RCWE	0,1,2,3,4	35,658	13,901	87,168	28,425	67	99	34	40	21	4	1	1
					121,461	Conventions	0,1,2,3	35,659	13,893	87,168	28,418	69	99	28	46	21	4		1
LAT	913567	Op	VH170170	DRC Spring 18	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
LAT	913567	Op	VH170170	DRC Spring 19	60,409	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,672	81	100	27	40	28	4	0	1
					60,409	Conventions	0,1,2,3	n/a	n/a	n/a	10,672	79	100	24	41	30	4		1
RST	982233	Op	VH060330	Pearson 2017 FT	1,500	RCWE	0,1,2,3,4	1,468	150	0	300	78	100	26	52	18	3	0	2
					1,500	Conventions	0,1,2,3	1,468	150	0	300	78	100	20	56	19	3		2
RST	982233	Op	VH060330	DRC Spring 19	62,117	RCWE	0,1,2,3,4	n/a	n/a	n/a	14,086	83	100	28	42	25	3	0	1
					62,117	Conventions	0,1,2,3	n/a	n/a	n/a	14,086	83	100	28	42	26	3		1

ELA Grade 5

Task	IDEAS ID	Spring 2021 Form	PARCC UIN	Source of IRR and SPD Data	Resp. 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 16	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 17	11,258	RCWE	0,1,2,3,4	11,045	1,127	0	2,231	87	100	80	13	1	0	0	6
					11,258	Conventions	0,1,2,3	11,045	1,127	0	2,231	82	99	65	25	4	0		6
LAT	801310	Op	VF821667	DRC Spring 19	61,201	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,486	77	100	55	37	7	1	0	0
					61,201	Conventions	0,1,2,3	n/a	n/a	n/a	12,486	75	100	46	43	10	1		0
RST	915510	Op	VH198972	Pearson 2016 FT	1,561	RCWE	0,1,2,3,4	n/a	n/a	n/a	332	69	100	39	41	16	4	0	1
					1,561	Conventions	0,1,2,3	n/a	n/a	n/a	332	70	99	28	43	23	5		1
RST	915510	Op	VH198972	DRC Spring 19	62,772	RCWE	0,1,2,3,4	n/a	n/a	n/a	15,458	80	100	32	36	25	5	1	0
					62,772	Conventions	0,1,2,3	n/a	n/a	n/a	15,458	80	100	32	36	25	6		0

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021

ELA Grade 6

Task	IDEAS ID	Spring 2021 Form	PARCC UIN	Source of IRR and SPD Data	Resp. 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 17	128,716	RCWE	0,1,2,3,4	36,320	13,240	93,042	29,065	73	99	32	35	25	6	1	1
					128,716	Conventions	0,1,2,3	36,320	13,240	93,042	29,065	71	99	32	33	26	8		1
RST	913715	Op	DD502035970	DRC Spring 19*	61,422	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,874	71	99	21	34	37	6	1	0
					61,422	Conventions	0,1,2,3	n/a	n/a	n/a	12,874	68	98	22	31	36	11		0
NWT	913694	Op	D1466	Pearson Spring 17	127,628	Expression	0,1,2,3,4	34,718	14,034	93,800	29,433	76	99	40	21	23	10	4	2
					127,628	Conventions	0,1,2,3	34,718	14,034	93,800	29,433	75	100	33	30	23	11		2
NWT	913694	Op	D1466	DRC Spring 18*	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
NWT	913694	Op	D1466	DRC Spring 19*	61,223	Expression	0,1,2,3,4	n/a	n/a	n/a	12,422	79	100	40	24	26	7	2	1
					61,223	Conventions	0,1,2,3	n/a	n/a	n/a	12,422	76	100	31	38	24	6		1

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021; an asterisk in the Source of IRR and SPD Data column (*) indicates previous AI scoring by Pearson

ELA Grade 7

Task	IDEAS ID	Spring 20121 Form	PARCC UIN	Source of IRR and SPD Data	Resp. 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 17	1,630	RCWE	0,1,2,3,4	n/a	n/a	n/a	345	76	99	31	33	23	8	3	3
					1,630	Conventions	0,1,2,3	n/a	n/a	n/a	345	76	100	31	33	23	11		3
RST	915582	Op	E1567	DRC Spring 19	57,944	RCWE	0,1,2,3,4	n/a	n/a	n/a	11,078	76	99	18	33	38	9	1	0
					57,944	Conventions	0,1,2,3	n/a	n/a	n/a	11,078	75	98	20	32	37	9		0
NWT	913842	Op	EE43013306	Pearson Spring 17	128,845	Expression	0,1,2,3,4	37,606	14,582	91,555	30,289	73	99	34	13	20	18	13	2
					128,845	Conventions	0,1,2,3	37,605	14,582	91,555	30,289	72	99	29	21	24	25		2
NWT	913842	Op	EE43013306	DRC Spring 18*	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
NWT	913842	Op	EE43013306	DRC Spring 19*	58,491	Expression	0,1,2,3,4	n/a	n/a	n/a	12,164	76	99	35	12	25	18	9	0
					58,491	Conventions	0,1,2,3	n/a	n/a	n/a	12,164	74	99	27	23	29	21		0

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021; an asterisk in the Source of IRR and SPD Data column (*) indicates previous AI scoring by Pearson

ELA Grade 8

Task	IDEAS ID	Spring 2021 Form	PARCC UIN	Source of IRR and SPD Data	Resp. 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 17	128,084	RCWE	0,1,2,3,4	36,606	4,234	89,633	19,154	70	100	26	31	26	11	3	2
					128,084	Conventions	0,1,2,3	36,606	4,234	89,634	19,154	72	100	23	31	29	15		2
LAT	913958	Op	F1460	DRC Spring 18*	57,038	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,090	73	99	18	32	35	15	1	0
					57,038	Conventions	0,1,2,3	n/a	n/a	n/a	12,090	76	100	14	31	39	15		0
LAT	913958	Op	F1460	DRC Spring 19*	57,108	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,678	75	99	18	30	36	13	1	1
					57,108	Conventions	0,1,2,3	n/a	n/a	n/a	12,678	74	99	15	28	40	16		1
RST	982327	Op	FF506834510	Pearson 2017 FT	1,625	RCWE	0,1,2,3,4	1,496	165	0	317	75	99	43	23	17	6	3	9
					1,625	Conventions	0,1,2,3	1,496	165	0	317	74	98	35	23	23	9		9
RST	982327	Op	FF506834510	DRC Spring 19	56,488	RCWE	0,1,2,3,4	n/a	n/a	n/a	11,422	75	99	28	42	23	5	1	0
					56,488	Conventions	0,1,2,3	n/a	n/a	n/a	11,422	75	99	32	32	27	9		0

Highlighted IDEAS ID indicates an item is being AI scored by Pearson in 2021; an asterisk in the Source of IRR and SPD Data column (*) indicates previous AI scoring by Pearson

Biology ERs and CRs

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
965124	ER	B_T	Spring 2018 FT	4842	Part A (0-3)	4842	69	29	2	9	34	36	20				0
				4842	Part B (0-6)	4842	61	26	14	36	19	21	8	10	2	3	0
965124	ER	B_T	Spring 2019	23,243	Part A (0-3)	6,054	82	17	1	9	40	29	15				6
				23,243	Part B (0-6)	6,054	80	15	6	31	20	19	10	8	2	3	6
965124	ER	B_T	Fall 2019	14,300	Part A (0-3)	4,498	92	7	0	16	49	16	6				13
				14,300	Part B (0-6)	4,498	89	9	2	40	21	14	5	4	1	1	13
965124	ER	B_T	Summer 2020	1,490	Part A (0-3)	630	98	2	0	18	54	8	0				20
				1,490	Part B (0-6)	630	98	2	0	47	23	9	1	0	0	0	20
965129	CR	B_T	Spring 2018 FT	1,566	0-2	332	81	19	1	58	29	10					3
965129	CR	B_T	Spring 2019	43,692	0-2	11,446	89	10	1	60	21	11					7
965129	CR	B_T	Fall 2019	14,665	0-2	5,340	95	4	0	59	13	7					20
965129	CR	B_T	Summer 2020	1,516	0-2	686	100	0	0	67	5	0					28
965237	CR	B_T	Spring 2018 FT	1,607	0-2	360	96	4	0	82	14	3					1
965237	CR	B_T	Spring 2019	42,922	0-2	9,751	94	5	0	82	10	4					4
965237	CR	B_T	Fall 2019	13,940	0-2	3,730	98	2	0	82	5	3					9
965237	CR	B_T	Summer 2020	1,396	0-2	424	100	0	0	86	1	0					13
965295	CR	B_T	Spring 2018 FT	1,622	0-2	318	76	23	1	57	33	10					1
965295	CR	B_T	Spring 2019	41,215	0-2	9,706	86	13	1	59	30	5					5
965295	CR	B_T	Fall 2019	14,386	0-2	4,754	95	5	0	61	21	3					14
965295	CR	B_T	Summer 2020	1,473	0-2	624	97	2	0	59	17	2					22
965286**	ER	A_T	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	7
				5,140	Part B (0-3)	5,140	84	13	3	36	30	12	16				7
965286	ER	A_T	Fall 2018	7,446	Part A (0-6)	1,588	87	10	3	55	13	13	14	2	1	1	1
				7,446	Part B (0-3)	1,588	85	14	1	41	35	11	11				1
965286	ER	A_T	Spring 2019	19,856	Part A (0-6)	4,689	88	10	2	44	14	14	19	2	1	2	4
				19,856	Part B (0-3)	4,689	88	10	2	34	34	11	16				4
965286	ER	A_T	Summer 2019	290	Part A (0-6)	126	95	5	0	57	13	3	3	0	0	0	24
				290	Part B (0-3)	126	100	0	0	51	21	2	3				24
965286	ER	A_T	Fall 2020	12,416	Part A (0-6)	3958	94	5	1	45	13	12	14	2	1	1	10
				12,416	Part B (0-3)	3958	95	4	1	35	32	10	12				10
965190	CR	A_T	Spring 2018 FT	1,626	0-2	324	84	15	1	65	20	14					1
965190	CR	A_T	Fall 2018	7,357	0-2	1,448	93	7	0	78	13	7					1
965190***	CR	A_T	Spring 2019 (supplemental FT)	1,673	0-2	414	87	11	2	66	18	12					3
965190	CR	A_T	Summer 2019	294	0-2	144	100	0	0	61	2	0					36
965190	CR	A_T	Fall 2020	12,369	0-2	4,024	95	4	1	64	12	8					15

Form Key: Form A_T = Administrative Error (AE), Form B_T = Operational

*Condition Code notes:
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs and CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

**ER 965286 FT item was re-scored in October 2018 using updated rubric.

***Spring 2019 – DRC conducted supplemental FT scoring for CR items 965190, 965222, and 965279

Biology ERs and CRs (continued)

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
965222	CR	A_T	Spring 2018 FT	1,592	0-2	350	93	7	0	64	28	4					3
965222	CR	A_T	Fall 2018	7,279	0-2	1,516	92	8	0	71	25	2					2
965222***	CR	A_T	Spring 2019 (supplemental FT)	1,664	0-2	430	93	7	0	60	29	2					8
965222	CR	A_T	Summer 2019	312	0-2	170	100	0	0	49	9	0					41
965222	CR	A_T	Fall 2020	12,573	0-2	4,696	98	2	0	58	20	1					21
965279	CR	A_T	Spring 2018 FT	1,625	0-2	316	75	25	1	46	31	22					1
965279	CR	A_T	Fall 2018	7,540	0-2	1,484	86	14	0	57	31	11					1
965279***	CR	A_T	Spring 2019 (supplemental FT)	1,700	0-2	448	88	12	0	53	28	14					4
965279	CR	A_T	Summer 2019	319	0-2	150	99	1	0	50	15	1					34
965279	CR	A_T	Fall 2020	12,830	0-2	4,386	96	3	0	49	22	11					18

Form Key: Form A_T = Administrative Error (AE), Form B_T = Operational

*Condition Code notes:
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs and CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

***Spring 2019 – DRC conducted supplemental FT scoring for CR items 965190, 965222, and 965279

Grade 3 Science

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
957382	ER	Op	Spring 2018 FT	2,768	0-6	536	78	18	4	63	13	15	4	4	0	0	1
957382	ER	Op	Spring 2019	59,772	0-6	13,444	86	9	5	68	10	10	2	3	0	0	7
957435	CR	Op	Spring 2018 FT	1,660	0-2	320	87	13	0	58	33	7					1
957435	CR	Op	Spring 2019	59,810	0-2	13,504	88	11	1	53	32	8					6
957418	CR	Op	Spring 2018 FT	1,661	0-2	322	88	12	0	36	61	2					0
957418	CR	Op	Spring 2019	58,744	0-2	11,366	87	13	0	40	53	4					4
957409	CR	Op	Spring 2018 FT	1,675	0-2	350	87	13	0	40	40	19					1
957409	CR	Op	Spring 2019	59,469	0-2	12,836	84	16	0	37	43	13					6

Grade 4 Science

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
957054	ER	Op	Spring 2018 FT	2,778	0-6	556	74	23	3	6	13	40	37	3	0	0	0
957054	ER	Op	Spring 2019	61,054	0-6	12,290	82	18	0	12	18	38	29	1	0	0	2
957144	CR	Op	Spring 2018 FT	1,668	0-2	326	88	12	0	83	14	1					2
957144	CR	Op	Spring 2019	61,131	0-2	12,556	93	7	0	80	13	1					5
957090	CR	Op	Spring 2018 FT	1,665	0-2	330	79	21	0	45	49	6					0
957090	CR	Op	Spring 2019	61,766	0-2	13,810	90	10	0	66	23	5					6
957099	CR	Op	Spring 2018 FT	1,657	0-2	314	96	4	0	71	25	3					1
957099	CR	Op	Spring 2019	61,028	0-2	12,332	96	4	0	70	21	6					3

*Condition Code notes:
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs and for CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

Grade 5 Science

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
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	0
959503	ER	Op	Spring 2019	62,020	0-9	14,565	78	18	3	38	10	8	8	9	7	6	3	2	1	8
959557	CR	Op	Spring 2018 FT	1,667	0-2	346	89	7	4	29	51	19								0
959557	CR	Op	Spring 2019	62,569	0-2	15,631	91	9	0	27	53	12								8
959548	CR	Op	Spring 2018 FT	1,658	0-2	324	96	4	1	69	12	19								0
959548	CR	Op	Spring 2019	61,742	0-2	13,757	94	6	0	58	17	20								5
959530	CR	Op	Spring 2018 FT	1,690	0-2	382	98	2	0	56	7	37								0
959530	CR	Op	Spring 2019	61,592	0-2	13,612	98	2	0	62	7	26								5

Grade *6 Science*

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
958421	ER	Op	Spring 2018 FT	4,988	Part A (0-3)	4,988	86	8	6	68	19	0	13				0
				4,988	Part B (0-3)	4,988	80	19	2	58	29	11	2				0
				4,988	Part C (0-3)	4,988	85	12	3	62	17	19	2				0
958421	ER	Op	Spring 2019	38,631	Part A (0-3)	9,622	90	6	4	67	17	0	9				7
				38,631	Part B (0-3)	9,622	88	11	1	60	24	8	1				7
				38,631	Part C (0-3)	9,622	89	10	1	60	21	11	1				7
958378	CR	Op	Spring 2018 FT	1,652	0-2	314	86	14	0	81	14	5					0
958378	CR	Op	Spring 2019	37,029	0-2	10,235	88	11	1	58	22	9					11
958308	CR	Op	Spring 2018 FT	1,653	0-2	316	88	11	1	67	29	3					0
958308	CR	Op	Spring 2019	34,456	0-2	8,129	90	10	0	56	35	2					7
958396	CR	Op	Spring 2018 FT	1,648	0-2	320	91	9	0	74	20	6					0
958396	CR	Op	Spring 2019	32,763	0-2	7,961	94	6	0	76	14	3					7

*Condition Code notes:
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs and for CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

Grade 7 Science

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
959635	ER	Op	Spring 2018 FT	4,952	Part A (0-3)	4,952	78	16	6	71	17	10	2				0
				4,952	Part B (0-4)	4,952	81	15	4	71	19	8	1	0			0
				4,952	Part C (0-2)	4,952	96	4	0	88	10	1					0
959635	ER	Op	Spring 2019	39,287	Part A (0-3)	9,890	93	7	0	72	11	6	2				8
				39,287	Part B (0-4)	9,890	93	7	0	68	14	7	2	1			8
				39,287	Part C (0-2)	9,890	98	2	0	80	8	3					8
959748	CR	Op	Spring 2018 FT	1,646	0-2	312	82	18	0	30	50	20					1
959748	CR	Op	Spring 2019	60,641	0-2	17,243	88	12	0	32	47	8					12
959697	CR	Op	Spring 2018 FT	1,651	0-2	332	92	8	0	39	42	19					0
959697	CR	Op	Spring 2019	48,504	0-2	48,504	95	5	0	38	37	15					10
959715	CR	Op	Spring 2018 FT	1,647	0-2	336	92	8	0	92	6	1					0
959715	CR	Op	Spring 2019	59,077	0-2	15,134	98	2	0	88	3	0					8

Grade 8 Science

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
959334	ER	Op	Spring 2018 FT	4,950	Part A (0-3)	4,950	62	30	8	28	28	28	15				0
				4,950	Part B (0-6)	4,950	49	32	20	12	13	20	21	17	10	7	0
959334	ER	Op	Spring 2019	58,461	Part A (0-3)	16,932	83	15	2	29	27	23	11				11
				58,461	Part B (0-6)	16,932	77	18	5	12	14	19	19	14	8	3	11
959309	CR	Op	Spring 2018 FT	1,656	0-2	324	90	10	0	87	11	1					0
959309	CR	Op	Spring 2019	56,634	0-2	12,774	90	9	1	77	13	3					7
959291	CR	Op	Spring 2018 FT	1,639	0-2	320	86	13	1	42	51	6					0
959291	CR	Op	Spring 2019	56,568	0-2	13,398	91	9	0	44	46	3					6
959221	CR	Op	Spring 2018 FT	1,648	0-2	326	88	12	0	76	20	3					0
959221	CR	Op	Spring 2019	56,039	0-2	11,798	91	8	0	70	21	5					5

*Condition Code notes:
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

U.S. History ERs and CRs

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
892955	ER	F_S	Spring 2017 FT	5,000	Content	0-4	5,000	65	32	3	34	29	25	9	3	0
				5,000	Claims	0-4	5,000	64	32	4	37	26	25	10	3	0
892955	ER	F_S	Spring 2018	10,506	Content	0-4	5,426	94	6	0	16	32	31	15	3	2
				10,506	Claims	0-4	5,426	93	7	0	21	28	30	15	3	2
892955	ER	F_S	Spring 2019	53,432	Content	0-4	29,250	93	7	0	22	35	22	10	2	10
				53,432	Claims	0-4	29,250	93	7	0	29	29	21	9	2	10
892955	ER	F_S	Fall 2019	13,883	Content	0-4	8,726	92	8	0	31	27	16	6	1	17
				13,883	Claims	0-4	8,726	92	8	0	38	21	15	6	2	17
892955	ER	F_S	Summer 2020	848	Content	0-4	482	97	3	0	53	15	0	0	0	31
				848	Claims	0-4	482	97	3	0	60	9	0	0	0	31
894271	CR	F_S	Spring 2017 FT	1,658		0-2	316	66	34	1	54	37	8			0
894271	CR	F_S	Spring 2018 FT	1,331		0-2	254	82	18	0	29	48	23			0
894271	CR	F_S	Spring 2019	44,402		0-2	11,662	88	12	0	31	39	21			10
894271	CR	F_S	Fall 2019	11,620		0-2	4,294	92	8	0	37	28	12			22
894271	CR	F_S	Summer 2020	800		0-2	368	96	4	0	40	24	2			35
957768	CR	F_S	Spring 2018 FT	1,557		0-2	294	86	14	0	48	27	25			0
957768	CR	F_S	Spring 2019	44,975		0-2	13,494	91	9	0	35	29	22			14
957768	CR	F_S	Fall 2019	12,016		0-2	5,338	96	4	0	35	21	11			31
957768	CR	F_S	Summer 2020	859		0-2	496	100	0	0	44	6	1			49

Form Key: Form C_S = Administrative Error (AE), Form F_S = Operational

*Condition Code notes:
Spring 2017 – B (Blank) was the only condition code in use for ERs and CRs.
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

U.S. History ERs and CRs (continued)

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
894104	ER	C_S	Spring 2017 FT	5,000	Content	0-4	5,000	62	33	5	31	34	22	9	4	0
				5,000	Claims	0-4	5,000	61	32	7	39	28	21	9	4	0
894104	ER	C_S	Fall 2017	7,649	Content	0-4	4028	90	9	0	36	34	20	6	2	1
				7,649	Claims	0-4	4028	89	11	0	45	30	17	6	1	1
894104	ER	C_S	Spring 2018	14,049	Content	0-4	9970	96	4	0	21	34	26	11	6	1
				14,049	Claims	0-4	9970	95	5	0	31	33	21	10	3	1
894104	ER	C_S	Sum 2018	215	Content	0-4	152	96	4	0	75	17	6	1	0	1
				215	Claims	0-4	152	99	1	0	83	12	3	1	0	1
894104	ER	C_S	Spring 2019 Senior	4,624	Content	0-4	3,516	97	2	0	39	24	6	1	1	28
				4,624	Claims	0-4	3,516	98	2	0	50	16	4	1	0	28
894104	ER	C_S	Fall 2020	15,023	Content	0-4	11,348	96	4	0	32	31	16	6	3	12
				15,023	Claims	0-4	11,348	96	4	0	43	25	13	5	2	12
894225	CR	C_S	Spring 2017 FT	1,660		0-2	320	71	29	0	44	34	22			0
894225	CR	C_S	Spring 2018	39,705		0-2	7600	80	19	0	55	24	21			0
894225	CR	C_S	Fall 2018	9,205		0-2	1,694	88	12	0	75	15	10			0
894225	CR	C_S	Summer 2019	3,405		0-2	1,560	99	1	0	59	3	1			37
894225	CR	C_S	Fall 2020	11,189		0-2	4,056	95	5	0	58	11	8			23
892994	CR	C_S	Spring 2017 FT	1,659		0-2	318	68	31	1	13	43	44			0
892994	CR	C_S	Spring 2018	39,867		0-2	7282	78	22	0	22	55	23			0
892994	CR	C_S	Fall 2018	9,375		0-2	1,728	80	20	0	43	39	18			0
892994	CR	C_S	Summer 2019	3,522		0-2	1,752	96	4	0	33	23	4			41
892994	CR	C_S	Fall 2020	11,197		0-2	3,750	93	7	0	23	42	16			19

Form Key: Form C_S = Administrative Error (AE), Form F_S = Operational

*Condition Code notes:
Spring 2017 – B (Blank) was the only condition code in use for ERs and CRs.
Spring 2018 – Condition codes B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – Condition codes B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

Social Studies Grade 3

IDEAS ID	Item Type	Spring 2021 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj %	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	*Cond Code %
801184	CR	Op	Spring 2016 FT	1,281		0-2	248	78	20	2	76	15	9	0
801184	CR	Op	Spring 2017	62,961		0-2	11,436	89	10	1	53	18	22	7
801184	CR	Op	Spring 2019	59,456		0-2	12,804	93	7	0	62	12	19	7
890683	CR	Op	Spring 2017 FT	1,654		0-2	308	81	18	2	42	38	16	3
890683	CR	Op	Spring 2019	59,278		0-2	12,484	86	14	0	56	29	10	5

Social Studies Grade 4

IDEAS ID	Item Type	Spring 2021 Form	Source of IRR and SPD Data	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj %	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	*Cond Code %
801539	CR	Op	Spring 2016 FT	1,654		0-2	308	71	25	3	29	37	30	4
801539	CR	Op	Spring 2017	62,340		0-2	11,406	82	17	1	40	36	20	3
801539	CR	Op	Spring 2019	61,163		0-2	12,630	81	19	0	34	35	26	4
890820	CR	Op	Spring 2017 FT	1,654		0-2	308	85	15	0	80	17	2	2
890820	CR	Op	Spring 2019	60,286		0-2	10,792	92	8	0	79	15	3	2

*Condition Code notes:
Spring 2016 and 2017 – B (Blank) was the only condition code in use for ERs and CRs.
Spring 2018 – B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

Social Studies Grade 5

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
807773	ER	Op	Spring 2016 FT	5,668	Content	0-4	5,668	78	21	1	62	25	12	2	0	0
				5,668	Claims	0-4	5,668	79	20	1	67	23	9	1	0	0
807773	ER	Op	Spring 2019	81,277	Content	0-4	53,370	92	8	0	39	31	15	3	0	12
				81,277	Claims	0-4	53,370	92	7	0	44	28	13	3	0	12
890885	CR	Op	Spring 2017 FT	1,650		0-2	300	76	23	1	54	39	6			0
890885	CR	Op	Spring 2019	63,111		0-2	17,222	88	12	0	37	41	10			11
890920	CR	Op	Spring 2017 FT	1,647		0-2	294	71	29	0	63	28	9			0
890920	CR	Op	Spring 2019	66,346		0-2	23,726	92	8	0	41	34	7			18

Social Studies Grade 6

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
804889	ER	Op	Spring 2016 FT	5,108	Content	0-4	5,108	67	32	1	42	44	12	1	0	0
				5,108	Claims	0-4	5,108	68	31	1	52	38	9	1	0	0
804889	ER	Op	Spring 2017	71,724	Content	0-4	39,110	93	6	0	56	32	10	2	0	0
				71,724	Claims	0-4	39,110	93	7	0	66	24	8	1	0	0
804889	ER	Op	Spring 2019	74,488	Content	0-4	39,812	91	8	0	36	36	13	3	0	12
				74,488	Claims	0-4	39,812	92	8	0	46	31	9	2	0	12
804851	CR	Op	Spring 2016 FT	1,632		0-2	320	73	28	0	46	50	5			0
804851	CR	Op	Spring 2017	56,842		0-2	10,362	80	20	0	41	53	6			0
804851	CR	Op	Spring 2019	62,768		0-2	16,484	88	12	0	27	51	11			11
949224	CR	Op	Spring 2018 FT	1,629		0-2	300	87	13	0	45	53	2			0
949224	CR	Op	Spring 2019	60,633		0-2	12,138	90	10	0	55	35	5			4

*Condition Code notes:
Spring 2016 and 2017 – B (Blank) was the only condition code in use for ERs and CRs.
Spring 2018 – B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

Social Studies Grade 7

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
805627	ER	Op	Spring 2016 FT	5,066	Content	0-4	5,066	73	25	2	45	41	12	2	0	0
				5,066	Claims	0-4	5,066	73	25	2	57	31	11	2	0	0
805627	ER	Op	Spring 2017	68,833	Content	0-4	34,732	91	9	0	48	34	13	4	1	0
				68,833	Claims	0-4	34,732	91	8	0	56	28	11	4	1	0
805627	ER	Op	Spring 2019	79,249	Content	0-4	54,932	94	5	0	34	35	12	4	1	14
				79,249	Claims	0-4	54,932	94	6	0	40	29	11	4	1	14
891266	CR	Op	Spring 2017 FT	1,648		0-2	296	75	25	0	43	43	14			0
891266	CR	Op	Spring 2019	59,206		0-2	14,880	87	13	0	46	35	9			10
805632	CR	Op	Spring 2016 FT	1,626		0-2	314	83	17	0	42	34	24			0
805632	CR	Op	Spring 2017	56,280		0-2	10,274	80	19	1	47	28	25			0
805632	CR	Op	Spring 2019	60,563		0-2	17,546	88	11	0	39	28	19			14

Social Studies Grade 8

IDEAS ID	Item Type	Spring 2021 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 %	*Cond Code %
808905	ER	Op	Spring 2016 FT	5,068	Content	0-4	5,068	65	33	2	30	36	25	7	2	0
				5,068	Claims	0-4	5,068	64	34	2	30	37	25	7	2	0
808905	ER	Op	Spring 2017	65,286	Content	0-4	30,674	89	11	1	32	30	25	9	3	0
				65,286	Claims	0-4	30,674	88	11	1	32	29	25	9	4	0
808905	ER	Op	Spring 2019	75,545	Content	0-4	50,970	92	8	0	17	32	27	10	4	10
				75,545	Claims	0-4	50,970	92	8	0	17	29	28	10	5	10
808955	CR	Op	Spring 2016 FT	1,623		0-2	320	79	21	0	39	40	21			0
808955	CR	Op	Spring 2017	54,395		0-2	10,174	77	22	0	32	51	17			0
808955	CR	Op	Spring 2019	56,385		0-2	12,610	80	20	0	24	53	17			6
892278	CR	Op	Spring 2017 FT	1,656		0-2	312	79	20	1	43	41	15			0
892278	CR	Op	Spring 2018	55,340		0-2	10,110	78	21	0	43	44	13			0
892278	CR	Op	Spring 2019	57,438		0-2	14,752	83	17	0	35	39	19			6

*Condition Code notes:
Spring 2016 and 2017 – B (Blank) was the only condition code in use for ERs and CRs.
Spring 2018 – B, F, I, and U (Blank, Foreign Language, Insufficient, and Unintelligible) were in use for ERs while B (Blank) was the only code in use for CRs.
Spring 2019 – B, F, I, N, R, and U (Blank, Foreign Language, Insufficient, "I don't know," Refusal, and Unintelligible) were in use for ERs and CRs. In addition, in Spring 2019, the definition of condition code "I" was broadened to include copied text response types that would have been scored as 0s in previous years.

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

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 repose 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.