2017-2018 LEAP 2025 High School Operational Technical Report English I, English II, Algebra I, and Geometry

Submitted to the
Louisiana Department of Education

July 2019

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 5
E.1 Overview of This Report 5
E.2 Administration 6
E.3 Student Performance 6
E.4 Validity and Test Scores 7
Chapter 1: Introduction 8
1.1. Background 8
1.2. Purpose of the LEAP 2025 8
1.3. Design of the LEAP 2025 9
Chapter 2: The Uses of Test Scores 12
2.1 Uses of Test Scores 12
2.2 Test-Level Scores 12
2.3 Scale Scores 13
2.4 Levels of Achievement 13
2.5 Use of Test-Level Scores 13
2.6 Claim-, Subclaim- and Subcategory-Level Subscores 14
2.7 Use of the Claim-, Subclaim- and Subcategory-Level Ratings 14
Chapter 3: Test Content Development 15
3.1 Defining the Specific Test Blueprint 17
3.2 Operational Test Selection 35
3.3 Universal Design 42
3.4 Accommodations and Designated Supports 44
3.5 Item and Task Specifications 45
3.6 Summary 45
Chapter 4: Test Administration 47
4.1 Return Material Forms and Guidelines 52
4.2 Security Checklists 52
4.3 Interpretive Guides 52
4.4 Test Security Measures 53
4.5 Test Administration 54

Copyright © 2019 by Louisiana Department of Education.

4.6 Summary 60
Chapter 5: Constructed-Response and Technology-Enhanced Scoring 62
5.1 Constructed-Response Item Scoring Process 62
5.2 Inter-Rater Reliability 69
5.3 Technology-Enhanced Item Scoring Process 79
5.4 Multiple-Choice and Multiple-Select Item Scoring Process 79
5.5 Summary 79
Chapter 6: Operational Data Analyses 81
6.1 Classical Item Statistics 81
6.2 Calibration Sample 86
6.3 Calibration and Linking 91
6.4 Comparability: Form Equating 108
6.5 Summary 109
Chapter 7: Test Results 110
7.1 Student Participation 110
7.2 Reports 122
7.3 Description of Each Type of Report 123
7.4 Summary 124
Chapter 8: Performance-Level Setting 126
8.1 PARCC Performance-Level Setting Process for English Language Arts and Mathematics 126
8.2 Cut Scores 126
8.3 Claim Cut Scores 127
8.4 Achievement-Level Definitions 127
8.5 Summary 128
Chapter 9: Evidence of Construct-Related Validity 129
9.1 Construct-Irrelevant Variance and Construct Underrepresentation 130
9.2 Reliability 130
9.3 Test Reliability 131
9.4 Standard Error of Measurement 132
9.5 Conditional Standard Error of Measurement 133
9.6 Classification Accuracy and Consistency 136
9.7 Convergent Validity 138
9.8 Principal Components Analysis 139

Copyright © 2019 by Louisiana Department of Education.

9.9 Analyses by Claims, Subclaims, and Subcategories 141
9.10 Correlations among Claims, Subclaims, and Subcategories 141
9.11 Reliability of Claims, Subclaims, or Subcategories 144
9.12 Standard Error of Measurement of Claims, Subclaims, or Subcategories 144
9.13 Divergent (Discriminant) Validity 148
9.14 Summary 148
Chapter 10: Fairness 150
10.1 Minimizing Bias through Careful Test Development 151
10.2 Evaluating Bias through Differential Item Functioning (DIF) Statistics 151
10.3 Evaluating Bias through Impact Analysis 157
10.4 Reliability 157
10.5 Effect Size 162
10.6 Summary 172
Appendix A – Transadaptation Process for Spanish Mathematics Forms 174
Appendix B – Accommodated Print Form Creation 177
Appendix C – Item Alignment Review Process 179
Appendix D – LEAP 2025 Spring 2018 Handscoring/AI Documentation 182
Appendix E. Classical Item Statistics 261
Appendix F – Student Participation Counts ..... **Error! Bookmark not defined.**
References 285

Copyright © 2019 by Louisiana Department of Education.

Executive Summary

This report is a technical summary of the 2017-2018 administrations of the Louisiana Educational Assessment Program (LEAP 2025) in English Language Arts (ELA) and mathematics for high school. The LEAP 2025 is a summative assessment in ELA and Mathematics administered in grades 3 through 8 and high school. These tests are designed to measure students' readiness for the next grade or course of study and proficiency in ELA and mathematics. The ELA and mathematics test forms were developed by Data Recognition Corporation (DRC) test development staff using the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium's item bank. Items taken from this bank were on pre-established item response theory (IRT) scales. This section provides a summary of the 2017-2018 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 the uses of the scores.

Test Content Development (Chapter 3)

Chapter 3 of the technical report provides a summary of the test development activities that occurred to create the 2017–2018 operational test forms.

Test Administration (Chapter 4)

Chapter 4 of the technical report describes the processes and activities 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 and activities for scoring constructed-response items. This chapter discusses how scorers are trained and the measures used for ensuring 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 analyses of the 2017-2018 LEAP 2025 ELA and mathematics assessments, which include the following major parts: the

Copyright © 2019 by Louisiana Department of Education.

classical item analysis; calibration, scaling, and linking using item response theory (IRT) models; and student scoring. This chapter also describes the demographics of the calibration samples and compares them to state census data. It reports the results of the classical item analysis and the results of the calibration, scaling, and linking processes.

Test Results (Chapter 7)

Chapter 7 of the technical report contains information on the results of the 2017-2018 LEAP 2025 assessments. Detailed summary statistics based on scale scores and achievement-level information are also provided. Finally, this chapter presents information on the score reports sent to school systems.

Performance-Level Setting (Chapter 8)

Chapter 8 of the technical report briefly discusses performance-level setting. It provides a brief overview of the PARCC performance-level setting procedures and derivation of 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 reliability and validity of the LEAP 2025 test scores. This chapter provides detailed results of the reliability of the tests as well as 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 to determine whether large differences exist between demographic groups in Louisiana. Results of the administration mode study are also summarized.

E.2 Administration

Louisiana administered the LEAP 2025 summative assessments in ELA and mathematics to high school students in 2017-2018. Computer-based tests (CBT) were administered during the following three testing windows: November 29 to December 13, 2017; April 23 to May 18, 2018; June 18-22, 2018. Test administration is discussed in Chapter 4 of this report.

Eighty-three school systems and twenty-two charter schools administered ELA and mathematics LEAP 2025 high school tests across the three administrations.

E.3 Student Performance

Table E.1 presents the percentage of students in 2017-2018 who were classified in each of the achievement levels for each subject.

Copyright © 2019 by Louisiana Department of Education.

Table E.1 Percentage of Students Classified in Achievement Levels Using 2017-2018 Census Data: English Language Arts and Mathematics

Administration	Subject	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
Fall	English I	8.7	16.8	29.3	37.8	7.5
	English II	34.7	19.2	17.0	21.8	7.2
	Algebra I	20.3	47.2	17.6	14.2	0.7
	Geometry	11.2	35.5	28.8	21.1	3.5
Spring	English I	10.1	17.7	27.7	36.4	8.1
	English II	16.2	17.1	22.4	32.3	12.0
	Algebra I	9.2	25.1	27.7	33.8	4.2
	Geometry	7.5	28.8	33.9	26.5	3.3
Summer	English I	58.5	29.4	10.4	1.7	0.0
	English II	64.3	23.9	9.9	1.8	0.0
	Algebra I	28.3	57.6	12.6	1.4	0.1
	Geometry	34.1	56.6	3.8	1.0	4.5

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 use of the LEAP 2025 test scores. Chapter 2 discusses the 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 2017-2018 administrations 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 performance-level setting procedures implemented by PARCC since 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 and the process for evaluating bias to ensure fairness of the LEAP 2025 for all examinees.

Copyright © 2019 by Louisiana Department of Education.

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 High School assessments administered in the 2017- 2018 academic year and presents evidence for the validity of the 2017-2018 LEAP 2025 ELA and mathematics High School 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. Beginning in 2015, students in grades 3-8 began taking these newly aligned assessments.

In 2016, Louisiana ELA and mathematics academic content standards underwent a review process resulting in the adoption of the Louisiana Student Standards in English language arts and mathematics. In spring 2017, ELA and math students in grades 3 through 8, except those qualifying for the LEAP Alternate Assessment Level 1 (LAA 1), took the LEAP 2025 assessments.

Beginning in 2017-2018 the Louisiana Department of Education (LDOE) transitioned to LEAP 2025 high school ELA and mathematics assessments aligned to the Louisiana Student Standards in ELA and mathematics. The five-performance-level LEAP 2025 high school assessments replaced the four-performance-level End-of-Course (EOC) tests. The implementation (use) of LEAP 2025assessments provides a coherent and consistent measure of student performance and growth from grades three through high school. Students in grades 9-12 enrolled in English I, English II, Algebra I, and Geometry took the corresponding LEAP 2025 high school assessments.

The information that follows describes the technical aspects of the 2017-2018 LEAP 2025 ELA and Mathematics assessments and provides information about how to read and interpret the data on the 2017-2018 assessment reports.

1.2. Purpose of the LEAP 2025

The Board of Elementary and Secondary Education (BESE) and the LDOE are committed to ensuring that every student is on track to be successful in postsecondary education and the workforce through their comprehensive plan, Louisiana Believes. The LEAP 2025 supports this vision by measuring the full range of student performance, including the performance of high- and low-performing students and providing

Copyright © 2019 by Louisiana Department of Education.

information for educators and parents about student readiness and whether students are “on track” for college and careers.

1.3. Design of the LEAP 2025

High School students were administered computer-based tests (CBTs) in both ELA and mathematics. Additionally, a braille form was available for each course and content area. Online tools allowed students to magnify assessment items, as needed. See Chapter 3, Section 3.4 for more information about the accommodations and designated supports available for LEAP 2025 testers. All mathematics assessments were translated into Spanish forms.

PARCC operational blueprints are the basis of the design of the 2017-2018 LEAP 2025 test blueprints and test design. The 2017-2018 LEAP 2025 test blueprints and test design for ELA and mathematics differ from PARCC in order to reduce testing time but maintain full coverage and include a variety of standards.

The 2017-2018 LEAP 2025 ELA blueprints kept a similar design as PARCC; including both performance-based tasks and standalone passage; however, only two of the three types of performance tasks—Research Simulation Task (RST) and Literary Analysis Task (LAT) or Narrative Writing Task (NWT)—are included on each of the HS ELA assessments. All three task types are represented across administrations, which encourages teachers to focus equally on all three writing types. Besides having two (instead of three) performance tasks, the 2017-2018 LEAP 2025 ELA blueprints are also different from PARCC 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 course.

The passages chosen for the 2017-2018 LEAP 2025 English I and English II assessments contain a variety of texts, including different genres, a balance of authors by gender and ethnicity, and texts that appeal to a diverse student population. Chosen passages are authentic, contain a variety of different types of text 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. This combination of criteria during passage selection allows students to demonstrate their ability to read and comprehend a range of complex texts and helps 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, that are essential to understanding a

Copyright © 2019 by Louisiana Department of Education.

particular text, and that include context that allows students to determine literal and figurative meanings.

- 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 Algebra I and Geometry assessments 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).

Copyright © 2019 by Louisiana Department of Education.

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

Copyright © 2019 by Louisiana Department of Education.

Chapter 2: The Uses of Test Scores

Validity is the central component 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 use of the test-level scores (i.e., scale scores and achievement levels), claim-level scores and achievement-level classifications, and subclaim 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. Test-level scores are reported at four reporting levels: the state, the school system, the school, and the student.

The LEAP 2025 high school ELA and mathematics test forms were developed by DRC's test development staff using the PARCC consortium's item bank. Items taken from this bank were on pre-established item

Copyright © 2019 by Louisiana Department of Education.

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, in addition to 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 to indicate a student's achievement on the LEAP 2025 assessments: (1) the scale score and (2) its associated level of achievement.

2.3 Scale Scores

A scale score indicates a student's total performance on the LEAP 2025 assessments. The overall scale score quantifies the achievement being measured by the 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 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) (see https://parcc-assessment.org/wp-content/uploads/2017/12/PARCC_PLS_TechReport_011316_toPARCC-final.pdf).

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 relative to the Louisiana Student Standards. Classroom teachers may use these scores as evidence of student achievement in English I, English II, Algebra I, and Geometry. 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

Copyright © 2019 by Louisiana Department of Education.

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 Claim-, Subclaim- and Subcategory-Level Subscores

A student's performance on the ELA claims (i.e., reading and writing) is reported by one of three ratings: *Strong, Moderate*, or *Weak*.

Additionally, subclaim subscores are reported at the student level for ELA and mathematics. ELA has three subclaims for reading and two subclaims for writing, as described in Table 3.1, ELA Claims and Subclaims. Mathematics has four subclaims, as described in Table 3.6, Overview of LEAP 2025 Mathematics Task Types and Reporting Categories. Subclaim 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 claim, subclaim, 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 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 Claim-, Subclaim- and Subcategory-Level Ratings

The purpose of reporting claim-, subclaim-, or subcategory-level subscores 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 claims and subclaims in ELA and the subclaims and subcategories in mathematics. 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 claim-, subclaim- or subcategory-level subscores. Chapter 9 of this technical report provides evidence of construct-related validity that further supports the use of these subscores.

Copyright © 2019 by Louisiana Department of Education.

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 section 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 2017–2018 LEAP 2025 high school 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 2017–2018 LEAP 2025 high school test specifications consisted of a blueprint and a design for each of the following tests: English I, English II, Algebra I, and Geometry. The 2017–2018 blueprints and test designs were closely aligned to the PARCC flagship blueprint that was used for the PARCC 2017–2018 test administrations. The specific course test blueprints for the 2017–2018 LEAP 2025 ELA assessments for high school were designed with the goal for all students to read, understand, and express understanding of complex, grade-level texts. The specific course test blueprints for the 2017–2018 LEAP 2025 mathematics assessments for high school 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 2017–2018 LEAP 2025 ELA and mathematics assessments for high school provide questions that have been reviewed by Louisiana educators to ensure their alignment to the Louisiana Student Standards and appropriateness for Louisiana students; measure the full range of student performance, including the performance of high- and low-performing

Copyright © 2019 by Louisiana Department of Education.

students; and inform educators and parents about student readiness in ELA and mathematics and whether students are "on track" for college and careers.

To construct the assessments following the LDOE-approved test blueprints and test designs, LDOE and DRC collaborated to use items from the PARCC item bank that were aligned to the Louisiana Student Standards. DRC contracted with PARCC and was provided access to the entire bank of items and related passages that could potentially be used on operational forms. The acquired items/passages made up the available item pool used for the 2017–2018 LEAP 2025 high school forms construction. Please refer to the PARCC Model Content Frameworks for ELA/Literacy (Grades 3-11) and the PARCC Model Content Frameworks for Mathematics (Grades 3-11) for additional information about the development of item specifications and blueprints for the PARCC assessments. These resources can be accessed via the Educator Resources page of the PARCC website. Once at the website, for ELA click on PARCC Model Content Frameworks for ELA/Literacy (Grades 3-11) and for mathematics click on PARCC Model Content Frameworks for Mathematics (Grades 3-11). 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 PARCC item bank prior to form selection. Educators reviewed alignment to content standards and appropriateness of item content.

The ELA and mathematics LEAP 2025 assessments for high school were developed based on the requirements of RFP #678PUR-LEAP 2016 Mathematics and ELA as follows:

The assessments shall be

- aligned to the ELA and mathematics Louisiana Student Standards;
- designed to be accessible for use by the widest possible range of students, including, but not limited to, students with disabilities and students with limited English proficiency;
- constructed to yield valid and reliable test results;
- constructed to report student performance using achievement level policy definitions and reporting categories which are comparable to a significant number of other states;
- developed to limit the amount of testing time required and to be in compliance with all state law regarding testing time;
- developed and reviewed with Louisiana educator involvement;
- 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.

The products of the above requirements are computer-based tests (CBTs) composed of PARCC test items aligned to the Louisiana Student Standards. Louisiana had access to the complete PARCC item bank from which to select for the alignment review more than enough items to build the forms needed for the 2017–2018 LEAP 2025 high school ELA and mathematics assessments. The items and passages approved by Louisiana educators during the item alignment review as appropriate for use on a Louisiana assessment (aligned to the Louisiana Student Standards and/or Louisiana Evidence Statements for mathematics and free of issues of bias, fairness, and sensitivity) became the available item pool used to construct the 2017–2018 forms. For each course, three test forms were selected from the available pool of items and/or passages. DRC and LDOE content experts scrutinized each final blueprint to ensure

Copyright © 2019 by Louisiana Department of Education.

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 in each course and are comparable to PARCC 2017–2018 test blueprints. The test specifications provide the numbers of items by reporting category, assessment focus, and item type, and they demonstrate the desired proportions within test delivery and available item pool constraints. These specifications can be found in the *2017-2018 LEAP 2025 High School English Language Arts and Mathematics Assessment Frameworks*. All assessments were fixed forms, in that all students who received the same form were administered the same set of items, as the forms were not adaptive.

The LEAP 2025 high school assessments are administered in fall, spring, and summer each school year. For each administration, two forms are administered: one as an operational form, the other as an administrative error form, which is used only if there is an administrative testing error (see Chapter 4 for additional details regarding the administrative error form). The forms are rotated such that the operational form and the administrative error form are not the same from administration to administration.

3.1 Defining the Specific Test Blueprint

The specific course 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.

English I and English II Test Blueprints and Test Designs

The English I and English II tests were administered during operational testing windows: November 29 to December 13, 2017; April 23 to May 18, 2018; June 18 to June 22, 2018. 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 tests. All three types were represented across administrations (fall, spring, summer), which encourages educators to focus on all three writing types. 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—are created for each course.

Student performance on the LEAP 2025 high school ELA assessments is reported by claim and subclaim as outlined in the following table.

Copyright © 2019 by Louisiana Department of Education.

Table 3.1 ELA Claims and Subclaims

Claim	Subclaim	Subclaim 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 provide parents and educators with valuable information about

- overall student performance, including readiness to continue further studies in English language arts;
- student performance broken down by subclaim, 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 and 3.3) are based on PARCC testing times proportioned to be comparable based on the passage type. The passage set that comes after the Narrative Writing Task or the Literary Analysis Task is designed to balance the reading load between the Narrative Writing Task or the Literary Analysis Task and the Research Simulation Task and to provide consistent timing in sessions one and two.

Copyright © 2019 by Louisiana Department of Education.

Table 3.2 English I and English II Test Design—Literary Analysis Task and Research Simulation Task

Section	Task/ Item Set	Number of Passages	Claims/Subclaims	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 subclaim)	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 of Language and 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 of Language and 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	2-3	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	2-3		10	20	0	0	10	20		
English I & II Totals		**8-9**	Reading: Reading Literary Text/Reading Vocabulary*	28	56	2	4	28	64	64	260
			Reading: Reading Informational Text/Reading Vocabulary*				4				
			Writing: Written Expression	0	0		24	2	24	30	
			Writing: Knowledge of Language and Conventions	0	0		6		6		
			Total	28	56	2	38	30	94	94	

*There must be at least eight points of reading vocabulary items on the test.

**The time in session 3 includes an additional passage set that is being field tested or acting as a placeholder.

Copyright © 2019 by Louisiana Department of Education.

Table 3.3 English I and English II Test Design—Research Simulation Task and Narrative Writing Task

Section	Task/ Item Set	Number of Passages	Claims/Subclaims	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 subclaim)	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 of Language and 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 of Language and 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	2-3	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 Vocabulary*			0	0			RI.1-3, 5, 7-10; vocabulary standards RI.4, L.4, L.5	
	Totals	2-3		10	20	0	0	10	20		
English I & II Totals		**7-9**	Reading: Reading Literary Text/Reading Vocabulary*	28	56	2	0	28	60	60	260
			Reading: Reading Informational Text/Reading Vocabulary*				4				
			Writing: Written Expression	0	0		24	2	24	30	
			Writing: Knowledge of Language and Conventions	0	0		6		6		
			Total	28	56	2	34	30	90	90	

*There must be at least eight points of reading vocabulary items on the test.

**The time in session 3 includes an additional passage set that is being field tested or acting as a placeholder.

Copyright © 2019 by Louisiana Department of Education.

The LEAP 2025 high school ELA assessments consist of tasks and reading passages. 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 then create a narrative related to the text (e.g., finish the story; retell the story in another narrative form, such as a journal entry).
    - This task focuses on a student's ability to use narrative elements (dialogue, description) when writing.
- Literary Analysis Task
    - This task provides students 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 a student's 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 2017–2018 LEAP 2025 high school ELA assessments:

- Selected-Response Items:
    - Evidence-based selected response – ESR: 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 opportunities to make explicit the evidence that supports their close analysis of a specific text.
    - Multiple select – MS: This item type requires a student to select more than one correct answer and may appear as a one-part question or as part of an ESR 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: ordering of ideas within a summary; ordering steps in a process; sorting, classifying, categorizing; matching two themes/ideas to their unique evidence, etc. The technology offers students additional ways to show understanding that parallels classroom instructional techniques teachers use to determine whether students are able to comprehend grade-level, complex text.
      TE Items may involve any of the following:

Copyright © 2019 by Louisiana Department of Education.

    - Highlighting text: requires a student to select text-based answer(s) from within a larger text
    - Drag and drop: requires a student 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)
    - Dropdown menu: requires a student to select from one or more dropdown menus to complete a phrase or sentence
    - Match interaction table: requires a student 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 of the tasks and asks students to create an extended and 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 knowledge of language and conventions.

A variety of item types allows for the measurement of the full range of student performance, including the performance of high- and low-performing students. Items and tasks should be clearly aligned with 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.

Copyright © 2019 by Louisiana Department of Education.

The following tables detail the number of items and points by session and item type for each of the English I and English II forms.

Table 3.4 Distribution of English I Items and Points by Session and Item Type

Form	Administration	Session	ESR		MS		TE		PCR		Total
			Items	Pts	Items	Pts	Items	Pts	Items	Pts	Pts
A	Fall (OP) Spring (AE) Summer (AE)	1. Research Simulation Task	6	12	0	0	2	4	1	19	90
		2. Narrative Writing Task/ Reading Passage	5	10	2	4	3	6	1	15	
		3. Reading Literary/ Informational Texts	8	16	1	2	1	2	0	0	
B	Spring (OP)	1. Literary Analysis Task/ Reading Passage	8	16	0	0	2	4	1	19	94
		2. Research Simulation Task	6	12	0	0	2	4	1	19	
		3. Reading Literary/ Informational Texts	6	12	1	2	3	6	0	0	
C	Fall (AE) Summer (OP)	1. Research Simulation Task	6	12	1	2	1	2	1	19	90
		2. Narrative Writing Task/ Reading Passage	5	10	3	6	2	4	1	15	
		3. Reading Literary/ Informational Texts	6	12	3	6	1	2	0	0	

OP: Operational

AE: Administrative Error

Copyright © 2019 by Louisiana Department of Education.

Table 3.5 Distribution of English II Items and Points by Session and Item Type

Form	Administration	Session	ESR		MS		TE		PCR		Total
			Items	Pts	Items	Pts	Items	Pts	Items	Pts	Pts
A	Fall (OP) Spring (AE) Summer (AE)	1. Research Simulation Task	5	10	2	4	1	2	1	19	90
		2. Narrative Writing Task/ Reading Passage	2	4	7	14	1	2	1	15	
		3. Reading Literary/ Informational Texts	5	10	4	8	1	2	0	0	
B	Spring (OP)	1. Literary Analysis Task/ Reading Passage	4	8	3	6	3	6	1	19	94
		2. Research Simulation Task	5	10	2	4	1	2	1	19	
		3. Reading Literary/ Informational Texts	4	8	4	8	2	4	0	0	
C	Fall (AE) Summer (OP)	1. Literary Analysis Task/ Reading Passage	8	16	0	0	2	4	1	19	94
		2. Research Simulation Task	5	10	2	4	1	2	1	19	
		3. Reading Literary/ Informational Texts	2	4	4	8	4	8	0	0	

OP: Operational

AE: Administrative Error

Mathematics Test Blueprints and Test Designs

The mathematics assessments were administered during operational testing windows: November 29 to December 13, 2017; April 23 to May 18, 2018; June 18 to June 22, 2018. The 2017–2018 mathematics assessments included three test sessions, and each test session included the four mathematics subclaims, using the three mathematics task types (see Table 3.6).

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

Copyright © 2019 by Louisiana Department of Education.

Table 3.6 Overview of LEAP 2025 Mathematics Task Types and Reporting Categories

Task Type	Description	Reporting Category	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 studies in mathematics;
- student performance broken down by mathematics subclaim, 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.

Copyright © 2019 by Louisiana Department of Education.

Tables 3.7 and 3.8 provide the distribution of operational points by reporting category, by form for each course.

Table 3.7 Distribution of Points by Reporting Category—Algebra

	Form		
Reporting Category	**A**	**B**	**C**
Major Content	28	28	28
Additional & Supporting Content	13*	14	14
Expressing Mathematical Reasoning	11	11	11
Modeling & Application	15	15	15
Total	**67**	**68**	**68**

* A one-point item within the Additional & Supporting Content reporting category was dropped from scoring in Form A.

Table 3.8 Distribution of Points by Reporting Category—Geometry

	Grade		
Reporting Category	**A**	**B**	**C**
Major Content	26	26	26
Additional & Supporting Content	16	16	16
Expressing Mathematical Reasoning	11	11	11
Modeling & Application	15	15	15
Total	**68**	**68**	**68**

Copyright © 2019 by Louisiana Department of Education.

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

Table 3.9 Major Content Subcategories by Course

Course	Major Content Subcategory
Algebra	• Interpreting Functions • Solving Algebraically • Solving Graphically/Rate of Change
Geometry	• Congruence Transformations/Similarity • Similarity in Trigonometry/Modeling & Applying

The resulting 2017–2018 LEAP 2025 mathematics test blueprints are shown in Tables 3.10 and 3.11.

Table 3.10 Algebra I Test Blueprint

Reporting Category	Major Content	Additional & Supporting Content	Expressing Mathematical Reasoning	Modeling & Application
Task Type	Type I: I.1 (24 items, 24 points) I.2 (7 items, 14 points) I.4 (1 item, 4 points) Total: 32 items, 42 points (62% of total)		Type II: II.3 (1 item, 3 points) II.4 (2 items, 8 points) Total: 3 items, 11 points (16% of total)	Type III: III.3 (3 items, 9 points) II.6 (1 item, 6 points) Total 4 items, 15 points
Total OP Points	28 (42% of total)	14 (22% of total)	11 (16 % of total)	15 (20% of total)
Assessable Content	A1: A-APR.A.1 A1: A-CED.A.3 A1: A-CED.A.4 A1: A-REI.B.3 A1: A-REI.B.4a A1: A-REI.B.4b A1: A-REI.D.10 A1: A-REI.D.11 A1: A-REI.D.12 A1: A-SSE.A.1 A1: A-SSE.A.2 A1: F-IF.A.1 A1: F-IF.A.2 A1: F-IF.B.4 A1: F-IF.B.5 A1: F-IF.B.6 LEAP.I.A1.1 LEAP.I.A1.2 LEAP.I.A1.3 LEAP.I.A1.4 LEAP.I.A1.5 LEAP.I.A1.6	A1: A-APR.A.3 A1: A-REI.C.6 A1: A-SSE.B.3a A1: A-SSE.B.3b A1: A-SSE.B.3c A1: F-BF.B.3 A1: F-IF.C.7a A1: F-IF.C.7b A1: F-IF.C.8a A1: F-IF.C.9 A1: F-LE.A.2 A1: S-ID.B.5 LEAP.I.A1.7	LEAP.II.A1.1 LEAP.II.A1.2 LEAP.II.A1.3 LEAP.II.A1.4 LEAP.II.A1.5 LEAP.II.A1.6 LEAP.II.A1.7 LEAP.II.A1.8 LEAP.II.A1.9 LEAP.II.A1.10	LEAP.III.A1.1 LEAP.III.A1.2 LEAP.III.A1.3 LEAP.III.A1.4

Copyright © 2019 by Louisiana Department of Education.

Table 3.11 Geometry Test Blueprint

Reporting Category	Major Content	Additional & Supporting Content	Expressing Mathematical Reasoning	Modeling & Application
Task Type	Type I: I.1 (24 items, 24 points) I.2 (7 items, 14 points) I.4 (1 item, 4 points) Total: 32 items, 42 points (62% of total)		Type II: II.3 (1 item, 3 points) II.4 (2 items, 8 points) Total: 3 items, 11 points (16% of total)	Type III: III.3 (1 item, 3 points) II.6 (2 items, 12 points) Total 4 items, 15 points (22% of total)
Total OP Points	26 (38% of total)	16 (24% of total)	11 (16% of total)	15 (22% of total)
Assessable Content	GM: G-CO.B.6 GM: G-GPE.B.6 GM: G-SRT.A.1a GM: G-SRT.A.1b GM: G-SRT.A.2 GM: G-SRT.A.2 GM: G-SRT.B.5 GM: G-SRT.C.6 GM: G-SRT.C.7 GM: G-SRT.C.8 LEAP.I.GM.1 LEAP.I.GM.2	GM: G-C.A.2 GM: G-CO.A.1 GM: G-CO.A.3 GM: G-CO.A.5 GM: G-GMD.A.1 GM: G-GMD.A.3 GM: G-GMD.B.4 GM: G-GPE.A.1 LEAP.I.GM.3 LEAP.I.GM.4 LEAP.I.GM.5	LEAP.II.GM.1 LEAP.II.GM.2 LEAP.II.GM.3 LEAP.II.GM.4	LEAP.III.GM.1 LEAP.III.GM.2 LEAP.III.GM.3 LEAP.III.GM.4 LEAP.III.GM.5

Copyright © 2019 by Louisiana Department of Education.

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 session 1a of the mathematics assessments, students are prohibited from using calculators, except for those students with a calculator accommodation. Calculators are allowed to be used by all students in sessions 1b, 2, and 3. The general test structures (see Tables 3.12 and 3.13) guided test form sequencing and design. The LEAP 2025 Calculator Policy provided the basis for calculator designation of tasks and items.

Table 3.12 Algebra I Testing Sessions

Reporting Category	Session 1a: No Calculator	Session 1b: Calculator	Session 2: Calculator	Session 3: Calculator	Totals
Major Content (points)	5	5	9	9	28
Additional & Supporting Content (points)	4	2	4	4	14
Expressing Mathematical Reasoning (points)	0	3	4	4	11
Modeling & Application (points)	0	3	6	6	15
Total Operational Points	9	13	23	23	68
Test Duration **(minutes)	25	55	80	80	240
# of Operational Items	I.1: 5 I.2: 2 I.4: 0 II.3: 0 II.4: 0 III.3: 0 III.6: 0	I.1: 3 I.2: 0 I.4: 1 II.3: 1 II.4: 0 III.3: 1 III.6: 0	I.1: 9 I.2: 2 I.4: 0 II.3: 0 II.4: 1 III.3: 0 III.6: 1	I.1: 7 I.2: 3 I.4: 0 II.3: 0 II.4: 1 III.3: 2 III.6: 0	I.1: 24 I.2: 7 I.4: 1 II.3: 1 II.4: 2 III.3: 3 III.6: 1

** 6 Embedded Field Test Items were included throughout the assessment; they are included in the total time.

Copyright © 2019 by Louisiana Department of Education.

Table 3.13 Geometry Testing Sessions

Reporting Category	Session 1a: No Calculator	Session 1b: Calculator	Session 2: Calculator	Session 3: Calculator	Totals
Major Content (points)	5	5	8	8	26
Additional & Supporting Content (points)	4	2	5	5	16
Expressing Mathematical Reasoning (points)	0	3	4	4	11
Modeling & Application (points)	0	3	6	6	15
Total Operational Points	9	13	23	23	68
Test Duration **(minutes)	25	55	80	80	240
# of Operational Items	I.1: 5 I.2: 2 I.4: 0 II.3: 0 II.4: 0 III.3: 0 III.6: 0	I.1: 3 I.2: 0 I.4: 1 II.3: 1 II.4: 0 III.3: 1 III.6: 0	I.1: 7 I.2: 3 I.4: 0 II.3: 0 II.4: 1 III.3: 0 III.6: 1	I.1: 9 I.2: 2 I.4: 0 II.3: 0 II.4: 1 III.3: 0 III.6: 1	I.1: 24 I.2: 7 I.4: 1 II.3: 1 II.4: 2 III.3: 1 III.6: 2

** 6 Embedded Field Test Items were included throughout the assessment; they are included in the total time.

Copyright © 2019 by Louisiana Department of Education.

The following item types were used in the 2017–2018 LEAP 2025 mathematics assessments:

- Multiple choice: This item type requires a student 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 a student 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 a student to enter a numeric response by typing from the keyboard. 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. Answers to SA items can be positive or negative and must be entered in integer or decimal form.
- Keypad input: This item type requires a student 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 a student 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 course) 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 a student to complete a bar graph or histogram by raising/lowering each bar to a value
    - Drag and drop: requires a student to move draggable elements into one or more drop boxes
    - Dropdown menu: requires a student to select from one or more dropdown menus to complete a sentence, phrase, or expression/equation/inequality

Copyright © 2019 by Louisiana Department of Education.

- Hot spot: requires a student to select one or more responses by choosing selectable areas on the screen
- Match interaction table: requires a student to select a checkbox in each row from two or more columns
- Graph input: requires a student 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 a student to complete a line plot with "X" as the input

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

The following table details the number of items by point value and task type as well as the number of points per task type for each form.

Table 3.14 Distribution of Mathematics Tasks and Points by Task Type

Form	Content Area	Type I				Type II			Type III			Total Points
		1 pt Tasks	2 pt Tasks	4 pt Tasks	Points	3 pt Tasks	4 pt Tasks	Points	3 pt Tasks	6 pt Tasks	Points	
A	Algebra	23*	7	1	41	1	2	11	3	1	15	67*
B	Algebra	24	7	1	42	1	2	11	3	1	15	68
C	Algebra	24	7	1	42	1	2	11	3	1	15	68
A	Geometry	24	7	1	42	1	2	11	1	2	15	68
B	Geometry	24	7	1	42	1	2	11	1	2	15	68
C	Geometry	24	7	1	42	1	2	11	1	2	15	68

* One Type I item worth 1 point was dropped from scoring in form A.

Copyright © 2019 by Louisiana Department of Education.

Item Development and Selection

The process 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 2017–2018 LEAP 2025 high school ELA and mathematics assessments came from the PARCC consortium's item bank.

The PARCC items selected for use on the 2017–2018 LEAP high school forms were used to equate to the PARCC scale. Operational forms were selected based on LEAP 2025 high school test blueprint specifications, which were supported by statistical data from PARCC operational testing.

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 during the development of the PARCC 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 Item Alignment Reviews.

Item Reviews

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

- text reviews of all passages (during which participants review and edit passages independently and then discuss content and bias concerns as a grade level group),
- item reviews (during which committees review and edit items for adherence to PARCC foundational documents, basic principles of Universal Design, PARCC Accessibility Guidelines, selected metadata fields, and the PARCC Style Guide),
- bias and sensitivity reviews (during which educators and community members review items and tasks to confirm the absence of issues relating to bias, fairness, and sensitivity to ensure that items and tasks do not unfairly advantage or disadvantage any student subgroup over another subgroup),
- editorial reviews (during which the review committee completes a copy edit review and records their comments), and

Copyright © 2019 by Louisiana Department of Education.

- data reviews (during which educators evaluate item-level statistics to determine eligibility of items and tasks to move forward to the operational assessments).

Additional information on PARCC's item review processes and procedures can be found at https://parcc-assessment.org/test-development/. Only items that have been approved by expert reviewers during text reviews (ELA only), item reviews, bias and sensitivity reviews, and editorial reviews are moved forward for field testing by PARCC affiliate states. Of the field tested items, only those determined to have acceptable statistics, either by having acceptable item parameters according to the data review flagging criteria or by being approved by expert reviewers during data review, are eligible for review by Louisiana educators for potential use on an operational assessment. These processes follow the *Standards*.

Standard 3.1 states the following:

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

Standard 3.2 states the following:

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

Independent studies of PARCC passages and items have found that the content being licensed assesses the skills that matter most and is rigorous, aligned to standards, and accessible to students with disabilities and English learners. For more information on the studies performed, refer to PARCC's website: https://parcc-assessment.org/independent-studies/.

Louisiana Item Alignment Review

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

DRC, with guidance from LDOE, conducted the Louisiana Item Alignment Reviews in June 2017 with committees of Louisiana educators. Course-specific committees met for two (mathematics) or three (ELA) days to provide feedback on the alignment and appropriateness of items that made up the PARCC item bank. To the extent possible, each committee included educators from different parts of Louisiana, who represent all Louisiana students (e.g., special education, English learners, students with disabilities, etc.). Committee members are also representative of the diverse demographics of the state.

As described in the preceding sections, items presented at this review had already been through a rigorous review process before and after the items were field tested by PARCC to ensure quality and appropriateness. They were then selected for inclusion in the form selection pool, imported into IDEAS, DRC's item banking system, and formatted for use on a Louisiana test form. The items were then placed on mock test forms to allow them to be reviewed as students would see the items, tasks, and passages.

Copyright © 2019 by Louisiana Department of Education.

Louisiana educators then reviewed these items to confirm that the items were acceptable for use in a Louisiana assessment. Educators reviewed items individually to verify that each item aligned to the Louisiana Student Standard(s) for that item prior to discussing the items as a group. In addition, educators reviewed item keys and discussed the cognitive complexity for each item and task. The group then came to a consensus regarding the status of each item: Accepted with Current Alignment, Accepted with Realignment, or Rejected. Items that were accepted were determined to appropriately measure the intended standard(s) and be free of issues of bias, fairness, or sensitivity that could impact student responses to the item. For a detailed description of the process followed during the Item Alignment Reviews, including results and descriptions of the demographics of each committee, see Appendix C, *Item Alignment Review Process*.

3.2 Operational Test Selection

Operational item selection for 2017–2018 took place from June 2017 through July 2017 by LDOE and DRC. The PARCC item pool was used to select fixed LEAP 2025 ELA and mathematics high school forms.

The LEAP 2025 high school assessments were given as computer-based tests (CBTs). For students unable to participate in a CBT administration, accommodated print forms were available for secure download and printing by authorized users. Test administrators then transcribed all student responses into the appropriate CBT test form. (See chapter 4 for additional details.)

English Language Arts Item and Passage Selection Process and Criteria

The item and passage selection process used for forms construction was a content-focused, collaborative process between the LDOE and DRC ELA content specialists, and it was followed by a psychometric evaluation of each selection. The critical psychometric consideration, other than individual item performance, was the degree to which the PARCC items reflected the 2018 LEAP 2025 targets, which were to match the Spring 2017 PARCC operational forms. Although the PARCC item pool was limited, items that were determined to be very difficult (IRT difficulty parameter B b > 2.0) and/or not discriminating (IRT A discrimination parameter a < 0.3) items were avoided when possible.

Item Selection Guidelines

- Using the PARCC pool of items, ELA content specialists select performance tasks and passage sets to match the blueprint. The performance tasks and passage sets include items that cover a range of standards and address the appropriate claims and subclaims.
- ELA content specialists verify that each item aligns to the content standards specified in the Louisiana English Language Arts Student Standards.
- ELA content specialists verify that each item meets psychometric guidelines for excellence as available item-performance data allows.
- ELA content specialists verify that each item meets technical quality requirements for well-crafted items, including
  - meeting the requirements of the item specifications;
  - clear and correct answer(s);
  - clear and concise language;
  - grammatical correctness;
  - appropriate range of difficulty; and
  - content that is not offensive, inappropriate, or biased.

Copyright © 2019 by Louisiana Department of Education.

ELA Content Review and Forms Development

After ELA performance tasks and passage sets were selected and reviewed by DRC and LDOE ELA content specialists, they were placed in forms. In constructing the forms, DRC and LDOE ELA content specialists used the guidelines provided in the following list.

Guidelines for Placing Tasks and Passage Sets into Forms

- Forms should include adequate content coverage, as required by the detailed test blueprint.
- No item in a form should "clue" another item on that same form (i.e., inadvertently provide or clue the answer for another item).
- Clang association should be avoided. Clang association occurs when a distractor can be associated with a stem word by sound (e.g., rhyming, alliteration) rather than meaning.
- Passage sets in forms should be diverse (i.e., contain a balance of authors by gender and ethnicity, and appeal to a diverse student population, etc.).
- Forms should include a wide range of topics and a variety of question types.
- Correct answer distributions should follow best practices (i.e., no more than three keys of the same answer option in a row).
- Forms *must not* contain any items that have been released to the public.

Review of the ELA Items and Forms

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

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

Copyright © 2019 by Louisiana Department of Education.

Guidelines for Reviewing Items Selected for Forms

A good item should

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

Mathematics Item Selection Process and Criteria

The item selection process used for forms construction was a content-focused, collaborative process between the LDOE and DRC mathematics content specialists, and it was followed by a psychometric evaluation of each selection. The critical psychometric consideration, other than individual item performance, was the degree to which the PARCC items reflected the 2018 LEAP 2025 targets, which were to match the Spring 2017 PARCC operational forms. Although the PARCC item pool was limited, items that were determined to be very difficult (IRT difficulty parameter $b > 2.0$) and/or not discriminating (IRT discrimination parameter $a < 0.3$) items were avoided when possible.

Copyright © 2019 by Louisiana Department of Education.

Item-Selection Guidelines

- Using the PARCC pool of items, mathematics content specialists select tasks to match the blueprint. The tasks include items that cover a range of standards and address the appropriate subclaims.
- Mathematics content specialists verify that each item aligns to the content standards specified in the Louisiana Student Standards for Mathematics.
- Mathematics content specialists verify that each item meets psychometric guidelines for excellence as available item performance data allows.
- Mathematics content specialists verify that each item meets technical quality requirements for well-crafted items, including
  - meeting the requirements of the item specifications;
  - clear and correct answer(s);
  - clear and concise language;
  - grammatical correctness;
  - appropriate range of difficulty; and
  - content that is not offensive, inappropriate, or biased.

Mathematics Content Review and Forms Development

After items have been selected and reviewed by DRC and LDOE mathematics content specialists for both psychometric excellence and technical quality, they were placed in forms. In constructing the forms, the mathematics content specialists used the guidelines provided in the following list.

Guidelines for Placing Items into Forms

- Forms should include adequate content coverage, as required by the detailed test blueprint.
- No item in a form should "clue" another item on that same form (i.e., inadvertently provide or clue the answer for another item). If it is necessary to use items with clueing issues due to item pool limitations, then these items should be positioned in separate testing sessions.
- Clang association should be avoided. Clang association occurs when a distractor can be associated with a stem word by sound (e.g., rhyming, alliteration) rather than meaning.
- Forms should be ethnically diverse, both in terms of artwork and in terms of names.
- Forms should target an equal representation of genders, both in terms of artwork and names.
- Forms should include a wide range of topics and a variety of question types.
- Correct answer distributions should follow best practices (i.e., no more than three keys of the same answer option in a row).
- Forms *must not* contain any items that have been released to the public.

Review of the Mathematics Items and Forms

DRC and LDOE mathematics content specialists also ensured the items were in compliance with the guidelines provided by LDOE, including alignment to the content standards and appropriateness for Louisiana students. As establishing content validity is one of the most important aspects in the legal defensibility of a test, the alignment of the item to the content standard must be reviewed and verified at every stage of the test development process. As a result, it is essential that an item selected for a

Copyright © 2019 by Louisiana Department of Education.

form link directly to the content standard that it purports to measure. The mathematics content specialists also verified all items against their designated content codes and metadata, both to evaluate the correctness of the coding and to ensure that the given item measures what it purports to measure.

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

Guidelines for Reviewing Items Selected for Forms

A good item should

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

Item-Selection Options for Special Cases

While every effort is made to select a test form that meets all psychometric guidelines for excellence, it may not be possible to comply with all the psychometric criteria for item/form difficulty due to item pool limitations. In these cases, critical psychometric guidelines are followed while allowing some

Copyright © 2019 by Louisiana Department of Education.

tolerance on less critical item-selection guidelines. The tolerance of meeting target characteristics, the relative exposure of previously used operational items, and other considerations (such as content coverage) may possibly be affected in such cases.

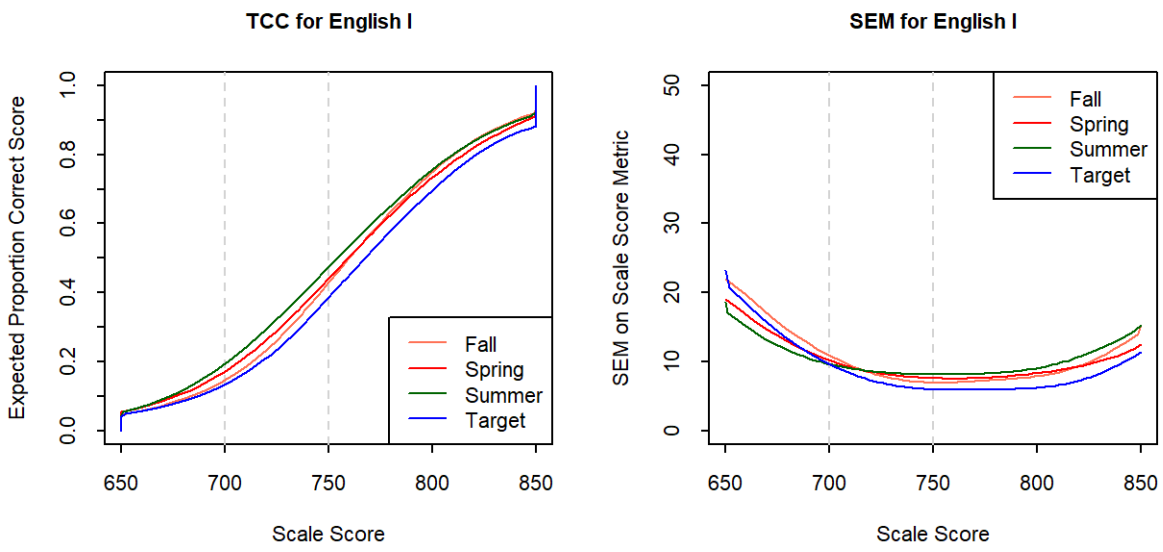
Psychometric Review

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

Selecting Targets

The PARCC 2017 operational form was used to build the new PARCC 2017 base scale, and, therefore, selected to be the target form in 2017–2018 LEAP 2025 form construction. The rationale for the choice of the targets was that each 2017–2018 LEAP 2025 form should be on the PARCC scale and be closely comparable to PARCC assessments. Figures 3.1 and Figure 3.2 for English I and II and Figure 3.3 and Figure 3.4 for Algebra I and Geometry show the test characteristic curves (TCCs) and standard errors of measurement (SEMs) of the final forms compared to those of the target forms. The left line graph displays the TCC of the target and the selected 2017-2018 forms, summarizing the expected proportion of the maximum raw score needed to achieve the raw score. The right line graph displays the SEM of the scale score of the target form and the selected 2017-2018 forms. This summarizes the amount of measurement error surrounding a scale score.

Figure 3.1 Spring English I Form Building Evaluation for 2017-2018 Administrations

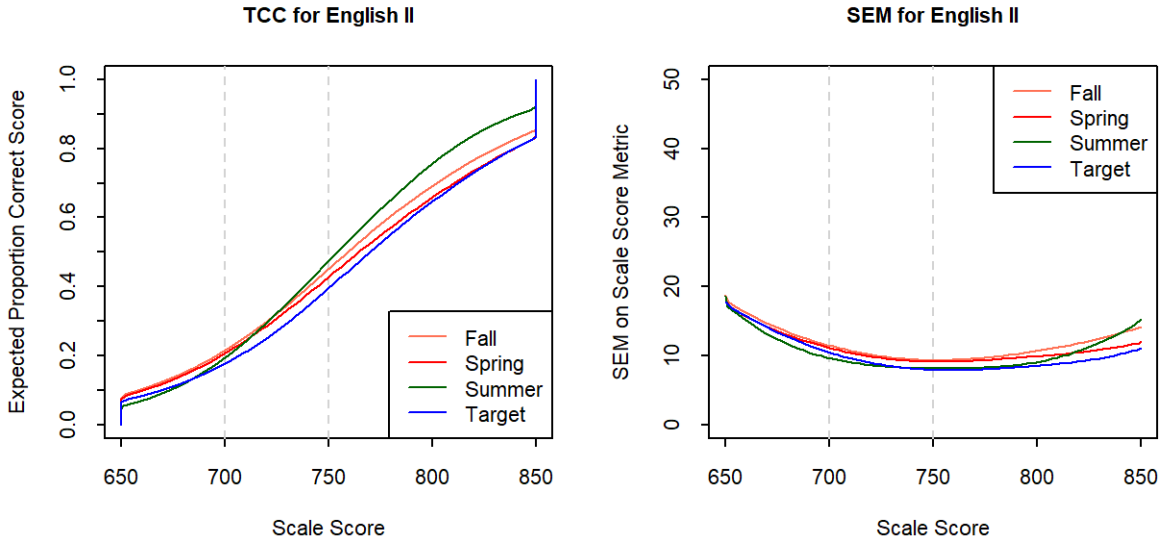


Notes:

- Target is the 2017 PARCC Spring test form.
- Fall, Spring, Summer are Fall 2017, Spring 2018, Summer 2018 LEAP 2025 HS test forms.

Copyright © 2019 by Louisiana Department of Education.

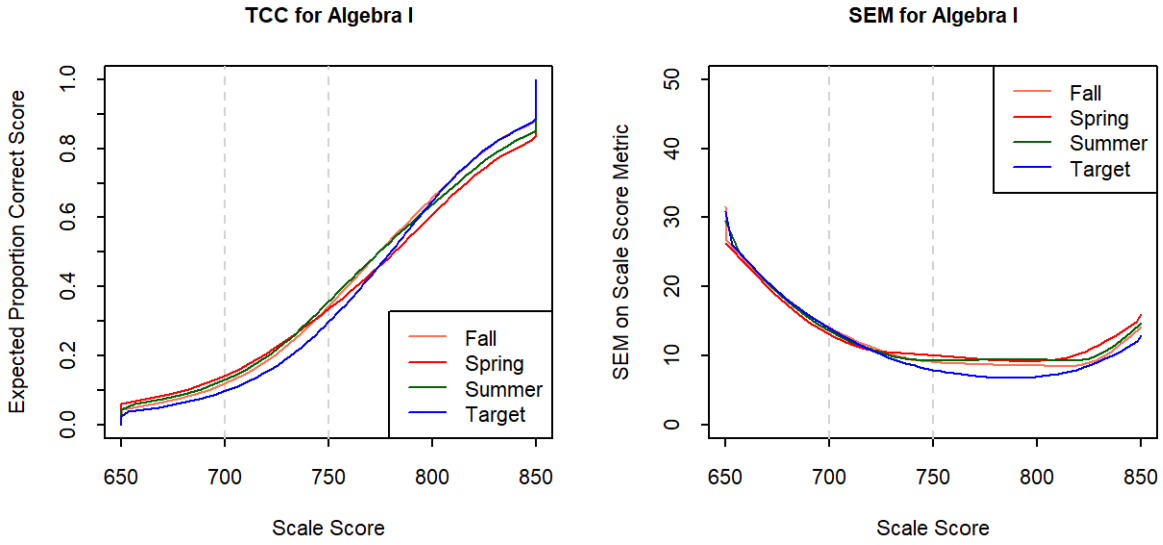
Figure 3.2 Spring English II Form Building Evaluation for 2017-2018 Administrations



Notes:

- Target is the 2017 PARCC Spring test form.
- Fall, Spring, Summer are Fall 2017, Spring 2018, Summer 2018 LEAP 2025 HS test forms.

Figure 3.3 Algebra I Form Building Evaluation for 2017-2018 Administrations

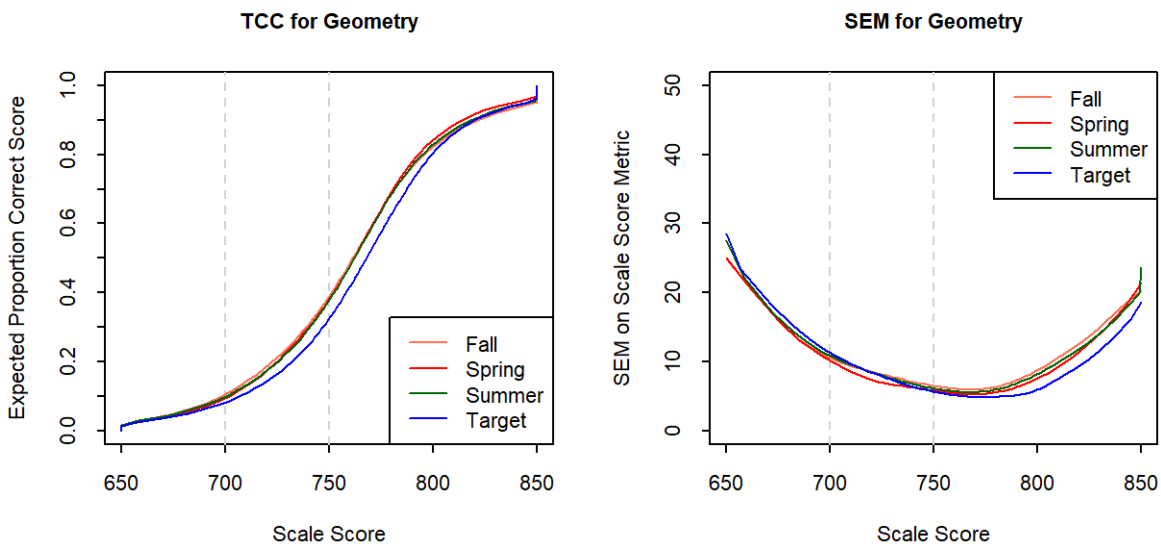


Notes:

- Target is the 2017 PARCC Spring test form.
- Fall, Spring, Summer are Fall 2017, Spring 2018, Summer 2018 LEAP 2025 HS test forms.

Copyright © 2019 by Louisiana Department of Education.

Figure 3.4 Geometry Form Building Evaluation for 2017-2018 Administrations



Notes:

- Target is the 2017 PARCC Spring test form.
- Fall, Spring, Summer are Fall 2017, Spring 2018, Summer 2018 LEAP 2025 HS test forms.

Selecting Anchors

Anchor sets used in the common item nonequivalent group design underwent considerable scrutiny due to the generally accepted guideline that the anchor set should mirror the total (or reference) test in terms of content and item characteristics. One of the critical psychometric considerations for an anchor set is the extent to which the TCC and SEM of the anchor set aligns to that of the total test.

3.3 Universal Design

Course-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 3.15 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 addresses how the elements of universal design were addressed in the construction of the 2017–2018 test forms in compliance with AERA, APA, & NCME (2014) Standard 3.1, which states the following:

Copyright © 2019 by Louisiana Department of Education.

> 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 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. Test directions at the beginning of each test session are clearly and simply stated, and the wording of such instructions is standardized as much as possible across tests to ensure clarity and consistency while being comparable to PARCC.

Table 3.15 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.

Copyright © 2019 by Louisiana Department of Education.

3.4 Accommodations and Designated Supports

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

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

Students with disabilities, students with 504 plans, and English learners (ELs) may be provided test administration accommodations based on their accommodation plan. More information on accommodations can be found in Chapter 4. Accommodation coding instructions can be found in the *Test Coordinators Manual*.

Accommodated Print forms were developed for the high school ELA and mathematics tests 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 B, *Accommodated Print Form Creation*.

Braille forms were constructed for each course to enable students with visual impairments to participate in the LEAP 2025 assessments. Braille forms were based on the Accommodated Print forms. There are no large-print versions of the Accommodated Print forms. Instead, students needing a large-print version use appropriate-size 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 high school assessments and the test administrator's notes that accompany the braille forms. The goal was to maximize the number of items on the braille form so it would be possible to transcribe all items 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).

The following accessibility features were available for all students with the particular need documented in their Individualized Education Program (IEP), Individual Accommodation Plan (IAP), English Learner (EL) plan, or Personal Needs Profile (PNP): 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).

Copyright © 2019 by Louisiana Department of Education.

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, mathematics Spanish read aloud, translated mathematics test, test read aloud (text-to-speech). For details on these assessment and accessibility features and accommodations, 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 A, *Forms Development Process for Spanish Translations Forms*.

3.5 Item and Task Specifications

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

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

The item and task specifications are designed to ensure that the assessment items measure the assessment's claims. The purpose of the item and task specifications is to define the characteristics of the items and tasks that will provide the evidence to support one or more claims. To do this, the item and task specifications delineate the types of evidence, or targets, that should be elicited for each reporting category within a grade level. Then, the specifications provide explicit guidance on how to write items to elicit the desired evidence. To address 2017–2018 LEAP 2025 high school assessments comparability goals with PARCC, PARCC claims, subclaims, and evidence statements, along with guidance provided by the *Louisiana Student Standards for ELA and Mathematics*, were used as item and task specifications.

The item and task specifications provide guidance on how to measure the targets (i.e., standards) first found in the content specifications. The item and task specifications provide 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 of the item types aligned to particular standards.

These item specifications were developed for each course 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 in working on a test question). The stimulus specifications also provide information on characteristics of stimuli or activities to avoid 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, 2017–2018 LEAP 2025 items were selected according to the elements of universal design.

3.6 Summary

In summary, the overall purpose of this chapter is to explicate the procedures used in the development of the 2017–2018 LEAP 2025 high school assessments. The efforts by LDOE and DRC in developing the

Copyright © 2019 by Louisiana Department of Education.

LEAP 2025 high school 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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 4: Test Administration

Chapter 4 of the technical report describes the processes and activities 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 2017-2018 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 adherence 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 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 the 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 the 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 school system adherence to the standardized administration of the tests.

The school system 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 administration of the LEAP 2025 High School assessments, DRC produced the following test administration manuals (TAMs): *High School Test Administration Manual: LEAP 2025 and End Of Course, Fall 2017; High School Test Administration Manual: LEAP 2025 and End Of Course, Spring 2018; High School Test Administration Manual: LEAP 2025 and End Of Course, Summer 2018.* DRC also produced the following district Test Coordinators Manuals (TCMs): *Test Coordinators Manual LEAP 2025 High School and End-of-Course (EOC) Fall 2017; Test Coordinators Manual LEAP 2025 High School and End-of-Course*

Copyright © 2019 by Louisiana Department of Education.

(EOC) Spring 2017; Test Coordinators Manual LEAP 2025 High School and End-of-Course (EOC) Summer 2017

. LDOE assessment administration and development staff review these manuals, provide feedback, and give final approval. The TCM includes both LEAP 2025 HS and EOC ELA, mathematics, U.S. history, and biology. It provides detailed instructions for school system and school test coordinators' responsibilities for distributing and collecting test materials and for returning them to DRC for scoring.

Test Coordinators Manual Table of Contents

1. Key Dates
2. Resources Available in eDIRECT
3. EOC and LEAP 2025 High School Alerts
4. Pre-Administration Oath of Security and Confidentiality Statement
5. Post-Administration Oath of Security and Confidentiality Statement
6. General Information
   6.1. eDIRECT and INSIGHT
7. EOC/LEAP2025 High School
   7.1. Testing Requirements
8. Test Security
   8.1.Key Definitions
   8.2.Violations of Test Security
   8.3.Testing Guidelines
   8.4.LEAP 2025 High School Testing Times
   8.5.End-of-Course suggested Testing Times
   8.6.Testing Conditions
   8.7.Alternate Testing Environments
   8.8.Testing Schedule
   8.9.Extend Time for Testing
   8.10.Extended Breaks
   8.11.Makeup Testing
9. Roles and Responsibilities
   9.1. District Test Coordinator
   9.2. School Test Coordinator
   9.3. Chief Technology Officer
10. Managing Test Tickets
    10.1. Student Transfers
    10.2. Locked Test Tickets
    10.3. Technical Issues
    10.4. Invalidating Test Tickets
11. Resources for Online Testing
    11.1. High School Test Administration Manual: LEAP 2025 and EOC
    11.2. eDIRECT User Guide

Copyright © 2019 by Louisiana Department of Education.

11.3. LEAP 2025 Accommodations and Accessibility User Guide
11.4. INSIGHT Technology User Guide
11.5. Student Tutorials
11.6. Online Tools Training (OTT)
12. Post-Administration Rescoring Process for EOC/LEAP 2025 Tests
13. Request for Rescoring
14. Void Notification

The TAMs provide detailed instructions for administering the LEAP 2025 assessments. The manuals include instructions for test security, test administrator responsibilities, test preparation, administration of tests, and post-test procedures. Information included in the TAMs is listed below.

Test Administrators Manual Table of Contents

1. Notes and Reminders
2. Pre-Administration Oath and Security Confidentiality Statement
3. Post-Administration Oath and Security Confidentiality Statement
4. Overview
5. Test Security
   5.1. Secure Test Materials
   5.2. Testing Irregularities and Security Breaches
   5.3. Testing Environment
   5.4. Violations of Test Security
   5.5. Voiding Student Tests
6. Test Administrator Responsibilities
   6.1. Software Tools and Features for Test Administrators
7. Test Administration Checklists
   7.1. Before Testing
   7.2. During Testing
   7.3. After Testing (Daily)
   7.4. After Testing (Last Day)
8. Test Materials
   8.1. Receipt of Test Materials
9. Testing Guidelines
   9.1. Testing Eligibility
   9.2. Testing Schedule
   9.3. LEAP 2025 Testing Time
   9.4. EOC Testing Time
   9.5. Extended Time for Testing
   9.6. Makeup Test Procedures
   9.7. Testing Conditions
   9.8. Accessibility Features

Copyright © 2019 by Louisiana Department of Education.

10. Special Populations and Accommodations
    10.1. IDEA Special Education Students
    10.2. Students with One or More Disabilities According to Section 504
    10.3. Gifted and Talented Special Education Students
    10.4. Test Accommodations for Special Education and Section 504 Students
    10.5. Special Considerations for Students who are Deaf or Hearing Impaired
    10.6. English Learners (ELs)
11. Directions for Administering the LEAP 2025 Tests
12. LEAP 2025 Testing Times
13. General Information for LEAP 2025
    13.1. LEAP 2025 English I and English II
    13.2. LEAP 2025 Algebra I
    13.3. LEAP 2025 Geometry
    13.4. LEAP 2025 U.S. History
14. Directions for Administering End-of-Course Tests
15. End-of-Course Suggested Testing Times
16. General Instructions for EOC
    16.1. End-of-Course English III
    16.2. End-of-Course Biology
    16.3. End-of-Course U.S. History
17. Post-Test Procedures
    17.1. Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    17.2. Returning Test Materials to the School Test Coordinator
18. Index

The *Standards* contain multiple references that are relevant to test administration. Information in the *LEAP 2025 Test Administration Manual* addresses these standards.

Directions for test administration found in the manual addresses 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 test administration manuals describe the following: general rules of online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

Copyright © 2019 by Louisiana Department of Education.

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 staff administer 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 Individualized Education Program (IEP); allowing accommodations for students who do not have an IEP; or defining terms on the test.

Each administration includes an administrative error retest, which provides an opportunity for students to retake a HS test that was voided during the regular test window (e.g., the student was not given enough time to complete the test, the student was not provided proper accommodations during the testing time, the teacher or administrator provided information or answers that resulted in the test being voided).

> *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.

Copyright © 2019 by Louisiana Department of Education.

- 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 braille test form, examiners are instructed to transcribe students' responses from the braille test form into the online testing system (INSIGHT) exactly as the responses appear in the original form.

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

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

4.1 Return Material Forms and Guidelines

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

4.2 Security Checklists

As soon as printed test materials are received by a school system, the district test coordinator ensures the first and last security barcodes on the tests match the packing list he or she received. The district test coordinator then packages the test materials to be sent to schools. District test coordinators are required to return Communication Assistance Scripts (CAS) and Braille test materials to DRC. School systems are required to document nonstandard situations, including lost, damaged, destroyed, extra, or missing materials. Any material not accounted for is placed on a missing materials list which is used by DRC and LDOE to follow up with all districts to ensure security of all materials.

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 *LEAP 2025 HS Interpretive Guide* is written for Louisiana

Copyright © 2019 by Louisiana Department of Education.

teachers and administrators who receive the LEAP 2025 score reports. More details about the guide can be found in Chapter 7.

4.4 Test Security Measures

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

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

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 the spurious. It is important to note that although the results may be used to identify potential problems within a school, identification is not an accusation of misconduct.

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

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

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.

Score Change 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. 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 administers' change

Copyright © 2019 by Louisiana Department of Education.

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

Web Monitoring

It is not expected that LEAP 2025 operational test content would 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 a threat is verified, steps are taken so that infringing content is removed quickly.

Plagiarism Detection

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

4.5 Test Administration

The 2017-2018 assessments were administered to students within the state testing windows of November 29 to December 13, 2017; April 23 to May 18, 2018; June 18-22, 2018. Each session of the LEAP 2025 assessments was required to be administered in one block of time.

Time

All sessions of the LEAP 2025 high school ELA and mathematics 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.

Copyright © 2019 by Louisiana Department of Education.

Table 4.1 LEAP 2025 Administration Schedule Timing by Session

Grade	Session	Minutes
English I	1	90
	2	90
	3	80
English II	1	90
	2	90
	3	80
Algebra I	1a	25
	1b	55
	2	80
	3	80
Geometry	1a	25
	1b	55
	2	80
	3	80

Copyright © 2019 by Louisiana Department of Education.

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 Manuals contain 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.

Tables 4.2 through 4.5 summarize the numbers of reportable students receiving accommodations or designated features by type for the 2017-2018 LEAP 2025 HS administrations. Accommodation assignment guidance is provided in the *LEAP 2025 Accommodations and Accessibility User Guide*. The analyses are based on census data and include only those students who received accommodations or designated features and received a scale score on the ELA or Mathematics LEAP 2025 high school assessments. The percentage represents the percentage of the census population receiving that accommodation or designated feature.

Table 4.2 Fall 2017 Number and Percentage of Students Receiving Accommodations by Accommodation/Designated Feature Type, as identified in eDIRECT

Accommodation/Designated Feature Type: Fall 2017			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
English I	Text-to-Speech	≥310	6.85
English I	Accommodated Print	<10	NR
English I	Human Read Aloud	<10	NR
English I	Native Language Word-to-Word Dictionary	≥30	0.65
English I	Directions in Native Language	≥10	0.39
English I	Communication Assistance	<10	NR
English I	Transferred Answers	≥10	0.30
English I	Answers Recorded	<10	NR
English I	Extended Time	≥630	13.74
English I	Individual/Small Group Administration	≥410	8.91

Copyright © 2019 by Louisiana Department of Education.

Accommodation/Designated Feature Type: Fall 2017			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
English II	Text-to-Speech	≥1,420	15.63
English II	Accommodated Print	<10	NR
English II	Human Read Aloud	≥100	1.16
English II	Native Language Word-to-Word Dictionary	≥440	4.82
English II	Directions in Native Language	≥160	1.81
English II	Communication Assistance	<10	NR
English II	Transferred Answers	<10	NR
English II	Answers Recorded	≥10	0.12
English II	Extended Time	≥2,370	26.00
English II	Individual/Small Group Administration	≥1,460	16.04
Algebra I	Text-to-Speech	≥1,710	20.49
Algebra I	Accommodated Print	<10	NR
Algebra I	Human Read Aloud	≥120	1.50
Algebra I	Native Language Word-to-Word Dictionary	≥300	3.64
Algebra I	Directions in Native Language	≥110	1.35
Algebra I	Communication Assistance	<10	NR
Algebra I	Transferred Answers	<10	NR
Algebra I	Answers Recorded	<10	NR
Algebra I	Calculator	≥1,760	21.11
Algebra I	Extended Time	≥2,480	29.76
Algebra I	Individual/Small Group Administration	≥1,640	19.72
Geometry	Text-to-Speech	≥750	11.17
Geometry	Accommodated Print	<10	NR
Geometry	Human Read Aloud	≥20	0.43
Geometry	Native Language Word-to-Word Dictionary	≥250	3.73
Geometry	Directions in Native Language	≥120	1.86
Geometry	Communication Assistance	<10	NR
Geometry	Transferred Answers	<10	NR
Geometry	Answers Recorded	<10	NR
Geometry	Calculator	≥640	9.62
Geometry	Extended Time	≥1,160	17.34
Geometry	Individual/Small Group Administration	≥630	9.45

Copyright © 2019 by Louisiana Department of Education.

Table 4.3 Spring 2018 Number and Percentage of Students Receiving Accommodations by Accommodation/Designated Feature Type, as identified in eDIRECT

Accommodation/Designated Feature Type: Spring 2018			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
English I	Text-to-Speech	≥3,720	8.08
English I	Accommodated Print	<10	NR
English I	Human Read Aloud	≥220	0.48
English I	Native Language Word-to-Word Dictionary	≥820	1.79
English I	Directions in Native Language	≥330	0.73
English I	Communication Assistance	≥20	0.05
English I	Transferred Answers	≥70	0.16
English I	Answers Recorded	≥60	0.15
English I	Extended Time	≥8,000	17.36
English I	Individual/Small Group Administration	≥5,050	10.96
English II	Text-to-Speech	≥3,730	8.42
English II	Accommodated Print	<10	NR
English II	Human Read Aloud	≥190	0.43
English II	Native Language Word-to-Word Dictionary	≥890	2.02
English II	Directions in Native Language	≥340	0.78
English II	Communication Assistance	≥30	0.07
English II	Transferred Answers	≥40	0.11
English II	Answers Recorded	≥40	0.09
English II	Extended Time	≥7,580	17.09
English II	Individual/Small Group Administration	≥4,880	11.01
Algebra I	Text-to-Speech	≥4,980	10.06
Algebra I	Accommodated Print	<10	NR
Algebra I	Human Read Aloud	≥250	0.52
Algebra I	Native Language Word-to-Word Dictionary	≥1,010	2.05
Algebra I	Directions in Native Language	≥370	0.76
Algebra I	Communication Assistance	≥30	0.07
Algebra I	Transferred Answers	≥70	0.14
Algebra I	Answers Recorded	≥60	0.14
Algebra I	Calculator	≥5,220	10.53
Algebra I	Extended Time	≥8,970	18.1
Algebra I	Individual/Small Group Administration	≥5,700	11.5
Geometry	Text-to-Speech	≥2,230	5.94
Geometry	Accommodated Print	<10	NR
Geometry	Human Read Aloud	≥100	0.28
Geometry	Native Language Word-to-Word Dictionary	≥520	1.39
Geometry	Directions in Native Language	≥220	0.61
Geometry	Communication Assistance	≥10	0.05
Geometry	Transferred Answers	≥30	0.09
Geometry	Answers Recorded	≥20	0.06

Copyright © 2019 by Louisiana Department of Education.

Accommodation/Designated Feature Type: Spring 2018			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
Geometry	Calculator	≥2,260	6.02
Geometry	Extended Time	≥4,530	12.04
Geometry	Individual/Small Group Administration	≥2,760	7.33

Table 4.4 Summer 2018 Number and Percentage of Students Receiving Accommodations by Accommodation/Designated Feature Type, as identified in eDIRECT

Accommodation/Designated Feature Type: Summer 2018			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
English I	Text-to-Speech	≥400	29.24
English I	Accommodated Print	<10	NR
English I	Human Read Aloud	≥20	1.93
English I	Native Language Word-to-Word Dictionary	≥50	3.72
English I	Directions in Native Language	≥20	1.43
English I	Communication Assistance	<10	NR
English I	Transferred Answers	<10	NR
English I	Answers Recorded	<10	NR
English I	Extended Time	≥600	43.10
English I	Individual/Small Group Administration	≥420	30.24
English II	Text-to-Speech	≥600	26.38
English II	Accommodated Print	<10	NR
English II	Human Read Aloud	≥30	1.65
English II	Native Language Word-to-Word Dictionary	≥90	4.27
English II	Directions in Native Language	≥30	1.48
English II	Communication Assistance	<10	NR
English II	Transferred Answers	<10	NR
English II	Answers Recorded	<10	NR
English II	Extended Time	≥930	40.53
English II	Individual/Small Group Administration	≥640	28.08
Algebra I	Text-to-Speech	≥430	27.39
Algebra I	Accommodated Print	<10	NR
Algebra I	Human Read Aloud	≥20	1.65
Algebra I	Native Language Word-to-Word Dictionary	≥30	2.16
Algebra I	Directions in Native Language	≥10	0.70
Algebra I	Communication Assistance	<10	NR
Algebra I	Transferred Answers	<10	NR
Algebra I	Answers Recorded	<10	NR
Algebra I	Calculator	≥470	29.93
Algebra I	Extended Time	≥630	40.46
Algebra I	Individual/Small Group Administration	≥440	28.03

Copyright © 2019 by Louisiana Department of Education.

Accommodation/Designated Feature Type: Summer 2018			
Content	**Accommodation/Designated Feature**	**Number**	**Percentage**
Geometry	Text-to-Speech	≥80	21.46
Geometry	Accommodated Print	<10	NR
Geometry	Human Read Aloud	<10	NR
Geometry	Native Language Word-to-Word Dictionary	≥10	2.53
Geometry	Directions in Native Language	<10	NR
Geometry	Communication Assistance	<10	NR
Geometry	Transferred Answers	<10	NR
Geometry	Answers Recorded	<10	NR
Geometry	Calculator	≥70	17.68
Geometry	Extended Time	≥110	28.03
Geometry	Individual/Small Group Administration	≥90	22.73

4.6 Summary

In summary, the overall purpose of each of the test administration trainings and the ancillary materials is to keep school systems informed about policies and procedures related to testing in general and the LEAP 2025 program in particular. The information imparted is clearly related to standardizing the administration of the LEAP 2025, maintaining the security of the assessment, allowing access to the assessments for special populations by clearly delineating appropriate accommodations, and maintaining integrity of the scores. These communication and training efforts by 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)

Copyright © 2019 by Louisiana Department of Education.

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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 5: Constructed-Response and Technology-Enhanced Scoring

In this chapter, the scoring process used for the 2017–2018 LEAP 2025 high school ELA and mathematics assessments is described, with a particular focus on the handscoring process 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 2017–2018 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 Artificial Intelligence (AI) scoring engine and/or human raters. Seven ELA items across grades English I and English II (noted in Table 5.1) were scored by an Artificial Intelligence (AI) scoring engine, Pearson's Intelligent Essay Assessor (IEA), using scoring models previously developed by Pearson. The human raters were trained by DRC. Human scorers provided second reads to 10% of these responses as well as handscoring supervisory reviews of all responses that IEA was not able to score. Handscoring and AI processing rules are detailed in Appendix D.

Table 5.1 Constructed-Response Operational Scoring

Administration	Course	Handscoring Only	AI Scoring	AI Vendor
Fall 2017	English I	902161, 906152	N/A	
	English II	902331, 902354	N/A	
Spring 2018	English I	N/A	902161, 902152	Pearson
	English II	N/A	902331, 906197	Pearson
Summer 2018	English I	902194	902203	Pearson
	English II	N/A	906181, 906190	Pearson

Copyright © 2019 by Louisiana Department of Education.

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 and describe 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 2017–2018 LEAP 2025 high school ELA and mathematics test responses had at least a four-year college degree in an appropriate field, such as a bachelor's degree in a STEM field or in English language arts.

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 also demonstrate his or her ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

ELA and Mathematics scorers hired by DRC meet the degree requirements as specified by PARCC and conveyed by Pearson described in the following two paragraphs:

Mathematics scorers had degrees in mathematics, science, engineering, education, or a related field; and/or teacher certification in a mathematics or science related field; or appropriate work experience that demonstrated proficiency in the subject area. Work experience outside of scoring was considered, as was college coursework. Three appropriate college-level courses (e.g., calculus, statistics, finance) were considered sufficient. Scorers of mathematics performance task responses had the mathematics knowledge needed to effectively score student responses.

ELA scorers had degrees in reading, education, history, psychology, journalism, or a related area and/or teacher certification or other work experience that enabled them to succeed in scoring the Literary Analysis Task (LAT), Research Simulation Task (RST), or Narrative Writing Task (NWT). Work experience outside

Copyright © 2019 by Louisiana Department of Education.

scoring was considered, as was college coursework. Following Pearson's practices for PARCC, three appropriate college-level courses with an emphasis on writing or literature were considered sufficient college-level experience for scoring ELA responses. Alternatively, successful completion of 250 hours of scoring was considered sufficient work experience for a scorer with insufficient college-level coursework to be hired to score student responses.

Security

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

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 PARCC-approved training materials provided by PARCC. These materials were developed according to processes described in PARCC Technical Reports and include the following:

- Passages, prompts, and associated stimuli
- Rubrics
- Anchor sets
- Practice sets
- Qualifying sets

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. 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.

Prototype Items

Some items included in the Louisiana forms were prototype items, meaning they had full sets of associated training materials, including anchor sets, practice sets, and qualifying sets. DRC started the training process

Copyright © 2019 by Louisiana Department of Education.

with a review of passages and items, rubrics, and anchor sets, followed by the scoring and discussion of practice sets and qualifying sets. Once this process was completed for a prototype item included on the Louisiana form, 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.[1] Readers who did not qualify on the prototype item training were not allowed to continue the training.

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

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

Table 5.2 Mathematics Training Set Composition

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

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).	Yes

[1] *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.*

Copyright © 2019 by Louisiana Department of Education.

		• Anchor Sets for abbreviated NWT item training included scores for Written Expression (WE).	
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 both traits appropriate to the task type	Yes
Practice Set 2	5 responses representing the range of responses for the Knowledge of Language and Conventions trait	10 responses representing the range of responses for both traits appropriate to the task type	Yes
Practice Set 3	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Practice Set 4	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Qualifying Set 1	10 responses comparable to the anchor set responses (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 of Language and Conventions trait, there were two mixed-prompt anchor sets per grade level (one for the narrative task and the other for 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 of Language and 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 of Language and 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.*

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.

Copyright © 2019 by Louisiana Department of Education.

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 (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 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 5.5. 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).

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)

Copyright © 2019 by Louisiana Department of Education.

The following section 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 all constructed-response items.

Reader Monitoring Procedures

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

Validity Sets and Inter-Rater Reliability

In addition to the feedback that supervisors provided to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducted validity scoring using PARCC-approved validity responses supplied by PARCC. Validity responses were inserted among the live student responses. DRC used the same validity responses that Pearson used, which included both paper-based test (PBT) and computer-based test (CBT) 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 as well as 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.

- **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

Copyright © 2019 by Louisiana Department of Education.

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

Table 5.7 Agreement Rate Requirements for Validity and Inter-Rater Reliability

Subject	Score Point Range	Perfect Agreement	Perfect Agreement + Adjacent
Mathematics	0–1	90%	100%
Mathematics	0–2	80%	95%
Mathematics	0–3	70%	95%
Mathematics	0–4	65%	95%
ELA	Multi-trait 0–3 or 0–4 (varies by grade and trait)	65% (each trait)	96% (each trait)

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

Calibration Sets

PARCC provided DRC with PARCC-approved calibration sets. 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 a reader scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and 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 provides LDOE with handscoring quality control reports for review throughout the scoring window. Further detail about these reports can be found in Appendix D.

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 is the case regardless of the first reader being human or AI. The statistics for the 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 score was examined.

A total of 51 items were scored by human readers across all LEAP 2025 high school ELA and mathematics assessments. The inter-rater reliability rates and the total numbers of reads are shown in Tables 5.8 and 5.9 for ELA items, Tables 5.10–5.12 for mathematics items, and Tables 5.11–5.13 for Spanish mathematics items.

Copyright © 2019 by Louisiana Department of Education.

As shown in Tables 5.8 and 5.9, raters demonstrated at least 99% perfect and adjacent agreement for all ELA handscored items. As shown in Table 5.10–5.12, raters demonstrated at least 99% perfect and adjacent agreement for mathematics items. As shown in Tables 5.13–5.15, raters demonstrated 100% perfect and adjacent agreement for Spanish mathematics items.

Table 5.8 Inter-Rater Agreement, English Language Arts Items, Fall 2017

Admin.	Course	Task Type	Question	Trait	Total Reads	Read 2x	Inter-Rater Reliability %		
							EX	AD	EX + AD
Fall 2017	English I	Research Simulation	902161	Reading Comprehension and Written Expression	≥5,100	≥980	78	21	99* (na = 0)
				Knowledge of Language and Conventions	≥5,100	≥980	78	21	99* (na = 0)
	English I	Narrative Writing	906152	Written Expression	≥5,040	≥1,070	81	18	99
				Knowledge of Language and Conventions	≥5,040	≥1,070	78	21	99
	English II	Research Simulation	902331	Reading Comprehension and Written Expression	≥10,070	≥2,020	79	21	101
				Knowledge of Language and Conventions	≥10,070	≥2,020	77	22	99
	English II	Narrative Writing	902354	Written Expression	≥9,970	≥2,090	81	19	100
				Knowledge of Language and Conventions	≥9,970	≥2,090	83	17	100

Copyright © 2019 by Louisiana Department of Education.

Table 5.9 Inter-Rater Agreement, English Language Arts Items, Spring 2018

Admin.	Course	Task Type	Question	Trait	Total Reads	Read 2x	Inter-Rater Reliability %		
							EX	AD	EX + AD
Spring 2018	English I	Research Simulation (AI)	902161	Reading Comprehension and Written Expression	≥50,790	≥10,120	81	19	100
				Knowledge of Language and Conventions	≥50,790	≥10,120	79	21	100
	English I	Literary Analysis (AI)	902152	Reading Comprehension and Written Expression	≥51,340	≥10,810	81	19	100
				Knowledge of Language and Conventions	≥51,340	≥10,810	82	18	100
	English II	Research Simulation (AI)	902331	Reading Comprehension and Written Expression	≥48,920	≥10,430	79	20	99* (na = 0)
				Knowledge of Language and Conventions	≥48,920	≥10,430	78	21	99* (na = 0)
	English II	Literary Analysis (AI)	906197	Reading Comprehension and Written Expression	≥48,960	≥10,460	77	22	99* (na = 0)
				Knowledge of Language and Conventions	≥48,960	≥10,460	78	22	100

**Total Ex + AD does not add up to 100% due to rounding*

Copyright © 2019 by Louisiana Department of Education.

Table 5.10 Inter-Rater Agreement, English Language Arts Items, Summer 2018

Admin.	Course	Task Type	Question	Trait	Total Reads	Read 2x	Inter-Rater Reliability %		
							EX	AD	EX + AD
Summer 2018	English I	Research Simulation	902194	Reading Comprehension and Written Expression	≥1,540	≥330	86	14	100
				Knowledge of Language and Conventions	≥1,540	≥330	82	18	100
	English I	Narrative Writing (AI)	902203	Written Expression	≥1,510	≥400	87	13	100
				Knowledge of Language and Conventions	≥1,510	≥400	93	7	100
	English II	Literary Analysis (AI)	906181	Reading Comprehension and Written Expression	≥2,630	≥860	90	10	100
				Knowledge of Language and Conventions	≥2,630	≥860	87	13	100
	English II	Research Simulation (AI)	906190	Reading Comprehension and Written Expression	≥2,440	≥630	93	6	99* (na = 0)
				Knowledge of Language and Conventions	≥2,440	≥630	87	13	100

**Total Ex + AD + Non-adjacent (na) does not add up to 100% due to rounding*

Copyright © 2019 by Louisiana Department of Education.

Table 5.11 Inter-Rater Agreement, Mathematics Items, Fall 2017

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
						EX	AD	EX + AD
Fall 2017	Algebra I	13	N/A	≥9,060	≥2,080	96	4	100
		15	Part A	≥9,090	≥1,940	99	1	100
			Part B	≥9,090	≥1,940	99	1	100
		28	Part A	≥9,460	≥2,500	100	0	100
			Part B	≥9,460	≥2,500	100	0	100
			Part C	≥9,460	≥2,500	98	2	100
		29	Part A	≥8,780	≥2,180	97	3	100
			Part B	≥8,780	≥2,180	99	1	100
		43	Part A	≥9,110	≥2,070	96	4	100
			Part B	≥9,110	≥2,070	99	1	100
		44	N/A	≥8,480	≥2,790	100	0	100
		45	Part A	≥8,680	≥2,250	94	6	100
			Part B	≥8,680	≥2,250	97	3	100
	Geometry	13	N/A	≥7,350	≥1,690	93	6	99* (na = 0)
		15	N/A	≥7,090	≥1,880	97	3	100
		27	N/A	≥7,180	≥1,820	95	5	100
		28	Part A	≥7,350	≥1,510	99	1	100
			Part B	≥7,350	≥1,510	96	3	99* (na = 0)
			Part C	≥7,350	≥1,510	97	2	99* (na = 0)
		43	Part A	≥7,200	≥1,630	98	2	100
			Part B	≥7,200	≥1,630	96	4	100
		44	Part A	≥7,110	≥1,760	95	5	100
			Part B	≥7,110	≥1,760	98	2	100

**Total Ex + AD does not add up to 100% due to rounding*

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

Copyright © 2019 by Louisiana Department of Education.

Table 5.12 Inter-Rater Agreement, Mathematics Items, Spring 2018

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
						EX	AD	EX + AD
Spring 2018	Algebra I	13	Part A	≥55,140	≥10,230	100	0	100
			Part B	≥55,140	≥10,230	91	9	100
		15	Part A	≥53,990	≥10,640	99	1	100
			Part B	≥53,990	≥10,640	97	3	100
		28	Part A	≥53,100	≥11,410	98	2	100
			Part B	≥53,100	≥11,410	96	4	100
			Part C	≥53,100	≥11,410	98	2	100
		29	Part A	≥51,420	≥12,160	98	2	100
			Part B	≥51,420	≥12,160	93	7	100
		43	N/A	≥52,470	≥11,900	92	8	100
		44	Part A	≥53,260	≥11,800	98	2	100
			Part B	≥53,260	≥11,800	93	7	100
		45	Part A	≥51,500	≥11,920	88	11	99* (na = 0)
			Part B	≥51,500	≥11,920	94	6	100
	Geometry	13	N/A	≥40,690	≥9,060	90	9	99* (na = 0)
		15	N/A	≥39,590	≥9,650	95	5	100
		27	N/A	≥40,090	≥9,510	94	5	99* (na = 0)
		28	Part A	≥40,480	≥8,510	96	3	99* (na = 0)
			Part B	≥40,480	≥8,510	97	3	100
			Part C	≥40,480	≥8,510	97	2	99
		43	N/A	≥40,560	≥9,550	96	4	100
		44	Part A	≥39,800	≥9,290	94	5	99
			Part B	≥39,800	≥9,290	97	3	100

**Total Ex + AD does not add up to 100% due to rounding*

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

Copyright © 2019 by Louisiana Department of Education.

Table 5.13 Inter-Rater Agreement, Mathematics Items, Summer 2018

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x	Inter-Rater Reliability %		
						EX	AD	EX + AD
Summer 2018	Algebra I	13	N/A	≥1,690	≥400	100	1	101* (na = 0)
		15	Part A	≥1,720	≥370	99	1	100
			Part B	≥1,720	≥370	99	1	100
		28	Part A	≥1,720	≥330	100	0	100
			Part B	≥1,720	≥330	100	0	100
			Part C	≥1,720	≥330	98	1	99
		29	Part A	≥1,630	≥410	97	2	99* (na = 0)
			Part B	≥1,630	≥410	99	1	100
		43	Part A	≥1,710	≥390	100	0	100
			Part B	≥1,710	≥390	100	0	100
		44	N/A	≥1,580	≥420	100	0	100
		45	Part A	≥1,650	≥390	92	8	100
			Part B	≥1,650	≥390	99	1	100
	Geometry	13	N/A	≥450	≥160	98	2	100
		15	N/A	≥420	≥140	99	1	100
		27	N/A	≥430	≥150	96	4	100
		28	Part A	≥430	≥110	100	0	100
			Part B	≥430	≥110	100	0	100
			Part C	≥430	≥110	100	0	100
		43	Part A	≥430	≥130	99	1	100
			Part B	≥430	≥130	100	0	100
		44	Part A	≥410	≥120	97	3	100
			Part B	≥410	≥120	98	2	100

**Total Ex + AD + Non-adjacent (na) does not add up to 100% due to rounding*

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

Copyright © 2019 by Louisiana Department of Education.

Table 5.14 Inter-Rater Agreement, Spanish Mathematics Items, Fall 2017

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x**	Inter-Rater Reliability %		
						EX	AD	EX + AD
Fall 2017	Algebra I	13	N/A	≥80	<10	100	0	100
		15	Part A	≥80	<10	100	0	100
			Part B	≥80	<10	100	0	100
		28	Part A	≥90	<10	100	0	100
			Part B	≥90	<10	100	0	100
			Part C	≥90	<10	100	0	100
		29	Part A	≥80	<10	100	0	100
			Part B	≥80	<10	100	0	100
		43	Part A	≥80	<10	100	0	100
			Part B	≥80	<10	100	0	100
		44	N/A	≥60	<10	100	0	100
		45	Part A	≥80	<10	100	0	100
			Part B	≥80	<10	100	0	100
	Geometry	13	N/A	≥20	≥10	100	0	100
		15	N/A	≥20	<10	100	0	100
		27	N/A	≥20	<10	100	0	100
		28	Part A	≥20	<10	100	0	100
			Part B	≥20	<10	100	0	100
			Part C	≥20	<10	100	0	100
		43	Part A	≥20	≥10	100	0	100
			Part B	≥20	≥10	100	0	100
		44	Part A	≥20	<10	100	0	100
			Part B	≥20	<10	100	0	100

**Total Ex + AD does not add up to 100% due to rounding*

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

Copyright © 2019 by Louisiana Department of Education.

Table 5.15 Inter-Rater Agreement, Spanish Mathematics Items, Spring 2018

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x**	Inter-Rater Reliability %		
						EX	AD	EX + AD
Spring 2018	Algebra I	13	Part A	≥150	<10	100	0	100
			Part B	≥150	<10	100	0	100
		15	Part A	≥160	≥40	100	0	100
			Part B	≥160	≥40	100	0	100
		28	Part A	≥130	<10	100	0	100
			Part B	≥130	<10	100	0	100
			Part C	≥130	<10	100	0	100
		29	Part A	≥160	≥50	96	4	100
			Part B	≥160	≥50	100	0	100
		43	N/A	≥130	<10	100	0	100
		44	Part A	≥150	≥50	100	0	100
			Part B	≥150	≥50	100	0	100
		45	Part A	≥130	<10	100	0	100
			Part B	≥130	<10	100	0	100
	Geometry	13	N/A	≥90	≥20	100	0	100
		15	N/A	≥80	<10	100	0	100
		27	N/A	≥80	<10	100	0	100
		28	Part A	≥80	<10	100	0	100
			Part B	≥80	<10	100	0	100
			Part C	≥80	<10	100	0	100
		43	N/A	≥70	<10	100	0	100
		44	Part A	≥70	<10	100	0	100
			Part B	≥70	<10	100	0	100

**Total Ex + AD does not add up to 100% due to rounding*

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

Copyright © 2019 by Louisiana Department of Education.

Table 5.16 Inter-Rater Agreement, Spanish Mathematics Items, Summer 2018

Admin.	Course	Question	Part(s)**	Total Reads	Read 2x***	Inter-Rater Reliability %		
						EX	AD	EX + AD
Summer 2018	Algebra I	13	N/A	≥10	<10	100	0	100
		15	Part A	≥10	<10	100	0	100
			Part B	≥10	<10	100	0	100
		28	Part A	≥10	<10	100	0	100
			Part B	≥10	<10	100	0	100
			Part C	≥10	<10	100	0	100
		29	Part A	≥10	<10	100	0	100
			Part B	≥10	<10	100	0	100
		43	Part A	≥10	<10	100	0	100
			Part B	≥10	<10	100	0	100
		44	N/A	≥10	<10	100	0	100
		45	Part A	≥10	<10	100	0	100
			Part B	≥10	<10	100	0	100
	Geometry	13	N/A	<10	<10	100	0	100
		15	N/A	<10	<10	100	0	100
		27	N/A	<10	<10	100	0	100
		28	Part A	<10	<10	100	0	100
			Part B	<10	<10	100	0	100
			Part C	<10	<10	100	0	100
		43	Part A	<10	<10	100	0	100
			Part B	<10	<10	100	0	100
		44	Part A	<10	<10	100	0	100
			Part B	<10	<10	100	0	100

**Total Ex + AD does not add up to 100% due to rounding*

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

*** Numbers Read 2x may be less than 10% of Total Reads for some items, because the smaller quantities of responses allowed scoring to be done directly by expert scorers/supervisors or via paired scoring between a supervisor and scorer. As such, fewer were routed through the 10% read-behind process.

Copyright © 2019 by Louisiana Department of Education.

5.3 Technology-Enhanced Item Scoring Process

All technology-enhanced items, as well as EBSR, MPSR, and SA items, were processed through DRC's autoscoring engine and scored according to the assigned scoring rules as established during content development by PARCC or DRC as applicable in conjunction with 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 any 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 designated specific information 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 test forms, student responses to items were transcribed into the online system by a test administrator.

5.4 Multiple-Choice and Multiple-Select Item Scoring Process

Responses to multiple-choice and multiple-select items were captured during test administration. In the case of braille forms, student responses to these items were transcribed into the online system by a test administrator. Additional information about this can be found in Chapter 6 under Test-Level Statistics.

5.5 Summary

The information presented in this chapter summarizes the scoring procedures for different types of items and the steps taken by DRC to ensure accuracy in the autoscoring and handscoring processes. The inter-rater reliability statistics presented in Section 5.4 demonstrate that the items are scored reliably. These efforts by

Copyright © 2019 by Louisiana Department of Education.

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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 6: Operational Data Analyses

This chapter of the LEAP 2025 High School 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, the IRT models used for calibrating the data are discussed and the purpose of data calibration and scaling for each content area is addressed. The calibration samples are then presented, followed by the data calibration results, including the model-data fit for the Louisiana student data. If the IRT models fit the empirical item response distributions for the population about which generalizations are to be made (i.e., Louisiana students), then the claim that the scores are valid indicators of an underlying ability is strengthened. The lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for the LEAP 2025 tests are presented.

Chapter 6 demonstrates adherence in the LEAP 2025 high school program 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 2017-2018 LEAP 2025 high school analyses, the Louisiana student population was used. In Section 6.3, the characteristics of calibration samples, such as subgroups, are discussed. Chapter 3 presents the test specifications. Information regarding reported data is discussed in detail in Chapter 7.

6.1 Classical Item Statistics

In this section, summary test statistics for each form and subject area of the LEAP 2025 high school ELA and mathematics tests are presented. These statistics are followed by item-level statistics for each subject area of LEAP 2025 test. These statistics were produced using census data with first time test takers. Students included in item-level statistics needed to meet at least the following psychometric analysis criteria (note that the criteria used to filter data for item statistics analyses are slightly different than what is used to produce students' performance statistics in this report):

- Student has total raw score in the data
- Student did not take administration error form
- Student did not take braille form
- Student did not take Spanish form
- Student's test score was not voided
- Student took the assessment for the first time (initial testers)
- Student finished all sessions
- Student's constructed-response items are scored

Copyright © 2019 by Louisiana Department of Education.

Test-Level Statistics

Table 6.1 presents the number of items, score points, mean and standard deviation of the raw scores, and the average form difficulty for each subject for the three administrations, 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 is similar with the fall and spring administrations. Average form difficulty for the summer administration was lower for all tests than the fall and spring administrations, likely due to the population taking the summer form as the summer form is a retest form. The average form difficulty for ELA ranged from 0.42 to 0.45 with the fall and spring administrations. The difficulty of the spring administration forms y was 0.29 and 0.25 for English I and English II, respectively. Average form difficulty for mathematics ranged from 0.33 to 0.34 for the fall and spring administrations. The spring administration of mathematics average form difficulty was 0.23 and 0.38 for Algebra I and Geometry, respectively. In general, the 2018 LEAP 2025 High School tests were relatively difficult tests, and mathematics tests were more difficult than the ELA tests.

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

Administration	Course	Total Items*	Total Points	Mean Raw Score (Std. Dev.)	Average Form Difficulty (Std. Dev.)
Fall	English I	33	90	38.23 (16.36)	0.43 (0.11)
	English II	33	90	40.15 (17.16)	0.45 (0.14)
	Algebra I	38	67	22.50 (11.23)	0.34 (0.19)
	Geometry	38	68	22.69 (12.79)	0.34 (0.19)
Spring	English I	34	94	39.22 (17.41)	0.42 (0.14)
	English II	34	94	40.01 (16.42)	0.43 (0.14)
	Algebra I	39	68	21.77 (10.26)	0.33 (0.18)
	Geometry	38	68	21.86 (12.47)	0.33 (0.19)
Summer	English I	33	90	25.75 (15.99)	0.29 (0.14)
	English II	34	94	21.99 (14.60)	0.25 (0.15)
	Algebra I	38	67	15.23 (10.35)	0.23 (0.14)
	Geometry	38	68	23.89 (21.16)	0.38 (0.11)

*With English I and English II, each writing prompt component is counted as one item. WE writing component is weighted in total points.

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 subject for each administration, respectively.

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

Copyright © 2019 by Louisiana Department of Education.

Table 6.2 LEAP 2025 High School *p*-Values and Item-Total Correlation (R_{it}) Descriptive Statistics

			Item *p*-Value				Average Item-Total Correlation			
Admin	**Course**	**N of Items**	**Mean**	**Std. Dev.**	**Min.**	**Max**	**Mean**	**Std. Dev.**	**Min.**	**Max**
Fall	English I	33	0.45	0.12	0.11	0.76	0.42	0.16	0.17	0.74
	English II	33	0.47	0.16	0.23	0.79	0.44	0.15	0.23	0.74
	Algebra I	38	0.38	0.18	0.03	0.74	0.37	0.15	0.09	0.72
	Geometry	38	0.40	0.20	0.07	0.82	0.45	0.14	0.15	0.70
Spring	English I	34	0.46	0.14	0.21	0.74	0.46	0.17	0.22	0.81
	English II	34	0.46	0.15	0.21	0.81	0.46	0.18	0.15	0.80
	Algebra I	39	0.36	0.15	0.05	0.75	0.34	0.16	0.08	0.62
	Geometry	38	0.38	0.19	0.06	0.85	0.46	0.13	0.14	0.72
Summer	English I	33	0.34	0.14	0.06	0.64	0.48	0.21	0.00	0.87
	English II	34	0.30	0.14	0.08	0.56	0.50	0.17	0.16	0.77
	Algebra I	38	0.28	0.15	0.02	0.60	0.38	0.21	-0.11	0.78
	Geometry	38	0.41	0.12	0.22	0.68	0.69	0.15	0.37	0.93

Item-Level Statistics

Appendix E tables present the item statistics for each operational item included in the regular forms by content and administration. The tables include item number, *p*-value, item-total correlation (R_{it}), omit rates, total N, adjusted N (adjusted N excludes omitted responses, responses that were not scored, or responses that received a non-score code), the percentage at each score point for polytomous items, and the percentage for each option for multiple-choice (MC) items by subject and administration. Summer administration population is not state representative and the number of students were very small, thus the interpretation of statistics in summer administration should be made with caution.

p-Value

The *p*-value is a measure of item difficulty. For an MC item, the *p*-value is calculated from the number of students who correctly responded to an item divided 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 from the average score for the item divided 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 in the mid-0.50s to low 0.70s. However, it is important to select items based on content rather than on purely statistical criteria when building a criterion-referenced test. As shown in Table 6.2, the average *p*-values range from 0.36 to 0.47 with fall and spring administrations. Average *p*-values range was lower with the summer administration: from 0.28 to 0.41. The average *p*-values of the English I and English II forms were higher than the Algebra I p-values.

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 the test to measure well throughout the range of skills present in a test form. That is, it is important that the items measure the performance of both low-scoring and high-scoring students, as well as students in the center of the distribution. Having a range of *p*-values also helps to prevent

Copyright © 2019 by Louisiana Department of Education.

floor and/or ceiling effects so that the test does not have large numbers of students at the minimum or maximum possible scores. The fall and spring English forms have items with *p*-values ranging from the 0.11 to the 0.81 (see Appendix E) and 0.06 to 0.64 with the summer administration. The *p*-values on the mathematics forms range from 0.03 to 0.85 (see Appendix E) for the fall and spring administration and 0.02 to 0.67 for the summer administration. Such a broad range of *p*-values, which indicates the items measure well throughout the range of skills at a given grade, supports the accuracy of the LEAP 2025 high school test scores.

Item-Total Correlations

An item-total correlation is the correlation between an item and the total test score, where the item score is not included in the total score. It indicates how well an item differentiates between low-scoring and high-scoring students. In general, items with correlations below 0.20 are said to be poorly discriminating. The majority of the items on the LEAP 2025 High School forms 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 as student test scores may not reflect their ability. Additionally, content validity may be threatened as the not-reached items 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 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 in Appendix E show that the percentage of omit for most of the items on the LEAP 2025 High School tests for fall and spring administration are less than 5%, suggesting that the majority of students were able to complete the test in the prescribed amount of time. With Algebra I and Geometry, there were a small number of items that exceeded the omit rate 5%. This is likely do to the difficulty of the items as seen with the low item *p*-values. 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 if everyone reaches at least 75% of the items. The reported omit rates fall within these ranges.

These item level statistics are reviewed at the beginning of the operational analyses process to ensure that items are not flawed and a careful review is given to determine that the answer key is correct.

Copyright © 2019 by Louisiana Department of Education.

An MC item is reviewed during the key check process if

- it has a *p*-value less than 0.25 or above .95,
- a greater number of high-performing students (top 20%) are choosing a distractor than the keyed response
- the item-total correlation is less than 0.20,
- any of the incorrect answer options yields a positive distractor-total correlation, or
- the percentage of 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 the

- *p*-value is less than 0.30 or above .80,
- percentage of students for any possible score point is less than 3%,
- item-total correlation is less than 0.30, or
- the percentage of omitting or not reaching the item is 15% or greater.

Item Response Theory (IRT)

Item parameters for items included in LEAP 2025 High School 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

$$P_j(\theta) = \frac{1}{1+\exp[-Da_j(\theta-b_j)]},$$

where *D* is 1.7, 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

$$P_{jx}(\theta) = \frac{exp[\sum_{k=0}^{x} Da_j(\theta-b_j+d_{jk})]}{\sum_{h=0}^{m_i} exp[\sum_{k=0}^{h} Da_j(\theta-b_j+d_{jk})]},$$

where d_{jk} is the relative difficulty of score category *x* of item *j*, and m_i is the maximum item score for item *j*.

The software PARSCALE (Muraki & Bock, 2003) was used for the IRT calibrations. PARSCALE is a multipurpose program that implements a variety of IRT models associated with mixed-item formats and associated statistics. PARSCALE has been used to calibrate large data sets, such as those of PARCC and Smarter Balanced assessments. The program implements MML estimation techniques for items and MLE estimation of theta.

Copyright © 2019 by Louisiana Department of Education.

6.2 Calibration Sample

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

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

Sample data was used for calibration with all subjects for the spring 2018 administration. Since full census data was not available, the intention was to use data files that had at least 20% of the test takers for a subject scored to completion. The sample was evaluated using Spring 2017 bio-demographic information and student performance to confirm that it was representative of the state student population. Tables 6.3 and 6.4 show the representativeness of the calibration samples compared to the census data of fall 2017 and spring 2018 administrations, including initial testers only. Calibration in both administrations excluded re-testers. The census data in these tables included initial testers who received a scale score. With LEAP 2025 High School tests, psychometric analyses such as item calibration and item statistics are computed excluding re-testers since students taking the exam again are not representative of the population. English I was not previously administered as part of EOC testing; however beginning in the fall of 2017 administration; it became part of LEAP 2025 testing. Therefore effort was made to include most English I students in the calibration. In both administrations, the calibration samples were representative of census data. The spring 2018 re-sampling strategy was successful in sampling the state representative of students.

Copyright © 2019 by Louisiana Department of Education.

Table 6.3 Summary of Calibration and Census Data: Fall Administration

Calibration and Census Data: Fall Administration						
		Calibration Sample		Census Data Initial Testers Only		
Course		N	%	N	%	(Calib % - Census %)
English I	All Students	≥4,590	100.00%	≥4,660	100.00%	0.00%
	Gender					
	Male	≥2,320	50.54%	≥2,380	51.07%	0.53%
	Female	≥2,270	49.46%	≥2,280	48.93%	(0.53%)
	Race Ethnicity					
	Hispanic/Latino	≥230	5.03%	≥230	5.11%	0.08%
	American Indian or Alaska Native	≥40	0.89%	≥40	0.88%	(0.01%)
	Asian	≥70	1.68%	≥70	1.67%	(0.00%)
	Black or African American	≥1,700	37.03%	≥1,730	37.12%	0.10%
	Native Hawaiian or Other Pacific	<10	NR	<10	NR	NR
	White	≥2,450	53.42%	≥2,480	53.26%	(0.16%)
	Two or More Races	≥80	1.89%	≥80	1.89%	(0.01%)
	Economically Disadvantaged	≥290	6.42%	≥300	6.63%	0.21%
	Not Economically Disadvantaged	≥4,290	93.58%	≥4,350	93.37%	(0.21%)
English II	All Students	≥6,000	100.00%	≥6,050	100.00%	0.00%
	Gender					
	Male	≥3,000	49.98%	≥3,030	50.17%	0.20%
	Female	≥3,000	50.02%	≥3,010	49.83%	(0.20%)
	Race Ethnicity					
	Hispanic/Latino	≥530	8.86%	≥530	8.86%	0.00%
	American Indian or Alaska Native	≥30	0.58%	≥30	0.58%	(0.00%)
	Asian	≥140	2.41%	≥140	2.40%	(0.02%)
	Black or African American	≥2,410	40.19%	≥2,430	40.31%	0.12%
	Native Hawaiian or Other Pacific	<10	NR	<10	NR	(0.00%)
	White	≥2,750	45.78%	≥2,760	45.68%	(0.10%)
	Two or More Races	≥120	2.13%	≥120	2.13%	0.00%
	Economically Disadvantaged	≥580	9.72%	≥590	9.78%	0.06%
	Not Economically Disadvantaged	≥5,420	90.28%	≥5,450	90.22%	(0.06%)

Table continues

Copyright © 2019 by Louisiana Department of Education.

Calibration and Census Data: Fall Administration (continued)						
		Calibration Sample		Census Data Initial Testers Only		
Course		N	%	N	%	(Calib % - Census %)
Algebra I	All Students	≥2,940	100.00%	≥2,960	100.00%	0.00%
	Gender					
	Male	≥1,560	53.01%	≥1,570	53.08%	0.08%
	Female	≥1,380	46.99%	≥1,390	46.92%	(0.08%)
	Race Ethnicity					
	Hispanic/Latino	≥180	6.18%	≥200	6.77%	0.59%
	American Indian or Alaska Native	≥20	0.85%	≥20	0.84%	(0.01%)
	Asian	≥60	2.21%	≥60	2.19%	(0.02%)
	Black or African American	≥1,170	39.86%	≥1,170	39.68%	(0.19%)
	Native Hawaiian or Other Pacific	<10	NR	<10	NR	NR
	White	≥1,440	49.03%	≥1,440	48.67%	(0.36%)
	Two or More Races	≥50	1.80%	≥50	1.79%	(0.01%)
	Economically Disadvantaged	≥410	14.09%	≥420	14.42%	0.32%
	Not Economically Disadvantaged	≥2,530	85.91%	≥2,540	85.58%	(0.32%)
Geometry	All Students	≥5,460	100.00%	≥5,470	100.00%	0.00%
	Gender					
	Male	≥2,610	47.81%	≥2,610	47.83%	0.02%
	Female	≥2,850	52.19%	≥2,850	52.17%	(0.02%)
	Race Ethnicity					
	Hispanic/Latino	≥480	8.88%	≥480	8.94%	0.06%
	American Indian or Alaska Native	≥30	0.68%	≥30	0.68%	(0.00%)
	Asian	≥150	2.82%	≥150	2.81%	(0.01%)
	Black or African American	≥2,120	38.91%	≥2,120	38.88%	(0.03%)
	Native Hawaiian or Other Pacific	<10	NR	<10	NR	NR
	White	≥2,540	46.53%	≥2,540	46.52%	(0.01%)
	Two or More Races	≥110	2.11%	≥110	2.10%	(0.00%)
	Economically Disadvantaged	≥490	8.99%	≥490	9.01%	0.02%
	Not Economically Disadvantaged	≥4,970	91.01%	≥4,970	90.99%	(0.02%)

Copyright © 2019 by Louisiana Department of Education.

Table 6.4 Summary of Calibration and Census Data: Spring Administration

Calibration and Census Data: Spring Administration						
		Calibration Sample		Census Data Initial Testers Only		
Course		N	%	N	%	(Calib % - Census %)
English I	All Students	≥16,650	100.00%	≥46,310	100.00%	0.00%
	Gender					
	Male	≥8,440	50.69%	≥23,500	50.75%	0.06%
	Female	≥8,210	49.31%	≥22,810	49.25%	(0.06%)
	Race Ethnicity					
	Hispanic/Latino	≥920	5.52%	≥2,840	6.14%	0.61%
	American Indian or Alaska Native	≥100	0.64%	≥290	0.63%	(0.01%)
	Asian	≥320	1.93%	≥830	1.79%	(0.14%)
	Black or African American	≥6,760	40.61%	≥20,110	43.43%	2.83%
	Native Hawaiian or Other Pacific	<10	NR	≥20	0.06%	NR
	White	≥8,260	49.62%	≥21,420	46.26%	(3.36%)
	Two or More Races	≥270	1.64%	≥780	1.69%	0.05%
	Economically Disadvantaged	≥9,480	56.94%	≥27,190	58.71%	1.77%
	Not Economically Disadvantaged	≥4,610	27.73%	≥13,450	29.06%	1.33%
English II	All Students	≥25,970	100.00%	≥42,030	100.00%	0.00%
	Gender					
	Male	≥12,710	48.95%	≥20,760	49.41%	0.46%
	Female	≥13,260	51.05%	≥21,260	50.59%	(0.46%)
	Race Ethnicity					
	Hispanic/Latino	≥1,380	5.34%	≥2,280	5.42%	0.08%
	American Indian or Alaska Native	≥220	0.87%	≥290	0.69%	(0.18%)
	Asian	≥500	1.95%	≥770	1.83%	(0.11%)
	Black or African American	≥10,870	41.86%	≥18,170	43.24%	1.38%
	Native Hawaiian or Other Pacific	≥10	0.07%	≥30	0.08%	0.00%
	White	≥12,540	48.29%	≥19,830	47.19%	(1.10%)
	Two or More Races	≥410	1.61%	≥640	1.54%	(0.07%)
	Economically Disadvantaged	≥14,790	56.94%	≥23,880	56.82%	(0.12%)
	Not Economically Disadvantaged	≥8,020	30.91%	≥13,190	31.40%	0.49%

Table continues

Copyright © 2019 by Louisiana Department of Education.

Calibration and Census Data: Spring Administration (continued)						
		Calibration Sample		Census Data Initial Testers Only		
Course		N	%	N	%	(Calib % - Census %)
Algebra I	All Students	≥14,300	100.00%	≥46,420	100.00%	0.00%
	Gender					
	Male	≥7,060	49.41%	≥23,050	49.66%	0.25%
	Female	≥7,230	50.59%	≥23,370	50.34%	(0.25%)
	Race Ethnicity					
	Hispanic/Latino	≥810	5.68%	≥2,940	6.35%	0.67%
	American Indian or Alaska Native	≥70	0.54%	≥300	0.65%	0.11%
	Asian	≥240	1.68%	≥830	1.80%	0.12%
	Black or African American	≥6,290	44.02%	≥19,800	42.67%	(1.36%)
	Native Hawaiian or Other Pacific	≥10	0.08%	≥30	0.08%	(0.01%)
	White	≥6,640	46.45%	≥21,720	46.80%	0.35%
	Two or More Races	≥220	1.55%	≥770	1.66%	0.11%
	Economically Disadvantaged	≥8,290	57.96%	≥26,800	57.73%	(0.23%)
	Not Economically Disadvantaged	≥4,170	29.18%	≥13,710	29.55%	0.37%
Geometry	All Students	≥15,190	100.00%	≥36,990	100.00%	0.00%
	Gender					
	Male	≥7,140	47.05%	≥17,560	47.47%	0.42%
	Female	≥8,040	52.95%	≥19,430	52.53%	(0.42%)
	Race Ethnicity					
	Hispanic/Latino	≥730	4.83%	≥2,060	5.57%	0.74%
	American Indian or Alaska Native	≥70	0.49%	≥190	0.52%	0.03%
	Asian	≥240	1.61%	≥710	1.94%	0.34%
	Black or African American	≥6,200	40.87%	≥15,760	42.61%	1.75%
	Native Hawaiian or Other Pacific	<10	NR	≥10	0.05%	(0.00%)
	White	≥7,690	50.67%	≥17,670	47.77%	(2.90%)
	Two or More Races	≥220	1.49%	≥560	1.53%	0.04%
	Economically Disadvantaged	≥8,340	54.91%	≥19,890	53.78%	(1.14%)
	Not Economically Disadvantaged	≥4,460	29.40%	≥12,450	33.67%	4.27%

Copyright © 2019 by Louisiana Department of Education.

6.3 Calibration and Linking

LEAP 2025 high school item calibration and linking were performed based on item response theory (IRT). Calibration and linking methodology used for the fall 2017 and spring 2018 LEAP 2025 High School administration closely followed most of the PARCC methods referenced in the PARCC document *Final Technical Report for 2015 Administration*. To maintain comparability to PARCC, the 2PL/GPC IRT model was applied to item calibration using the software PARSCALE (Muraki & Bock, 2003). To avoid local independence between traits, the writing traits written expression (WE) and written knowledge and use of language (WKL) were separately calibrated using the sparse matrix method.

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

Descriptions of the PARCC calibration and equating approach can be found in the PARCC documents *Final Technical Report for 2015 Administration* (see https://parcc-assessment.org/wp-content/uploads/2018/02/PARCC-2015-Tech-Report.pdf) and *Final Technical Report for 2016 Administration* (see https://parcc-assessment.org/wp-content/uploads/2018/02/PARCC-2016-Tech-Report.pdf).

There was one CBT test form per course for each administration. Table 6.5 summarizes the student count and item count by administration for the courses.

The following two steps were taken to place the LEAP 2025 tests on the 2016 PARCC baseline scale with both fall 2017 and spring 2018 administrations:

1.1.1 Calibrate the LEAP 2025 High School tests.
1.1.2 Link LEAP 2025 High School tests to the 2016 PARCC baseline scale under the non-equivalent common item design.

PARCC established a new baseline scale using 2016 PARCC Spring tests. Fall 2017 and spring 2018 LEAP 2025 High School tests were directly linked to this PARCC 2016 baseline scale using PARCC item parameters as anchor item parameters. Therefore, fall 2017 and spring 2018 LEAP 2025 High School tests were placed on the PARCC scale. Since fall 2017 and spring 2018 LEAP 2025 High School tests were calibrated with Louisiana students, the scale for fall 2017 and spring 2018 LEAP 2025 High School tests will be referred to as LEAP 2025 scale, although its scale was placed on PARCC scales built with PARCC associated states' data. Spring 2019 LEAP 2025 forms will be linked to the LEAP 2025 scale using LEAP 2025 items, which will be administered on LEAP 2025 forms in 2018-2019, as anchors by the Stocking & Lord procedure. Since the 2019 anchor items are on PARCC scale, future LEAP 2025 forms continue to be considered on the PARCC scale.

Copyright © 2019 by Louisiana Department of Education.

Calibration of 2017-2018 LEAP 2025 Tests

For the LEAP 2025 item calibration, the 2PL/GPC IRT model was applied to the Louisiana students' calibration samples using the software PARSCALE (Muraki & Bock, 2003). Table 6.5 shows the number of students in the calibration samples and number of items. Spring calibration included population representative samples of students. For English I and II, reading items (RL/RI) in the writing prompts are not counted in the N-Items columns because the calibration does not include reading item scores; it only includes WE item scores. A RL/RI score and a WE item score for the same writing prompt are the same. There were 32 English items and 38 or39 mathematics items in both administrations.

Table 6.5 Summary of Student Count and Item Count

Administration	Course	N-Students	N-Items
Fall	English I	≥4,590	32
	English II	≥6,000	32
	Algebra I	≥2,940	39
	Geometry	≥5,460	38
Spring	English I	≥16,650	32
	English II	≥25,970	32
	Algebra I	≥14,300	39
	Geometry	≥15,190	38

Separate Calibration for ELA Prose Constructed-Response Tasks

For English I and II, the same sample repeated design was applied to the WE and WKL calibration to address the issue of local independence for ELA prose-constructed response (PCR) tasks. For English I, for example, two datasets of responses were generated. One calibration dataset included two WE responses and the responses of the other items, but excluded the WKL responses. The other calibration dataset included two WKL responses and the same responses of the other items, but excluded the WE responses. Therefore, these datasets were the same except for WE and WKL items, and each dataset included either WE or WKL responses. After each dataset was separately calibrated, the item parameters with WKL responses were equated to those with WE responses using all common items as anchor items. Table 6.6 illustrates this design.

Table 6.6 Calibration Data Structure for ELA WE and WKL Traits with the Same Sample Repeated Design

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

Copyright © 2019 by Louisiana Department of Education.

IRT Item Fit

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

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

After convergence was achieved for each IRT data set, an item characteristic curve (ICC) for each item was plotted together with empirical students' performances from theta ability -4 to 4. No items were suppressed from calibration due to poor fit. No items were removed from the anchor sets due to poor fit with the LEAP 2025 High School tests.

After calibration, the IRT model fit was evaluated by reviewing item chi-square values from PARSCALE (Muraki & Bock, 2003) and calculating adjusted fit values and flagging them if > 0.45 (Pearson, 2015).

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

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

One limitation of the PARSCALE (Muraki & Bock, 2003) output is that when a chi-square value is greater than 9,999.99, PARSCALE does not print the value in the phase two output. Instead, it prints asterisks (*****). When that happens, the chi-square value or adjusted fit is not available.

Tables 6.7 and 6.8 show adjusted item fit using PARSCALE chi-square values and calibration sample sizes for the English and mathematic content (Muraki & Bock, 2003). The average adjusted fit ranged from 0.16 to 0.18 for the fall administration and 0.11 to 0.15 for the spring administration. Items were not excluded based on model fit statistics because the adjusted item fits for all items were lower than the criterion value of 0.45, as can be seen in the maximum values for both administrations. The largest adjusted fit value was 0.36 for geometry in the fall administration.

Table 6.7 Summary of Adjusted Fit for Fall Administration

Course	N-Items	Mean	Std. Dev.	Min.	Max.	N-Flagged Items	N-Not Available
English I	32	0.16	0.03	0.10	0.23	0	0
English II	32	0.17	0.04	0.11	0.26	0	0
Algebra I	39	0.16	0.05	0.10	0.28	0	0
Geometry	38	0.18	0.07	0.10	0.36	0	0

Copyright © 2019 by Louisiana Department of Education.

Table 6.8 Summary of Adjusted Fit for Spring Administration

Course	N-Items	Mean	Std. Dev.	Min.	Max.	N- Flagged Items	N-Not Available
English I	32	0.15	0.04	0.09	0.28	0	0
English II	32	0.14	0.04	0.08	0.22	0	0
Algebra I	39	0.11	0.04	0.06	0.22	0	0
Geometry	38	0.12	0.05	0.06	0.28	0	0

Linking 2017-2018 LEAP 2025 to PARCC Scale

This section explains the linking procedure used to place the LEAP 2025 High School spring 2018 administration onto the PARCC scale. The same procedure explained in this section also applies to the fall 2017 administration linking. The 2017 fall administration was the first administration of the LEAP 2025 High School tests and the Stocking & Lord procedure (1983) was used to link the LEAP 2025 tests to the PARCC scale using intact PARCC items embedded in the test forms. This yielded item parameters on the PARCC scale, however, the equated Louisiana item parameters were based on only Louisiana students' responses from the fall 2017 administration while intact PARCC item parameters were estimated based on PARCC associated states' responses. To distinguish these sets of two item parameters, item parameters based on only Louisiana students' responses will be called LEAP 2025 item parameters and its scale as the LEAP 2025 scale.

The 2018 spring administration was the first spring administration of LEAP 2025 High School tests. As the fall and spring administration populations were not similar, the 2018 spring administration was equated to the PARCC scale with intact PARCC parameters.

Although it was pre-determined that the spring 2018 administration was going to be linked to PARCC with intact PARCC parameters, it was approved by the Technical Advisory Council (TAC) in March 2018 to use three anchor sets in the 2018 Spring LEAP 2025 High School assessments equating process. Additionally, the use of multiple anchor sets will assist in establishing comparability between the LEAP 2025 scale and the PARCC scale for the current and future administrations.

Anchor 1 items are intact PARCC items embedded in the spring 2018 LEAP 2025 forms, and Anchor 1 results were the final item parameters for the 2018 spring LEAP 2025 High School tests. Anchor 2 items were common items between the fall 2017 and the spring 2018 LEAP 2025 High School spring forms. Anchor 3 item parameters consisted of all Anchor 2 item parameters, and Anchor 1 item parameters were used for the remaining items. Table 6.9 summarizes the number and score points of the initial anchor item selection before equating. Table 6.9 also summarizes the number and score points of the final anchor item selections. The difference between the initial number of anchor items and the final number of anchor items are the number of anchor items dropped.

Copyright © 2019 by Louisiana Department of Education.

Table 6.9 The Number of Linking Items for Linking LEAP 2025 Spring 2018 to PARCC

Admin	Course		Anchor 1		Anchor 2		Anchor 3	
			N. of Items	Score Points	N. of Items	Score Points	N. of Items	Score Points
Spring	English I	Initial	32	70	14	31	32	70
		Final	23	49	14	31	21	45
	English II	Initial	31	68	14	31	31	68
		Final	28	61	12	26	26	57
	Algebra I	Initial	39	68	15	20	39	68
		Final	30	57	9	12	28	54
	Geometry	Initial	38	68	15	27	38	68
		Final	27	52	13	23	31	57

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

Table 6.10 presents the slope and intercept equating constants for each anchor set for each subject. The constants from Anchor 1 were used to bring the 2018 LEAP 2025 estimated parameters onto the PARCC IRT scale.

Table 6.10 Stocking & Lord Transformation Constants for Linking LEAP 2025 Spring 2018 to PARCC

Admin	Course	Anchor 1		Anchor 2		Anchor 3	
		Slope	Intercept	Slope	Intercept	Slope	Intercept
Spring	English I	0.897237	0.170078	0.916910	0.044801	0.885096	0.080349
	English II	0.880115	0.246457	0.865672	0.153463	0.864477	0.168286
	Algebra I	0.917254	0.125604	0.898884	0.136905	0.917397	0.168600
	Geometry	0.911517	0.115516	0.928992	0.156748	0.922761	0.105248

Figures 6.1 to 6.4 are test characteristic curves for the anchor items, the corresponding LEAP 2025 Spring 2018 estimated anchor items before equating (ANC) and after equating (EQ_ANC), intact PARCC parameters (Pre_EQ), and all 2018 LEAP 2025 estimated items (EQ_ALL) for LEAP 2025 High School Spring 2018 after applying the Stocking & Lord equating procedure. The blue solid line illustrates the pre-equated anchor items, the red dotted line is the 2018 LEAP 2025 equated anchor items, the black dash line is the PARCC parameters, and the green dotted line is for all 2018 LEAP 2025 High School equated items. Anchor items for each anchor set 1, 2, and 3 are different as mentioned above. For all tests, the TCCs for anchor items and the corresponding 2018 estimated anchor items were overlapped across most ability levels.

When anchor (blue) and pre-equated (black) TCCs are close to each other, it means the anchor value in the condition is similar to the TCC with PARCC parameters' pre-equated form. The anchor and pre-equated TCCs are closely overlapped with most tests with the Anchor 1 condition. With the Anchor 2 condition, the anchor TCC used the 2017 fall administration LEAP 2025 parameters and were far from the pre-equated PARCC TCC, especially with English II and geometry.

Pre-equated (black) and equated (green) TCCs overlapped closely with all tests with the Anchor 1 condition. Under the Anchor 2 condition, those TCCs were slightly far from each other with English I and English II, indicating that equating results could differ if fall 2017 administration LEAP 2025 parameters were used. The number of anchor items with Anchor 2 conditions are small, thus it is likely that the parameter difference as

Copyright © 2019 by Louisiana Department of Education.

well as the limited number of anchor items resulted in the TCCs difference. To clarify, the results from Anchor 1 was used as the final equating results for Spring 2018 LEAP 2025 High School tests.

Figures 6.5 to 6.8 present scatter plots of slope item parameters and difficulty item parameters and their correlation after linking 2018 LEAP 2025 to the PARCC 2016 scale.

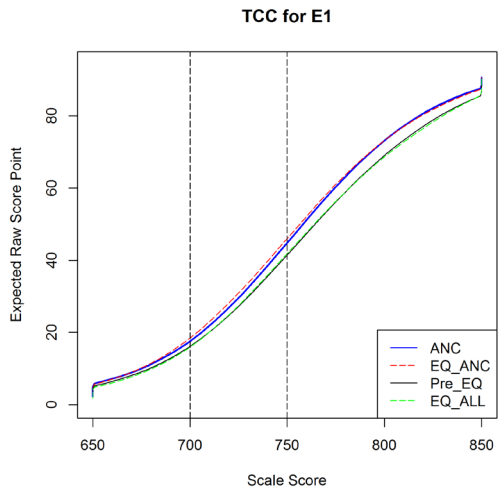
The parameters were mostly close to the identity line with Anchor 1, and the correlation ranged from 0.94 to 0.98. Most parameters were around the identity line with Anchor 2, as the anchor parameters were from LA students, except the writing items with English I and English II. The correlation for Anchor 2 ranged from 0.97 to 0.99.

Compared to English I and English II, Algebra I and Geometry item discrimination parameters were a little scattered from the identity line. It is usual to find higher correlations for difficulty parameters than those for slope parameters.

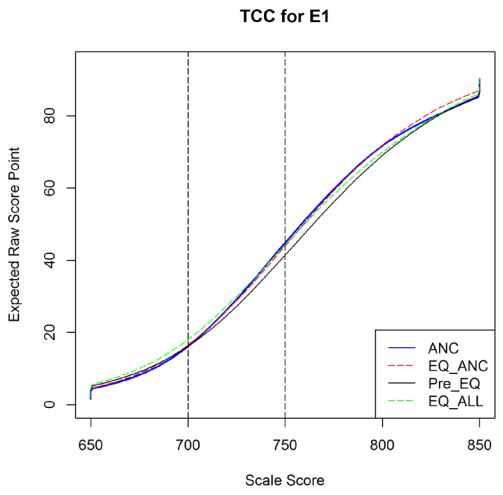
Copyright © 2019 by Louisiana Department of Education.

Figure 6.1 English I TCC between Anchor, 2018 LEAP 2025 Equated Anchor, PARCC, and ALL 2018 LEAP 2025 Items

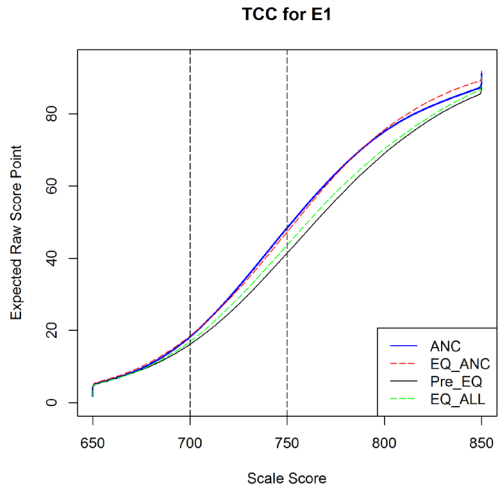
Anchor 1



Anchor 2



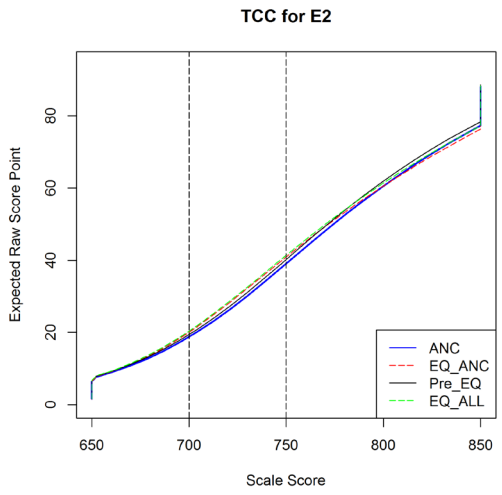
Anchor 3



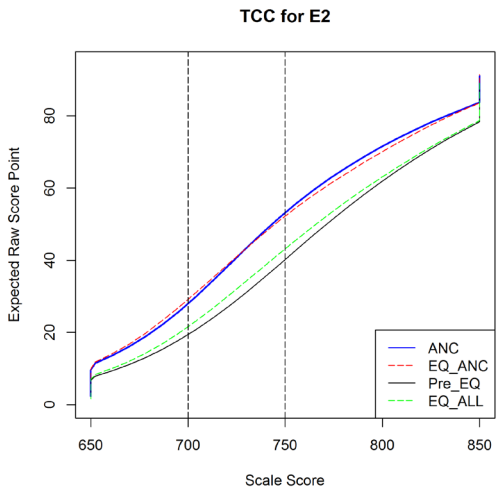
Copyright © 2019 by Louisiana Department of Education.

Figure 6.2 English II TCC between Anchor, 2018 LEAP 2025 Equated Anchor, PARCC, and ALL 2018 LEAP 2025 Items

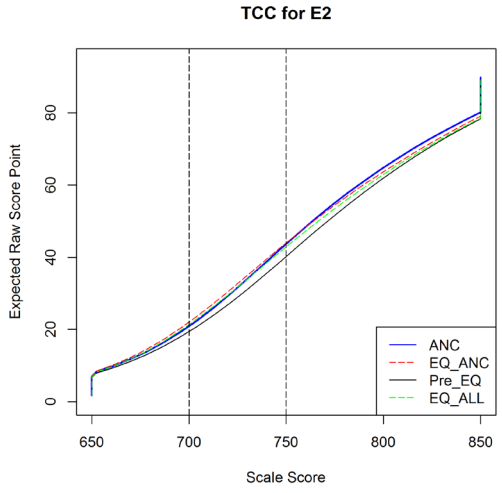
Anchor 1



Anchor 2



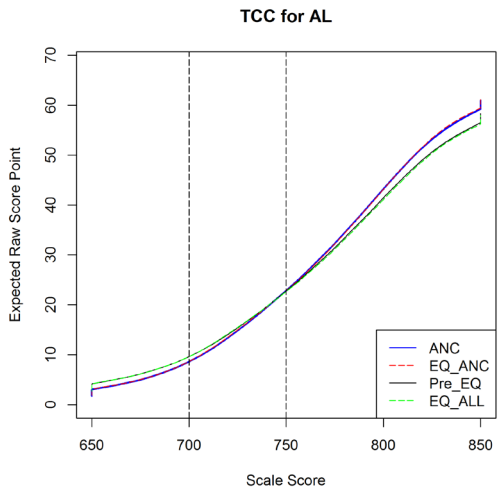
Anchor 3



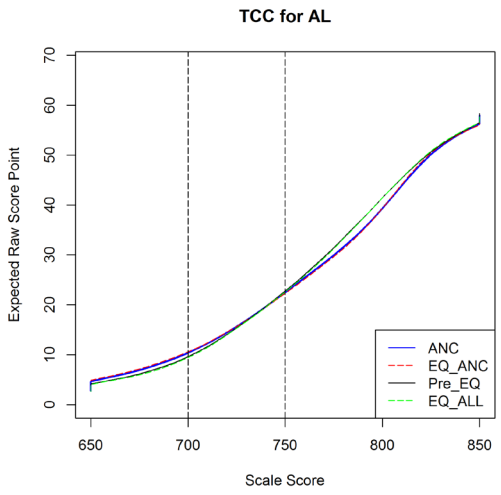
Copyright © 2019 by Louisiana Department of Education.

Figure 6.3 Algebra I TCC between Anchor, 2018 LEAP 2025 Equated Anchor, PARCC, and ALL 2018 LEAP 2025 Items

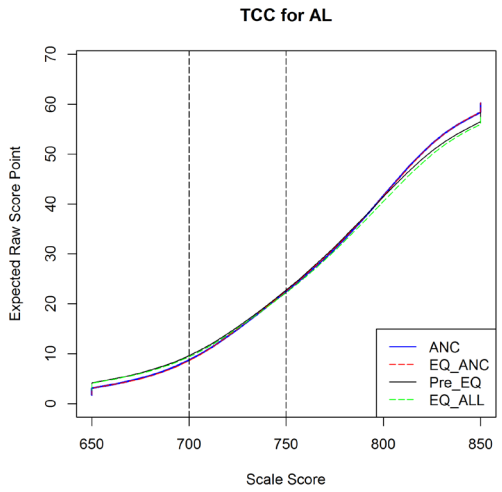
Anchor 1



Anchor 2



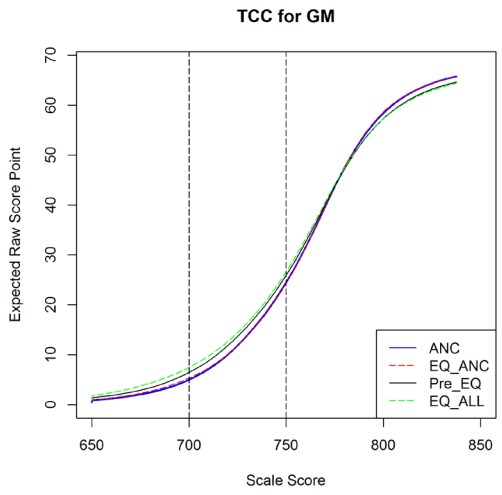
Anchor 3



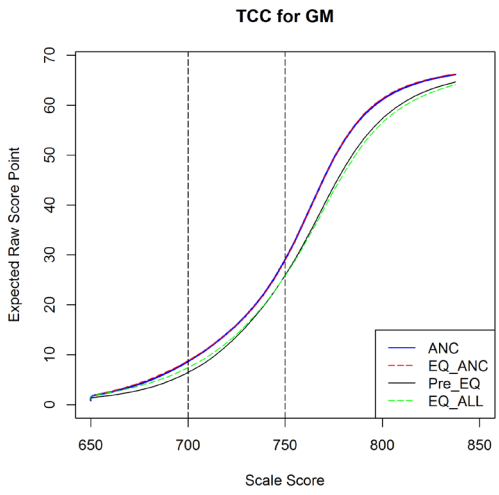
Copyright © 2019 by Louisiana Department of Education.

Figure 6.4 Geometry TCC between Anchor, 2018 LEAP 2025 Equated Anchor, PARCC, and ALL 2018 LEAP 2025 Items

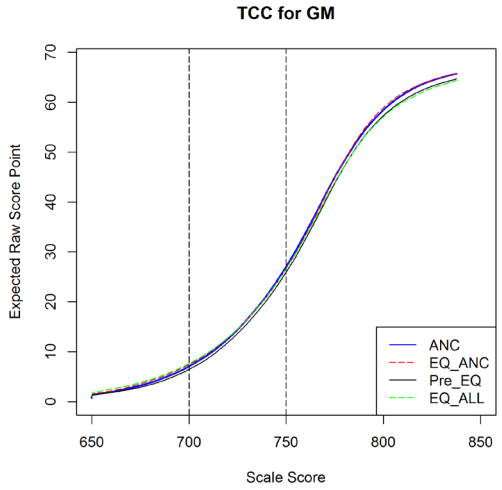
Anchor 1



Anchor 2

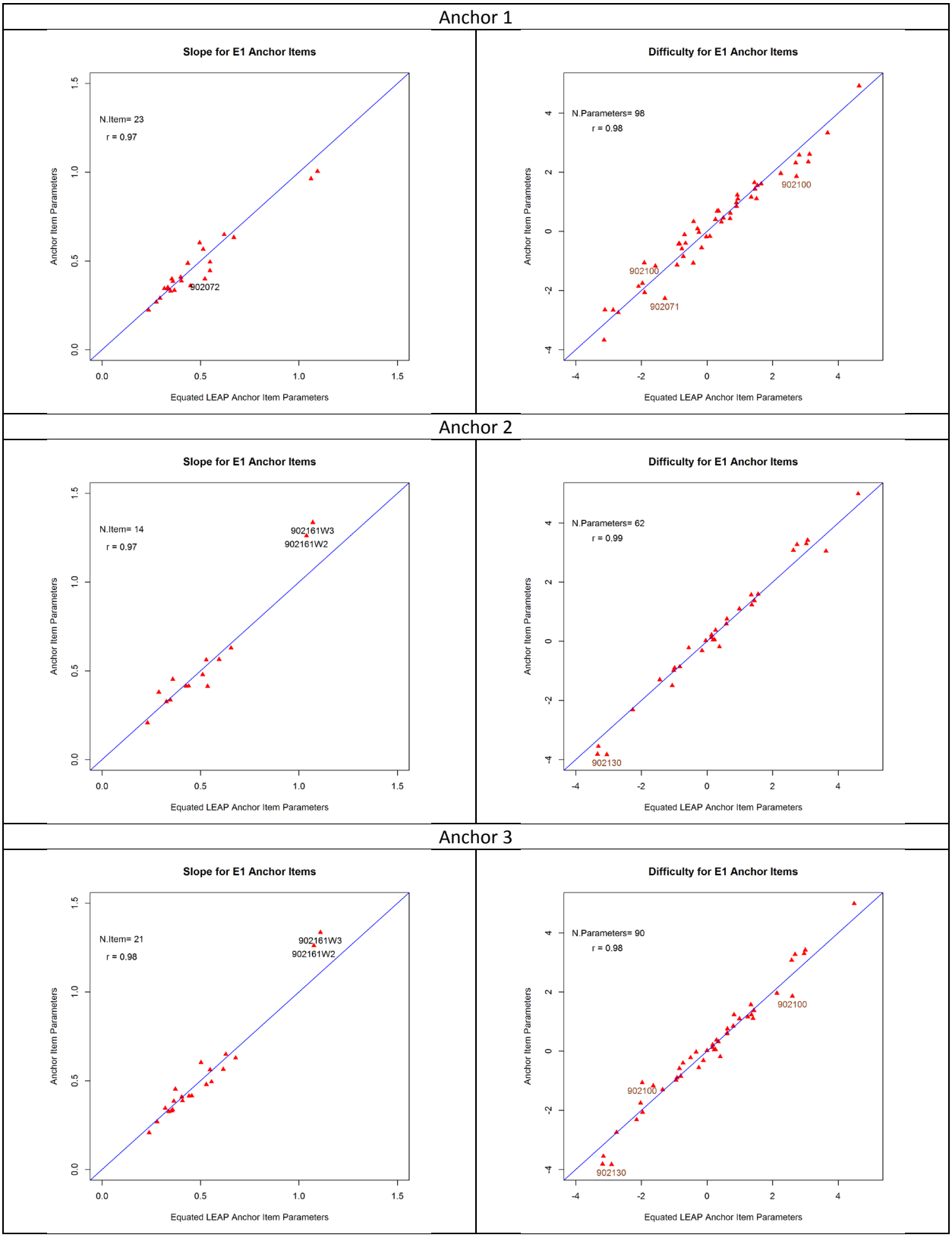


Anchor 3



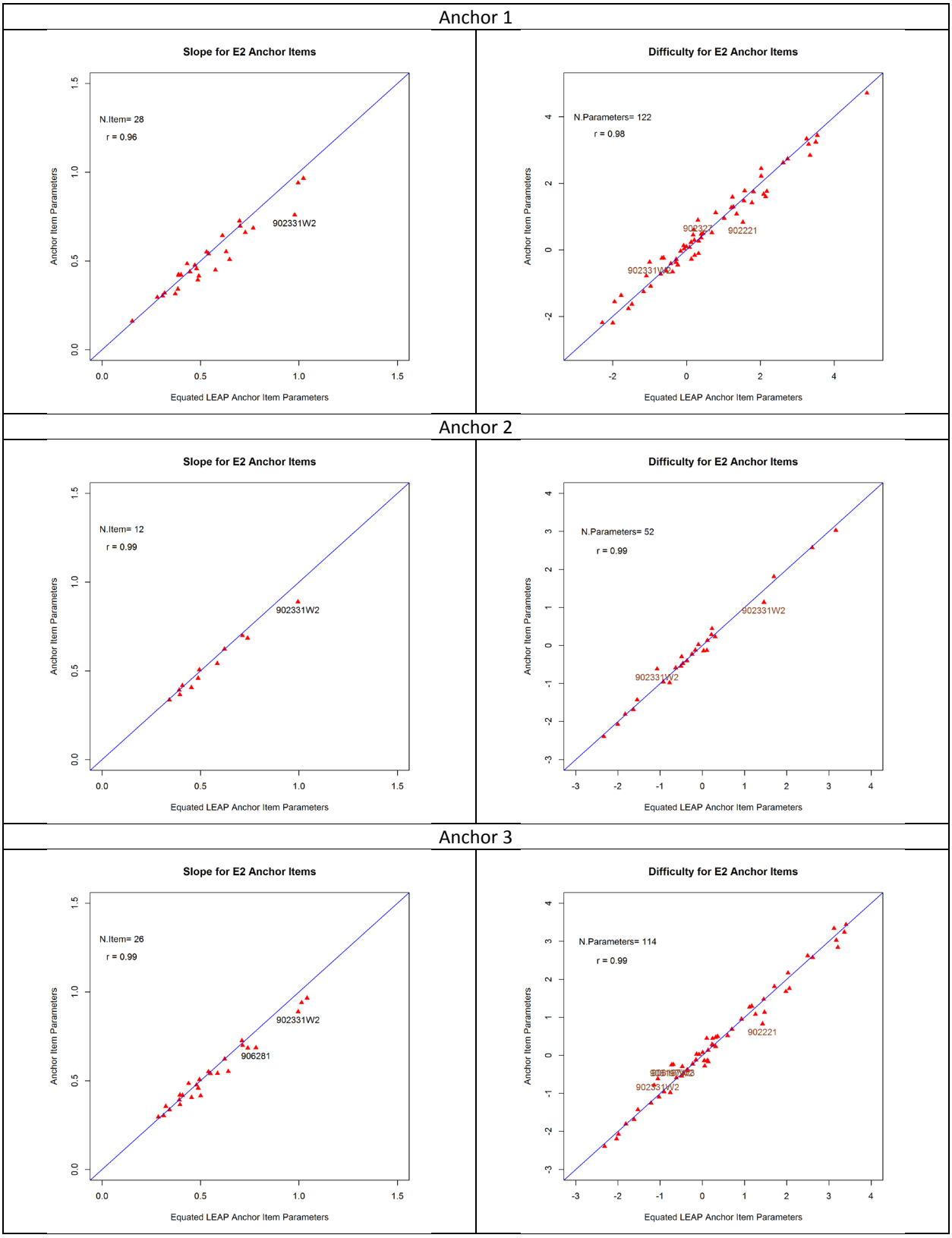
Copyright © 2019 by Louisiana Department of Education.

Figure 6.5 English I Anchor Slope and Difficulty between Equated Anchor Item parameters and Anchor Item Parameters



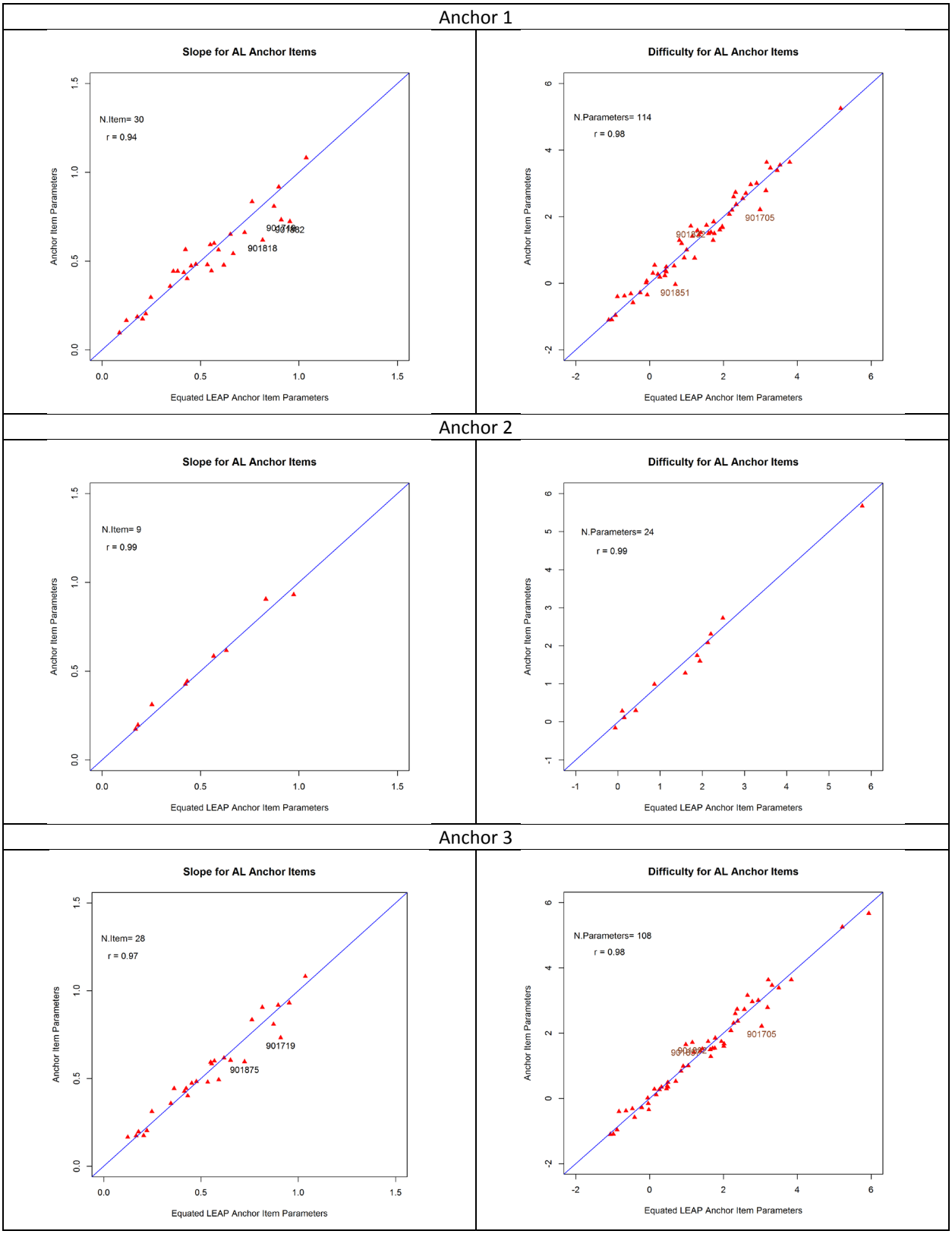
Copyright © 2019 by Louisiana Department of Education.

Figure 6.6 English II Anchor Slope and Difficulty between Equated Anchor Item parameters and Anchor Item Parameters



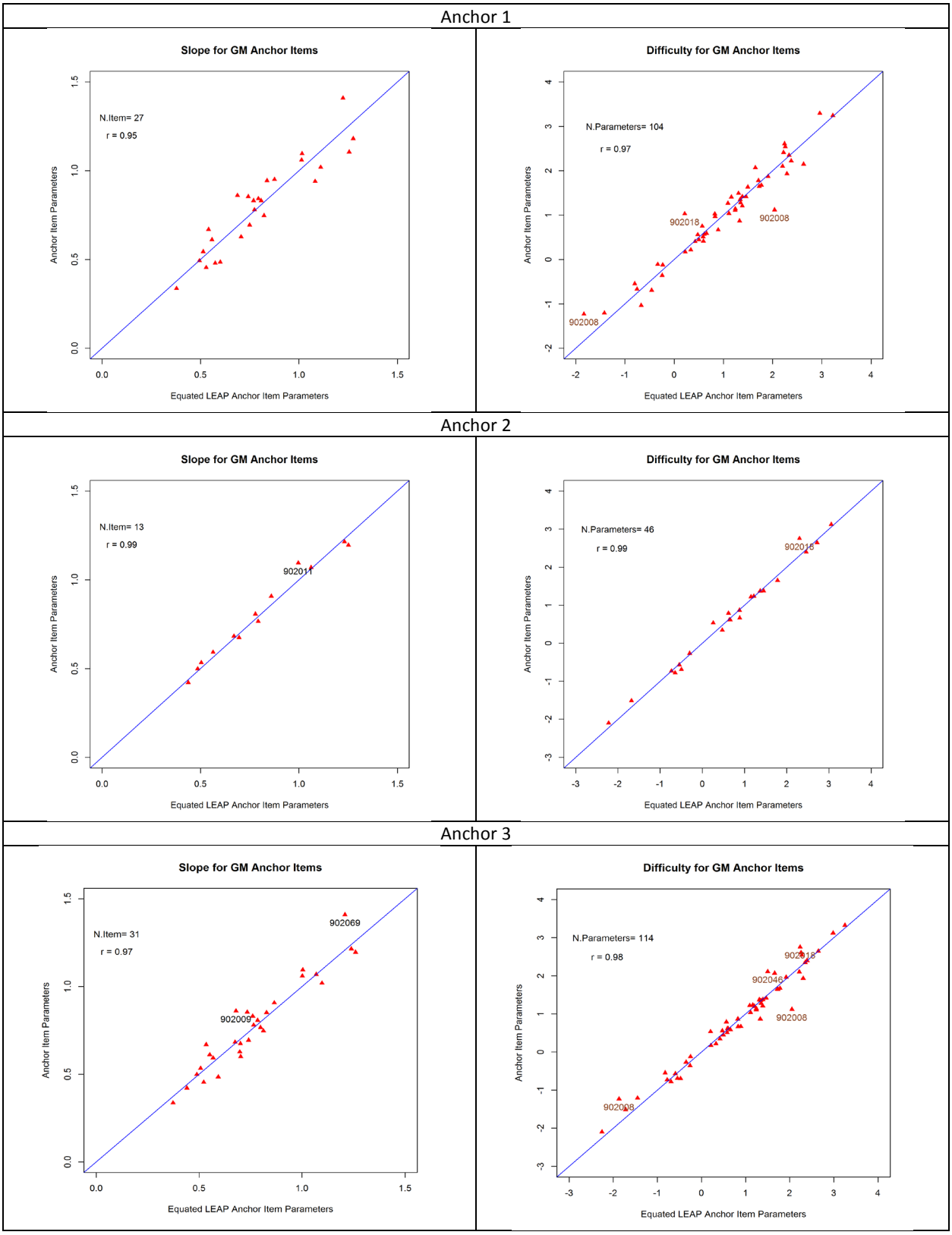
Copyright © 2019 by Louisiana Department of Education.

Figure 6.7 Algebra Anchor Slope and Difficulty between Equated Anchor Item parameters and Anchor Item Parameters



Copyright © 2019 by Louisiana Department of Education.

Figure 6.8 Geometry Anchor Slope and Difficulty between Equated Anchor Item parameters and Anchor Item Parameters



Copyright © 2019 by Louisiana Department of Education.

Evaluation of Anchor Item Stability

Standard 5.15 requires information about the anchors, stating the following:

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

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

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

- Exclude CR items from anchor set if categories were collapsed due to small sample size.
- Exclude items with content or parameter estimation issues.
- Run Robust Z method and remove flagged items from anchor set using the criterion value of |1.96|
- Run STUIRT using the remaining items after Robust Z is applied, and flag items for further inspection if the WRMSD was greater than the values in Table 6.25. If the items were flagged, then they were removed from the anchor set and the ICC was reviewed with the WRMSD (Kim & Kolen, 2004).
- Flag outliers using the plots of slope and difficulty item parameters with their correlations (Kolen & Brennan, 2014).
- Check score points and the numbers of items by claim and subclaim before and after dropping an anchor item.

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

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

Copyright © 2019 by Louisiana Department of Education.

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

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

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

where *Q* represents a quadrature point (i.e., node), *W* represents its weight given quadrature point *Q* from PARSCALE PH2 output, *INTACT* represents intact PARCC item parameters, and *EST* represents estimated item parameters corresponding to intact PARCC item parameters. Table 6.11 summarizes WRMSD flagging criteria for inspection and possible removal of linking items.

Table 6.11 PARCC WRMSD Flagging Criteria

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

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 by 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 course, are called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). All LEAP 2025 High School content areas in 2018 used the same LOSS and HOSS values, as established by PARCC. The LOSS value was 650, and the HOSS value was 850.

Claim-Level, Subclaim-Level, and Subcategory-Level Subscores

A student's performance on the ELA claims (i.e., reading and writing) is reported in one of three ratings: *Strong*, *Moderate*, or *Weak*.

Copyright © 2019 by Louisiana Department of Education.

Additionally, subclaim ratings are reported at the student level. English I and English II have three subclaims for reading (i.e., literary text, informational text, and vocabulary) and two subclaims for writing (i.e., written expression and knowledge and use of language conventions).

Algebra I and Geometry have four subclaims. The subclaim A, Major Content, is further reported at three subcategories in algebra I and two subcategories in Geometry. Subclaim and subcategory performance is reported in one of three ratings of achievement: *Strong*, *Moderate*, or *Weak*.

Although the performance ratings are determined only by the items included within a claim, subclaim, or subcategory, the level of knowledge and ability needed to achieve a performance rating is connected to the level of knowledge and ability required by the subject-level achievement 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.

For all scoring tables, the TCC inverse method was applied to generate raw-to-scale-score tables following the PARCC approach.

The 2017-2018 LEAP 2025 High School reporting categories are summarized in Table 6.12.

Table 6.12 LEAP 2025 High School Reporting Categories

Course	Claim	Subclaim	Subcategory
English I/II	1. Reading 2. Writing	1. Reading Informational Texts—RI 2. Reading Literature—RL 3. Reading Vocabulary—RV 4. Written Expression—WE 5. Written Knowledge of Language—WKL	
Algebra I		1. Subclaim A – Major Content 2. Subclaim B – Additional and Supporting Content 3. Subclaim C – Expressing Mathematical Reasoning 4. Subclaim D – Modeling and Application	1. Subcategory A.1 – Interpreting Functions 2. Subcategory A.2 – Solving Algebraically 3. Subcategory A.3 – Solving Graphically/Rate of Change
Geometry		1. Subclaim A – Major Content 2. Subclaim B – Additional and Supporting Content 3. Subclaim C – Expressing Mathematical Reasoning 4. Subclaim D – Modeling and Application	1. Subcategory A.1 – Congruence Transformations/Similarity 2. Subcategory A.2 – Similarity in Trigonometry/Modeling and Applying

Copyright © 2019 by Louisiana Department of Education.

Reading and writing claim scores were produced for English I and English II. The reading claim score range is 10–90 and the writing claim score range is 10–60. The method for scaling the claims scores followed the PARCC methodology (Pearson, 2017). For the reading claim, two theta score points corresponding to English I and English II scale scores of 725 and 750 were used for scaling. Linear transformation constants mapping the two theta points to the scale score points of 40 and 50 were calculated for reading claim. After these transformation values were applied to item parameters belonging to the reading claim, a scoring table was generated using the TCC inverse method. A similar approach was applied to scale the writing claim, 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 claim (see Table 6.13 for cut scores for summatives, claims, and subclaims).

For subclaims, only performance-level ratings were reported. Therefore, there is no need to scale them. Using the item parameters belonging to a given subclaim, a raw-score-to-theta scoring table is generated by applying the TCC inverse method. The two raw scores corresponding to Θ_{L3} and Θ_{L4} are cut scores for the subclaim.

For subcategory, mathematics subclaim A is split as subcategories, A.1, A.2, and A.3. As like subclaim, only performance-level ratings were reported. The same scoring procedure and the cuts for subclaims are applied to generate a raw-score-to-theta scoring tables.

Table 6.13 Cut Scores for Summative, Claim, Subclaim, and Subcategory

Performance Level	Summative Test	Claim		Subclaim/Subcategory
		Reading	Writing	
1				
2	700	30	25	
3	725	40	30	Θ_{L3}
4	750	50	35	Θ_{L4}
5	Around 800			

*Subclaim thetas are those from summative tests (i.e., 725 & 750).
**Yellow highlight shows cut scores for claim and subclaim.

6.4 Comparability: Form Equating

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. 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.

Copyright © 2019 by Louisiana Department of Education.

In the spring of 2018, the forms used were post-equated forms linking to PARCC scale. The equating was conducted using the test characteristic transformation function method in the common-item non-equivalent-groups design (Stocking & Lord, 1983). The fall 2017 administration was also post-equated to link to the PARCC scale. The summer 2018 forms were either intact forms with established scale scores or pre-equated forms using PARCC parameters. Post-equating in summer does not occur because the small population who are administered the form are largely retesters. Since 2017-2018 was the first administration with the LEAP 2025 High School tests, the comparison of scale scores at selected percentiles will be examined in 2018-2019 technical report. Comparison among same seasonal administrations will be made (e.g., spring 2018 and spring 2019) since the population of test takers for fall, spring and summer are different. The test characteristic curves (TCCs) and conditional SEM for across years will also be examined in the 2018-2019 technical report.

6.5 Summary

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. It also helps 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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 7: Test Results

This chapter of the technical report contains information on the results of the Spring LEAP 2025 High School administration of English I, English II, Algebra I, and Geometry. 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: *Advanced*, *Mastery*, *Basic*, *Approaching Basic*, and *Unsatisfactory*.

While the scale score provides an essential quantitative reference to 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 course.

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 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.

Results presented 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 reliability and validity of the scores from the LEAP 2025 high school ELA and mathematics assessments and should not be used for state accountability purposes.

7.1 Student Participation

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

The number of attempted students, reportable students, and in the students in the technical report sample are summarized by grade in Table 7.1. Attempted students include all the students who attempted at least one item on the assessment. Reportable students are students who finished all sections in the assessment, which includes students in private school and home-study programs. The category “Technical Report Sample”

Copyright © 2019 by Louisiana Department of Education.

represents the sample of students included in the analyses for this report, and they are the students who finished all sections of the assessment and counted toward the state total score; students in private school and home-study programs were excluded from this sample.

Table 7.1 Count of Students who Attempted, were Reportable, and Included in the Technical Report Sample: English I

Administration	Group	Grade							
		6	7	8	9*	10	11	12	Total
Fall	Attempted	<10	<10	<10	≥4,160	≥410	≥70	<10	≥4,660
	Reportable	<10	<10	<10	≥4,120	≥400	≥70	<10	≥4,610
	Technical Report	<10	<10	<10	≥4,120	≥400	≥70	<10	≥4,610
Spring	Attempted	<10	<10	≥2,070	≥42,650	≥1,630	≥200	≥40	≥46,620
	Reportable	<10	<10	≥2,060	≥42,510	≥1,620	≥200	≥40	≥46,460
	Technical Report	<10	<10	≥2,040	≥42,210	≥1,620	≥200	≥40	≥46,130
Summer	Attempted	<10	<10	≥10	≥1,310	≥80	≥20	<10	≥1,430
	Reportable	<10	<10	≥10	≥1,280	≥80	≥20	<10	≥1,400
	Technical Report	<10	<10	≥10	≥1,280	≥80	≥20	<10	≥1,390

* Grade 9 includes grade that is coded as "T9."

Table 7.2 Count of Students who Attempted, were Reportable, and Included in the Technical Report Sample: English II

Administration	Group	Grade							
		6	7	8	9*	10	11	12	Total
Fall	Attempted	<10	<10	<10	≥630	≥6,660	≥1,560	≥430	≥9,290
	Reportable	<10	<10	<10	≥620	≥6,570	≥1,510	≥410	≥9,130
	Technical Report	<10	<10	<10	≥620	≥6,570	≥1,510	≥410	≥9,130
Spring	Attempted	<10	<10	<10	≥2,800	≥39,150	≥2,350	≥570	≥44,890
	Reportable	<10	<10	<10	≥2,780	≥39,000	≥2,310	≥560	≥44,670
	Technical Report	<10	<10	<10	≥2,750	≥38,720	≥2,310	≥560	≥44,350
Summer	Attempted	<10	<10	<10	≥130	≥1,710	≥420	≥90	≥2,370
	Reportable	<10	<10	<10	≥130	≥1,660	≥410	≥90	≥2,300
	Technical Report	<10	<10	<10	≥130	≥1,650	≥410	≥90	≥2,290

* Grade 9 includes grade that is coded as "T9."

Copyright © 2019 by Louisiana Department of Education.

Table 7.3 Count of Students who Attempted, were Reportable, and included in the Technical Report Sample Algebra I

Administration	Group	Grade							
		6	7	8	9*	10	11	12	Total
Fall	Attempted	<10	<10	≥10	≥4,540	≥2,610	≥1,010	≥420	≥8,610
	Reportable	<10	<10	≥10	≥4,440	≥2,520	≥970	≥400	≥8,350
	Technical Report	<10	<10	≥10	≥4,440	≥2,520	≥970	≥400	≥8,340
Spring	Attempted	<10	≥280	≥7,910	≥34,540	≥5,970	≥1,320	≥310	≥50,360
	Reportable	<10	≥260	≥7,820	≥34,400	≥5,900	≥1,290	≥300	≥50,000
	Technical Report	<10	≥260	≥7,710	≥34,120	≥5,890	≥1,280	≥300	≥49,590
Summer	Attempted	<10	<10	≥20	≥1,030	≥390	≥140	≥20	≥1,620
	Reportable	<10	<10	≥20	≥1,020	≥360	≥140	≥20	≥1,580
	Technical Report	<10	<10	≥20	≥1,010	≥360	≥140	≥20	≥1,570

* Grade 9 includes grade that is coded as "T9."

Table 7.4 Count of Students who Attempted, were Reportable, and Included in the Technical Report Sample: Geometry

Administration	Group	Grade							
		6	7	8	9*	10	11	12	Total
Fall	Attempted	<10	<10	<10	≥1,290	≥3,850	≥1,260	≥410	≥6,830
	Reportable	<10	<10	<10	≥1,280	≥3,810	≥1,220	≥390	≥6,720
	Technical Report	<10	<10	<10	≥1,280	≥3,810	≥1,220	≥390	≥6,720
Spring	Attempted	<10	≥10	≥220	≥6,440	≥26,030	≥4,770	≥630	≥38,130
	Reportable	<10	≥10	≥210	≥6,430	≥25,970	≥4,730	≥610	≥37,980
	Technical Report	<10	≥10	≥210	≥6,370	≥25,740	≥4,690	≥610	≥37,650
Summer	Attempted	<10	≥10	<10	≥30	≥210	≥120	≥20	≥410
	Reportable	<10	≥10	<10	≥20	≥200	≥110	≥20	≥390
	Tech Report	<10	≥10	<10	≥20	≥200	≥110	≥20	≥390

* Grade 9 includes grade that is coded as "T9."

The counts and percentages of students in demographic groups by grade for the group of students comprising the technical report sample for the spring 2018 administration are summarized in Tables 7.5 through Table 7.12. The same information regarding the technical report samples for the fall 2017 and summer 2018 administrations can be found in Appendix F.

Copyright © 2019 by Louisiana Department of Education.

Table 7.5 Count of Students taking the Spring Administration: English I

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	<10	<10	≥2,040	≥42,210	≥1,620	≥200	≥40	≥46,130
Gender								
Female	<10	<10	≥1,170	≥20,830	≥580	≥60	≥10	≥22,670
Male	<10	<10	≥870	≥21,380	≥1,030	≥130	≥20	≥23,450
Ethnicity								
Hispanic/Latino	<10	<10	≥170	≥2,340	≥250	≥50	<10	≥2,830
American Indian or Alaska Native	<10	<10	<10	≥270	≥10	<10	<10	≥290
Asian	<10	<10	≥150	≥650	≥10	<10	<10	≥820
Black or African American	<10	<10	≥650	≥18,440	≥840	≥90	≥30	≥20,070
Native Hawaiian or Other Pacific	<10	<10	<10	≥20	<10	<10	<10	≥20
White	<10	<10	≥990	≥19,770	≥460	≥40	≥10	≥21,280
Two or More Races	<10	<10	≥50	≥690	≥20	<10	<10	≥770
Education Classification								
Regular	<10	<10	≥1,500	≥36,370	≥1,340	≥160	≥40	≥39,430
Special	<10	<10	≥30	≥3,420	≥270	≥30	<10	≥3,780
Gifted	<10	<10	≥490	≥2,400	<10	<10	<10	≥2,910
Economic Status								
Economically Disadvantaged	<10	<10	≥880	≥24,990	≥1,200	≥150	≥30	≥27,270
Not Economically Disadvantaged	<10	<10	≥880	≥12,320	≥230	≥20	<10	≥13,470
English Learner Status								
Non-EL	<10	<10	≥2,040	≥41,390	≥1,400	≥140	≥40	≥45,010
EL	<10	<10	<10	≥820	≥220	≥50	<10	≥1,110
Migrant Status								
Nonmigrant	<10	<10	≥2,040	≥42,160	≥1,610	≥200	≥40	≥46,080
Migrant	<10	<10	<10	≥40	<10	<10	<10	≥50
Section 504 Status								
Non-Section 504	<10	<10	≥1,950	≥38,730	≥1,420	≥180	≥40	≥42,330
Section 504	<10	<10	≥90	≥3,470	≥200	≥20	<10	≥3,790

Copyright © 2019 by Louisiana Department of Education.

Table 7.6 Reportable Percentage of Students taking the Spring Administration: English I

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	0.00	0.00	4.43	91.50	3.52	0.44	0.10	100.00
Gender								
Female	0.00	0.00	2.54	45.16	1.27	0.15	0.04	49.16
Male	0.00	0.00	1.89	46.35	2.25	0.29	0.06	50.84
Ethnicity								
Hispanic/Latino	0.00	0.00	0.38	5.08	0.56	0.12	0.01	6.15
American Indian or Alaska Native	0.00	0.00	0.02	0.60	0.02	0.00	0.00	0.64
Asian	0.00	0.00	0.34	1.42	0.03	0.01	0.00	1.80
Black or African American	0.00	0.00	1.43	39.99	1.84	0.21	0.07	43.53
Native Hawaiian or Other Pacific	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06
White	0.00	0.00	2.15	42.86	1.00	0.10	0.03	46.14
Two or More Races	0.00	0.00	0.12	1.51	0.06	0.00	0.00	1.69
Education Classification								
Regular	0.00	0.00	3.27	78.86	2.91	0.36	0.09	85.49
Special	0.00	0.00	0.08	7.43	0.59	0.08	0.01	8.20
Gifted	0.00	0.00	1.08	5.21	0.02	0.01	0.00	6.31
Economic Status								
Economically Disadvantaged	0.00	0.00	1.92	54.17	2.61	0.33	0.08	59.12
Not Economically Disadvantaged	0.00	0.00	1.92	26.72	0.50	0.05	0.02	29.21
English Learner Status								
Non-EL	0.00	0.00	4.43	89.72	3.04	0.31	0.09	97.59
EL	0.00	0.00	0.01	1.78	0.48	0.13	0.02	2.41
Migrant Status								
Nonmigrant	0.00	0.00	4.43	91.41	3.51	0.44	0.10	99.89
Migrant	0.00	0.00	0.00	0.09	0.01	0.00	0.00	0.11
Section 504 Status								
Non-Section 504	0.00	0.00	4.24	83.96	3.08	0.39	0.10	91.77
Section 504	0.00	0.00	0.20	7.54	0.44	0.05	0.01	8.23

Copyright © 2019 by Louisiana Department of Education.

Table 7.7 Count of Students taking the Spring Administration: English II

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	<10	<10	<10	≥2,750	≥38,720	≥2,310	≥560	≥44,350
Gender								
Female	<10	<10	<10	≥1,350	≥19,610	≥840	≥200	≥22,020
Male	<10	<10	<10	≥1,390	≥19,100	≥1,460	≥350	≥22,320
Ethnicity								
Hispanic/Latino	<10	<10	<10	≥240	≥1,940	≥320	≥120	≥2,630
American Indian or Alaska Native	<10	<10	<10	≥10	≥270	≥10	<10	≥310
Asian	<10	<10	<10	≥120	≥620	≥30	≥10	≥790
Black or African American	<10	<10	<10	≥1,220	≥16,790	≥1,280	≥340	≥19,640
Native Hawaiian or Other Pacific	<10	<10	<10	<10	≥30	<10	<10	≥30
White	<10	<10	<10	≥1,080	≥18,460	≥610	≥80	≥20,250
Two or More Races	<10	<10	<10	≥60	≥570	≥20	<10	≥670
Education Classification								
Regular	<10	<10	<10	≥2,240	≥33,610	≥1,730	≥410	≥38,010
Special	<10	<10	<10	≥210	≥2,870	≥550	≥140	≥3,780
Gifted	<10	<10	<10	≥300	≥2,220	≥20	<10	≥2,550
Economic Status								
Economically Disadvantaged	<10	<10	<10	≥1,590	≥22,060	≥1,650	≥300	≥25,610
Not Economically Disadvantaged	<10	<10	<10	≥770	≥12,350	≥300	≥40	≥13,470
English Learner Status								
Non-EL	<10	<10	<10	≥2,680	≥37,990	≥1,980	≥440	≥43,110
EL	<10	<10	<10	≥70	≥720	≥320	≥120	≥1,240
Migrant Status								
Nonmigrant	<10	<10	<10	≥2,750	≥38,670	≥2,300	≥560	≥44,290
Migrant	<10	<10	<10	<10	≥40	<10	<10	≥50
Section 504 Status								
Non-Section 504	<10	<10	<10	≥2,550	≥35,880	≥2,020	≥500	≥40,970
Section 504	<10	<10	<10	≥200	≥2,830	≥280	≥50	≥3,380

Copyright © 2019 by Louisiana Department of Education.

Table 7.8 Reportable Percentage of Students taking the Spring Administration: English II

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	0.00	0.00	0.00	6.21	87.30	5.21	1.27	100.00
Gender								
Female	0.00	0.00	0.00	3.06	44.23	1.91	0.46	49.67
Male	0.00	0.00	0.00	3.15	43.07	3.30	0.80	50.33
Ethnicity								
Hispanic/Latino	0.00	0.00	0.00	0.55	4.38	0.73	0.27	5.94
American Indian or Alaska Native	0.00	0.00	0.00	0.03	0.62	0.04	0.00	0.70
Asian	0.00	0.00	0.00	0.28	1.41	0.08	0.02	1.79
Black or African American	0.00	0.00	0.00	2.75	37.87	2.89	0.77	44.30
Native Hawaiian or Other Pacific	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.07
White	0.00	0.00	0.00	2.45	41.64	1.39	0.19	45.67
Two or More Races	0.00	0.00	0.00	0.15	1.31	0.07	0.01	1.53
Education Classification								
Regular	0.00	0.00	0.00	5.06	75.79	3.92	0.93	85.71
Special	0.00	0.00	0.00	0.47	6.49	1.24	0.33	8.54
Gifted	0.00	0.00	0.00	0.68	5.02	0.05	0.01	5.76
Economic Status								
Economically Disadvantaged	0.00	0.00	0.00	3.60	49.74	3.74	0.68	57.75
Not Economically Disadvantaged	0.00	0.00	0.00	1.75	27.84	0.69	0.09	30.38
English Learner Status								
Non-EL	0.00	0.00	0.00	6.05	85.66	4.48	1.00	97.20
EL	0.00	0.00	0.00	0.16	1.64	0.73	0.27	2.80
Migrant Status								
Nonmigrant	0.00	0.00	0.00	6.21	87.20	5.19	1.27	99.87
Migrant	0.00	0.00	0.00	0.00	0.10	0.02	0.00	0.13
Section 504 Status								
Non-Section 504	0.00	0.00	0.00	5.75	80.91	4.57	1.14	92.37
Section 504	0.00	0.00	0.00	0.46	6.39	0.64	0.13	7.63

Copyright © 2019 by Louisiana Department of Education.

Table 7.9 Count of Students taking the Spring Administration: Algebra I

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	<10	≥260	≥7,710	≥34,120	≥5,890	≥1,280	≥300	≥49,590
Gender								
Female	<10	≥130	≥4,310	≥17,030	≥2,490	≥550	≥130	≥24,670
Male	<10	≥130	≥3,400	≥17,080	≥3,390	≥730	≥160	≥24,920
Ethnicity								
Hispanic/Latino	<10	≥10	≥450	≥2,020	≥510	≥130	≥30	≥3,180
American Indian or Alaska Native	<10	<10	≥40	≥210	≥40	<10	<10	≥310
Asian	<10	≥40	≥320	≥390	≥40	≥10	<10	≥840
Black or African American	<10	≥40	≥2,100	≥15,290	≥3,410	≥750	≥190	≥21,800
Native Hawaiian or Other Pacific	<10	<10	<10	≥20	<10	<10	<10	≥30
White	<10	≥150	≥4,630	≥15,590	≥1,780	≥360	≥50	≥22,580
Two or More Races	<10	<10	≥140	≥570	≥70	≥10	<10	≥810
Education Classification								
Regular	<10	≥130	≥5,900	≥30,140	≥4,670	≥910	≥220	≥41,980
Special	<10	<10	≥130	≥2,660	≥1,160	≥360	≥70	≥4,400
Gifted	<10	≥120	≥1,680	≥1,320	≥50	≥10	<10	≥3,210
Economic Status								
Economically Disadvantaged	<10	≥50	≥2,840	≥20,710	≥4,550	≥930	≥160	≥29,260
Not Economically Disadvantaged	<10	≥180	≥3,610	≥9,350	≥910	≥190	≥20	≥14,270
English Learner Status								
Non-EL	<10	≥260	≥7,680	≥33,380	≥5,440	≥1,160	≥270	≥48,220
EL	<10	<10	≥30	≥730	≥440	≥120	≥20	≥1,360
Migrant Status								
Nonmigrant	<10	≥260	≥7,710	≥34,080	≥5,870	≥1,280	≥290	≥49,530
Migrant	<10	<10	<10	≥30	≥10	<10	<10	≥60
Section 504 Status								
Non-Section 504	<10	≥260	≥7,460	≥31,270	≥5,120	≥1,130	≥260	≥45,530
Section 504	<10	<10	≥250	≥2,850	≥760	≥150	≥30	≥4,060

Copyright © 2019 by Louisiana Department of Education.

Table 7.10 Reportable Percentage of Students taking the Spring Administration: Algebra I

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	0.02	0.54	15.56	68.80	11.88	2.60	0.60	100.00
Gender								
Female	0.01	0.27	8.69	34.35	5.04	1.12	0.28	49.75
Male	0.01	0.27	6.87	34.45	6.84	1.48	0.33	50.25
Ethnicity								
Hispanic/Latino	0.00	0.03	0.92	4.09	1.04	0.28	0.07	6.43
American Indian or Alaska Native	0.00	0.00	0.09	0.44	0.09	0.02	0.01	0.64
Asian	0.02	0.09	0.66	0.80	0.09	0.02	0.01	1.70
Black or African American	0.00	0.10	4.25	30.83	6.88	1.52	0.40	43.97
Native Hawaiian or Other Pacific	0.00	0.00	0.01	0.04	0.01	0.00	0.00	0.07
White	0.00	0.31	9.35	31.44	3.60	0.73	0.11	45.54
Two or More Races	0.00	0.01	0.29	1.16	0.16	0.03	0.01	1.64
Education Classification								
Regular	0.01	0.27	11.90	60.77	9.42	1.84	0.45	84.65
Special	0.00	0.01	0.27	5.37	2.36	0.73	0.15	8.87
Gifted	0.01	0.26	3.39	2.67	0.11	0.03	0.00	6.47
Economic Status								
Economically Disadvantaged	0.00	0.10	5.73	41.77	9.19	1.88	0.34	59.00
Not Economically Disadvantaged	0.00	0.37	7.28	18.85	1.84	0.39	0.05	28.79
English Learner Status								
Non-EL	0.02	0.54	15.49	67.32	10.98	2.35	0.55	97.24
EL	0.00	0.00	0.07	1.48	0.90	0.25	0.06	2.76
Migrant Status								
Nonmigrant	0.02	0.54	15.55	68.73	11.85	2.59	0.60	99.88
Migrant	0.00	0.00	0.01	0.07	0.03	0.01	0.00	0.12
Section 504 Status								
Non-Section 504	0.02	0.53	15.04	63.05	10.33	2.29	0.54	91.81
Section 504	0.00	0.01	0.51	5.75	1.55	0.31	0.07	8.19

Copyright © 2019 by Louisiana Department of Education.

Table 7.11 Count of Students taking the Spring Administration: Geometry

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	<10	≥10	≥210	≥6,370	≥25,740	≥4,690	≥610	≥37,650
Gender								
Female	<10	<10	≥100	≥3,400	≥13,560	≥2,290	≥290	≥19,680
Male	<10	<10	≥100	≥2,960	≥12,180	≥2,390	≥310	≥17,970
Ethnicity								
Hispanic/Latino	<10	<10	<10	≥290	≥1,360	≥390	≥80	≥2,130
American Indian or Alaska Native	<10	<10	<10	≥30	≥130	≥10	<10	≥190
Asian	<10	<10	≥20	≥280	≥360	≥40	<10	≥720
Black or African American	<10	<10	≥40	≥1,880	≥11,070	≥2,880	≥380	≥16,280
Native Hawaiian or Other Pacific	<10	<10	<10	<10	≥10	<10	<10	≥10
White	<10	<10	≥130	≥3,750	≥12,380	≥1,300	≥120	≥17,710
Two or More Races	<10	<10	<10	≥110	≥400	≥50	<10	≥580
Education Classification								
Regular	<10	<10	≥120	≥5,070	≥23,260	≥4,110	≥510	≥33,080
Special	<10	<10	<10	≥100	≥1,330	≥420	≥90	≥1,950
Gifted	<10	<10	≥90	≥1,190	≥1,150	≥160	<10	≥2,610
Economic Status								
Economically Disadvantaged	<10	<10	≥40	≥2,570	≥14,210	≥3,300	≥400	≥20,540
Not Economically Disadvantaged	<10	≥10	≥140	≥2,960	≥8,370	≥1,010	≥90	≥12,600
English Learner Status								
Non-EL	<10	≥10	≥210	≥6,330	≥25,360	≥4,440	≥540	≥36,910
EL	<10	<10	<10	≥40	≥380	≥250	≥60	≥740
Migrant Status								
Nonmigrant	<10	≥10	≥210	≥6,370	≥25,720	≥4,690	≥610	≥37,620
Migrant	<10	<10	<10	<10	≥20	<10	<10	≥30
Section 504 Status								
Non-Section 504	<10	≥10	≥200	≥6,140	≥24,070	≥4,320	≥560	≥35,320
Section 504	<10	<10	≥10	≥220	≥1,670	≥370	≥50	≥2,330

Copyright © 2019 by Louisiana Department of Education.

Table 7.12 Percentage of Students taking the Spring Administration: Geometry

Group	Grade							
	6	7	8	9	10	11	12	Total
All Students	0.00	0.04	0.57	16.92	68.37	12.48	1.63	100.00
Gender								
Female	0.00	0.01	0.28	9.05	36.02	6.10	0.79	52.27
Male	0.00	0.02	0.28	7.87	32.35	6.37	0.83	47.73
Ethnicity								
Hispanic/Latino	0.00	0.00	0.02	0.77	3.62	1.04	0.22	5.66
American Indian or Alaska Native	0.00	0.00	0.00	0.10	0.36	0.04	0.01	0.52
Asian	0.00	0.02	0.06	0.75	0.96	0.11	0.01	1.92
Black or African American	0.00	0.00	0.12	5.01	29.42	7.66	1.03	43.25
Native Hawaiian or Other Pacific	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.05
White	0.00	0.02	0.35	9.97	32.90	3.48	0.34	47.05
Two or More Races	0.00	0.00	0.01	0.31	1.07	0.14	0.02	1.56
Education Classification								
Regular	0.00	0.01	0.32	13.47	61.79	10.92	1.35	87.86
Special	0.00	0.00	0.00	0.28	3.53	1.12	0.25	5.19
Gifted	0.00	0.02	0.25	3.17	3.05	0.43	0.02	6.95
Economic Status								
Economically Disadvantaged	0.00	0.00	0.11	6.84	37.75	8.78	1.08	54.56
Not Economically Disadvantaged	0.00	0.04	0.39	7.87	22.23	2.70	0.24	33.46
English Learner Status								
Non-EL	0.00	0.04	0.57	16.81	67.35	11.81	1.44	98.02
EL	0.00	0.00	0.00	0.11	1.02	0.67	0.18	1.98
Migrant Status								
Nonmigrant	0.00	0.04	0.57	16.92	68.30	12.46	1.63	99.92
Migrant	0.00	0.00	0.00	0.00	0.07	0.01	0.00	0.08
Section 504 Status								
Non-Section 504	0.00	0.03	0.53	16.32	63.92	11.49	1.50	93.79
Section 504	0.00	0.01	0.03	0.60	4.45	0.99	0.13	6.21

Tables 7.13 through 7.16 summarize the mean scale score and standard deviation as well as the percentage of students in each achievement level based on the state population for the 2017-2018 administration of the LEAP 2025 high school ELA and mathematics assessments. All three administrations are presented. Previous year's results are not shown because 2017-2018 was the first school year of the LEAP 2025 high school ELA and mathematics assessments were administered.

Copyright © 2019 by Louisiana Department of Education.

Table 7.13 Comparison of Percentage of Students in Each Achievement Level: English I

				Scale Score		Achievement Level				
	Year	Administration	N	Mean	SD	1	2	3	4	5
All	2017	Fall	≥4,610	744.58	31.35	8.7	16.8	29.3	37.8	7.5
	2018	Spring	≥46,130	744.04	34.41	10.1	17.7	27.7	36.4	8.1
	2018	Summer	≥1,390	696.12	23.18	58.5	29.4	10.4	1.7	0.0
First-Time Testers	2017	Fall	≥4,610	744.58	31.35	8.7	16.8	29.3	37.8	7.5
	2018	Spring	≥45,850	744.27	34.32	9.9	17.6	27.8	36.6	8.1
	2018	Summer	≥60	713.22	33.15	35.0	31.7	13.3	20.0	0.0
Retesters	2017	Fall	<10	NR	NR	NR	NR	NR	NR	NR
	2018	Spring	≥170	694.27	19.98	59.4	34.9	5.7	0.0	0.0
	2018	Summer	≥1,290	694.59	21.87	60.9	29.1	9.4	0.7	0.0
Previously Passed	2017	Fall	<10	NR	NR	NR	NR	NR	NR	NR
	2018	Spring	≥90	724.99	25.46	13.4	36.1	39.2	10.3	1.0
	2018	Summer	≥40	715.63	25.27	24.5	34.7	34.7	6.1	0.0

Levels: 1 = Unsatisfactory, 2 = Approaching Basic, 3 = Basic, 4=Mastery, 5 = Advanced

Table 7.14 Comparison of Percentage of Students in Each Achievement Level: English II

				Scale Score		Achievement Level				
	Year	Administration	N	Mean	SD	1	2	3	4	5
All	2017	Fall	≥9,130	723.96	44.60	34.7	19.2	17.0	21.8	7.2
	2018	Spring	≥44,350	743.63	43.21	16.2	17.1	22.4	32.3	12.0
	2018	Summer	≥2,290	688.67	27.62	64.3	23.9	9.9	1.8	0.0
First-Time Testers	2017	Fall	≥6,010	743.02	40.74	15.2	17.7	23.6	32.5	10.9
	2018	Spring	≥41,570	746.95	41.87	13.3	16.5	23.2	34.1	12.8
	2018	Summer	≥70	706.04	41.93	45.2	19.2	19.2	15.1	1.4
Retesters	2017	Fall	≥2,840	684.36	21.15	76.5	21.0	2.3	0.3	0.0
	2018	Spring	≥2,110	686.42	25.60	69.7	23.0	6.3	0.9	0.0
	2018	Summer	≥2,190	687.86	26.75	65.3	23.9	9.5	1.3	0.0
Previously Passed	2017	Fall	≥260	717.33	29.84	29.1	33.6	25.4	10.8	1.1
	2018	Spring	≥650	717.85	33.00	29.4	32.6	22.7	13.7	1.5
	2018	Summer	≥30	703.38	29.30	44.1	35.3	11.8	8.8	0.0

Levels: 1 = Unsatisfactory, 2 = Approaching Basic, 3=Basic, 4=Mastery, 5 = Advanced

Copyright © 2019 by Louisiana Department of Education.

Table 7.15 Comparison of Percentage of Students in Each Achievement Level: Algebra I

				Scale Score		Achievement Level				
	Year	Administration	N	Mean	SD	1	2	3	4	5
All	2017	Fall	≥8,340	719.51	28.81	20.3	47.2	17.6	14.2	0.7
	2018	Spring	≥49,590	741.14	33.91	9.2	25.1	27.7	33.8	4.2
	2018	Summer	≥1,570	707.31	18.87	28.3	57.6	12.6	1.4	0.1
First-Time Testers	2017	Fall	≥2,930	742.75	30.41	6.5	24.3	29.4	37.8	2.0
	2018	Spring	≥45,840	743.26	33.55	7.8	23.6	28.3	35.8	4.5
	2018	Summer	≥70	720.75	30.46	19.0	45.6	20.3	13.9	1.3
Retesters	2017	Fall	≥4,980	705.55	16.83	29.1	61.0	9.3	0.5	0.0
	2018	Spring	≥1,750	704.16	20.84	38.3	47.9	11.8	2.0	0.0
	2018	Summer	≥1,390	706.34	17.78	29.1	58.3	11.9	0.7	0.0
Previously Passed	2017	Fall	≥430	722.52	23.85	12.7	43.8	31.8	11.8	0.0
	2018	Spring	≥1,990	725.10	27.94	14.7	39.8	28.5	15.8	1.2
	2018	Summer	≥100	710.31	17.54	23.8	58.4	16.8	1.0	0.0

Levels: 1 = Unsatisfactory, 2 = Approaching Basic, 3=Basic, 4=Mastery, 5 = Advanced

Table 7.16 Comparison of Percentage of Students in Each Achievement Level: Geometry

				Scale Score		Achievement Level				
	Year	Administration	N	Mean	SD	1	2	3	4	5
All	2017	Fall	≥6,720	730.62	27.50	11.2	35.5	28.8	21.1	3.5
	2018	Spring	≥37,650	735.74	25.87	7.5	28.8	33.9	26.5	3.3
	2018	Summer	≥390	708.56	25.73	34.1	56.6	3.8	1.0	4.5
First-Time Testers	2017	Fall	≥5,440	736.91	25.88	5.5	30.5	33.8	25.9	4.3
	2018	Spring	≥36,560	736.45	25.66	6.9	28.2	34.4	27.2	3.4
	2018	Summer	≥70	735.74	45.41	18.6	42.9	8.6	4.3	25.7
Retesters	2017	Fall	≥1,160	702.28	14.20	38.0	57.3	4.5	0.2	0.0
	2018	Spring	≥710	706.56	17.05	34.7	52.4	11.9	1.0	0.0
	2018	Summer	≥310	702.46	13.03	37.5	59.9	2.2	0.3	0.0
Previously Passed	2017	Fall	≥110	720.45	18.93	12.0	49.6	34.2	4.3	0.0
	2018	Spring	≥370	721.80	23.38	17.9	39.2	32.0	9.9	1.1
	2018	Summer	≥10	708.71	15.75	35.7	50.0	14.3	0.0	0.0

Levels: 1 = Unsatisfactory, 2 = Approaching Basic, 3=Basic, 4=Mastery, 5 = Advanced

7.2 Reports

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:

Copyright © 2019 by Louisiana Department of Education.

> 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 related to the test scores are disseminated in two ways: (1) the individual score report and (2) the *LEAP Interpretive Guide* .

In addition to providing interpretation, it is important that the information is understandable by 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)

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 relevant information to understand their child's test score. For more information about the test, parents are provided A Parent Guide to the LEAP 2025 Student Reports. In the 2017-2018 administration year, student reports for each school were posted by subject, then downloaded and printed from eDIRECT by the school system and school. eDIRECT is DRC's secure online system that provides schools and districts access to manage student tests and reports.

7.3 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 individual student achievement and groups of students. Standard 12.18 states:

Copyright © 2019 by Louisiana Department of Education.

> 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 high school ELA and mathematics assessments, is available to school systems and schools through eDIRECT. Total test scores and achievement-level indicators are shown for the test of interest. Claim, subclaim, and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by claim, subclaim, and subcategory are summarized. More details can be found in the LEAP 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 the parents. At the top of the page, overall student performance is reported by scale scores and achievement level. In the middle of the page, claim, subclaim, and subcategory performance indicators are reported. When a student does not receive a scale score, his or her achievement level will be left blank. ISRs for students whose scores were invalidated display a blank scale score for a given course.

A data file referred to as Louisiana Department of Education Student File (LDESTD) was provided to 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 Interpretive Guide was written for Louisiana school system and school administrators, teachers, parents, and the general public to better understand the LEAP 2025 ELA and Mathematics tests. The *LEAP 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 Interpretive Guidehas 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.

7.4 Summary

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)

Copyright © 2019 by Louisiana Department of Education.

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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 8: Performance-Level Setting

This chapter briefly describes LEAP High School 2025 performance-level setting and presents the cut scores established and the achievement-level descriptors derived from the performance-level setting. Since the LDOE used PARCC cut scores for the LEAP 2025 High school 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 2017-2018 administrations, LEAP 2025 High School assessments measuring different content and constructs than 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 the new scales, and the 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 to which the LEAP 2025 High School assessments were aligned.

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, educator judgment about what students should know and be able to demonstrate, and research 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 LEAP 2025 High School assessments. Tables 8.1 and 8.2 summarize the cut scores.

Copyright © 2019 by Louisiana Department of Education.

Table 8.1 LEAP 2025 Highs School Cut Scores

Content	Cut Scores			
	Approaching Basic	Basic	Mastery	Advanced
English I	700	725	750	791
English II	700	725	750	794
Algebra I	700	725	750	805
Geometry	700	725	750	783

8.3 Claim Cut Scores

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

8.4 Achievement-Level Definitions

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

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

Table 8.2 summarizes the LEAP 2025 High School scale-score ranges for each level of achievement.

Copyright © 2019 by Louisiana Department of Education.

Table 8.2 Achievement-Level Scale-Score Ranges

Achievement Level	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English I	650–699	700–724	725–749	750–790	791–850
English II	650–699	700–724	725–749	750–793	794–850
Algebra I	650–699	700–724	725–749	750–804	805–850
Geometry	650–699	700–724	725–749	750–782	783–850

8.5 Summary

This chapter presented a brief overview of PARCC's performance-level setting process, which set the cut scores used by LDOE for reporting student performance on the LEAP 2025 High School 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)

Copyright © 2019 by Louisiana Department of Education.

Chapter 9: Evidence of Construct-Related Validity

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

Table 9.1 Evidence of Validity and the Corresponding Technical Report Chapter

Source of Validity	Related Information	Related Chapter/Source
Evidence Based on Test Content	Item Development Process	Chapter 3 2017-2018 LEAP 2025 High School ELA and Mathematics Assessment Frameworks
	Test Blueprint and Item Alignment to Curriculum and Standards	Chapter 3 2017-2018 LEAP 2025 High School 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	2017-2018 LEAP 2025 High School 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
Evidence Based on the Consequences of Testing	Scale Score and Performance Level Information	Chapter 7
	Test Interpretive Guide	Chapter 4

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

Copyright © 2019 by Louisiana Department of Education.

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 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

Copyright © 2019 by Louisiana Department of Education.

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 9.3 of this chapter. The subscore reliabilities and SEMs are presented in Sections 9.11 and 9.4. The SEM of the total score is discussed in Section 9.4.

> *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 9.3 and is reported in raw score units. The conditional SEM is discussed in Section 9.4 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 9.3 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.

9.3 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

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

Copyright © 2019 by Louisiana Department of Education.

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; a longer test can be expected to be more reliable than a shorter test.

The reliability coefficients for the LEAP 2025 HS assessments are reported in Table 9.2. English I and English II have one writing component (RI or RL) that is the same score of another component (WE). The item score for the RI/RL writing component was excluded from the reliability computation. The reliability statistics ranged from 0.87 to 0.91 and from 0.87 to 0.91, for the fall and spring administration, respectively. The two administrations had very similar reliability. These results indicate acceptable reliability coefficients for the LEAP 2025 HS tests.

Table 9.2 Reliability

Administration	Course	Number of Items	Number of Score Points	SEM	Cronbach's Alpha	N-Count
Fall	English I	33	90	5.57	0.88	≥4,630
	English II	33	90	5.45	0.90	≥6,020
	Algebra I	38	67	4.07	0.87	≥2,920
	Geometry	38	68	3.85	0.91	≥5,430
Spring	English I	34	94	5.60	0.90	≥46,180
	English II	34	94	5.29	0.90	≥41,880
	Algebra I	39	68	3.75	0.87	≥46,010
	Geometry	38	68	3.73	0.91	≥36,800

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

9.4 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:

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

Copyright © 2019 by Louisiana Department of Education.

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. With English I and English II, the raw score used to calculate the SD included the RI/RL component and weighting of WE.

9.5 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):

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

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. The CSEMs at the four cut scores that define the performance levels are presented in Table 9.3.

Copyright © 2019 by Louisiana Department of Education.

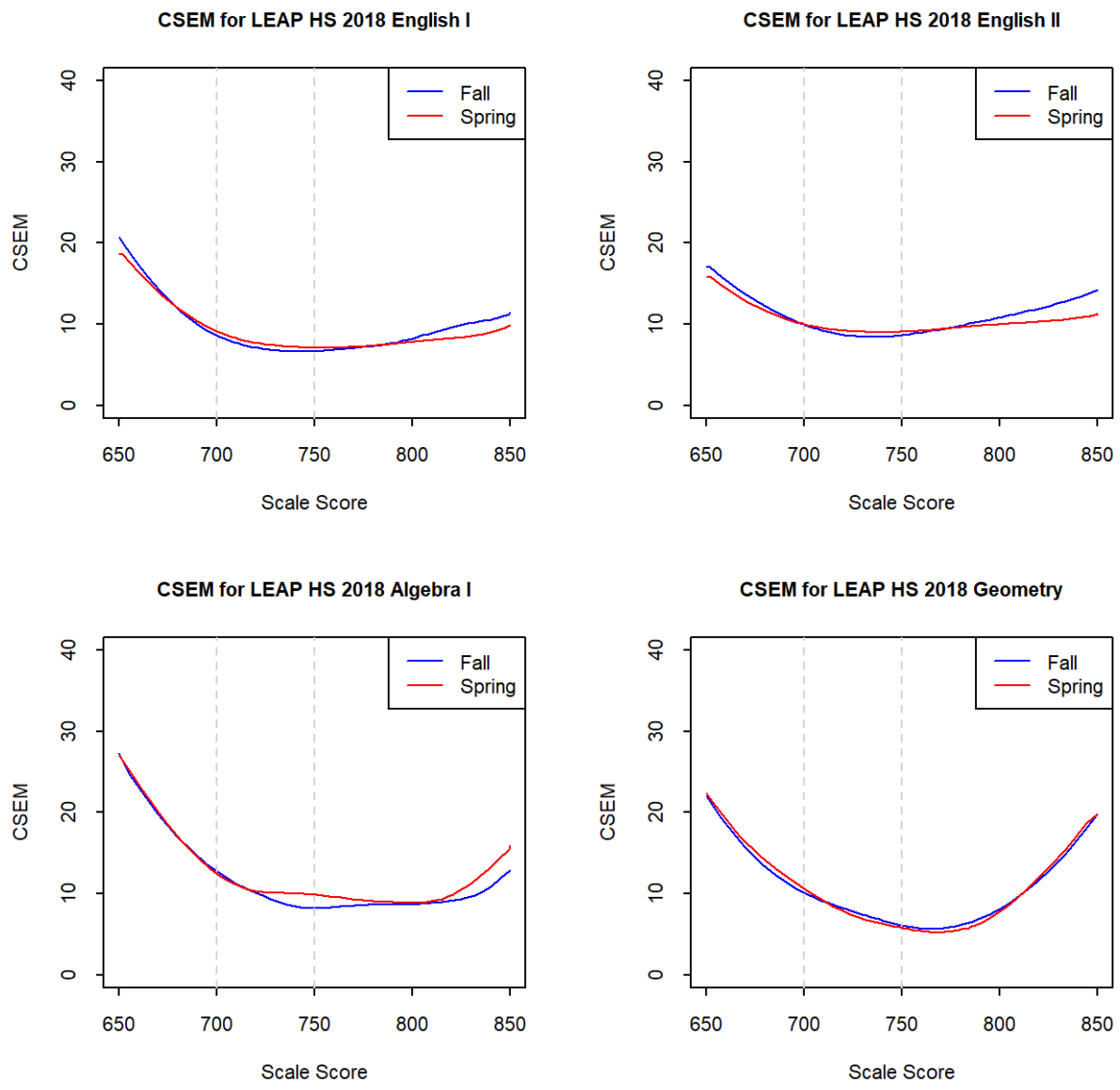
Table 9.3 Conditional Standard Errors of Measurement at the *Approaching Basic*, *Basic*, *Mastery*, and *Advanced* Cut Scores

		Approaching Basic		*Basic*		*Mastery*		*Advanced*	
Administration	**Course**	**Cut Score**	**CSEM**	**Cut Score**	**CSEM**	**Cut Score**	**CSEM**	**Cut Score**	**CSEM**
Fall	English I	700	9	725	7	750	7	791	8
	English II	700	10	725	9	750	9	794	11
	Algebra I	700	13	725	10	750	8	805	9
	Geometry	700	10	725	8	750	6	783	6
Spring	English I	700	9	725	8	750	7	791	8
	English II	700	10	725	9	750	9	794	10
	Algebra I	700	12	725	10	750	10	805	9
	Geometry	700	11	725	7	750	6	783	6

Figure 9.1 displays the CSEM curves for each subject area. The estimates of measurement error tend to be higher at the low and high ends of the scale score range. The measurement error increases when there are few observations at a particular ability level. 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. Figure 9.1 demonstrates that the tests are designed so that measurement error is minimized in the middle of the scale range, where the majority of students are located.

Copyright © 2019 by Louisiana Department of Education.

Figure 9.1 CSEM Curves



Copyright © 2019 by Louisiana Department of Education.

9.6 Classification Accuracy and Consistency

Classification Accuracy

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

Classification Consistency

Decision consistency, or classification consistency, is defined as the extent to which the classifications of students in a particular achievement level agree on the basis of two independent administrations of the test 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 decision consistency from one administration of a test.

The Livingston-Lewis (1995) methodology was used to calculate decision accuracy statistics based on the spring 2018 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 decision accuracy and consistency are based on census data.

Classification consistency and classification accuracy conditioned on performance 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 2018 LEAP 2025 high school ELA and mathematics assessments 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 0.00 to 0.86. Classification consistency (see Table 9.5) conditioned on achievement level ranged from 0.24 to 0.78. For some Geometry tests, classification accuracy and consistency conditioned on the *Unsatisfactory* level were very low. A possible reason for these relatively low *Unsatisfactory* level values is the fact that there were not enough easy items to distinguish the *Unsatisfactory* level from other performance levels.

Copyright © 2019 by Louisiana Department of Education.

Table 9.4 Classification Accuracy Conditioned on Level of Achievement

Classification Accuracy						
Administration	**Course**	**Unsatisfactory**	**Approaching Basic**	**Basic**	**Mastery**	**Advanced**
Fall	English I	0.67	0.64	0.73	0.81	0.55
	English II	0.74	0.67	0.63	0.78	0.70
	Algebra I	0.35	0.66	0.70	0.86	0.31
	Geometry	0.00	0.86	0.69	0.84	0.69
Spring	English I	0.69	0.68	0.70	0.81	0.67
	English II	0.73	0.63	0.65	0.77	0.71
	Algebra I	0.13	0.74	0.58	0.81	0.70
	Geometry	0.00	0.82	0.74	0.84	0.67

Table 9.5 Classification Consistency Conditioned on Level of Achievement

Classification Consistency						
Administration	**Course**	**Unsatisfactory**	**Approaching Basic**	**Basic**	**Mastery**	**Advanced**
Fall	English I	0.63	0.52	0.59	0.72	0.53
	English II	0.72	0.50	0.51	0.68	0.67
	Algebra I	0.40	0.52	0.56	0.77	0.37
	Geometry	0.24	0.63	0.61	0.77	0.68
Spring	English I	0.67	0.54	0.56	0.72	0.63
	English II	0.71	0.48	0.49	0.67	0.68
	Algebra I	0.32	0.50	0.46	0.74	0.66
	Geometry	0.31	0.58	0.63	0.78	0.66

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 performance levels is collapsed into a dichotomized distribution around that specific cut point. As an example, the 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 High School cut scores. Table 9.6 shows that classification accuracy at achievement cut points ranged from 0.87 to 0.98. Classification consistency (see Table 9.7) conditioned at achievement cut points ranged from 0.82 to 0.98. Classification consistency and accuracy at achievement cut points tend to be higher values than those conditioned on performance level.

Copyright © 2019 by Louisiana Department of Education.

The classification accuracy statistics are at or above 0.87, while the classification consistency statistics are at or above 0.82. These results suggest that consistent and accurate performance-level classifications are being made for students in Louisiana based on the LEAP 2025 High School assessments.

Table 9.6 Classification Accuracy at Achievement Cut Points

Classification Accuracy					
Administration	**Course**	***Unsatisfactory/ Approaching Basic***	***Approaching Basic/ Basic***	***Basic/ Mastery***	***Mastery/ Advanced***
Fall	English I	0.96	0.92	0.90	0.95
	English II	0.94	0.92	0.90	0.95
	Algebra I	0.94	0.89	0.90	0.98
	Geometry	0.95	0.89	0.93	0.98
Spring	English I	0.96	0.92	0.90	0.96
	English II	0.95	0.92	0.90	0.94
	Algebra I	0.92	0.87	0.89	0.98
	Geometry	0.93	0.89	0.93	0.98

Table 9.7 Classification Consistency at Achievement Cut Points

Classification Consistency					
Administration	**Course**	***Unsatisfactory/ Approaching Basic***	***Approaching Basic/ Basic***	***Basic/ Mastery***	***Mastery/ Advanced***
Fall	English I	0.94	0.89	0.86	0.93
	English II	0.92	0.88	0.87	0.93
	Algebra I	0.91	0.85	0.86	0.97
	Geometry	0.91	0.84	0.90	0.97
Spring	English I	0.94	0.88	0.86	0.94
	English II	0.92	0.88	0.86	0.92
	Algebra I	0.89	0.82	0.85	0.97
	Geometry	0.90	0.85	0.90	0.98

9.7 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 geometry test is designed to measure a single overall construct—geometry achievement; therefore, the items comprising the LEAP 2025 Geometry test should measure only geometry, not language or reading.

Copyright © 2019 by Louisiana Department of Education.

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). The reliability shown in Table 9.2 are all above 0.80, indicating items on the 2018 LEAP 2025 High School assessments measure the same construct or the same content domain. In order 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.

9.8 Principal Components Analysis

As another measure of construct validity, DRC examined the unidimensionality of each subject-level LEAP 2025 test. One of the underlying assumptions of the IRT models used to scale LEAP 2025 is that the tests being calibrated are unidimensional; that is, items comprising LEAP 2025 in each subject area measure a single content domain. For example, Algebra I items should measure algebra 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 LEAP 2025 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., measures one underlying construct).

A PCA was conducted for each subject of LEAP 2025. 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 the subjects of the LEAP 2025, the ratio of the variance accounted for by the first factor to the second is sufficiently large, indicating that the unidimensionality assumption holds. All the LEAP 2025 High School tests exhibit first principal components accounting for more than 20% of the test variance (Table 9.8 through Table 9.11), except for the algebra I Spring 2018 administration. To further investigate the unidimensionality of the assessments, the ratio of the first eigenvalue to the second 1111eigenvalue was explored and is summarized in the row below the second component in each table. These ratios show that the first eigenvalue is at least four times as large as the second eigenvalue for assessments. This substantial difference in magnitude indicates that one factor appears to be dominant and that the LEAP 2025 High School 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.

Copyright © 2019 by Louisiana Department of Education.

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: English I

Administration	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Fall	First Component	7.63	23.85	23.85
Fall	Second Component	1.19	3.72	27.57
Fall	Ratio (First/Second)	6.41	—	—
Spring	First Component	8.07	25.23	25.23
Spring	Second Component	1.49	4.65	29.88
Spring	Ratio (First/Second)	5.42	—	—

Table 9.9 Principal Component Analysis: English II

Administration	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Fall	First Component	8.01	25.04	25.04
Fall	Second Component	1.37	4.28	29.32
Fall	Ratio (First/Second)	5.85	—	—
Spring	First Component	8.01	25.04	25.04
Spring	Second Component	1.56	4.89	29.93
Spring	Ratio (First/Second)	5.12	—	—

Table 9.10 Principal Component Analysis: Algebra I

Administration	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Fall	First Component	7.51	19.76	19.76
Fall	Second Component	1.40	3.70	23.46
Fall	Ratio (First/Second)	5.35	—	—
Spring	First Component	7.08	18.15	18.15
Spring	Second Component	1.43	3.66	21.81
Spring	Ratio (First/Second)	4.95	—	—

Table 9.11 Principal Component Analysis: Geometry

Administration	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Fall	First Component	9.75	25.67	25.67
Fall	Second Component	1.56	4.10	29.77

Copyright © 2019 by Louisiana Department of Education.

Administration	Components	Eigenvalue	Percentage of Variance Explained	Cumulative Percentage of Variance Explained
Fall	Ratio (First/Second)	6.26	—	—
Spring	First Component	10.06	26.48	26.48
Spring	Second Component	1.50	3.95	30.44
Spring	Ratio (First/Second)	6.70	—	—

9.9 Analyses by Claims, Subclaims, and Subcategories

Three sets of analyses were conducted at the claim and subclaim level for ELA content and at the subclaim and subcategory for mathematics content in another attempt to assess the construct validity of LEAP 2025. First, correlation coefficients that measure the relationship between the claim scores and subclaim scores were computed. Second, the reliability of each claim and subclaim was computed. Finally, the SEM was computed for each reportable claim and subclaim.

9.10 Correlations among Claims, Subclaims, and Subcategories

This section reports the strength of the interrelationships among the claims, subclaims, or subcategories by computing the correlation between them. Table 9.12 through Table 9.17 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 claims, subclaims, 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 Table 9.18 through Table 9.25.

The uncorrected PPM in Table 9.12 through Table 9.17 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 claim or subclaim is moderately to strongly related to performance on another claim or subclaim within the same grade and content area. The value of the correlation coefficients will be affected by the limited number of items measuring each claim or subclaim. Therefore, caution should be used when comparing the PPM coefficients measuring the relationships between claims, subclaims, or subcategories to those measuring the relationships between content areas. A more modest relationship (i.e., smaller correlation coefficients) is expected to be reported between the claims, subclaims, or subcategories as a consequence of the lower number of items measuring each of the reporting categories. The PPM between two claims subscores, for example, 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

Copyright © 2019 by Louisiana Department of Education.

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 for the other claim or GLE strand.

English shows moderate relationships between reading and writing claims, indicating these two claims measure some different traits. Across all tables, the CAPPM indicates moderate or strong relationships between the subclaims. The CAPPM for reading vocabulary, written expression, and knowledge and use of language were 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 Table 9.12 through Table 9.17 are further evidence of the validity of the test construct. Since the overall content area is comprised of the claim or subclaim subscores and the content area is expected to measure a single dimension, these subscores are expected to be moderately or highly related.

Table 9.12 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Claims: English I

Administration	No.	Claim	N Items	1	2
Fall	1	Reading	29	—	0.86
	2	Writing	4	0.75	—
Spring	1	Reading	30	—	0.81
	2	Writing	4	0.73	—

Copyright © 2019 by Louisiana Department of Education.

Table 9.13 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Subclaims: English I

Subclaim Uncorrected and Corrected Correlation Coefficients: English I								
Administration	**No.**	**Subclaim**	**N Items**	**1**	**2**	**3**	**4**	**5**
Fall	1	Reading Literary Text	7	—	0.98	0.91	0.91	0.88
	2	Reading Information Text	16	0.68	—	0.96	0.95	0.92
	3	Reading Vocabulary	6	0.43	0.50	—	0.79	0.80
	4	Written Expression	2	0.63	0.73	0.41	—	1.18
	5	Written Knowledge & Use of Language	2	0.62	0.72	0.43	0.92	—
Spring	1	Reading Literary Text	9	—	1.04	1.03	0.92	0.90
	2	Reading Information Text	15	0.74	—	1.02	0.86	0.85
	3	Reading Vocabulary	6	0.57	0.63	—	0.74	0.74
	4	Written Expression	2	0.67	0.70	0.47	—	1.11
	5	Written Knowledge & Use of Language	2	0.66	0.70	0.48	0.95	—

Table 9.14 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Claims: English II

Administration	No.	Claim	N Items	1	2
Fall	1	Reading	29	—	0.85
	2	Writing	4	0.75	—
Spring	1	Reading	30	—	0.80
	2	Writing	4	0.72	—

Table 9.15 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Subclaims: English II

Subclaim Uncorrected and Corrected Correlation Coefficients: English II								
Administration	**No.**	**Subclaim**	**N Items**	**1**	**2**	**3**	**4**	**5**
Fall	1	Reading Literary Text	6	—	0.97	0.94	0.85	0.84
	2	Reading Information Text	16	0.61	—	0.99	0.94	0.91
	3	Reading Vocabulary	7	0.51	0.68	—	0.85	0.86
	4	Written Expression	2	0.53	0.74	0.58	—	1.20
	5	Written Knowledge & Use of Language	2	0.54	0.73	0.59	0.94	—
Spring	1	Reading Literary Text	7	—	1.09	1.04	1.02	1.01
	2	Reading Information Text	14	0.71	—	0.98	0.85	0.84
	3	Reading Vocabulary	9	0.61	0.68	—	0.69	0.70
	4	Written Expression	2	0.70	0.69	0.51	—	1.11
	5	Written Knowledge & Use of Language	2	0.69	0.69	0.52	0.95	—

Copyright © 2019 by Louisiana Department of Education.

Table 9.16 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Subclaims: Algebra I

Administration	No.	Subclaim	N Items	1	2	3	4
Fall	1	Major Content	21	—	1.02	1.03	1.11
	2	Additional & Supporting Con	10	0.63	—	1.11	1.03
	3	Expressing Mathematical Rea	3	0.60	0.52	—	1.14
	4	Modeling & Application	4	0.76	0.57	0.58	—
Spring	1	Major Content	21	—	1.03	0.88	1.00
	2	Additional & Supporting Con	11	0.66	—	0.92	0.94
	3	Expressing Mathematical Rea	3	0.56	0.54	—	0.86
	4	Modeling & Application	4	0.71	0.62	0.56	—

Table 9.17 Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Subclaims: Geometry I

Administration	No.	Subclaim	N Items	1	2	3	4
Fall	1	Major Content	20	—	1.00	0.97	1.04
	2	Additional & Supporting Con	12	0.77	—	0.92	1.02
	3	Expressing Mathematical Rea	3	0.75	0.66	—	1.07
	4	Modeling & Application	3	0.75	0.68	0.72	—
Spring	1	Major Content	20	—	—	1.01	0.99
	2	Additional & Supporting Con	12	0.76	—	1.02	1.01
	3	Expressing Mathematical Rea	3	0.75	0.68	—	1.12
	4	Modeling & Application	3	0.75	0.68	0.74	—

9.11 Reliability of Claims, Subclaims, or Subcategories

Raw score summary statistics (i.e., mean and standard deviation), Cronbach's (1951) coefficient alpha, and SEM were computed for each of the claims, subclaims, or subcategories by subject using the census data. These statistics are presented in Tables 9.18–9.22. Reliability indices, such as Cronbach's coefficient alpha (and resulting SEM), are a function of the number of test items, the average covariance between item-pairs, and the variance of the total score. In general, it is expected that the coefficient alpha would be lower for a claim or subclaim assessed by a small number of items compared to a claim or subclaim assessed by a larger number of items.

9.12 Standard Error of Measurement of Claims, Subclaims, or Subcategories

This chapter also reports the SEM associated with each of the claims, subclaims, and subcategories in Table 9.18 through Table 9.25. These SEMs are reported in the raw score metric.

Copyright © 2019 by Louisiana Department of Education.

Table 9.18 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English I Claims

Administration	Claim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Reading	29	60	27.11	10.73	4.33	0.84
	Writing	4	30	10.95	6.71	1.98	0.91
Spring	Reading	30	64	29.81	12.04	4.47	0.86
	Writing	4	30	9.41	6.55	1.62	0.94

Table 9.19 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English I Subclaims

Administration	Subclaim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Reading Literary Text	7	14	6.58	3.22	1.99	0.62
	Reading Information Text	16	34	15.62	6.77	3.24	0.77
	Reading Vocabulary	6	12	4.91	2.44	1.96	0.36
	Written Expression	2	24	8.26	5.09	2.44	0.77
	Knowledge & Use of Language	2	6	2.69	1.72	0.76	0.80
Spring	Reading Literary Text	9	20	7.50	4.00	2.43	0.63
	Reading Information Text	15	32	15.77	6.78	3.09	0.79
	Reading Vocabulary	6	12	6.54	2.72	1.96	0.48
	Written Expression	2	24	6.97	4.98	1.97	0.84
	Knowledge & Use of Language	2	6	2.44	1.63	0.63	0.85

Table 9.20 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English II Claims

Administration	Claim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Reading	29	60	28.15	10.30	3.89	0.86
	Writing	4	30	11.98	7.99	2.30	0.92
Spring	Reading	30	64	29.20	10.89	4.08	0.86
	Writing	4	30	10.82	6.71	1.59	0.94

Copyright © 2019 by Louisiana Department of Education.

Table 9.21 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of English II Subclaims

Administration	Subclaim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Reading Literary Text	6	12	5.38	2.43	1.71	0.51
	Reading Information Text	16	34	14.12	6.38	2.91	0.79
	Reading Vocabulary	7	14	8.65	2.89	1.85	0.59
	Written Expression	2	24	8.93	6.14	2.87	0.78
	Knowledge & Use of Language	2	6	3.05	1.94	0.87	0.80
Spring	Reading Literary Text	7	16	6.57	3.06	2.06	0.55
	Reading Information Text	14	30	12.16	5.56	2.61	0.78
	Reading Vocabulary	9	18	10.47	3.63	2.22	0.63
	Written Expression	2	24	8.17	5.10	1.95	0.85
	Knowledge & Use of Language	2	6	2.65	1.67	0.62	0.86

Table 9.22 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Algebra I Subclaims

Administration	Subclaim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Major Content	21	28	12.21	5.45	2.56	0.78
	Additional & Supporting Content	10	13	4.41	2.21	1.57	0.50
	Expressing Mathematical Reasoning	3	11	0.85	1.32	1.00	0.43
	Modeling & Application	4	15	4.81	3.78	2.37	0.61
Spring	Major Content	21	28	12.27	4.70	2.58	0.70
	Additional & Supporting Content	11	14	4.73	2.55	1.63	0.59
	Expressing Mathematical Reasoning	3	11	1.04	1.51	0.98	0.58
	Modeling & Application	4	15	3.72	3.16	1.65	0.73

Copyright © 2019 by Louisiana Department of Education.

Table 9.23 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Algebra I Subcategories

Administration	Subcategories	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	A1	8	10	4.80	2.26	1.45	0.59
	A2	5	9	3.53	2.23	1.62	0.47
	A3	4	4	1.91	1.23	0.89	0.48
Spring	A1	5	7	2.90	1.70	1.25	0.46
	A2	6	10	5.22	2.06	1.74	0.29
	A3	6	6	2.62	1.40	1.09	0.40

Table 9.24 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Geometry Subclaims

Administration	Subclaim	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	Major Content	20	26	11.10	5.59	2.31	0.83
	Additional & Supporting Content	12	16	7.38	3.00	1.62	0.71
	Expressing Mathematical Reasoning	3	11	1.69	2.40	1.26	0.72
	Modeling & Application	3	15	2.59	3.22	1.95	0.63
Spring	Major Content	20	26	11.27	5.58	2.19	0.85
	Additional & Supporting Content	12	16	6.24	2.79	1.60	0.67
	Expressing Mathematical Reasoning	3	11	1.87	2.29	1.35	0.65
	Modeling & Application	3	15	2.47	3.19	1.81	0.68

Table 9.25 Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Geometry Subcategories

Administration	Subcategories	Number of Items	Number of Score Points	Mean Raw Score	Raw Score Std. Dev.	SEM	Cronbach's Alpha
Fall	A1	12	16	6.64	3.67	1.82	0.75
	A2	8	10	4.46	2.36	1.43	0.64
Spring	A1	13	17	7.72	3.71	1.76	0.77
	A2	7	9	3.55	2.26	1.30	0.67

Copyright © 2019 by Louisiana Department of Education.

9.13 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 High School assessments, correlations were computed between the English I, English II, Algebra I and Geometry total scores for students who took more than one subject test in 2018. These correlations are based on the census data, and the results are shown in Table 9.26 and Table 9.27 for fall 2017 and spring 2018 administrations. The correlation coefficients ranged from 0.60 to 0.85. The lowest correlation was observed between English II and algebra I, and the highest correlation was between English I and English II and also between algebra I and geometry in fall 2017 administration. Similar patterns were observed in fall 2017 and spring 2018 administrations. 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 but still measure some different traits as planned.

Table 9.26 Inter-Correlation of HS Contents Scale Scores with Fall Administration

	English I	English II	Algebra I	Geometry
English I	—	0.85	0.74	0.72
English II	0.85 (33)*	—	0.60	0.71
Algebra I	0.74(849)	0.60 (206)	—	0.85
Geometry	0.72 (406)	0.71 (1,432)	0.85 (34)	—

*The count of observations in the analysis is in parenthesis

Table 9.27 Inter-Correlation of HS Contents Scale Scores with Spring Administration

	English I	English II	Algebra I	Geometry
English I	—	0.84	0.70	0.73
English II	0.84 (745)*	—	0.65	0.69
Algebra I	0.70 (31,708)	0.65 (4,251)	—	0.80
Geometry	0.73 (5,227)	0.69 (23,985)	0.80 (350)	—

*The count of observations in the analysis is in parenthesis

9.14 Summary

In summary, the overall purpose of establishing construct validity is to ensure that the meaning of test scores is supported. Evidence of validity is necessary to justify the use of the LEAP 2025 test scores. This

Copyright © 2019 by Louisiana Department of Education.

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)

Copyright © 2019 by Louisiana Department of Education.

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 differences in test performance among subgroups does not mean that a test is unfair—it simply means that groups perform differently on the 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 is particularly relevant 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 differing in meaning is 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, LDOE assessment research and development experts, along with 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 conducted differential item functioning (DIF) studies of their items prior to PARCC operational administrations. Items are typically evaluated for possible DIF in the field test phase of the test development, and any items flagged for DIF are further examined to determine possible bias. During the ELA and mathematics test development, DRC content experts tried to avoid including PARCC operational items flagged for DIF. Section 10.2 of this chapter explains the steps taken to evaluate LEAP 2025 items through the use of DIF to adhere to this standard.

In addition, standardized test administration and extensive training of 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)

Copyright © 2019 by Louisiana Department of Education.

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 section 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 (see http://parcconline.org).

The DIF and reliability analyses in this section are based on the CIA data described in Chapter 6. The impact analysis (scale score mean and standard deviation) are based on the technical report sample described in Chapter 7.

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 at least these same people and was also subjected to an independent review by 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 (see Tables 10.1 and 10.2) indicate the degree to which members of a particular subgroup perform

Copyright © 2019 by Louisiana Department of Education.

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 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. 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.

Copyright © 2019 by Louisiana Department of Education.

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:

1. Moderate DIF: Significant MH chi-square statistic ($p < 0.05$) and $1.0 \leq |\text{MH D-DIF}| < 1.5$
2. Large DIF: Significant MH chi-square statistic ($p < 0.05$) and $|\text{MH D-DIF}| \geq 1.5$

For constructed-response items, an effect size (ES) statistic based on the MH chi-square will be used. The ES is obtained by dividing the SMD statistics by the standard deviation of the item. The SMD is an effect size index of DIF, which is relatively easy to interpret. The SMD compares the mean of the reference and focal group, adjusting for the distribution of reference and focal group members on the conditioning variable, which for these analyses is the LEAP 2025 raw score. The SMD is computed as follows (Zwick et al., 1993):

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

Copyright © 2019 by Louisiana Department of Education.

where p_{Fk} = 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 $|ES|$ is between 0.17 and 0.25
- Large DIF: If the MH statistic is significant ($p < .05$) and $|ES| \geq 0.25$

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. 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.

English Learner 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 differences should they exist.

Tables 10.1 and 10.2 summarize the number of DIF flags for the fall 2017 and spring 2018 administrations by subject for each focal group that included at least 200 students. Results are not reported (NR) for groups with an insufficient number of students. The analyses were conducted by test form. The numbers of initial testers in summer 2018 were small, and therefore, DIF was not computed.

The fall 2017 administration DIF table (see Table 10.1) can be considered an example. On the English I form, a total of five items were flagged for DIF for the female subgroup: two of the five items had moderate negative DIF, one of the five exhibited moderate positive DIF, and other two had large positive DIF. Three items were flagged for the special education subgroups; two had moderate negative DIF and one showed large negative DIF.

Table 10.1 Fall 2017 Administration DIF Statistics: Number of Flagged Items

DIF Statistics				Count of Items at DIF Magnitude			
				Moderate		Large	
Content	Number of Items	Category	Group	B-	B+	C-	C+
English I	33	Gender	Female	2	1	0	2
		Ethnicity	Hispanic/Latino	0	0	0	0

Copyright © 2019 by Louisiana Department of Education.

DIF Statistics				Count of Items at DIF Magnitude			
				Moderate		Large	
Content	Number of Items	Category	Group	B-	B+	C-	C+
		Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
		Ethnicity	Asian	NR	NR	NR	NR
		Ethnicity	Black or African American	0	0	0	0
		Ethnicity	Two or More Races	NR	NR	NR	NR
		Education Classification	Special Education	2	0	1	0
		Education Classification	Gifted or Talented	0	0	0	0
		English Learner Status	EL	NR	NR	NR	NR
		Economic Status	Economically Disadvantaged	NR	NR	NR	NR
		Section 504 Status	Section 504	0	0	0	0
English II	33	Gender	Female	0	5	0	0
		Ethnicity	Hispanic/Latino	1	0	0	0
		Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
		Ethnicity	Asian	NR	NR	NR	NR
		Ethnicity	Black or African American	1	0	0	0
		Ethnicity	Two or More Races	NR	NR	NR	NR
		Education Classification	Special Education	4	0	0	0
		Education Classification	Gifted or Talented	0	2	0	0
		English Learner Status	EL	NR	NR	NR	NR
		Economic Status	Economically Disadvantaged	NR	NR	NR	NR
		Section 504 Status	Section 504	0	0	0	0
Algebra I	38	Gender	Female	1	1	0	0
		Ethnicity	Hispanic/Latino	NR	NR	NR	NR
		Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
		Ethnicity	Asian	NR	NR	NR	NR
		Ethnicity	Black or African American	1	0	2	0
		Ethnicity	Two or More Races	NR	NR	NR	NR
		Education Classification	Special Education	0	0	2	0
		Education Classification	Gifted or Talented	NR	NR	NR	NR
		English Learner Status	EL	NR	NR	NR	NR
		Economic Status	Economically Disadvantaged	NR	NR	NR	NR
		Section 504 Status	Section 504	NR	NR	NR	NR
Geometry	38	Gender	Female	1	0	0	0
		Ethnicity	Hispanic/Latino	2	0	1	0
		Ethnicity	American Indian or Alaska Native	NR	NR	NR	NR
		Ethnicity	Asian	NR	NR	NR	NR
		Ethnicity	Black or African American	1	0	1	0
		Ethnicity	Two or More Races	NR	NR	NR	NR

Copyright © 2019 by Louisiana Department of Education.

DIF Statistics				Count of Items at DIF Magnitude			
				Moderate		Large	
Content	Number of Items	Category	Group	B-	B+	C-	C+
		Education Classification	Special Education	0	1	2	0
		Education Classification	Gifted or Talented	0	1	0	0
		English Learner Status	EL	NR	NR	NR	NR
		Economic Status	Economically Disadvantaged	NR	NR	NR	NR
		Section 504 Status	Section 504	0	0	0	0

Table 10.2 Spring 2018 Administration DIF Statistics: Number of Flagged Items

DIF Statistics				Count of Items at DIF Magnitude			
				Moderate		Large	
Content	Number of Items	Category	Group	B-	B+	C-	C+
English I	34	Gender	Female	2	6	1	0
		Ethnicity	Hispanic/Latino	0	0	0	0
		Ethnicity	American Indian or Alaska Native	0	0	0	0
		Ethnicity	Asian	1	2	0	0
		Ethnicity	Black or African American	1	0	0	0
		Ethnicity	Two or More Races	0	0	0	0
		Education Classification	Special Education	0	0	0	0
		Education Classification	Gifted or Talented	0	1	0	0
		English Learner Status	EL	3	0	0	0
		Economic Status	Economically Disadvantaged	0	0	0	0
		Section 504 Status	Section 504	0	0	0	0
English II	34	Gender	Female	3	6	0	0
		Ethnicity	Hispanic/Latino	1	0	0	0
		Ethnicity	American Indian or Alaska Native	1	0	0	0
		Ethnicity	Asian	0	0	0	0
		Ethnicity	Black or African American	2	0	0	0
		Ethnicity	Two or More Races	0	0	0	0
		Education Classification	Special Education	1	0	0	0
		Education Classification	Gifted or Talented	0	0	0	0
		English Learner Status	EL	3	0	1	0
		Economic Status	Economically Disadvantaged	0	0	0	0
		Section 504 Status	Section 504	0	0	0	0
Algebra I	39	Gender	Female	1	1	0	0
		Ethnicity	Hispanic/Latino	0	0	1	0

Copyright © 2019 by Louisiana Department of Education.

DIF Statistics				Count of Items at DIF Magnitude			
				Moderate		Large	
Content	Number of Items	Category	Group	B-	B+	C-	C+
		Ethnicity	American Indian or Alaska Native	0	0	0	0
		Ethnicity	Asian	0	3	0	0
		Ethnicity	Black or African American	2	0	1	0
		Ethnicity	Two or More Races	0	0	0	0
		Education Classification	Special Education	1	0	2	0
		Education Classification	Gifted or Talented	0	1	0	1
		English Learner Status	EL	1	0	1	0
		Economic Status	Economically Disadvantaged	1	0	0	0
		Section 504 Status	Section 504	0	0	0	0
Geometry	38	Gender	Female	0	0	0	0
		Ethnicity	Hispanic/Latino	1	0	0	0
		Ethnicity	American Indian or Alaska Native				
		Ethnicity	Asian	0	3	1	0
		Ethnicity	Black or African American	1	0	0	0
		Ethnicity	Two or More Races	0	0	0	0
		Education Classification	Special Education	1	0	0	0
		Education Classification	Gifted or Talented	0	0	0	0
		English Learner Status	EL	1	2	1	0
		Economic Status	Economically Disadvantaged	0	0	0	0
		Section 504 Status	Section 504	0	0	0	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 10.3–10.16 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.

10.4 Reliability

Tables 10.3–10.6 show the regular test form reliability coefficients and SEM by student gender, ethnicity, education classification, economic status, LEP status, migrant status and 504 status. The reliability coefficients for English I and II forms ranged from 0.79 to 0.91 and from 0.82 to 0.93 for the fall 2017 and spring 2018 administrations, respectively. For algebra I and geometry, the reliability coefficients ranged from 0.80 to 0.92, and from 0.79 to 0.93 for the fall 2017 and spring 2018 administrations, respectively. This analysis shows that the test reliability is of acceptable magnitude for all the subgroups. Note that the reliability coefficients are NR for subgroups smaller than 10 students.

Copyright © 2019 by Louisiana Department of Education.

Table 10.3 English I Reliability and SEM by Subgroup

	Fall 2017			Spring 2018		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥4,630	0.88	5.57	≥46,180	0.90	5.60
Gender						
Female	≥2,270	0.88	5.56	≥22,760	0.89	5.61
Male	≥2,360	0.88	5.45	≥23,420	0.90	5.46
Ethnicity						
Hispanic/Latino	≥230	0.89	5.65	≥2,820	0.92	5.43
American Indian or Alaska Native	≥40	0.88	5.49	≥290	0.88	5.65
Asian	≥70	0.87	5.66	≥830	0.90	5.69
Black or African American	≥1,710	0.87	5.43	≥20,030	0.88	5.50
Native Hawaiian or Other Pacific	<10	NR	NR	≥20	0.82	5.91
White	≥2,470	0.87	5.61	≥21,390	0.88	5.71
Two or More Races	≥80	0.90	5.42	≥780	0.89	5.78
Education Classification						
Regular Education	≥3,960	0.87	5.58	≥39,540	0.88	5.63
Special Education	≥330	0.84	4.86	≥3,720	0.86	4.92
Gifted or Talented	≥330	0.82	5.49	≥2,910	0.85	5.58
Economic Status						
Economically Disadvantaged	<10	NR	NR	≥27,090	0.88	5.54
Not Economically Disadvantaged	≥4,330	0.88	5.57	≥13,440	0.88	5.72
English Learner Status						
Non-EL	≥4,600	0.88	5.57	≥45,070	0.89	5.62
EL	≥30	0.79	4.81	≥1,110	0.85	4.81
Migrant Status						
Migrant	≥4,630	0.88	5.57	≥46,130	0.90	5.60
Not Migrant	<10	NR	NR	≥50	0.89	5.27
Section 504 Status						
Non-Section 504	≥4,290	0.88	5.58	≥42,430	0.90	5.61
Section 504	≥340	0.85	5.21	≥3,750	0.88	5.38

Copyright © 2019 by Louisiana Department of Education.

Table 10.4 English II Reliability and SEM by Subgroup

	Fall 2017			Spring 2018		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥6,020	0.90	5.45	≥41,880	0.90	5.29
Gender						
Female	≥3,000	0.90	5.45	≥21,200	0.89	5.27
Male	≥3,020	0.90	5.35	≥20,680	0.90	5.16
Ethnicity						
Hispanic/Latino	≥530	0.91	5.39	≥2,260	0.92	5.29
American Indian or Alaska Native	≥30	0.88	5.50	≥290	0.87	5.21
Asian	≥140	0.91	5.52	≥770	0.92	5.23
Black or African American	≥2,420	0.88	5.36	≥18,070	0.87	5.26
Native Hawaiian or Other Pacific	<10	NR	NR	≥30	0.93	5.07
White	≥2,750	0.89	5.41	≥19,800	0.89	5.32
Two or More Races	≥120	0.89	5.44	≥640	0.89	5.39
Education Classification						
Regular Education	≥5,190	0.89	5.45	≥36,270	0.88	5.31
Special Education	≥380	0.87	4.79	≥3,080	0.86	4.88
Gifted or Talented	≥450	0.87	5.13	≥2,520	0.87	5.17
Economic Status						
Economically Disadvantaged	<10	NR	NR	≥23,760	0.88	5.28
Not Economically Disadvantaged	≥5,430	0.90	5.44	≥13,170	0.89	5.30
English Learner Status						
Non-EL	≥5,820	0.90	5.45	≥40,980	0.89	5.30
EL	≥200	0.80	4.58	≥900	0.86	5.02
Migrant Status						
Migrant	≥6,010	0.90	5.44	≥41,830	0.90	5.29
Not Migrant	<10	NR	NR	≥40	0.91	5.28
Section 504 Status						
Non-Section 504	≥5,620	0.90	5.45	≥38,860	0.89	5.30
Section 504	≥400	0.89	5.22	≥3,020	0.88	5.20

Copyright © 2019 by Louisiana Department of Education.

Table 10.5 Algebra I Reliability and SEM by Subgroup

	Fall 2017			Spring 2018		
Group	**N Count**	**Cronbach's Alpha**	**SEM**	**N Count**	**Cronbach's Alpha**	**SEM**
All Students	≥2,920	0.87	4.07	≥46,010	0.87	3.75
Gender						
Female	≥1,370	0.87	4.12	≥23,170	0.86	3.76
Male	≥1,540	0.87	4.01	≥22,840	0.87	3.73
Ethnicity						
Hispanic/Latino	≥180	0.88	4.15	≥2,800	0.87	3.74
American Indian or Alaska Native	≥20	0.80	4.10	≥300	0.86	3.80
Asian	≥60	0.89	4.25	≥820	0.90	4.20
Black or African American	≥1,150	0.84	3.81	≥19,610	0.81	3.46
Native Hawaiian or Other Pacific	<10	NR	NR	≥30	0.87	4.11
White	≥1,430	0.86	4.10	≥21,660	0.86	3.85
Two or More Races	≥50	0.86	4.23	≥760	0.87	3.81
Education Classification						
Regular Education	≥2,570	0.86	4.07	≥39,250	0.85	3.69
Special Education	≥200	0.84	3.32	≥3,680	0.79	3.22
Gifted or Talented	≥130	0.87	4.07	≥3,080	0.88	4.16
Economic Status						
Economically Disadvantaged	<10	NR	NR	≥26,500	0.83	3.58
Not Economically Disadvantaged	≥2,510	0.86	4.09	≥13,660	0.87	3.92
English Learner Status						
Non-EL	≥2,860	0.87	4.07	≥44,960	0.87	3.75
EL	≥60	0.86	3.68	≥1,050	0.82	3.31
Migrant Status						
Migrant	≥2,910	0.87	4.07	≥45,960	0.87	3.75
Not Migrant	<10	NR	NR	≥50	0.90	3.67
Section 504 Status						
Non-Section 504	≥2,720	0.87	4.08	≥42,390	0.87	3.76
Section 504	≥200	0.84	3.69	≥3,620	0.84	3.48

Copyright © 2019 by Louisiana Department of Education.

Table 10.6 Geometry Reliability and SEM by Subgroup

	Fall 2017			Spring 2018		
Group	N Count	Cronbach's Alpha	SEM	N Count	Cronbach's Alpha	SEM
All Students	≥5,430	0.91	3.85	≥36,800	0.91	3.73
Gender						
Female	≥2,830	0.91	3.87	≥19,350	0.91	3.77
Male	≥2,590	0.91	3.83	≥17,450	0.91	3.68
Ethnicity						
Hispanic/Latino	≥480	0.92	3.69	≥1,970	0.91	3.66
American Indian or Alaska Native	≥30	0.89	3.75	≥190	0.90	3.92
Asian	≥150	0.92	4.40	≥710	0.92	4.35
Black or African American	≥2,110	0.88	3.33	≥15,680	0.88	3.23
Native Hawaiian or Other Pacific	<10	NR	NR	≥10	0.93	3.65
White	≥2,530	0.90	4.08	≥17,640	0.90	3.97
Two or More Races	≥110	0.90	3.93	≥560	0.91	3.89
Education Classification						
Regular Education	≥4,620	0.90	3.77	≥32,430	0.90	3.67
Special Education	≥290	0.86	3.06	≥1,780	0.88	2.93
Gifted or Talented	≥510	0.90	4.37	≥2,590	0.91	4.30
Economic Status						
Economically Disadvantaged	<10	NR	NR	≥19,750	0.90	3.46
Not Economically Disadvantaged	≥4,950	0.91	3.91	≥12,440	0.91	4.02
English Learner Status						
Non-EL	≥5,250	0.91	3.87	≥36,180	0.91	3.74
EL	≥180	0.82	2.98	≥620	0.90	3.19
Migrant Status						
Migrant	≥5,430	0.91	3.85	≥36,770	0.91	3.73
Not Migrant	<10	NR	NR	≥30	0.91	3.28
Section 504 Status						
Non-Section 504	≥5,130	0.91	3.87	≥34,570	0.91	3.75
Section 504	≥300	0.90	3.55	≥2,220	0.90	3.38

Copyright © 2019 by Louisiana Department of Education.

10.5 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 10.7–10.10. Results are NR for subgroups with fewer than ten students. For all subjects across both the fall and spring administrations, small differences in test scores were seen between females and males, with females slightly outperforming males. Mean scale scores and ES show that Asian and white students tend to outperform other ethnicity groups across subjects. For most ELA and mathematics tests, there were clear performance differences between regular and special education students in education classification and non-LEP and LEP students in LEP status

Copyright © 2019 by Louisiana Department of Education.

Table 10.7 Impact Analysis: English I

	Fall 2017				Spring 2018			
Group	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
All Students	≥4,610	744.58	31.35		≥46,130	744.04	34.41	
Gender								
Male	≥2,350	737.62	30.80		≥23,450	737.50	34.40	
Female	≥2,260	751.83	30.27	-0.47	≥22,670	750.79	33.10	-0.39
Ethnicity								
White	≥2,460	753.09	29.69		≥21,280	754.84	32.20	
Hispanic/Latino	≥230	741.84	32.95	0.38	≥2,830	735.01	38.66	0.60
American Indian or Alaska Native	≥40	739.15	31.02	0.47	≥290	747.94	31.63	0.21
Asian	≥70	759.64	30.09	-0.22	≥820	771.90	36.05	-0.53
Black or African American	≥1,700	731.91	29.03	0.72	≥20,070	732.34	31.47	0.71
Native Hawaiian or Other Pacific	<10	NR	NR	NR	≥20	759.04	25.52	-0.13
Two or More Races	≥80	747.90	32.30	0.17	≥770	751.56	33.93	0.10
Education Classification								
Regular Education	≥3,950	744.62	29.03		≥39,430	744.62	32.02	
Special Education	≥320	710.41	26.78	1.19	≥3,780	709.53	29.22	1.10
Gifted or Talented	≥330	777.61	25.80	-1.15	≥2,910	780.98	29.13	-1.14
Economic Status								
Not Economically Disadvantaged	≥4,320	745.89	30.98		≥13,470	759.71	32.07	
Economically Disadvantaged	<10	NR	NR	NR	≥27,270	736.16	32.42	0.73
English Learner Status								
Non-EL	≥4,570	744.93	31.15		≥45,010	745.08	33.88	
EL	≥30	699.71	25.48	1.45	≥1,110	701.75	28.74	1.28
Migrant Status								
Nonmigrant	≥4,610	744.58	31.35		≥46,080	744.05	34.41	
Migrant	<10	NR	NR	NR	≥50	730.92	31.58	0.38
Section 504 Status								
Non-Section 504	≥4,270	746.30	31.01		≥42,330	745.64	34.20	
Section 504	≥330	722.96	27.36	0.76	≥3,790	726.14	31.61	0.57

Copyright © 2019 by Louisiana Department of Education.

Table 10.8 Impact Analysis: English II

	Fall 2017				Spring 2018			
Group	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
All Students	≥9,130	723.96	44.60		≥44,350	743.63	43.21	
Gender								
Male	≥4,990	716.36	42.81		≥22,320	734.58	43.17	
Female	≥4,140	733.12	44.98	-0.38	≥22,020	752.79	41.28	-0.43
Ethnicity								
White	≥3,370	745.22	44.43		≥20,250	758.00	41.26	
Hispanic/Latino	≥900	710.39	45.27	0.78	≥2,630	730.51	48.22	0.65
American Indian or Alaska Native	≥40	727.15	40.93	0.41	≥310	747.24	38.86	0.26
Asian	≥170	747.15	50.76	-0.04	≥790	772.87	48.53	-0.36
Black or African American	≥4,460	709.13	36.61	0.90	≥19,640	728.94	38.40	0.73
Native Hawaiian or Other Pacific	<10	NR	NR	NR	≥30	756.73	50.88	0.03
Two or More Races	≥160	739.06	45.24	0.14	≥670	754.09	41.82	0.09
Education Classification								
Regular Education	≥7,350	727.28	41.56		≥38,010	745.13	40.16	
Special Education	≥1,300	684.20	26.98	1.09	≥3,780	698.02	34.87	1.19
Gifted or Talented	≥470	782.27	40.27	-1.33	≥2,550	788.87	38.19	-1.09
Economic Status								
Not Economically Disadvantaged	≥6,010	740.63	41.75		≥13,470	762.90	41.07	
Economically Disadvantaged	<10	NR	NR	NR	≥25,610	733.81	40.25	0.72
English Learner Status								
Non-EL	≥8,580	726.67	44.20		≥43,110	745.04	42.62	
EL	≥540	681.01	24.17	1.06	≥1,240	694.57	34.10	1.19
Migrant Status								
Nonmigrant	≥9,100	724.04	44.58		≥44,290	743.64	43.20	
Migrant	≥20	693.25	42.77	0.69	≥50	731.82	47.73	0.27
Section 504 Status								
Non-Section 504	≥8,340	725.74	45.00		≥40,970	745.41	43.07	
Section 504	≥780	705.18	34.90	0.46	≥3,380	722.07	38.91	0.55

Copyright © 2019 by Louisiana Department of Education.

Table 10.9 Impact Analysis: Algebra I

	Fall 2017				Spring 2018			
Group	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
All Students	≥8,340	719.51	28.81		≥49,590	741.14	33.91	
Gender								
Male	≥4,620	718.34	28.10		≥24,920	739.43	34.23	
Female	≥3,720	720.95	29.60	-0.09	≥24,670	742.86	33.50	-0.10
Ethnicity								
White	≥2,750	731.77	32.15		≥22,580	753.06	32.85	
Hispanic/Latino	≥640	716.78	28.43	0.48	≥3,180	736.62	35.04	0.50
American Indian or Alaska Native	≥40	724.31	31.60	0.23	≥310	748.05	33.08	0.15
Asian	≥90	748.24	40.36	-0.51	≥840	775.42	39.63	-0.68
Black or African American	≥4,670	711.91	22.95	0.74	≥21,800	727.81	28.66	0.82
Native Hawaiian or Other Pacific	<10	NR	NR	NR	≥30	760.27	36.61	-0.22
Two or More Races	≥120	723.73	31.92	0.25	≥810	746.12	34.06	0.21
Education Classification								
Regular Education	≥6,610	722.42	28.63		≥41,980	741.27	31.44	
Special Education	≥1,540	702.67	18.69	0.73	≥4,400	711.84	26.73	0.95
Gifted or Talented	≥190	754.01	35.63	-1.09	≥3,210	779.70	34.66	-1.21
Economic Status								
Not Economically Disadvantaged	≥3,530	736.40	31.35		≥14,270	757.06	34.20	
Economically Disadvantaged	<10	NR	NR	NR	≥29,260	732.97	30.48	0.76
English Learner Status								
Non-EL	≥7,930	720.14	29.02		≥48,220	741.88	33.77	
EL	≥410	707.31	20.99	0.45	≥1,360	715.12	28.36	0.80
Migrant Status								
Nonmigrant	≥8,320	719.52	28.79		≥49,530	741.14	33.91	
Migrant	≥20	715.52	33.64	0.14	≥60	740.55	36.74	0.02
Section 504 Status								
Non-Section 504	≥7,510	720.63	29.23		≥45,530	742.48	33.91	
Section 504	≥820	709.28	22.20	0.40	≥4,060	726.10	30.17	0.49

Copyright © 2019 by Louisiana Department of Education.

Table 10.10 Impact Analysis: Geometry

	Fall 2017				Spring 2018			
Group	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size	N	Scale Score Mean	Scale Score Std. Dev.	Effect Size
All Students	≥6,720	730.62	27.50		≥37,650	735.74	25.87	
Gender								
Male	≥3,320	729.62	27.83		≥17,970	734.68	26.41	
Female	≥3,390	731.59	27.14	-0.07	≥19,680	736.70	25.32	-0.08
Ethnicity								
White	≥2,760	743.35	25.46		≥17,710	745.84	23.95	
Hispanic/Latino	≥590	725.75	27.30	0.68	≥2,130	731.98	26.80	0.57
American Indian or Alaska Native	≥40	732.76	22.85	0.42	≥190	743.47	24.34	0.10
Asian	≥160	759.17	29.48	-0.62	≥720	758.97	28.29	-0.54
Black or African American	≥3,030	718.15	22.45	1.05	≥16,280	723.96	22.05	0.95
Native Hawaiian or Other Pacific	<10	NR	NR	NR	≥10	730.32	30.09	0.65
Two or More Races	≥130	737.32	26.90	0.24	≥580	740.28	25.11	0.23
Education Classification								
Regular Education	≥5,630	730.38	25.40		≥33,080	735.16	24.40	
Special Education	≥550	704.80	20.42	1.02	≥1,950	711.72	22.45	0.96
Gifted or Talented	≥530	760.30	26.62	-1.17	≥2,610	760.97	25.62	-1.05
Economic Status								
Not Economically Disadvantaged	≥5,260	737.20	25.93		≥12,600	746.32	24.89	
Economically Disadvantaged	<10	NR	NR	NR	≥20,540	729.12	23.80	0.71
English Learner Status								
Non-EL	≥6,450	731.43	27.50		≥36,910	736.17	25.73	
EL	≥260	710.71	18.72	0.76	≥740	714.43	23.71	0.85
Migrant Status								
Nonmigrant	≥6,710	730.60	27.49		≥37,620	735.74	25.87	
Migrant	<10	NR	NR	NR	≥30	725.97	25.98	0.38
Section 504 Status								
Non-Section 504	≥6,280	731.47	27.48		≥35,320	736.42	25.79	
Section 504	≥430	718.35	24.74	0.48	≥2,330	725.35	24.78	0.43

Additional data for mean scale scores are provided in Tables 10.11 and 10.14. These tables report the number of students, mean scale scores, and standard deviations for special education classification. Groups that have fewer than 10 students are not reported (NR) in the tables.

Copyright © 2019 by Louisiana Department of Education.

Table 10.11 Special Education Classification Scale-Score Means and Standard Deviations: English I

Special Education Classification Scale-Score Means and Standard Deviations: English I							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
Fall 2017	Gifted	≥150	788.56	21.33	≥4,460	743.11	30.55
	Talented	≥180	768.74	25.74	≥4,420	743.57	31.16
	Autism	<50	NR	NR	≥4,600	744.61	31.34
	Deaf-Blindness	<50	NR	NR	≥4,610	744.59	31.35
	Developmental Delay	<50	NR	NR	≥4,610	744.59	31.35
	Emotional Disturbance	<50	NR	NR	≥4,600	744.69	31.30
	HI—Deaf	<50	NR	NR	≥4,610	744.59	31.35
	HI—Hard-of-Hearing	<50	NR	NR	≥4,610	744.61	31.34
	Mild Mental Disability	<50	NR	NR	≥4,600	744.73	31.26
	Moderate Mental Disability	<50	NR	NR	≥4,610	744.58	31.35
	Orthopedic Impairment	<50	NR	NR	≥4,610	744.57	31.35
	Other Health Impairment	≥60	708.54	25.09	≥4,550	745.07	31.15
	Specific Learning Disability	≥180	705.30	22.36	≥4,430	746.21	30.60
	Speech or Language Impairment	<50	NR	NR	≥4,590	744.65	31.34
	Traumatic Brain Injury	<50	NR	NR	≥4,610	744.59	31.35
	Visual Impairment	<50	NR	NR	≥4,600	744.59	31.35
	Other	<50	NR	NR	≥4,610	744.58	31.35
Spring 2018	Gifted	≥1,550	792.43	24.87	≥44,570	742.35	33.46
	Talented	≥1,350	767.88	28.13	≥44,770	743.31	34.33
	Autism	≥190	722.64	31.99	≥45,930	744.13	34.40
	Deaf-Blindness	<50	NR	NR	≥46,130	744.04	34.41
	Developmental Delay	<50	NR	NR	≥46,130	744.04	34.41
	Emotional Disturbance	≥140	702.83	31.70	≥45,990	744.16	34.35
	HI—Deaf	<50	NR	NR	≥46,110	744.05	34.41
	HI—Hard-of-Hearing	≥60	722.52	37.45	≥46,060	744.07	34.40
	Mild Mental Disability	≥150	686.03	21.34	≥45,970	744.23	34.29
	Moderate Mental Disability	<50	NR	NR	≥46,120	744.04	34.41
	Orthopedic Impairment	≥60	735.39	33.73	≥46,060	744.05	34.41
	Other Health Impairment	≥780	712.75	30.94	≥45,340	744.58	34.22
	Specific Learning Disability	≥2,110	706.06	25.16	≥44,010	745.86	33.74
	Speech or Language Impairment	≥180	728.33	31.78	≥45,940	744.10	34.41
	Traumatic Brain Injury	<50	NR	NR	≥46,120	744.04	34.41
	Visual Impairment	<50	NR	NR	≥46,090	744.05	34.41
	Other	<50	NR	NR	≥46,120	744.04	34.41
Summer 2018	Gifted	<50	NR	NR	≥1,390	696.10	23.16
	Talented	<50	NR	NR	≥1,390	696.00	23.12
	Autism	<50	NR	NR	≥1,380	696.23	23.17
	Deaf-Blindness	<50	NR	NR	≥1,390	696.12	23.18
	Developmental Delay	<50	NR	NR	≥1,390	696.12	23.18

Copyright © 2019 by Louisiana Department of Education.

Special Education Classification Scale-Score Means and Standard Deviations: English I							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Emotional Disturbance	<50	NR	NR	≥1,370	696.26	23.16
	HI—Deaf	<50	NR	NR	≥1,390	696.16	23.18
	HI—Hard-of-Hearing	<50	NR	NR	≥1,390	696.13	23.19
	Mild Mental Disability	<50	NR	NR	≥1,370	696.47	23.19
	Moderate Mental Disability	<50	NR	NR	≥1,390	696.12	23.18
	Orthopedic Impairment	<50	NR	NR	≥1,390	696.12	23.15
	Other Health Impairment	≥80	691.85	21.38	≥1,310	696.40	23.27
	Specific Learning Disability	≥240	686.85	19.56	≥1,150	698.08	23.42
	Speech or Language Impairment	<50	NR	NR	≥1,380	696.09	23.22
	Traumatic Brain Injury	<50	NR	NR	≥1,390	696.12	23.18
	Visual Impairment	<50	NR	NR	≥1,390	696.10	23.16
	Other	<50	NR	NR	≥1,390	696.12	23.18

Table 10.12 Special Education Classification Scale-Score Means and Standard Deviations: English II

Special Education Classification Scale-Score Means and Standard Deviations: English II							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
Fall 2017	Gifted	≥200	805.19	32.86	≥8,930	722.13	43.10
	Talented	≥270	765.21	36.69	≥8,860	722.70	44.22
	Autism	<50	NR	NR	≥9,080	724.10	44.56
	Deaf-Blindness	<50	NR	NR	≥9,130	723.96	44.60
	Developmental Delay	<50	NR	NR	≥9,130	723.96	44.60
	Emotional Disturbance	<50	NR	NR	≥9,100	724.08	44.57
	HI—Deaf	<50	NR	NR	≥9,120	723.99	44.59
	HI—Hard-of-Hearing	<50	NR	NR	≥9,110	724.03	44.58
	Mild Mental Disability	≥90	670.24	16.80	≥9,030	724.53	44.45
	Moderate Mental Disability	<50	NR	NR	≥9,120	723.99	44.58
	Orthopedic Impairment	<50	NR	NR	≥9,110	724.03	44.59
	Other Health Impairment	≥250	687.03	26.85	≥8,870	725.04	44.55
	Specific Learning Disability	≥770	683.20	24.12	≥8,350	727.73	44.17
	Speech or Language Impairment	<50	NR	NR	≥9,080	724.09	44.60
	Traumatic Brain Injury	<50	NR	NR	≥9,120	723.97	44.60
	Visual Impairment	<50	NR	NR	≥9,120	723.96	44.60
	Other	<50	NR	NR	≥9,120	724.00	44.59
Spring 2018	Gifted	≥1,220	807.22	30.11	≥43,130	741.83	42.15
	Talented	≥1,330	772.00	37.04	≥43,020	742.75	43.09
	Autism	≥170	717.08	42.56	≥44,180	743.73	43.18
	Deaf-Blindness	<50	NR	NR	≥44,350	743.63	43.21
	Developmental Delay	<50	NR	NR	≥44,350	743.63	43.21

Copyright © 2019 by Louisiana Department of Education.

Special Education Classification Scale-Score Means and Standard Deviations: English II							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Emotional Disturbance	≥110	699.45	41.01	≥44,240	743.74	43.16
	HI—Deaf	<50	NR	NR	≥44,320	743.65	43.20
	HI—Hard-of-Hearing	≥50	708.82	39.80	≥44,290	743.67	43.20
	Mild Mental Disability	≥180	675.91	22.50	≥44,170	743.91	43.06
	Moderate Mental Disability	<50	NR	NR	≥44,340	743.64	43.20
	Orthopedic Impairment	≥50	721.63	49.36	≥44,290	743.66	43.20
	Other Health Impairment	≥730	702.03	35.82	≥43,620	744.33	42.98
	Specific Learning Disability	≥2,210	693.81	30.50	≥42,130	746.25	42.18
	Speech or Language Impairment	≥160	719.68	40.62	≥44,180	743.72	43.20
	Traumatic Brain Injury	<50	NR	NR	≥44,340	743.64	43.21
	Visual Impairment	<50	NR	NR	≥44,310	743.64	43.22
	Other	<50	NR	NR	≥44,350	743.63	43.21
Summer 2018	Gifted	<50	NR	NR	≥2,290	688.63	27.58
	Talented	<50	NR	NR	≥2,280	688.57	27.58
	Autism	<50	NR	NR	≥2,280	688.69	27.66
	Deaf-Blindness	<50	NR	NR	≥2,290	688.67	27.62
	Developmental Delay	<50	NR	NR	≥2,290	688.67	27.62
	Emotional Disturbance	<50	NR	NR	≥2,280	688.71	27.63
	HI—Deaf	<50	NR	NR	≥2,290	688.69	27.63
	HI—Hard-of-Hearing	<50	NR	NR	≥2,290	688.66	27.65
	Mild Mental Disability	<50	NR	NR	≥2,260	688.81	27.65
	Moderate Mental Disability	<50	NR	NR	≥2,290	688.67	27.62
	Orthopedic Impairment	<50	NR	NR	≥2,290	688.68	27.63
	Other Health Impairment	≥100	685.05	27.36	≥2,180	688.84	27.63
	Specific Learning Disability	≥330	678.48	24.63	≥1,960	690.38	27.73
	Speech or Language Impairment	<50	NR	NR	≥2,270	688.76	27.63
	Traumatic Brain Injury	<50	NR	NR	≥2,290	688.68	27.62
	Visual Impairment	<50	NR	NR	≥2,290	688.67	27.63
	Other	<50	NR	NR	≥2,290	688.67	27.62

Copyright © 2019 by Louisiana Department of Education.

Table 10.13 Special Education Classification Scale-Score Means and Standard Deviations: Algebra I

Special Education Classification Scale-Score Means and Standard Deviations: Algebra I							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
Fall 2017	Gifted	≥60	782.05	29.02	≥8,280	719.05	28.31
	Talented	≥130	741.46	30.90	≥8,210	719.15	28.63
	Autism	≥50	709.14	29.44	≥8,280	719.58	28.79
	Deaf-Blindness	<50	NR	NR	≥8,340	719.50	28.81
	Developmental Delay	<50	NR	NR	≥8,340	719.51	28.81
	Emotional Disturbance	<50	NR	NR	≥8,300	719.58	28.84
	HI—Deaf	<50	NR	NR	≥8,330	719.53	28.81
	HI—Hard-of-Hearing	<50	NR	NR	≥8,320	719.53	28.82
	Mild Mental Disability	≥110	698.92	14.38	≥8,230	719.80	28.85
	Moderate Mental Disability	<50	NR	NR	≥8,340	719.51	28.81
	Orthopedic Impairment	<50	NR	NR	≥8,330	719.53	28.82
	Other Health Impairment	≥290	704.00	19.56	≥8,050	720.06	28.93
	Specific Learning Disability	≥890	701.49	17.72	≥7,450	721.67	29.12
	Speech or Language Impairment	≥60	706.81	21.73	≥8,280	719.60	28.83
	Traumatic Brain Injury	<50	NR	NR	≥8,340	719.51	28.81
	Visual Impairment	<50	NR	NR	≥8,330	719.53	28.81
	Other	<50	NR	NR	≥8,330	719.52	28.81
Spring 2018	Gifted	≥1,610	797.18	29.03	≥47,970	739.25	32.42
	Talented	≥1,590	761.90	30.64	≥48,000	740.45	33.80
	Autism	≥190	724.61	32.89	≥49,390	741.21	33.90
	Deaf-Blindness	<50	NR	NR	≥49,590	741.14	33.91
	Developmental Delay	<50	NR	NR	≥49,590	741.14	33.91
	Emotional Disturbance	≥150	711.77	29.23	≥49,440	741.23	33.89
	HI—Deaf	<50	NR	NR	≥49,570	741.15	33.91
	HI—Hard-of-Hearing	≥70	720.47	33.18	≥49,520	741.17	33.91
	Mild Mental Disability	≥200	696.09	18.75	≥49,380	741.33	33.84
	Moderate Mental Disability	<50	NR	NR	≥49,590	741.15	33.91
	Orthopedic Impairment	≥60	729.23	33.15	≥49,530	741.16	33.91
	Other Health Impairment	≥890	713.98	27.84	≥48,700	741.64	33.81
	Specific Learning Disability	≥2,510	708.49	22.51	≥47,080	742.88	33.53
	Speech or Language Impairment	≥210	733.98	36.11	≥49,370	741.17	33.90
	Traumatic Brain Injury	<50	NR	NR	≥49,580	741.15	33.91
	Visual Impairment	<50	NR	NR	≥49,550	741.15	33.91
	Other	<50	NR	NR	≥49,590	741.14	33.91
Summer 2018	Gifted	<50	NR	NR	≥1,570	707.20	18.58
	Talented	<50	NR	NR	≥1,560	707.26	18.81
	Autism	<50	NR	NR	≥1,550	707.36	18.78

Copyright © 2019 by Louisiana Department of Education.

Special Education Classification Scale-Score Means and Standard Deviations: Algebra I							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Deaf-Blindness	<50	NR	NR	≥1,570	707.31	18.87
	Developmental Delay	<50	NR	NR	≥1,570	707.31	18.87
	Emotional Disturbance	<50	NR	NR	≥1,560	707.32	18.83
	HI—Deaf	<50	NR	NR	≥1,570	707.26	18.87
	HI—Hard-of-Hearing	<50	NR	NR	≥1,570	707.36	18.86
	Mild Mental Disability	<50	NR	NR	≥1,550	707.50	18.88
	Moderate Mental Disability	<50	NR	NR	≥1,570	707.31	18.87
	Orthopedic Impairment	<50	NR	NR	≥1,570	707.32	18.88
	Other Health Impairment	≥70	703.58	17.78	≥1,500	707.50	18.91
	Specific Learning Disability	≥240	700.25	15.21	≥1,330	708.61	19.19
	Speech or Language Impairment	<50	NR	NR	≥1,560	707.31	18.88
	Traumatic Brain Injury	<50	NR	NR	≥1,570	707.30	18.86
	Visual Impairment	<50	NR	NR	≥1,570	707.33	18.86
	Other	<50	NR	NR	≥1,570	707.32	18.87

Table 10.14 Special Education Classification Scale-Score Means and Standard Deviations: Geometry

Special Education Classification Scale-Score Means and Standard Deviations: Geometry							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
Fall 2017	Gifted	≥210	776.63	18.88	≥6,510	729.12	26.42
	Talented	≥320	749.48	25.48	≥6,400	729.67	27.26
	Autism	<50	NR	NR	≥6,700	730.65	27.50
	Deaf-Blindness	<50	NR	NR	≥6,720	730.62	27.50
	Developmental Delay	<50	NR	NR	≥6,720	730.62	27.50
	Emotional Disturbance	<50	NR	NR	≥6,710	730.64	27.50
	HI—Deaf	<50	NR	NR	≥6,720	730.62	27.50
	HI—Hard-of-Hearing	<50	NR	NR	≥6,710	730.63	27.50
	Mild Mental Disability	<50	NR	NR	≥6,680	730.84	27.43
	Moderate Mental Disability	<50	NR	NR	≥6,720	730.63	27.49
	Orthopedic Impairment	<50	NR	NR	≥6,710	730.63	27.51
	Other Health Impairment	≥100	705.66	19.86	≥6,620	731.00	27.43
	Specific Learning Disability	≥320	702.58	18.45	≥6,390	732.04	27.12
	Speech or Language Impairment	≥30	720.00	28.34	≥6,690	730.67	27.49
	Traumatic Brain Injury	<50	NR	NR	≥6,720	730.62	27.50
	Visual Impairment	<50	NR	NR	≥6,710	730.62	27.50
	Other	<50	NR	NR	≥6,720	730.63	27.49
Spring 2018	Gifted	≥1,380	773.29	20.95	≥36,270	734.31	24.94
	Talented	≥1,230	747.21	23.25	≥36,420	735.35	25.86
	Autism	≥110	727.64	28.98	≥37,540	735.76	25.85

Copyright © 2019 by Louisiana Department of Education.

Special Education Classification Scale-Score Means and Standard Deviations: Geometry							
Administration	Group	Yes			No		
		N	Mean	Std. Dev.	N	Mean	Std. Dev.
	Deaf-Blindness	<50	NR	NR	≥37,650	735.74	25.87
	Developmental Delay	<50	NR	NR	≥37,650	735.74	25.87
	Emotional Disturbance	≥60	709.34	21.61	≥37,590	735.78	25.85
	HI—Deaf	<50	NR	NR	≥37,640	735.74	25.87
	HI—Hard-of-Hearing	<50	NR	NR	≥37,620	735.75	25.87
	Mild Mental Disability	≥60	694.14	14.40	≥37,590	735.81	25.82
	Moderate Mental Disability	<50	NR	NR	≥37,650	735.74	25.87
	Orthopedic Impairment	≥50	727.40	25.36	≥37,600	735.75	25.87
	Other Health Impairment	≥430	714.20	22.45	≥37,220	735.98	25.80
	Specific Learning Disability	≥1,030	707.49	18.76	≥36,620	736.54	25.59
	Speech or Language Impairment	≥100	726.53	27.66	≥37,550	735.76	25.86
	Traumatic Brain Injury	<50	NR	NR	≥37,650	735.74	25.87
	Visual Impairment	<50	NR	NR	≥37,630	735.75	25.87
	Other	<50	NR	NR	≥37,650	735.74	25.87
Summer 2018	Gifted	<50	NR	NR	≥380	706.08	21.18
	Talented	<50	NR	NR	≥390	708.09	24.67
	Autism	<50	NR	NR	≥390	708.59	25.78
	Deaf-Blindness	<50	NR	NR	≥390	708.56	25.73
	Developmental Delay	<50	NR	NR	≥390	708.56	25.73
	Emotional Disturbance	<50	NR	NR	≥380	708.60	25.95
	HI—Deaf	<50	NR	NR	≥390	708.56	25.73
	HI—Hard-of-Hearing	<50	NR	NR	≥390	708.56	25.73
	Mild Mental Disability	<50	NR	NR	≥390	708.62	25.89
	Moderate Mental Disability	<50	NR	NR	≥390	708.56	25.73
	Orthopedic Impairment	<50	NR	NR	≥390	708.59	25.76
	Other Health Impairment	<50	NR	NR	≥380	708.60	25.83
	Specific Learning Disability	<50	NR	NR	≥350	709.83	26.44
	Speech or Language Impairment	<50	NR	NR	≥390	708.51	25.79
	Traumatic Brain Injury	<50	NR	NR	≥390	708.58	25.76
	Visual Impairment	<50	NR	NR	≥390	708.54	25.76
	Other	<50	NR	NR	≥390	708.58	25.76

10.6 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:

Copyright © 2019 by Louisiana Department of Education.

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

Copyright © 2019 by Louisiana Department of Education.

Appendix A – Transadaptation Process for Spanish Mathematics Forms

For English Learners, the LDOE offers the mathematics assessments in Spanish for computer-based tests (CBT) to mirror the English language forms, as well as text-to-speech (TTS). The Spanish versions of the test were developed through transadaptation. Transadaptation takes into consideration grade-level appropriateness of the words and sentence structures used, as well as the linguistic and cultural differences that exist between speakers of two different languages. Accounting for these differences allows experts to ensure that the item will measure the same construct as the English-language version of the item at the same level of rigor and therefore be 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 PARCC Spanish-language form). 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 and 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 as well as 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;
- had not introduced 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. The Spanish item was associated with the corresponding English item in the item bank, and the item was formatted and graphics finalized for review. The final approved transadapation was then compared to the Spanish item in the DRC assessment system as well as the English version of the item, and all changes were verified.

Copyright © 2019 by Louisiana Department of Education.

DRC's Spanish Test Development team then used the final, approved Communication Assistance Scripts in English to transadapt descriptions of graphics as necessary. These graphic descriptions were used when preparing the TTS for review. Scripting of TTS and reviews of Spanish forms in their final forms was conducted by native Spanish speakers at DRC prior to their submission to LDOE for a translation review by third-party translation vendor.

The vendor reviewed the transadapted forms and provided feedback to LDOE and DRC. Experienced DRC Spanish Test Development team members (who are all native Spanish speakers) and the translation vendor resolved any issues and DRC made modifications as necessary. The form was then approved by both DRC and the LDOE translation vendor.

Copyright © 2019 by Louisiana Department of Education.

Appendix B – Accommodated Print Form Creation

Guidelines for Building Accommodated Print Forms

Careful consideration is given to all items selected for a test form, which are to be 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 in such a way as 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. For example:

- 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 (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). The directions are modified to ask the student to write the correct answer in its corresponding box. Students are also able to circle and draw arrows to where the text should be placed or to 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 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 clicking on the selected text, whereby the selected answer (a word, phrase, or sentence) is highlighted. 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 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 allow students to type their 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 system allow 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 system require students 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 in the online item, including titles, axes labels, and keys, and are asked to complete the graph.

Copyright © 2019 by Louisiana Department of Education.

Displaying items similarly in both accommodated print and online (and allowing the student to interact with the items in a similar manner) maintains item integrity by assessing a similar construct in a similar manner. This provides students who are unable to access the assessment online with an assessment at the same level of rigor as the online test.

The 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.

Copyright © 2019 by Louisiana Department of Education.

Appendix C – Item Alignment Review Process

June 2017 Item Alignment Review Process

1. Committee members attended a general session and received an overview of the process, which included:
    a. Background on items and the review process that took place prior to items being field-tested on forms
    b. Overview of purpose of the alignment review:
        i. To confirm content alignment of each item to designated Louisiana State Standard
        ii. To confirm grade appropriateness of each item
        iii. To confirm items were free of issues of bias, fairness, or sensitivity that could impact student responses to item
    c. Overview of process itself:
        i. Committee members reviewed each item individually and decided status of each item.
            1. Status could be either "accepted with current alignment," "accepted with realignment," or "rejected."
        ii. Group then came to consensus regarding status of each item.
        iii. Items that were accepted were determined to appropriately measure the intended standard and were free of issues of bias, fairness, or sensitivity that could impact student responses to item.
2. After the general session, committee members broke out into assigned groups.
3. The facilitator went over the review process again by first walking the group through the use of the review spreadsheet.
    a. The ELA spreadsheet contained prepopulated columns of the following information for each item: session, sequence, item ID, passage, key, primary, secondary, and tertiary standards (as applicable).
    b. The Math spreadsheet contained prepopulated columns of the following information for each item: session, sequence, PARCC ID, IDEAS ID, maximum points, LSSM/LEAP evidence statement, secondary and tertiary standards (as applicable), key, and item type.
        i. The facilitator reviewed each column in the spreadsheet so committee members could see what data was prepopulated and what data needed to be filled in by them.
            1. The facilitator went over what each column represented to make sure all participants understood the metadata being presented.
            2. The facilitator went over the columns to be filled in individually by each committee member as they reviewed an item (accept with current alignment, accept with realignment, reject), and any comments they might have about the item. If committee members did not agree with an alignment, they were asked to propose a new alignment, if possible.

Copyright © 2019 by Louisiana Department of Education.

ii. The facilitator then reviewed the process to be used with each item.
    1. For the first couple of items, the facilitator and committee members worked through reviewing the item and its alignment(s) together. They then came to group consensus as to the status of the items.
    2. Once the group had been through a couple of items and understood the process, committee members reviewed each item individually (facilitator indicated how many items to do before getting back together for group discussion of each item) and deciding if they agree with the content alignment.
    3. Once every member finished reviewing the assigned set of items, individuals shared the status they gave each item. Items were discussed one at a time until a status was assigned through group consensus to each item in a given set.
    4. This process continued for all items taken to the item alignment review.
    5. Committee members were provided with an evaluation at the end of the meeting in order that DRC and LDOE could make improvements with future item alignment reviews.

Table C.1 LEAP 2025 June 2017 Item Alignment Review Totals

Content Area	Course	Total Items Reviewed	Accepted with Current Alignment	Accepted with Realignment	Rejected
ELA	English I	221	169	39	13
ELA	English II	188	145	37	6
Math	Algebra I	165	130	10	25
Math	Geometry	150	121	15	14

Table C.2 LEAP 2025 June 2017 English I Item Alignment Review Committee Make-Up

Member #	Gender	Race/Ethnicity	Background
1	Male	White	Teacher
2	Female	White	Teacher
3	Female	African American/ Native American	Teacher
4	Female	White	Teacher
5	Male	African American	Special Education (SPED) Teacher
6	Female	White	Supervisor of English Learners (EL)
7	Female	White	Instructional Coach
8	Male	African American	Teacher
9	Female	White	Teacher of Hearing Impaired (HI)

Copyright © 2019 by Louisiana Department of Education.

Table C.3 LEAP 2025 June 2017 English II Item Alignment Review Committee Make-Up

Member #	Gender	Race/Ethnicity	Background
1	Female	African American	Teacher (SPED)
2	Female	African American	Teacher
3	Female	White	Supervisor
4	Female	White	Teacher of Visually Impaired (VI)
5	Male	African American	Teacher
6	Male	White	Teacher
7	Male	Hispanic	Teacher (EL)
8	Female	White	Teacher (SPED)
9	Female	White	Teacher

Table C.4 LEAP 2025 June 2017 Algebra I Item Alignment Review Committee Make-Up

Member #	Gender	Race/Ethnicity	Background
1	Female	White	Instructional Lead/Supervisor
2	Female	African American	Teacher (SPED – Gifted)
3	Female	White	Teacher
4	Female	White	Instructional Lead/Supervisor
5	Female	White	Teacher
6	Male	White	Administrator
7	Female	African American	Curriculum/Instructional Specialist
8	Female	White	Teacher (SPED – Gifted)
9	Female	White	Teacher (SPED – HI/VI)
10	Female	White	Teacher
11	Male	White	Teacher
12	Female	White	Instructional Lead/Supervisor
13	Female	Asian	Teacher (SPED – HI/VI)

Copyright © 2019 by Louisiana Department of Education.

Table C.5 LEAP 2025 June 2017 Geometry Item Alignment Review Committee Make-Up

Member #	Gender	Race/Ethnicity	Background
1	Female	White	Instructional Lead/Supervisor
2	Female	White	Teacher (SPED – Gifted)
3	Female	White	Teacher (SPED – HI/VI)
4	Female	White	Teacher
5	Female	African American	Teacher
6	Male	White	Teacher
7	Male	White	Teacher
8	Female	White	Instructional Lead/Supervisor
9	Male	White	Teacher
10	Female	Asian	Teacher (SPED – HI/VI)
11	Female	White	Teacher

Copyright © 2019 by Louisiana Department of Education.

Appendix D – LEAP 2025 Spring 2018 Handscoring/AI Documentation

Copyright © 2019 by Louisiana Department of Education.

LEAP 2025 SPRING 2018
HANDSCORING/AI DOCUMENTATION

LEAP 2025 GRADES 3-8

ELA, Math, and Social Studies

LEAP 2025 HIGH SCHOOL AND EOC

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

Table of Contents

Schedule, Locations, and Staffing 1
Training and Scoring Schedule 1
Scorer Degree Requirements 2
Training 2
Training Materials 3
EOC Biology, LEAP 2025 and EOC U.S. History, and LEAP 2025 Grades 3-8 Social Studies 3
EOC English III 4
LEAP 2025 Algebra I, Geometry, English I, English II, and Grades 3-8 ELA and Math 4
Algebra I, Geometry, and Grades 3-8 Math Training Set Composition 5
English I, English II, and Grades 3-8 ELA Training Set Composition 6
English I Prose Constructed Response (PCR) Items and Associated Training Materials 7
English II Prose Constructed Response (PCR) Items and Associated Training Materials 7
Grades 3-8 ELA PCR Items and Associated Training Materials 8
Algebra I Items and Associated Training Materials 9
Geometry Items and Associated Training Materials 10
Grade 3 Math Items and Associated Training Materials 11
Grade 4 Math Items and Associated Training Materials 11
Grade 5 Math Items and Associated Training Materials 12
Grade 6 Math Items and Associated Training Materials 12
Grade 7 Math Items and Associated Training Materials 13
Grade 8 Math Items and Associated Training Materials 13
Qualifying 14
EOC Biology, English III, and U.S. History 14
Biology 14
English III 14
U.S. History 14
LEAP 2025 Constructed-Response and Extended-Response Items 15
English I, English II, and Grades 3-8 ELA 15

Algebra I, Geometry, and Grades 3-8 Math 15
U.S. History and Grades 3-8 Social Studies 16
Spring 2018 Scoring Plan 16
LEAP 2025 High School and EOC 16
LEAP 2025 Grades 3-8 16
Handscoring Rules 17
AI Scoring 17
Handscoring 17
Reader Monitoring Procedures 18
Team Leader Read-Behinds 18
Validity Sets and Inter-Rater Reliability 18
Calibration Sets 21
Handscoring Quality Control Reports 22
Scoring Summary Report Sample 22
Scoring Summary Report Sample with AI (Reader ID #3) 22
Reader Feedback Logs 23
Handling Unusual Responses 23
Nonscore Codes and Definitions 23
Nonscore Code Definitions 24
Nonscore Codes by Grade/Course 24
Alerts 24
Artificial Intelligence Scoring 25
AI Scoring – Measurement, Inc. 25
Model Building 26
Evaluation Metric 27
Scoring Responses with the AI Engine 28
Quality Control of the AI Engine 29
Identifying Responses for Human Review 29
Alert Detection System 30

Identification of Nonscoreables 31
Identifying Copied Text and Plagiarism with the AI Engine 31
AI Scoring – Pearson 34
The Intelligent Essay Assessor 34
How the Intelligent Essay Assessor was Trained 35
Quality Monitoring 37
Scoring (DRC) 38
Rescores (High School Tests Only) 39
Auto Rescores 39
Parent-Requested Rescores 40
Appendix A 42
DRC-MI Streaming Scoring Documentation 42
SECTION 1 – General Information 43
SECTION 2 – SCHEMA SUPPLEMENT 44
SECTION 3 – STATUS CODE INFORMATION 47
Appendix B 48
AI Model Building – English III (Spring 2018) 48
Quadratic Weighted Kappa (QWK) and Inter-rater Reliability (IRR) 48
Score Point Distribution (SPD) 48
AI Model Building – EOC U.S. History ER (Spring 2018) 49
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 49
AI Model Building – LEAP 2025 U.S. History ERs (Spring 2018) 50
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 50
AI Model Building – Social Studies Grades 5-8 ERs (Spring 2018) 51
Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD) 51
PCR Performance – ELA Grades 5-8, English I, and English II (Spring 2018) 52
Reading Comprehension and Written Expression or Written Expression Trait 52
Conventions Trait 52
Spring 2018 LEAP 2025 and EOC Items – IRR and SPD from Previous Administrations 53

Algebra I 53
Geometry 54
Math Grade 3 55
Math Grade 4 56
Math Grade 5 57
Math Grade 6 58
Math Grade 7 59
Math Grade 8 60
English I 61
English II 62
English III (All Report Data Scored by DRC and MI [AI]) 63
ELA Grade 3 64
ELA Grade 4 65
ELA Grade 5 66
ELA Grade 6 66
ELA Grade 7 67
ELA Grade 8 67
Biology 68
U.S. History ERs (EOC) 69
U.S. History ERs and CRs (LEAP 2025) 70
Social Studies Grades 3-8 71

Schedule, Locations, and Staffing

All reader training and handscoring for Spring 2018 LEAP 2025 Grades 3-8 and High School tests and End-of-Course (EOC) High School tests will take place at the DRC scoring center locations noted in the table below.

Training and Scoring Schedule

DRC's Spring 2018 LEAP 2025 reader training and scoring schedule is based on the spring testing windows of April 9, 2018 – May 4, 2018 (LEAP 2025 grades 3-8) and April 23, 2018 – May 18, 2018 (LEAP 2025 and EOC high school). Reader training and scoring locations and the anticipated dates for each are noted below.

Grade/Subject/ Course	DRC Scoring Center Location	Anticipated Staffing	2018 Reader Training and Scoring Window
3 ELA	Columbus, OH	1 Scoring Director, 5 Team Leaders, 55 Readers	May 14 – June 4
4 ELA	Madison, WI	1 Scoring Director, 6 Team Leaders, 57 Readers	May 14 – June 4
5 ELA	Madison, WI	1 Scoring Director, 4 Team Leaders, 34 Readers	April 16 – May 11
6 ELA	Lake Mary, FL	1 Scoring Director, 4 Team Leaders, 35 Readers	April 12 – May 11
7 ELA	Atlanta, GA	1 Scoring Director, 1 Team Leader, 10 Readers	April 16 – May 11
8 ELA	Atlanta, GA	1 Scoring Director, 4 Team Leaders, 35 Readers	April 16 – May 11
3 Math	Plymouth, MN	2 Scoring Directors, 4 Team Leaders, 36 Readers	May 14 – June 4
4 Math	Plymouth, MN	2 Scoring Directors, 4 Team Leaders, 38 Readers	May 14 – June 4
5 Math	Sharonville, OH	2 Scoring Directors, 4 Team Leaders, 38 Readers	April 16 – May 11
6 Math	Atlanta, GA	2 Scoring Directors, 5 Team Leaders, 48 Readers	April 16 – May 11
7 Math	Woodbury, MN	2 Scoring Directors, 5 Team Leaders, 51 Readers	April 16 – May 11
8 Math	Sharonville, OH	2 Scoring Directors, 5 Team Leaders, 51 Readers	April 16 – May 11
3 SS	Atlanta, GA	1 Scoring Director, 4 Team Leaders, 43 Readers	May 14 – June 4
4 SS (CRs)	Atlanta, GA	1 Scoring Director, 3 Team Leaders, 25 Readers	May 14 – June 4
4 SS (ER)	Plymouth, MN	1 Scoring Director, 2 Team Leaders, 18 Readers	May 14 – June 4
5 SS	Plymouth, MN	1 Scoring Director, 2 Team Leaders, 15 Readers	April 9 – May 11
6 SS	Plymouth, MN	1 Scoring Director, 2 Team Leaders, 15 Readers	April 9 – May 11
7 SS	Plymouth, MN	1 Scoring Director, 2 Team Leaders, 15 Readers	April 9 – May 11
8 SS	Plymouth, MN	1 Scoring Director, 2 Team Leaders, 15 Readers	April 9 – May 11
LEAP 2025 Algebra I	Sharonville, OH	2 Scoring Directors, 4 Team Leaders, 37 Readers	April 24 – May 25
LEAP 2025 Geometry	Sharonville, OH	2 Scoring Directors, 2 Team Leaders, 23 Readers	April 24 – May 25
LEAP 2025 English I	Plymouth, MN	1 Scoring Director, 1 Team Leader, 10 Readers	April 23 – May 25
LEAP 2025 English II	Plymouth, MN	1 Scoring Director, 1 Team Leader, 8 Readers	April 23 – May 25
EOC English III	Plymouth, MN	1 Scoring Director, 1 Team Leader, 8 Readers	April 23 – May 25
EOC Biology	Plymouth, MN	1 Scoring Director, 1 Team Leader, 6 Readers	April 19 – May 25
EOC U.S. History	Plymouth, MN	1 Scoring Director, 3 Team Leaders	April 23 – May 25
LEAP 2025 U.S. History	Plymouth, MN	1 Scoring Director, 3 Team Leaders, 21 Readers	
For U.S. History, the same core group of scoring supervisors will be involved with all EOC and LEAP 2025 scoring. In addition, the same team of 21 readers will be involved in the LEAP 2025 ER and CR scoring.			

Each DRC scoring center is a secure facility. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials

may leave the Scoring Center and must sign legally binding confidentiality agreements before work begins. DRC will retain these agreements for the duration of the contract. To prevent the unauthorized duplication of secured materials, cell phone/camera use within the scoring rooms is strictly forbidden. Readers only have access to student responses they are qualified to score. Each scorer is assigned a unique username and password to access DRC's imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

Scorers will be divided by subjects as detailed in the previous table. Depending on the overall progress of the project, more scorers may be added to some groups. Additionally, depending on the overall progress of the project, some groups may subdivide and work on different items.

Scorer Degree Requirements

DRC readers scoring for Louisiana have at least a four-year college degree. DRC has a Human Resources Director dedicated solely to recruiting and retaining our handscoring staff. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. During personal interviews, reader candidates are asked to demonstrate their own proficiency in writing by responding to a DRC writing topic and in mathematics by solving word problems with correct work shown. All of this results in a highly educated and diverse workforce. Our personnel files for readers and Team Leaders include evaluations for each project completed. We use these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performance on similar projects at DRC.

Training

In preparation for the scoring of all Spring 2018 LEAP 2025 and EOC items, DRC scoring supervisors will train readers using the same training materials that were used by previous vendors for prior administrations of the same items. These training materials originated from the sources noted below.

Reader training materials for the following were provided to DRC by Pacific Metrics and previously approved by LDOE.

- EOC Biology, English III, and U.S. History

Reader training materials for the following were developed by DRC in conjunction with LDOE:

- LEAP 2025 grades 3-8 Social Studies
- LEAP 2025 U.S. History

Reader training materials for the following were provided to DRC by the Partnership for Assessment of Readiness for College and Careers (PARCC) and were approved by PARCC:

- LEAP 2025 grades 3-8 ELA and Math
- LEAP 2025 Algebra I, Geometry, English I, and English II

The materials include:

- Passages, items/prompts, associated stimuli for applicable content areas and item types;
- Rubrics;
- Anchor Sets;
- Training Sets (or Practice Sets); and
- Qualifying Sets.

DRC will start the training with a review of passages, items/prompts, rubrics, and anchor responses, followed by the scoring and discussion of Training/Practice Sets and the scoring and discussion of Qualifying Sets. Once this process has been completed for an item or prompt, qualified readers will be able to start scoring live student responses. A group of scorers will score responses for a particular item until the scoring for that item is complete. Then they may move on to score a different item. Depending on the overall progress of the project and the current quantity of responses available to score for each item, some groups may subdivide and work on different items. Additionally, depending on the overall progress of the project, more scorers may be added to some groups when the groups are ready to score new items.

The following tables detail the composition of the training materials for all of the Spring 2018 LEAP 2025 Grades 3-8 and High School and EOC tests.

Training Materials

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

Reader training for EOC Biology and U.S. History is conducted using item-specific anchor sets, training sets, and qualifying sets provided by Pacific Metrics. The LEAP 2025 U.S. History and grades 3-8 Social Studies item-specific training materials were developed by DRC in conjunction with LDOE.

Set Type	**Biology, U.S. History, and Grades 3-8 Social Studies Training Materials**	**Annotated**
Anchor Set	Each item-specific Anchor Set contains at least two responses per score point (with at least one of each at the top score points).	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 one or more 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 may not be present in some Training or Qualifying Sets as there may have been few or no examples found during rangefinding or subsequent field test scoring. In such cases, DRC Scoring Directors will endeavor to find examples of these scores during live scoring to supplement reader training.		

EOC English III

For English III, the Content and Style dimensions will be trained using prompt-specific scoring guides, training sets, and qualifying sets provided by Pacific Metrics.

Set Type	Content and Style Dimension Training Materials	Annotated
Content and Style Anchor Set*	20 responses representing all numeric score points and including nonscore condition codes.	Yes
Content and Style Training Sets 1 – 3	30 responses across the three training sets. All numeric score points are represented in each set. Readers will score each response for both Content and Style.	No
Content and Style Qualifying Sets 1 – 3	40 responses across the three qualifying sets. All numeric score points are represented in each set. Readers will score each response for both Content and Style.	No
* Some responses appear more than once in the Anchor Set to illustrate both a Content and a Style score.		

English III training materials covering the four elements of the Conventions dimension (Sentence Formation [F], Usage [U], Mechanics [M], and Spelling [S]) are made up of sets that include examples from multiple writing prompts. Roughly half of the responses in each set are a score point 0 for each element, and the other half are a score point 1 for each element.

Set Type	Conventions (F, U, M, S) Element Training Materials	Annotated
Conventions Anchor Set*	6 responses for each Conventions element representing both numeric score points (three 0s and three 1s for each of the four elements).	Yes
Conventions Training Sets 1 – 2	20 responses across both training sets. Both numeric score points are represented in each set. Readers will score each response for all four elements.	No
Conventions Qualifying Sets 1 – 2	20 responses across both training sets. Both numeric score points are represented in each set. Readers will score each response for all four elements.	No
* Some responses appear more than once in the Conventions Anchor Set to illustrate scores in more than one element.		

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

For LEAP 2025 Algebra I, Geometry, English I, English II, and grades 3-8 ELA and math, DRC will use the PARCC-approved training and qualifying materials provided by PARCC. Training materials for each item can be grouped into one of two categories: “prototype” item materials or “abbreviated” item materials.

Prototype Item Materials

Some items included in the Louisiana forms are prototype items, meaning they have full sets of associated training materials, including Anchor Sets, Practice Sets, and Qualifying Sets. DRC will start the training with a review of passages/items, rubrics, and anchor responses, followed by the scoring and discussion of Practice Sets and the scoring and discussion of Qualifying Sets. Once this process has been completed for a prototype item included on the Louisiana form, qualified readers will start scoring live student responses for that item.

Abbreviated Item Materials

Unlike prototype items, abbreviated item training materials have only two item-specific Practice Sets and no Qualifying Sets; therefore, abbreviated items require a two-step training/qualifying process. First, scorers will train and qualify as described in the Prototype Item Materials section above using PARCC-approved training materials for an associated prototype item that is similar to the abbreviated one they will be scoring on the Louisiana form.[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 they are going to score. This consists of an item-specific Anchor Set and two item-specific Practice Sets. After completing the abbreviated item's training, readers may begin scoring live responses for the abbreviated item.

The following tables detail the composition of the training materials provided by PARCC for math and ELA:

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

Set Type	Mathematics Prototype Item Training	Mathematics Abbreviated Item Training	Annotated
Anchor Set	3 responses per score point (Composite items will have 3 responses per composite score)	3 responses per score point (Composite items will have 3 responses per composite score)	Yes
Practice Set 1	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Practice Set 2	10 responses representing the range of responses	10 responses representing the range of responses	Yes
Qualifying Set 1	10 responses comparable to the anchor set responses		No
Qualifying Set 2	10 responses comparable to the anchor set responses		No
Qualifying Set 3	10 responses comparable to the anchor set responses		No

[1] 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 English items, this determination by PARCC and Pearson was based on grade/course and task type.

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

Set Type	English Prototype Item Training	English Abbreviated 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).	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	5 responses representing the range of responses for ● the RCWE trait (for LAT and RST items) ● the WE trait (for NWT items)	10 responses representing the range of responses for both traits appropriate to the task type	Yes
Practice Set 2	5 responses representing the range of responses for the Knowledge of Language and Conventions trait	10 responses representing the range of responses for both traits appropriate to the task type	Yes
Practice Set 3	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Practice Set 4	10 responses representing the range of responses for both traits appropriate to the task type		Yes
Qualifying Set 1	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Qualifying Set 2	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Qualifying Set 3	10 responses comparable to the anchor set responses (includes both traits appropriate to the task type)		No
Direct Copy Set*	3-5 responses composed entirely or partially of text copied from passage or passages (includes both traits appropriate to the task type)	3-5 responses composed entirely or partially of text copied from passage or passages (includes both traits appropriate to the task type)	Yes
*The PARCC-approved Direct Copy sets provide additional annotated sample responses that explain the scoring rationale for responses composed entirely or partially of text copied from the source passage(s) associated with an item. DRC scoring supervisors review these item-specific sets with the readers prior to scoring the associated item.			

Set Type	English Prototype and Abbreviated Item Training	Annotated
Anchor Set (for the Knowledge of Language and 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 of Language and 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 of Language and Conventions trait during each prototype item qualifying session.	Yes

English I Prose Constructed Response (PCR) Items and Associated Training Materials

Question	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
7	LAT	902152	VH017536_2T	abbreviated	QCE09_L_VH037763_2T
20	RST	902161	VH017542_2T	abbreviated	QCE09_R_VH017542_2T
9 [AE form]	RST	902161	VH017542_2T	abbreviated	QCE09_R_VH017542_2T
14 [AE form]	NWT	906152	VH084830	abbreviated	QCE09_N_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 Prose Constructed Response (PCR) Items and Associated Training Materials

Question	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
7	LAT	906197	HH428127697	abbreviated	QCE10_L_7233_2T
20	RST	902331	VH004490	abbreviated	QCE10_R_7121_2T
9 [AE form]	RST	902331	VH004490	abbreviated	QCE10_R_7121_2T
14 [AE form]	NWT	902354	7064	abbreviated	QCE10_N_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 PCR Items and Associated Training Materials

Grade	Question	Task	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
3	7	RST	868500	E03_R_VF581237	abbreviated	E03_R_VF906000
	12	NWT	913497	E03_N_AA431426588	abbreviated	E03_N_VF910093
4	7	LAT	913567	E04_L_VH170170	abbreviated	E04_L_VF925727
	20	RST	802342	E04_R_1805	abbreviated	E04_R_VF653524
5	9	RST	868590	E05_R_C1381	abbreviated	E05_R_2208
	14	NWT	913663	E05_N_CC434542756	abbreviated	E05_N_VH003158
6	9	RST	870622	E06_R_VH082653	abbreviated	E06_R_3538
	14	NWT	913694	E06_N_D1466	abbreviated	E06_N_VH000592
7	9	RST	868704	E07_R_VH069361	abbreviated	E07_R_VH014400
	14	NWT	913842	E07_N_EE430133306	abbreviated	E07_N_4284
8	7	LAT	913958	E08_L_F1460	abbreviated	E08_L_5271
	20	RST	868741	E08_R_5013	abbreviated	E08_R_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.

Algebra I Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item
12	901832	3031-M44083P	Abbreviated	3003_M43111*
13	901882	VH196970	Abbreviated	VH046614*
25	901687	2407-M41752	Prototype	N/A
26	906043	2362-M41568	Abbreviated	VH046614*
37	901851	M41726	Abbreviated	3003_M43111*
38	901705	VF883359	Abbreviated	VH046614*
39	901857	VH046479	Abbreviated	2407_M41752*
12 [AE form]	901832	3031 M44083P	Abbreviated	3003_M43111*
13 [AE form]	901882	VH196970	Abbreviated	VH046614*
26 [AE form]	906043	2362 M41568	Abbreviated	VH046614*
28 [AE form]	901859	3003 M43111	Prototype	N/A
37 [AE form]	901851	M41726	Abbreviated	3003_M43111*
39 [AE form]	901857	VH046479	Abbreviated	2407_M41752*
44 [AE form]	901848	M47287	Abbreviated	M41686*
*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.				

Geometry Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item
13	902012	M41169	Abbreviated	VF935309*
15	902046	M46668	Abbreviated	3042_M44133*
27	902027	M43233	Abbreviated	VH001716*
28	902042	3020-M44058	Abbreviated	3042_M44133*
43	902062	VH150384	Abbreviated	VF613786*
44	902018	3042-M44133	Prototype	N/A
13 [AE form]	902012	M41169	Abbreviated	VF935309*
15 [AE form]	902046	M46668	Abbreviated	3042_M44133*
27 [AE form]	902027	M43233	Abbreviated	VH001716*
28 [AE form]	902036	2904-M43021	Prototype	N/A
43 [AE form]	902047	VH150404	Abbreviated	VF935309*
44 [AE form]	902018	3042-M44133	Prototype	N/A
*An item ID listed in the Associated Prototype column indicates that readers must be qualified on that prototype prior to reviewing the Abbreviated training materials described in the cells to the left.				

Grade 3 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	914008	VH054861	abbreviated	QCM03_M01883
18	914048	M05158	abbreviated	QCM03_M00848
32	798301	VF658085	abbreviated	QCM03_M02345
33	868619	M00848	prototype	N/A
48	914039	M02527	abbreviated	QCM03_M00848
49	914037	4003-M03007	abbreviated	QCM03_M00848
*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.				

Grade 4 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	914084	4112-M03491P	abbreviated	QCM04_0081_M00445
18	914086	M04133	abbreviated	QCM04_M03436
32	870723	M02081	abbreviated	QCM04_M02470
33	870317	M02255	abbreviated	QCM04_M02470
48	870722	0220-M00114P	abbreviated	QCM04_0081_M00445
49	870724	0083-M00453	abbreviated	QCM04_M02470
*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.				

Grade 5 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
17	914193	VH055372	abbreviated	QCM05_VH084803
18	914148	M03888	abbreviated	QCM05_VH141466
32	870325	4106-M03478	abbreviated	QCM05_VH084803
33	868655	M03555	prototype	N/A
48	914195	0154-M00796	abbreviated	QCM05_VH084803
49	868654	4134-M03661	abbreviated	QCM05_M03555
*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.				

Grade 6 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
31	914286	5047-M25308	abbreviated	QCM06_VH122131
35	868818	M22225	abbreviated	QCM06_M21577
36	868798	1298-M21432	abbreviated	QCM06_M21577
37	870825	M21477	abbreviated	QCM06_VF655921
47	914231	1740-M23030	abbreviated	QCM06_VH122131
48	914244	5012-M25073	abbreviated	QCM06_M21577
49	914281	M25152	abbreviated	QCM06_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.				

Grade 7 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
30	914362	VH083535	abbreviated	QCM07_VF643181
35	914333	1300-M21438	abbreviated	QCM07_VH151385
36	868849	M22018	prototype	N/A
37	878705	VH095906	abbreviated	QCM07_M22018
47	914336	M25731	abbreviated	QCM07_M20598
48	914339	VH151385	prototype	N/A
49	914329	M25580	abbreviated	QCM07_M20598
*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.				

Grade 8 Math Items and Associated Training Materials

Question	DRC Item ID	PARCC UIN	Material Type	Associated Prototype Item*
29	892626	VH095800	prototype	N/A
34	800273	M20195	abbreviated	QCM08_M20587
35	914376	VH049778	abbreviated	QCM08_M21063
36	878156	VH057654	abbreviated	QCM08_M20198
46	914381	M25425	abbreviated	QCM08_M21063
47	868882	VH081259	abbreviated	QCM08_M20198
48	914400	M21215	abbreviated	QCM08_VH095800
*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.				

As part of their training, readers in all content areas are also instructed on how to recognize alerts such as disturbing content and suspected plagiarism. Alert processing procedures are described in a later section of this document.

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 Spring 2018 LEAP 2025 and EOC items as were established by the vendors who scored these items previously.

EOC Biology, English III, and U.S. History

Descriptions of the qualifying standards for the Biology, English III, and U.S. History item types are below. These standards were established by Pacific Metrics.

Biology

Course	Qualifying Standard
Biology	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.

English III

Course	Qualifying Standard
English III (Content and Style)	English III scorers will first qualify on the Content and Style dimensions before moving on to qualify in the Conventions dimension. Each reader will complete at least two qualifying sets, and a score of 70% exact agreement or higher is required in each dimension in order to qualify. Since readers complete two sets, they may qualify on one dimension in the first set and the other dimension in the second set.
English III (Conventions)	Once qualified for Content and Style, readers must then qualify for each of the four elements (F, U, M, and S) that make up the Conventions dimension. An exact agreement rate of 80% or higher is required once on each of the individual Conventions elements. A scorer may qualify on some elements in the first qualifying set and the remaining elements in the second qualifying set.

U.S. History

Course	Qualifying Standard
U.S. History	Readers must qualify with 70% exact agreement or higher on one or more of the qualifying sets in order to score student responses.

LEAP 2025 Constructed-Response and Extended-Response Items

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

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

Course	Qualifying Standard	
English I, English II, and Grades 3-8 ELA	Perfect Agreement	Perfect Plus Adjacent Agreement
	70% average for both traits on two of three qualifying sets	96% across the three qualifying sets combined on both traits
	70% on each trait at least once across three qualifying sets	

Readers of English I, English II, and grades 3-8 ELA responses are required to meet all three of the qualifications listed in the table. Perfect Plus Adjacent Agreement of 96% means that out of the entire pool of scores that a reader gives across the three qualifying sets for an item, no more than 4% of those scores can be non-adjacent. In other words, no more than 2 of the 60 applied scores can be non-adjacent (3 sets x 10 responses/set x 2 traits = 60 applied score).

Algebra I, Geometry, and Grades 3-8 Math

Course	Qualifying Standard		
Algebra I, Geometry, and Grades 3-8 Math	Comprehensive	Perfect Agreement	Perfect Plus Adjacent Agreement
	0, 1, 2 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

Course	Qualifying Standard		
Algebra I, Geometry, and Grades 3-8 Math	Composite (multi-part) Items*	Perfect Agreement	Perfect Plus Adjacent Agreement
	0, 1 Rubric	90% on two of three sets	100% on two of three sets
	0, 1, 2 Rubric	80% on two of three sets	96% on two of three sets
	0, 1, 2, 3 Rubric	70% on two of three sets	96% on two of three sets
	0, 1, 2, 3, 4 Rubric	70% on two of three sets	95% on two of three sets

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

U.S. History and Grades 3-8 Social Studies

Course and Item Type	Qualifying Standard
Grades 3-8 Social Studies and U.S. History 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.
Grade 3 Social Studies 4 point, single dimension ER (Content & Claims)	Scorers must qualify with 70% exact agreement or higher in the Content & Claims trait on one or more of the qualifying sets in order to score student responses.
Grades 4-8 Social Studies and U.S. History 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.

Spring 2018 Scoring Plan

The charts below provide an overview of the Spring 2018 LEAP 2025 and EOC scoring plan, detailing the types of scoring that will be done for each course/grade.

LEAP 2025 High School and EOC

Course	Handscoring Only	AI Scoring	AI Vendor
LEAP 2025 English I	PCRs (AE form)	PCRs (operational)	Pearson
LEAP 2025 English II	PCRs (AE form)	PCRs (operational)	Pearson
EOC English III	Writing Prompt (AE form)	Writing Prompt (operational)	Measurement Inc.
LEAP 2025 Algebra I	All CRs	N/A	
LEAP 2025 Geometry	All CRs	N/A	
EOC U.S. History	ER (AE form)	ER (operational)	Measurement Inc.
LEAP 2025 U.S. History	All CRs, ER (AE form)	ERs (operational)	Measurement Inc.
EOC Biology	ERs (operational and AE form)	N/A	
Note: All Administrative Error [AE] form items are handscored by DRC scoring supervisors.			

LEAP 2025 Grades 3-8

Course	Handscoring Only	AI Scoring	AI Vendor
ELA grade 3	Both PCRs	N/A	
ELA grade 4	Both PCRs	N/A	
ELA grade 5	RST_C1381	NWT_CC434542756	Pearson
ELA grade 6	RST_VH082653	NWT_D1466	Pearson
ELA grade 7	N/A	Both PCRs	Pearson
ELA grade 8	RST_5013	LAT_F1460	Pearson
Social Studies grades 3 and 4	All CRs and ERs	N/A	
Social Studies grades 5-8	All CRs	All ERs	Measurement Inc.
Math grades 3-8	All CRs	N/A	

Handscoring Rules

AI Scoring

For EOC English III, EOC U.S. History, and the LEAP 2025 U.S. History and grades 5-8 Social Studies ER items, Measurement Incorporated's (MI) Project Essay Grade (PEG) AI scoring system will provide the first score (the score of record). For LEAP 2025 English I, English II, and select PCRs in grades 5-8 ELA (see Spring 2018 Scoring Plan), 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. (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 2018 Scoring Plan above) will be provided by DRC's handscoring team. The first score will be the score of record. Ten percent of the responses will be scored twice to monitor and maintain inter-rater reliability. Scoring Directors will review all nonscores and alerts.

In addition, per PARCC/Pearson rules for ELA and math, if the first two scores are nonadjacent (e.g., 0, 2), a third, independent reading by a Team Leader or Scoring Director will be conducted for additional quality control monitoring. In the unlikely event that a response receives three nonadjacent scores (i.e., 0, 2, 4), a Scoring Director or Project Manager will review the response and provide retraining as needed.

Calculating the Final Score:

- The first score is always the score of record, regardless of how many subsequent scores are applied.
- After handscoring, when the final score processing for the ELA items takes place, the Written Expression trait score is multiplied by 3 (for the Narrative Writing Task). The Reading Comprehension and Written Expression (RCWE) trait score is multiplied by 4 (for the Literary Analysis and Research Simulation tasks), and one fourth of this weighted score will be assigned as the Reading Comprehension score, and three fourths of this weighted score will be assigned as the Written Expression score. The Knowledge of Language and Conventions score is not weighted. (For more information, please see section *2.7.3 Scoring Rules* within the *LEAP 2025 High School and EOC Processing Rules – Scoring* document.)

Reader Monitoring Procedures

Team Leader Read-Behinds

Throughout the handscoring process, DRC Project Managers, Scoring Directors, and Team Leaders will review the statistics that are generated on a daily basis. DRC will assign one Team Leader for every 10–12 readers. If scoring patterns are apparent among individual scorers, Team Leaders will handle these issues on an individual basis. If a scorer appears to need clarification of the scoring rules, DRC supervisors typically monitor one out of five of the scorer's readings, making adjustments to that ratio as needed. If a supervisor disagrees with a reader's scores during monitoring, he or she will correct the score and provide retraining in the form of direct feedback to the reader, using rubric language and applicable training responses. The supervisor's corrected score becomes the score of record; it is not a second read.

DRC will also monitor the inter-rater reliability, which is to be based on the 10% of responses that receive second reads. If a scorer falls below the expected rate of agreement, the Team Leader or Scoring Director will retrain the scorer. If a scorer fails to improve after retraining and feedback, DRC will remove the scorer from the project. In this situation, DRC will remove all unreported scores that were assigned by the scorer during the period in question. These unreported responses with dropped scores will then be re-dealt and rescored.

Validity Sets and Inter-Rater Reliability

In addition to the feedback that supervisors provide to readers based on regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC will also conduct validity scoring. For LEAP 2025 Algebra I, Geometry, English I, English II, and grades 3-8 math and ELA, DRC will utilize the same validity responses that Pearson is using. These PARCC-approved validity responses were supplied by PARCC and may include both paper- and computer-based test responses.

Since the previous vendor of the EOC items did not provide validity responses, and because the Biology item is operational at DRC for the first time in the spring of 2018, DRC's Scoring Directors and Team Leaders will need to identify validity responses for Biology during live scoring. DRC scoring supervisors will also identify validity responses during live scoring for all newly operational LEAP 2025 U.S. History and grades 3-8 Social Studies ERs and CRs. DRC Education Project Management staff will post these validity responses and their scores onto the secure LDOE SFTP site for LDOE content staff to review and approve. Validity responses previously identified and approved for English III will be reused.

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 room drift as well as individual scorer drift. This data will be used to make decisions regarding the retraining and/or release of scorers, as well as the rescoring of responses.

To monitor inter-rater reliability, DRC will produce handscoring quality control reports on a daily basis (see the sample on page 22) 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). For English III and the U.S. History and grades 5-8 Social Studies ERs, MI's PEG AI scoring system will provide the first scores (the scores of record). For LEAP 2025 English I, English II, and select PCRs in grades 5-8 ELA (see Spring 2018 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 25. 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 inter-rater and validity exact agreement percentages based on those of previous administrations for the same items. These rates should be at or above the rates specified below for each content area. When a reader's validity or inter-rater agreement falls 5% or more below these expectations, or when non-adjacent percentages exceed 5%, 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 need to 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 Biology, English III, U.S. History, and grades 3-8 Social Studies items are shown below:

Agreement Rate Requirements for Validity and Inter-Rater Reliability – Biology, English III, U.S. History, and Grades 3-8 Social Studies			
Subject	**Score Point Range**	**Perfect Agreement**	**Perfect Agreement + Adjacent**
EOC Biology	0-4	80%	95%
EOC English III (Style and Content)	1-4 (each domain)	70%	95%
EOC English III (F, U, M, S)	0-1 (each domain)	80%	95%
EOC U.S. History 4 point ERs	0-4	70%	95%
LEAP 2025 U.S. History, Grades 4-8 Social Studies 8 point, 2 dimension ERs	0-4 (each trait)	70%	95%
LEAP 2025 Grade 3 Social Studies 4 point ERs	0-4	70%	95%
LEAP 2025 U.S. History, Grades 3-8 Social Studies 2 point CRs	0-2	80%	95%

The validity and inter-rater reliability expectations for PARCC mathematics and English items (provided by Pearson) are shown in the following table:

Agreement Rate Requirements for Validity and Inter-Rater Reliability – Algebra I, Geometry, English I, English II, and Grades 3-8 ELA (PARCC Items)			
Subject	**Score Point Range**	**Perfect Agreement**	**Perfect Agreement + Adjacent**
Algebra I, Geometry, Grades 3-8 Math	0-1	90%	95%
Algebra I, Geometry, Grades 3-8 Math	0-2	80%	95%
Algebra I, Geometry, Grades 3-8 Math	0-3	70%	95%
Algebra I, Geometry, Grades 3-8 Math	0-4	65%	95%
English I, English II, Grades 3-8 ELA	Multi-trait	65% (each trait)	96% (each trait)

Each reader will be required to maintain an acceptable level of exact agreement on validity responses and on inter-rater reliability as described above. Additionally, readers will be required 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" on the tables 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 are chosen by DRC scoring supervisors during live scoring (for English III, Biology, U.S. History, and grades 3-8 Social Studies items) or supplied by Pearson (for Algebra I, Geometry, English I, English II, and grades 3-8 ELA and math). After the readers take one of these calibration sets (usually 5-10 responses), the Scoring Director will review it from the front of the room using rubric language and the anchor responses to explain the reasoning behind each response's score. These sets do not have a passing requirement but are designed to help refocus readers on how to properly use the scoring guidelines to score responses. The Scoring Director or Team Leaders will provide individual feedback to any scorers in need of additional clarification based on their performance.

Handscoring Quality Control Reports

Scoring Summary Report Sample

EOC Biology Q000

Totals	Inter-rater Reliability				Score Point Distribution									
	2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%4	%B	%F	%R	%U
Current Handscore	**11,834**	**87**	**12**	**1**	**58,111**	**47**	**32**	**12**	**6**	**3**	**0**	**0**	**0**	**0**
7435	500	84	14	1	2,437	42	34	11	7	6	0	0	0	0
7443	512	87	13	0	2,751	50	37	10	3	0	0	0	0	0
7762	624	88	11	1	2,968	51	26	7	4	2	0	0	0	0
8103	812	87	13	0	4,237	47	30	10	7	6	0	0	0	0
8339	447	88	11	1	2,410	52	32	9	6	1	0	0	0	0

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

G05 Writing Q00 **Conventions**

Totals	Inter-rater Reliability				Score Point Distribution										
	2X	%EX	%AD	%NA	Total	%0	%1	%2	%3	%B	%F	%N	%R	%T	%U
Current Handscore	**11,704**	**68**	**30**	**1**	**61,310**	**34**	**38**	**23**	**4**	**0**	**0**	**0**	**0**	**0**	**0**
3	5,852	68	30	1	55,458	33	39	24	4	0	0	0	0	0	0
27188	6	100	0	0	6	67	33	0	0	0	0	0	0	0	0
61011	383	68	34	1	383	39	40	18	3	0	0	0	0	0	0
64018	503	69	31	1	503	36	37	22	5	0	0	0	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				●			●		

In addition to the Reader Feedback Logs, DRC will continue to provide the LDOE with handscoring quality control reports (the same cumulative scoring reports that we have provided in the past; samples are provided on page 22). The Scoring Summary reports show inter-rater reliability data and score point distribution information for each item (by part where appropriate).

Handling Unusual Responses

Nonscore Codes and Definitions

Responses that cannot be assigned a score based on the rubric will be assigned a nonscore code. When readers apply nonscore codes, the responses are automatically routed to DRC handscoring supervisors for validation. Responses that receive a nonscore code count as zero points toward student scores that display on reports. The nonscore code will display in the response string that is included in the file provided to the LDOE.

If readers suspect plagiarism but have no concrete evidence, they score the response and alert it for suspected plagiarism. These responses are sent to supervisors for additional investigation. When supervisors find evidence of student-student plagiarism, each of the associated responses is scored according to rubric requirements and processed as an alert. Responses with proven student-internet

plagiarism receive a score of 0 and are also processed as alerts. If supervisors cannot find proof of plagiarism, the scores are left as is and the response is sent on as an alert (see *Alerts* section below).

The nonscore codes that will be used for Spring 2018 scoring are the same codes that were established and used during previous administrations. The non-score codes and the grades/courses to which they apply are described below:

Nonscore Code Definitions

Nonscore Code	Explanation
B	Blank/no response
F	Response is not written in English (Math responses from Spanish forms will be scored by a Spanish-qualified math scorer.)
I	Insufficient amount of original writing to score
N	Don't understand/know
R	Refusal to respond
T	Off-topic
U	Incoherent, unintelligible, or undecipherable

Nonscore Codes by Grade/Course

Grade/Course	B	F	I	N	R	T	U
LEAP 2025 Algebra I, English I, English II, Geometry, and grades 3-8 ELA and math	✓	✓	N/A	✓	✓	✓	✓
EOC Biology	✓	✓	N/A	N/A	✓	N/A	✓
EOC English III	✓	✓	✓	N/A	N/A	N/A	✓
LEAP 2025 and EOC U.S. History, and grades 3-8 social studies ERs	✓	✓	✓	N/A	N/A	N/A	✓
LEAP 2025 and EOC U.S. History, and grades 3-8 social studies CRs	✓	N/A	N/A	N/A	N/A	N/A	N/A

Alerts

Scorers have the ability to apply an alert flag to specific student responses. These are responses that may indicate the possibility of teacher interference, plagiarism, or disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). After setting the alert flag, which states the reason for the alert, and providing a brief description (as necessary), the reader will score the response according to the specific scoring guidelines for that item.

Likewise, PEG and IEA have the ability to detect specific alerts (described in detail later in the *Artificial Intelligence Scoring* section of this document). All alerted responses (whether identified by a human reader or by AI) are automatically routed to the Scoring Director who reviews the score and forwards

appropriate responses (including grade, subject, 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. 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 student extended responses to items in English III, U.S. History, and grades 5-8 Social Studies. Pearson's Intelligent Essay Assessor (IEA) is used to score student responses in English I, English II, and grades 5-8 ELA. Both MI and Pearson are recognized leaders in the field of performance assessment scoring (human scoring) and automated essay scoring.

AI Scoring – Measurement, Inc.

The items in the following table will be AI scored by MI during the Spring 2018 administration. AI scoring models were built for each of these items by MI in the fall of 2016 or fall of 2017 and followed the model-building process described below. (Model-building data for all items included on the spring 2018 test may be found in the Appendix.)

Course	Item Type	IDEAS ID
EOC English III	Writing Prompt	851368
EOC U.S. History	ER	848632
LEAP 2025 U.S. History	ER	894256
LEAP 2025 U.S. History	ER	894104
LEAP 2025 U.S. History	ER	892955
LEAP 2025 U.S. History	ER	894047
LEAP 2025 Grade 5 Social Studies	ER	807728
LEAP 2025 Grade 6 Social Studies	ER	804944
LEAP 2025 Grade 7 Social Studies	ER	805569
LEAP 2025 Grade 8 Social Studies	ER	809006

Model Building

For each model built, PEG analyzed a set of inputs that were randomly pulled from the training set itself, which is made up of approximately 2,500 examples of student field test responses scored by expert human scorers. Specifically, the training set was divided into two independent pieces:

- One set of response data was used to train the AI engine and produce the scoring model. This attributed to 85% of the training set (~2,125 responses).
- The remaining 15% of the training set (~375 responses) was then used to validate the resulting model.

A regression model was built by choosing a set of variables (e.g., grammar, punctuation, style, etc.) and using least squares Linear Regression to find a best-fit relationship based on the training set. An algorithm chose the initial set of variables and added to the set as needed to produce a good fit, by taking into account correlation statistics and multicollinearity. Once the model was built, it was then run against the validation set, so that it could be evaluated for accuracy. Training was complete once PEG's validation set scores agreed with the human scores; however, if this level of accuracy was not met, then further iterations of training (which may involve new parameterizations or new algorithms) were used to produce a different model with higher accuracy. This process was completed for each trait that needed to be scored.

To further understand the importance of the validation set, consider that one of the risks inherent in machine learning is over-fitting the data. This means that it is possible to home in on particular elements of the responses in training data in such a way that the model does not generalize well to unseen data. To mitigate this risk, PEG uses a hold-out validation strategy[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, so, is complete for any model in production. The essential element of the process is that the models are trained on a larger subset of the training sample (approximately 85%), then validated against an entirely separate smaller subset of the training sample (approximately 15%). What is critical about this process and all validation schemes used in PEG training is that the AI's agreement is always based upon samples the AI has not encountered during training. Put another way, the samples used to train are never the same as the samples used to validate. This maximizes generalizability and minimizes the chance for over-fitting.

---

[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.

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).

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, with the exception of the binary (i.e. zero-or-one) scores for some English traits. Analysis in this case showed not only that human-human quadratic weighted kappas in the training set were low, but, more to point, that random sets of such binary scoring showed similarly low quadratic weighted kappas. In this case, the low quadratic weighted kappa was simply an artifact of the definition of quadratic weighted kappa itself and no further R1 vs. R2 analysis was necessary.

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 prompt/item, and a FERPA-compliant student ID that uniquely identifies either the student or the student-test combination. The table on page 4 of the "DRC – Streaming Scoring" document (see Appendix) also contains information on identifiers.

When PEG receives the file, it processes the batch of responses and records the scores. Each record is specific to a student-test-item combination and will contain the item's score or a reason why it could not be scored (most commonly because the response is too short, or does not contain English). After the batch is processed, the scored records will be returned to DRC for reporting.

DRC will send files to MI daily. Scored files will typically be returned to DRC in 2 to 3 days; however, these timeframes are not definite, because they are dependent on numerous variables involved (e.g. number of responses submitted, number of different items, number of traits per item, the average response length, the standard deviation of response lengths, number of unique words submitted in each response, etc.).

Regardless of whether responses are scored by humans or machines, it is inevitable that scoring anomalies requiring human intervention will occur. Built into MI's automated scoring engine are a variety of triggers for identifying alert papers and responses in which it has low confidence. This is detailed later under "Identifying Responses for Human Review."

Quality Control of the AI Engine

The guidelines below are purposefully general as they have proven to be the best practice for training the PEG engine. The PEG team followed this standard procedure in the DRC/Louisiana project and attempted to maximize human-machine quadratic weighted kappa among all holdout sets.

PEG holds out a 15% set of training data for use in validation. This holdout set is not seen by the AI during training. Instead, once training is complete, the holdout set is submitted for test evaluation and PEG's output is compared to the known, human-expert scores. As discussed in "Evaluation Metric" above, the quadratic weighted kappa has proven to be the most valuable agreement metric in PEG's recent history; however, others (e.g., exact, adjacent, and any host of others) are also applicable.

This evaluation was performed along with model building prior to operational scoring, and the results were shared with LDOE and the TAC to demonstrate sufficient scoring accuracy by PEG. For details on these results, please see pages 46-50 in the Appendix.

Once training and model building is complete, the performance of any given model is essentially deterministic (so, for a precise, given input, the output is expected to be identical). The PEG team monitors the services for unexpected events (for instance physical damage to its cloud infrastructure), and handles any data flow issues (for instance, if the client was using a different item number during live scoring than was used during training) but the AI itself does not change during live scoring. When read-behind data becomes available to the PEG team (typically this is on an annual basis), it can be used to re-evaluate and, if necessary, retrain the existing models prior to the next season of use, but such changes do not happen during live scoring. As part of our continuous improvement cycle, the analysis of this data is on-going with no current end date (i.e., items are being reviewed on a rolling basis).

Identifying Responses for Human Review

Built into MI's automated scoring engine are a variety of triggers for identifying responses that require human review, including potential alerts (suspected plagiarism included) and potential nonscoreable

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 31 for detailed information about these custom thresholds.)

Please note that all responses that are identified in the sections below for human review will be automatically forwarded to a DRC Scoring Director who will determine the correct score or nonscore code to apply to the response. The Scoring Director will provide the final, reported score (or nonscore) for these responses. If the Scoring Director needs assistance in determining the correct score or nonscore, DRC will work with LDOE content staff to ensure that the response is scored correctly.

Alert Detection System

PEG has a robust system for detecting potential alerts, which is described in detail in this section. When PEG detects the presence of alert language, this alone does not indicate that a response is unscorable. Therefore, unless the response is unscorable for some other reason, PEG will return scores as well as the alert status code of 500 (in cases of unscorable alerts, the status code is in the range of 501-599, inclusive). Regardless of the alert flag, any responses returned with a flag to DRC will be evaluated by the handscoring supervisory team who will determine if the response needs to be processed as an alert as described previously in this document (see *Handling Unusual Responses – Alerts*). When it is concluded that a response does warrant an alert, DRC Project Management will contact the LDOE with the student's LASID and post the response information to the SFTP site for LDOE's review.

PEG's Alert flagging system is a pattern-matching system, targeting phrases suggestive of violence towards self or others, drug or alcohol abuse, feelings of anxiety or depression or the use of weapons. This system is rules-based. It responds to concentrations of "alert language" detected within submissions. Typically, these are word counts of particularly violent or profane language often found in actionable alerts. (Such language may also be found in non-alert submissions, but PEG does not attempt to determine "intent" in these cases, rather it flags only the presence of detected verbiage.) PEG currently tracks two types of alert language that differ only in severity (e.g., a statement regarding a person "killing" is considered more severe than a statement regarding a person "beating up," but both are counted as forms of alert language). By default, PEG issues an alert flag if it encounters one instance of severe alert language or two instances of less-severe language. PEG may also issue an alert flag if high counts of profanity are found. By default, this is three instances of severely profane or five instances of less profane verbiage. Although this means that non-actionable alerts may also certainly be flagged, PEG's default settings are purposefully kept highly sensitive to alert language. These levels are configurable, however, so if the rate of return is too high or too low, adjustments can be made. For the responses that it cannot score, PEG returns a condition code to the test delivery system indicating why the response could not be scored (i.e., the response receives a tentative nonscore code that is reviewed by a Scoring Director and corrected if needed). The test delivery system can then route the flagged responses to DRC's performance assessment handscoring system. DRC will perform human handscoring for the limited number of responses that cannot be scored by AI.

With regards to the process and timing, the alerts detection is typically run in series with other essay analysis, so it is no slower (or faster) than a regular scoring. A batch of individually identified extended responses are posted to PEG's Streaming Scoring service, and at that point a response may be flagged as a potential alert. This flag takes the form of a "status code."

The rules are purposefully over-sensitive (they are more likely to give false positives than false negatives), so it is likely that the great majority of ER's flagged with a "5##" status code will not require actual intervention with the student; however, PEG is in no way capable of diagnosing this. Instead PEG just follows rules designed to sense and flag the use of language which has, in the past, been associated with alerts.

Identification of Nonscoreables

PEG's nonscoreable 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 non-scorable parameters which can be configured in this way. Each non-scorable 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 control 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 predominately 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 there is 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, English III responses having an insufficient amount

of original writing to score will receive a nonscore of "I." For U.S. History, any response having an insufficient amount of original writing to score, because it is made up entirely or almost entirely of text copied from the directions or reference sources, will receive a score of 0 in the case of EOC ERs, or in the case of LEAP 2025 ERs, a 0 in Content and a 0 in Claims (unless the item-specific rubric makes exceptions for the use of relevant copied text).

Text that a student extracts and uses from a source external to the test itself is considered plagiarized. When PEG detects these responses, 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 English III, this is the prompt and any associated reading material provided with each test item. For the U.S. History and Social Studies ERs, this includes the prompt and any appropriate source material (including MC/MS items) provided with each test item. In addition to external sources of plagiarism previously provided by LDOE based on results from past administrations, DRC will pre-identify other websites that may be likely sources of external plagiarism. These may include Wikipedia's pages relevant to the topic and/or other "top hit" websites. These external sources will be used by the AI engine to identify potentially plagiarized responses. All of these text references will be added to the appropriate scoring models for each related item.

Upon receiving a response, PEG conducts a high-speed sequence scan of both the reference text and the response. Each sequence is evaluated for both the length and density of copied/plagiarized text. Length is a direct character count, and density is a measure of similarity between sequences. A verbatim copy has a density of 1.0, and a copy that contains some substitutions, additions, or deletions would likely have a density in the ~0.6 - 0.4 range. The product of these two numbers provides a value that is used to flag responses requiring human review due to large amounts of copied/plagiarized text. Clients can configure two thresholds for a low and high flag. For example, the default values for these are 50 and 100 respectively. So, a verbatim copy of 72 characters (~12 prompt words) would be reported as a low match, whereas a verbatim copy of 100 characters (roughly 16 words) would be flagged as a high match. Similarly, a copy (even with some substitutions) of 40 words would still be reported as a high match in the default setting example. The low and high matches will be flagged with status codes. This is similar to the alert flagging above. There will be a three-digit code for low-match (status code 605) and a three-digit code for high-match (status code 610).

Custom thresholds for copied text, plagiarism, and insufficient responses have been established by DRC in consultation with LDOE and were based on recommendations from MI. They are described below:

1. When PEG scans responses for copied text/plagiarism, any text copied from the supplied reference texts (regardless of whether it is contained within quotations marks) will be considered when determining if a response meets or exceeds the thresholds required for it to be routed to DRC for human review. These threshold configurations are noted in 2a–4b below.

2. English III
    a. Copied text thresholds
        i. Low flag (status 605) – 125 characters
        ii. High flag (status 610) – 200 characters
    b. MIN_WORDS (status 200) – 45 words or fewer

3. Grades 5-8 Social Studies
    a. Copied text thresholds
        i. Low flag (status 605) – 125 characters
        ii. High flag (status 610) – 200 characters
    b. MIN_WORDS (status 200) – 25 words or fewer

4. 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 original numeric scores and apply the appropriate numeric scores or nonscore as necessary, depending on the content area scoring rules. For Social Studies and U.S. History, flagged responses composed entirely of text copied from item source material (or all copied text combined with an insufficient amount of original student work) are scored as 0s. For English III, such responses are given a nonscore of "I" (Insufficient).

Less frequently, responses will be flagged as potential nonscores for having too little written to be evaluated at all (status code 200). Just as DRC requires all nonscores given by human readers to be reviewed by scoring supervisors, this same requirement holds true when PEG flags responses as potential nonscores. For example, if the DRC supervisor reviews a response flagged by PEG and agrees with PEG's assessment that the response has too little writing to be assessed, the supervisor will validate the AI score of "I," and this nonscore code will be submitted as the final score for that response. On the other hand, if DRC's supervisor reviews the response, and based on the training responses provided in the handscoring training materials, he or she feels that 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 2018 LEAP 2025 administration. AI scoring models for each of these items were previously built and used in operational scoring by Pearson during PARCC operational scoring in the spring of 2017. (Model-building data for all items included on the spring 2018 test may be found in the Appendix.)

Course	Item Type	IDEAS ID	PARCC UIN
English I	PCR	902152	LAT_VH017536
English I	PCR	902161	RST_VH017542
English II	PCR	906197	LAT_HH428127697
English II	PCR	902331	RST_VH004490
Grade 5 ELA	PCR	913663	NWT_CC434542756
Grade 6 ELA	PCR	913694	NWT_1466
Grade 7 ELA	PCR	868704	RST_VH069361
Grade 7 ELA	PCR	913842	NWT_EE430133306
Grade 8 ELA	PCR	913958	LAT_F1460

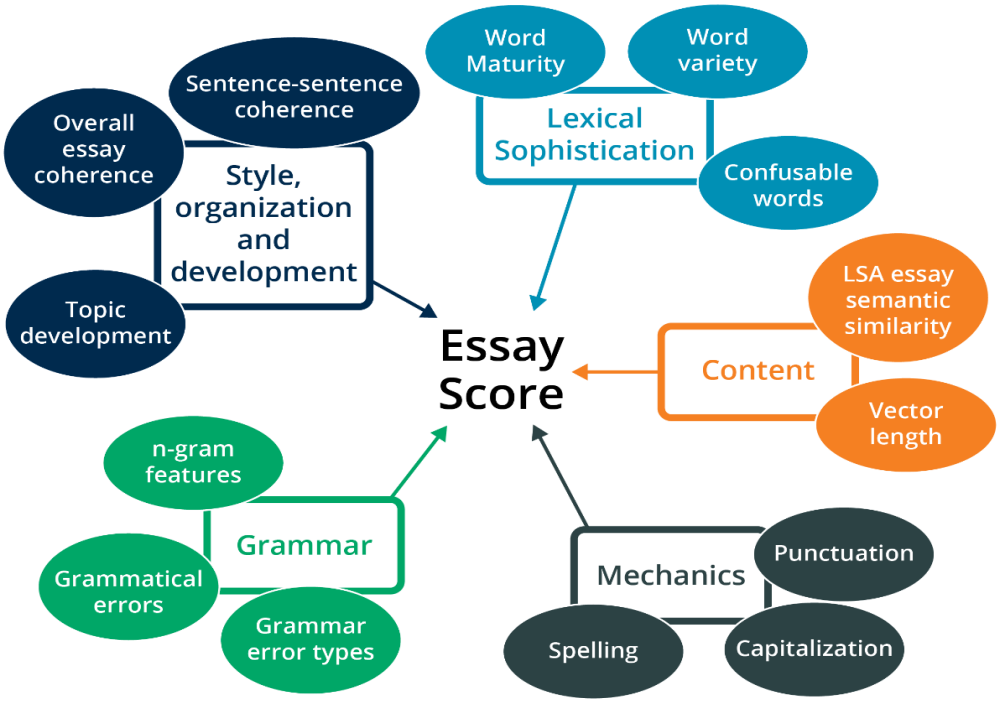
The Intelligent Essay Assessor

Pearson's Intelligent Essay Assessor (IEA) uses a range of machine learning and natural language processing technologies to learn to score based on human-scored responses. One of the hallmarks of IEA is its ability to score constructed responses in content domains using Pearson's unique implementation of Latent Semantic Analysis (LSA), an approach that generates semantic similarity of words and passages by analyzing large bodies of relevant text. LSA can then "understand" the meaning of text much the same as a human scorer.

IEA's background knowledge of English is derived from a collection of texts equivalent to what students are likely to have encountered over the course of their academic career (about 12 million words). Because LSA operates over the semantic representation of texts, rather than at the individual word level, it can evaluate similarity even when texts have few words in common. For example, LSA finds the following two sentences to have a high degree of semantic similarity:

- Surgery is often performed by a team of doctors.
- On many occasions, several physicians are involved in an operation.

The following figure illustrates some of the features used in IEA and how they relate to specific constructs of student writing performance.



Example features used in the Intelligent Essay Assessor. *Like human scorers, IEA evaluates essays for ideas, organization, development, and various grammatical and mechanics errors.*

IEA is trained to associate features extracted from each essay with scores assigned by human scorers. A machine learning-based approach is used to determine the optimal set of features, and the weights for each of those features, to best model the scores for each essay. From these comparisons, IEA derives a prompt- and trait-specific scoring model that predicts the scores human scorers would assign to any new responses.

The automated scoring process mimics the approach that human scorers take when evaluating essays. Human scorers train based on anchors of annotated student responses with agreed-upon scores. Human scorers compare new responses against the anchor set of two to three examples per score point to determine the appropriate score. IEA scores essays similarly, but makes comparisons against a much larger set of examples. Rather than comparing a new essay against the 16-24 examples in an anchor set, it compares against the set of hundreds or thousands of responses on which it was trained.

How the Intelligent Essay Assessor was Trained

For the ELA prompts that will be used by Louisiana, IEA was trained based on operational PARCC responses using Pearson's Continuous Flow approach to training and scoring. When these prompts were first administered, student responses flowed to IEA even before human scoring started. IEA then selected a sample of responses for humans to score first to expedite the creation of automated scoring models. The sample included responses that represented different demographic subgroups to ensure

equity in scoring, as well as responses that were algorithmically selected to likely span the score range. As the human-scored responses flowed back to IEA, the engine automatically built potential scoring models, evaluating them against the industry standards for performance criteria included in the table below.

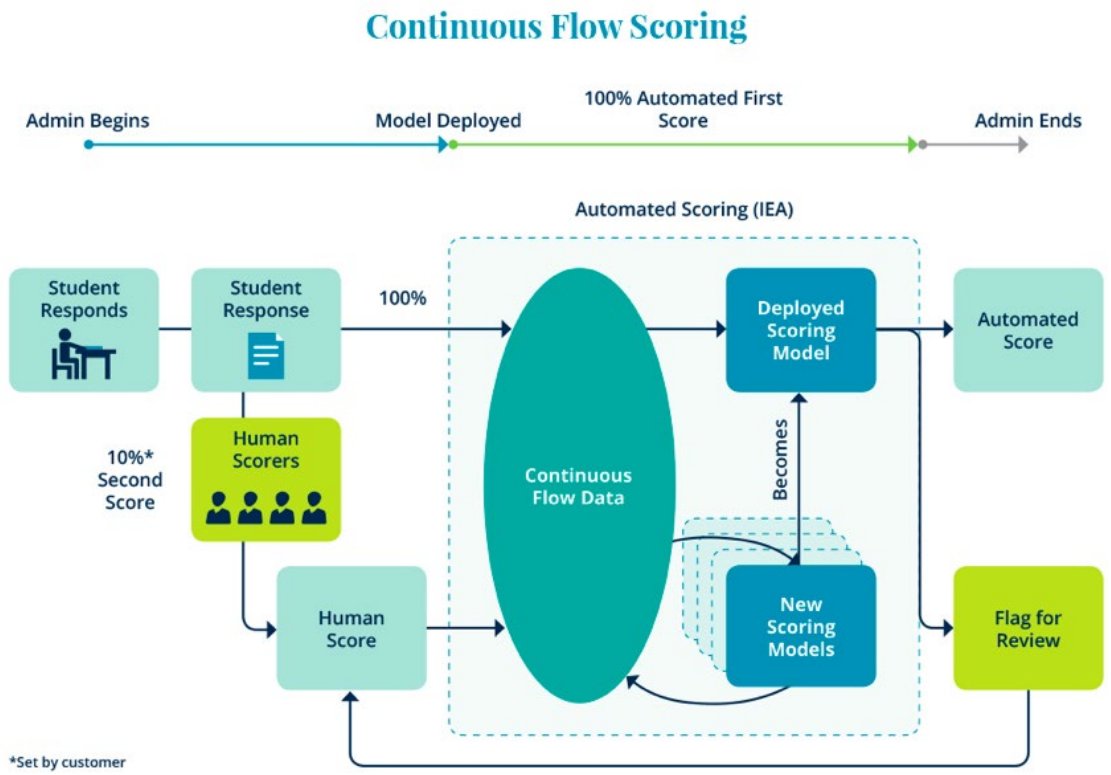
Evaluation of Automated Scoring Systems	
Criterion	**Threshold**
Quadratic weighted kappa (QWK)	Greater than or equal to 0.70
Pearson correlation (r)	Greater than or equal to 0.70
Standardized mean difference (SMD) between human and automated scoring	Less than or equal to|0.15|
Difference in QWK or r from human-human rates	Less than or equal to 0.10
Difference in exact agreement from human-human rates	Less than or equal to 0.05

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

While the engine was being trained, scoring and psychometrics teams met daily to review progress, quality, and next steps. When IEA met or exceeded the performance criteria for a given constructed response item, it took over as the first scorer for that item.

Responses for which IEA is less confident in its score are routed for additional human scoring. This "smart routing" of responses by the scoring engine occurs when responses fall in a particular score range for which the engine has lower agreement with human scorers, or for responses that are highly unusual or creative.

The figure 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. Child in danger alerts are also scanned for out of band and flagged for human review.

Quality Monitoring

Human scorers play a key role in maintaining quality throughout the scoring process starting with IEA learning to score based on their scores. Since the models for the 2018 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, Team Leaders 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 subsets of student responses. Time required for alterations of the AI models varies based upon the extent of the necessary correction. A configuration change (for instance in a non-scorable parameter) can typically be deployed within 1-2 days. A complete model retraining (for instance, if it were discovered that all original human training data were invalid) could require 1-2 weeks from receipt of the new training data.

Rescores (High School Tests Only)

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. If the rescore results in a final score that falls in the *Fair* achievement level for EOC tests or the Approaching Basic achievement level for LEAP 2025 HS tests, the rescore results are sent to LDOE content specialists for review. If LDOE agrees with the rescore, the original score is replaced with the rescore and the score report is produced and sent. Note: The score reports are always held until the rescore process is complete.

Auto Rescores

In Window: Automatic rescores will occur per the criteria listed below when all sessions for a test have been completed.

Post Window: Automatic rescores will occur per the criteria listed below regardless of the number of sessions completed.

Biology, U.S. History (EOC)

Students are eligible for an automatic rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade = '12'
2	Subject Scale Score within 10 points below the Fair (F) Achievement Level cut score

Algebra I, Geometry, U.S. History (LEAP 2025)

Students are eligible for an automatic rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade = '12'
2	Subject Scale Score within 10 points below the Approaching Basic (APP) Achievement Level cut score

English III

Students are eligible for an automatic rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade = '12'
2	Subject Scale Score within 20 points below the Fair (F) Achievement Level cut score

English I and English II

Students are eligible for an automatic rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade = '12'
2	Subject Scale Score within 20 points below the Approaching Basic (APP) Achievement Level cut score

Students that are automatically rescored will be flagged on the LDESTD file as "<Subject> Rescore Flag" = "P" while the rescore is being processed and "R" after the rescore has been completed. Refer to the **Rescore Flag** section of the LEAP 2025 High School and EOC Processing Rules - Scoring document for the rules to set the Rescore Flag. The original scale score will be maintained in the student record for all rescores that did not result in an achievement level change. The score, original or rescore, approved by LDOE after reviewing the Rescore File will be maintained on the student record for all rescores that resulted in an achievement level change.

All tests that require an automatic rescore will be held back from reporting until the rescore process is either complete with no achievement level change or the rescore process resulted in an achievement level change which was reviewed and approved by LDOE.

Students whose scores have been voided are not eligible for an automatic rescore.

Parent-Requested Rescores

Post Window: Parent-requested rescores are processed by DRC project management.

Biology, U.S. History (EOC)

Students are eligible for a parent-requested rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade <> '12'
2	Subject Scale Score within 10 points below the Fair (F) Achievement Level cut score

Algebra I, Geometry, U.S. History (LEAP 2025)

Students are eligible for a parent-requested rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade <> '12'
2	Subject Scale Score within 10 points below the Approaching Basic(APP) Achievement Level cut score

English III

Students are eligible for a parent-requested rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade <> '12'
2	Subject Scale Score within 20 points below the Fair (F) Achievement Level cut score

English I and English II

Students are eligible for a parent-requested rescore when both of the following are true for the subject:

Ref. #	Rule/Requirement
1	Reporting Grade <> '12'
2	Subject Scale Score within 20 points below the Approaching Basic (APP) Achievement Level cut score

Student responses that are rescored will be flagged on the LDESTD file as "<Subject> Rescore Flag" = "P" while the rescore is being processed and "R" after the rescore has completed. Refer to the **Rescore Flag** section of the LEAP 2025 High School and EOC Processing Rules - Scoring document for the rules to set the Rescore Flag. The original scale score will be maintained in the student record for all rescores that did not result in an achievement level change. The score, original or rescored, approved by LDOE after reviewing the Rescore File will be maintained on the student record for all rescores that resulted in an achievement level change.

Students whose scores have been voided are not eligible for a parent-requested rescore.

Appendix A

DRC-MI Streaming Scoring Documentation

DRC – MI STREAMING SCORING SUBMIT SERVICE DOCUMENTATION

NOTICE: The contents of this document and any references to external resources are intended for review only by representatives of Data Recognition Corporation, Measurement Incorporated, and LDOE, and are considered private. Technical specifications are subject to change.

REVISED: 2015-11-23; *created*

CONTENTS:

SECTION 1 – General Information 43

SECTION 2 – SCHEMA SUPPLEMENT 44-46

SECTION 3 – STATUS CODE INFORMATION 47

SECTION 1 – General Information

1.1 PURPOSE: Submit Service accepts groups ("batches") of constructed responses for processing by the MI Streaming Scoring product.

1.2 SERVICE TYPE: The Submit Service uses a standard SOAP web service interface.

1.3 INTEGRATION: Application-generated service definition (WSDL 1.1) document is available; WCF (Windows Community Foundation) client integration is also possible. The WSDL and WCF URLs for each environment are as follows:

DEVELOPMENT
- **WSDL:**
- **WCF:**

STAGING
- **WSDL:**
- **WCF:**

PRODUCTION:
- **WSDL:**
- **WCF:**

1.4 SERVICE SIGNATURE: The Submit Service provides a single operation **SubmitBatch**. The operation signature – request and response structure – is defined in the WSDL. The structure of each complex type, with field descriptions and expected value ranges is described below.

SECTION 2 – SCHEMA SUPPLEMENT

2.1.1 SUPPLEMENTAL SCHEMA DOCUMENTATION: The following tables are supplemental to the schema for the Submit Service, but are not, themselves, the schema. The service schema is contained within the WSDL, and may be emitted from that source to an XML schema document (XSD) through various means, though this will likely be unnecessary. To reduce confusion in terminology, the following tables will be referred to as the "supplement" or "schema supplement".

2.1.2 TABLE STRUCTURE: Each table documents a specific complex type defined by the Submit Service WSDL, with each row in a table representing a field of that complex type. Column definitions are provided here.

- **Name:** Name of field; note that for complex type fields, the name of the field and the name of the type may, or may not be the same.
- **Type:** Field type; this may be a simple type (string, integer, etc.) or another complex type, which is described in another table.
- **Min:** Minimum expected occurrences (minOccurs). This value 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

AI Model Building – English III (Spring 2018)

Quadratic Weighted Kappa (QWK) and Inter-rater Reliability (IRR)

Course	Form	Item #	Rater	Content				Style				Sentence Formation			Usage			Mechanics			Spelling		
				QWK	EX	ADJ	NON-	QWK	EX	ADJ	NON-	QWK	EX	ADJ	QWK	EX	ADJ	QWK	EX	ADJ	QWK	EX	ADJ
English III	U	851368	H-H*	0.78	65	33	2	0.74	65	34	1	0.26	76	24	0.44	74	26	0.53	77	23	0.47	85	15
			AI-H**		71	29	0		66	33	0		76	24		74	26		82	18		85	15

*Human to human (H-H) inter-rater metrics are from Pacific Metrics EFT scoring.
**Human to AI (AI-H) inter-rater metrics are from the MI 2016 model building results.

Score Point Distribution (SPD)

Course	Form	Item #	Rater	Content				Style				Sentence Formation		Usage		Mechanics		Spelling	
				1%	2%	3%	4%	1%	2%	3%	4%	0%	1%	0%	1%	0%	1%	0%	1%
English III	U	851368	H	12	42	37	9	4	39	46	10	20	79	34	66	26	74	16	83
			AI	12	42	38	8	7	37	46	10	22	78	36	64	26	74	17	83

AI Model Building – EOC U.S. History ER (Spring 2018)

Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD)

Course	Form	Item #	QWK Score Agreement	IRR Comparison	Inter-Rater Agreement %			SPD Group	Score Point Distribution %				
					EX	ADJ	NON-ADJ		0s	1s	2s	3s	4s
USH	E	848632	0.75	Human to Human*	60	35	4	Human	39	42	15	3	1
				AI to Human**	60	37	4	AI	39	41	15	3	2

*Human to human inter-rater metrics are from Pacific Metrics EFT scoring.
**AI to human inter-rater metrics are from the MI 2016 model building results.

AI Model Building – LEAP 2025 U.S. History ERs (Spring 2018)

Quadratic Weighted Kappa (QWK), Inter-rater Reliability (IRR), and Score Point Distribution (SPD)

Course	IDEAS Item #	# of Responses	Content											Claims										
			QWK	Inter-Rater Agreement %				Score Point Distribution %						QWK	Inter-Rater Agreement %				Score Point Distribution %					
				Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s	4s		Comparison	Exact	Adjacent	Nonadjacent	SPD Group	0s	1s	2s	3s	4s
USH	894104	2500	0.86	H to H	62	33	5	Human	31	34	22	9	4	0.84	H to H	61	32	7	Human	39	28	21	9	4
		15%		AI to H	70	29	1	AI	30	38	21	8	3		AI to H	63	36	1	AI	39	28	21	8	3
USH	894256	2500	0.85	H to H	66	32	2	Human	28	37	24	8	3	0.86	H to H	64	33	3	Human	34	33	21	8	3
		15%		AI to H	71	29	1	AI	24	37	29	8	2		AI to H	69	30	1	AI	34	34	22	8	3
USH	892955	2500	0.88	H to H	65	32	3	Human	34	29	25	9	3	0.88	H to H	64	32	4	Human	37	26	25	10	3
		15%		AI to H	74	26	0	AI	31	34	24	9	2		AI to H	72	28	0	AI	37	28	22	10	3
USH	894047	2500	0.86	H to H	66	31	3	Human	39	29	19	9	4	0.86	H to H	66	30	4	Human	40	29	19	9	4
		15%		AI to H	71	28	1	AI	40	30	20	7	4		AI to H	73	26	1	AI	44	29	18	5	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 models were 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 2018)

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	807728	2599	0.87	H to H[1]	76	23	1	Human	61	23	13	3	0	0.87	H to H[1]	78	20	2	Human	66	20	11	2	0
		≈500		H to H[2]	65	32	2	Human	19	36	34	10	1		H to H[2]	67	31	3	Human	25	32	35	8	1
		15%		AI to H	70	29	1	AI	32	32	27	7	1		AI to H	70	29	1	AI	39	29	24	7	1
6	804944	2975	0.85	H to H[1]	73	25	1	Human	61	32	7	1	0	0.83	H to H[1]	75	24	1	Human	64	31	5	1	0
		≈500		H to H[2]	68	29	2	Human	19	42	34	5	1		H to H[2]	69	28	3	Human	23	43	29	4	1
		15%		AI to H	77	23	0	AI	39	39	17	3	0		AI to H	75	24	1	AI	44	39	14	3	1
7	805569	2610	0.89	H to H[1]	68	31	1	Human	32	42	21	4	0	0.90	H to H[1]	66	32	1	Human	35	41	20	4	0
		≈500		H to H[2]	76	24	0	Human	1	30	53	15	1		H to H[2]	74	26	1	Human	4	33	50	13	1
		15%		AI to H	75	25	0	AI	30	30	30	9	1		AI to H	74	26	0	AI	32	31	28	7	1
8	809006	2543	0.84	H to H	63	25	2	Human	27	38	24	8	2	0.86	H to H	62	36	2	Human	33	33	24	8	2
		15%		AI to H	69	31	1	AI	26	38	27	7	3		AI to H	71	29	0	AI	34	33	25	8	1

H to H[1] – Human scored 2016 Field Test sample of ≈ 2500 responses per item.

H to H[2] – Human scored 2018 targeted sample of ≈ 500 responses per item used to augment and retrain the AI models. This sample came from spring 2018 operational responses and was intended to find high score points to add to the existing AI models for the purpose of retraining the models prior to spring 2018 operational scoring.

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.

The existing AI scoring model for grade 8 (809006) from 2016 was deemed acceptable; therefore, no augmentation was necessary for spring 2018.

PCR Performance – ELA Grades 5-8, English I, and English II (Spring 2018)

Reading Comprehension and Written Expression or Written Expression Trait

		IEA-Human Agreement						Human-Human Agreement					
Prompt	**Grade**	**Exact**	**SP0**	**SP1**	**SP2**	**SP3**	**SP4**	**Exact**	**SP0**	**SP1**	**SP2**	**SP3**	**SP4**
E05_N_CC434542756	5	**69.9**	**74.5**	**62.2**	**66.5**	**80.8**		67.5	84.6	56.6	61.6	62.2	
E06_N_D1466	6	**74.4**	88.3	**61.3**	**73.6**	**63.5**	**66.0**	67.3	89.9	55.6	68.5	55.4	54.3
E07_R_VH069361	7	**77.9**	**84.4**	**80.9**	**74.7**	60.4	**68.2**	73.5	81.8	73.2	69.1	61.4	66.7
E07_N_EE430133306	7	**73.0**	**88.2**	59.5	**65.9**	**68.0**	**76.7**	66.1	87.4	63.7	57.6	48.4	68.0
E08_L_F1460	8	**68.7**	**79.6**	**68.7**	**61.6**	**65.0**	**56.3**	59.6	71.4	58.1	56.9	48.3	45.7
VH017542_2T	9	**73.9**	**80.7**	**80.1**	**64.9**	**55.9**	**71.9**	64.6	80.5	70.6	52.9	43.0	58.7
VH017536_2T	9	**68.5**	**78.2**	**75.9**	**60.9**	**40.9**	**70.8**	61.9	75.3	67.0	58.1	31.9	51.6
VH004490	10	**74.4**	81.5	**76.9**	**71.7**	**57.9**	**63.1**	70.1	81.7	70.9	66.9	47.1	53.8
HH428127697	10	**73.6**	**83.5**	**75.4**	**71.8**	**59.8**	**60.5**	64.4	81.3	59.1	65.7	51.0	24.0

- Green indicates IEA agreement higher than human agreement
- Red indicates IEA agreement lower than human agreement by > 5.25%
- Source - Pearson

Conventions Trait

		IEA-Human Agreement					Human-Human Agreement				
Prompt	**Grade**	**Exact**	**SP0**	**SP1**	**SP2**	**SP3**	**Exact**	**SP0**	**SP1**	**SP2**	**SP3**
E05_N_CC434542756	5	**64.9**	78.6	**55.8**	**60.7**	**67.9**	60.3	82.2	53.7	47.6	53.8
E06_N_D1466	6	**77.6**	**88.4**	**70.3**	**73.2**	**76.9**	70.3	85.5	64.3	61.6	67.7
E07_R_VH069361	7	**79.4**	**85.8**	**78.6**	**73.3**	**76.9**	72.2	80.3	67.9	70.2	69.2
E07_N_EE430133306	7	**69.5**	**79.6**	**53.0**	**63.7**	**81.2**	65.5	74.4	62.5	48.9	74.3
E08_L_F1460	8	**70.5**	**76.2**	**69.9**	**64.2**	**73.9**	62.8	65.3	59.2	58.5	72.9
VH017542_2T	9	**74.5**	**80.7**	**74.7**	**57.2**	**83.4**	68.6	80.5	62.9	51.1	74.2
VH017536_2T	9	**67.9**	**70.9**	**69.2**	58.5	**78.2**	62.8	69.4	57.3	60.9	67.6
VH004490	10	**74.9**	**85.1**	**71.9**	**67.9**	**73.4**	67.0	84.3	63.1	58.5	53.2
HH428127697	10	**74.8**	82.7	**72.9**	**72.5**	**72.5**	68.7	84.9	63.6	63.1	63.9

- Green indicates IEA agreement higher than human agreement
- Red indicates IEA agreement lower than human agreement by > 5.25%
- Source - Pearson

Spring 2018 LEAP 2025 and EOC Items – IRR and SPD from Previous Administrations

Algebra I

IDEAS ID	Spring 2018 Form	PARCC UIN	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 %	Report Source
901882	Op, AE	VH196970	9,586	Part A	0,1	1,950	97.7%	99.9%	71.1%	12.5%				16.4%	Pearson Spring 2016
901882	Op, AE	VH196970	9,586	Part B	0,1,2	1,950	89.8%	97.1%	66.2%	6.7%	3.9%			23.1%	Pearson Spring 2016
901882	Op, AE	VH196970	8,522	Part A	0,1	1,940	99.0%	100.0%	94.0%	3.0%				4.0%	DRC Fall 2017, Op
901882	Op, AE	VH196970	8,522	Part B	0,1,2	1,940	99.0%	100.0%	94.0%	2.0%	1.0%			4.0%	DRC Fall 2017, Op
901859	AE	3003-M43111	253,395	Part C	0,1,2,3	48,917	92.1%	99.3%	43.7%	5.4%	14.9%	25.9%		10.1%	Pearson Spring 2016
901859	AE	3003-M43111	8,485	Part C	0,1,2,3	2,504	98.0%	100.0%	73.0%	4.0%	7.0%	13.0%		2.0%	DRC Fall 2017, Op
901832	Op, AE	3031-M44083P	95,907	Part B	0,1,2	18,835	91.3%	99.9%	29.9%	45.1%	11.9%			13.1%	Pearson Spring 2016
901848	AE	M47287	17,146	Overall	0,1,2,3,4	3,335	97.2%	99.6%	70.1%	9.2%	0.9%	0.4%	0.2%	19.1%	Pearson Spring 2016
901848	AE	M47287	8,445	Overall	0,1,2,3,4	2,796	100.0%	100.0%	78.0%	4.0%	0.0%	0.0%	0.0%	17.0%	DRC Fall 2017, Op
906043	Op, AE	2362-M41568	77,509	Part A	0,1,2	14,705	97.4%	99.9%	78.2%	3.4%	5.9%			12.5%	Pearson Spring 2017
906043	Op, AE	2362-M41568	77,509	Part B	0,1,2	14,705	91.2%	99.8%	28.0%	49.8%	4.5%			17.7%	Pearson Spring 2017
901687	Op	2407-M41752	83,197	Part A	0,1,2	15,543	96.7%	99.5%	63.3%	4.3%	23.3%			9.0%	Pearson Spring 2017
901687	Op	2407-M41752	83,197	Part B	0,1,2	15,543	92.6%	99.7%	68.3%	10.6%	7.0%			14.1%	Pearson Spring 2017
901687	Op	2407-M41752	83,197	Part C	0,1,2	15,543	95.5%	99.5%	71.4%	5.8%	4.1%			18.7%	Pearson Spring 2017
901857	Op, AE	VH046479	78,418	Part A	0,1,2	13,963	88.2%	99.8%	51.2%	36.3%	2.9%			9.6%	Pearson Spring 2017
901857	Op, AE	VH046479	78,418	Part B	0,1	13,963	91.8%	99.7%	68.9%	19.0%				12.1%	Pearson Spring 2017
901857	Op, AE	VH046479	8,686	Part A	0,1,2	2,258	94.0%	100.0%	77.0%	13.0%	1.0%			9.0%	DRC Fall 2017, Op
901857	Op, AE	VH046479	8,686	Part B	0,1	2,258	97.0%	100.0%	86.0%	5.0%				9.0%	DRC Fall 2017, Op
901851	Op, AE	M41726	77,612	Comp-rehensive	0,1,2,3	14,776	88.4%	99.1%	52.4%	10.8%	19.3%	9.9%		7.6%	Pearson Spring 2017
901705	Op	VF883359	147,898	Part A	0,1,2,3	29,794	96.8%	99.6%	83.9%	2.5%	1.5%	1.2%		10.9%	Pearson Spring 2015
901705	Op	VF883359	147,898	Part B	0,1	29,794	91.1%	99.8%	68.3%	12.3%				19.4%	Pearson Spring 2015

Geometry

IDEAS ID	Spring 2018 Form	PARCC UIN	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 %	Report Source
902036	AE	2904-M43021	42,708	Part A	0,1,2	8,216	95.9%	99.5%	53.9%	6.8%	33.2%			6.1%	Pearson Spring 2016
902036	AE	2904-M43021	42,708	Part B	0,1,2	8,216	94.7%	99.2%	61.0%	6.5%	24.1%			8.5%	Pearson Spring 2016
902036	AE	2904-M43021	42,708	Part C	0,1,2	8,216	94.9%	98.4%	75.2%	3.5%	4.2%			17.2%	Pearson Spring 2016
902036	AE	2904-M43021	6,800	Part A	0,1,2	1,518	99.0%	100.0%	69.0%	9.0%	20.0%			2.0%	DRC Fall 2017, Op
902036	AE	2904-M43021	6,800	Part B	0,1,2	1,518	96.0%	99.0%	68.0%	10.0%	19.0%			2.0%	DRC Fall 2017, Op
902036	AE	2904-M43021	6,800	Part C	0,1,2	1,518	97.0%	99.0%	91.0%	3.0%	3.0%			2.0%	DRC Fall 2017, Op
902042	Op	3020-M44058	45,304	Part A	0,1,2,3	8,509	94.5%	99.7%	47.9%	29.7%	7.3%	4.1%		11.0%	Pearson Spring 2016
902042	Op	3020-M44058	45,304	Part B	0,1	8,509	96.1%	99.8%	61.4%	21.9%				16.7%	Pearson Spring 2016
902042	Op	3020-M44058	45,304	Part C	0,1,2	8,509	94.8%	97.7%	61.2%	4.7%	12.2%			21.9%	Pearson Spring 2016
902047	AE	VH150404	47,576	Part A	0,1,2	8,713	97.9%	99.6%	64.1%	6.6%	2.8%			26.5%	Pearson Spring 2016
902047	AE	VH150404	47,576	Part B	0,1,2	8,713	92.1%	99.7%	52.6%	19.2%	7.0%			21.2%	Pearson Spring 2016
902047	AE	VH150404	6,775	Part A	0,1,2	1,636	98.0%	100.0%	81.0%	10.0%	4.0%			5.0%	DRC Fall 2017, Op
902047	AE	VH150404	6,775	Part B	0,1,2	1,636	96.0%	100.0%	74.0%	15.0%	6.0%			5.0%	DRC Fall 2017, Op
902012	Op, AE	M41169	90,471	Overall	0,1,2,3	16,723	87.1%	98.7%	46.2%	12.3%	14.6%	7.0%		19.9%	Pearson Spring 2016
902046	Op, AE	M46668	42,630	Overall	0,1,2,3	7,622	92.9%	98.9%	70.0%	8.5%	5.2%	0.5%		15.8%	Pearson Spring 2016
902046	Op, AE	M46668	6,821	Overall	0,1,2,3	1,880	97.0%	100.0%	78.0%	9.0%	3.0%	0.0%		9.0%	DRC Fall 2017, Op
902062	Op	VH150384	2,581	Overall	0,1,2,3,4	542	88.6%	97.4%	56.6%	6.1%	4.1%	1.5%	0.8%	30.9%	Pearson Spring 2016
902018	Op, AE	3042-M44133	48,083	Part A	0,1,2,3	9,072	86.1%	98.2%	45.8%	4.1%	17.4%	17.3%		15.3%	Pearson Spring 2017
902018	Op, AE	3042-M44133	48,083	Part B	0,1,2,3	9,072	91.6%	98.2%	62.9%	5.0%	5.1%	6.4%		20.7%	Pearson Spring 2017
902018	Op, AE	3042-M44133	6,775	Part A	0,1,2,3	1,762	95.0%	100.0%	61.0%	4.0%	13.0%	16.0%		7.0%	DRC Fall 2017, Op
902018	Op, AE	3042-M44133	6,775	Part B	0,1,2,3	1,762	98.0%	2.0%	82.0%	4.0%	4.0%	4.0%		7.0%	DRC Fall 2017, Op
902027	Op, AE	M43233	84,614	Overall	0,1,2,3,4	15,944	88.2%	97.7%	51.8%	12.5%	9.5%	5.0%	5.4%	15.9%	Pearson Spring 2017

Math Grade 3

Grade	IDEAS ID	PARCC UIN	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 %	Report Source
3	914037	4003-M03007	7,194	Part A	0,1,2	668	92.0%	99.7%	34.6%	41.4%	20.8%			3.2%	Pearson Spring 2017
3	914037	4003-M03007	7,194	Part B	0,1,2	668	97.4%	100.0%	83.2%	4.5%	7.0%			5.3%	Pearson Spring 2017
3	914037	4003-M03007	7,194	Part C	0,1,2	668	97.5%	98.2%	0.7%	0.4%	91.5%			7.4%	Pearson Spring 2017
3	914008	VH054861	103,690	Part A	0,1,2	9,200	84.1%	94.7%	42.2%	13.3%	43.5%			1.0%	Pearson Spring 2017
3	914008	VH054861	103,690	Part B	0,1	9,207	94.4%	100.0%	87.0%	11.3%				1.7%	Pearson Spring 2017
3	914039	M02527	7,113	Comprehensive	0,1,2,3	699	92.7%	98.7%	37.9%	30.2%	23.4%	1.7%		6.8%	Pearson Spring 2017
3	914048	M05158	79,640	Comprehensive	0,1,2,3	7,819	92.4%	99.3%	52.8%	18.5%	11.5%	15.7%		1.5%	Pearson Spring 2017
3	798301	VF658085	57,049	Part C	0,1	11,648	87.0%	100.0%	80.0%	18.0%				2.0%	DRC Spring 2017
3	798301	VF658085	57,699	Part C	0,1	11,918	87.0%	100.0%	60.0%	37.0%				2.0%	DRC Spring 2016
3	798301	VF658085	55,234	Part C	0,1	10,885	88.5%	99.8%	78.7%	14.4%				6.9%	Pearson Spring 2015
3	868619	M00848	57,049	Comprehensive	0,1,2,3	11,716	93.0%	99.0%	64.0%	9.0%	5.0%	13.0%		8.0%	DRC Spring 2017

Math Grade 4

Grade	IDEAS ID	PARCC UIN	Responses Available	Trait	Score Points	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Condition Code %	Report Source
4	870722	0220-M00114P	45,036	Part B	0,1,2	5	91.3%	99.6%	82.1%	10.3%	2.6%			5.0%	Pearson Spring 2017
4	870722	0220-M00114P	1,965	Part B	0,1,2	472	100.0%	100.0%	60.0%	41.0%	0.0%			0.0%	DRC 2017 [online]
4	870722	0220-M00114P	54,550	Part B	0,1,2	11,387	95.0%	100.0%	89.0%	7.0%	2.0%			2.0%	DRC 2017 [paper]
4	914084	4112-M03491P	383,723	Part C	0,1,2	37,737	94.9%	99.9%	65.0%	28.5%	2.5%			4.0%	Pearson Spring 2017
4	870317	M02255	191,643	Overall	0,1,2,3	38,068	82.4%	99.1%	46.0%	21.9%	25.6%	5.7%		0.8%	Pearson Spring 2016
4	870317	M02255	56,517	Overall	0,1,2,3	11,474	85.0%	99.0%	43.0%	20.0%	18.0%	14.0%		4.0%	DRC Spring 2017
4	914086	M04133	107,359	Comprehensive	0,1,2,3	10,670	91.1%	99.4%	53.1%	23.9%	7.4%	14.9%		0.7%	Pearson Spring 2017
4	870724	0083-M00453	79,205	Part C	0,1,2,3,4	16,182	94.4%	99.4%	56.7%	18.1%	14.7%	3.6%	2.6%	4.3%	Pearson Spring 2015
4	870723	M02081	79,454	Comprehensive	0,1,2,3	15,743	87.0%	98.1%	58.2%	13.1%	13.5%	10.8%		4.5%	Pearson Spring 2015

Math Grade 5

Grade	IDEAS ID	PARCC UIN	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 %	Report Source
5	870325	4106-M03478	320,279	Part A	0,1	31,015	87.8%	100.0%	34.3%	64.6%				1.1%	Pearson Spring 2015
5	870325	4106-M03478	320,279	Part B	0,1,2	31,015	94.0%	99.8%	73.3%	16.2%	6.1%			4.4%	Pearson Spring 2015
5	870325	4106-M03478	53,396	Part A	0,1	11,104	79.0%	100.0%	45.0%	54.0%				0.0%	DRC Spring 2017
5	870325	4106-M03478	53,396	Part B	0,1,2	11,104	93.0%	100.0%	75.0%	19.0%	5.0%			0.0%	DRC Spring 2017
5	868654	4134-M03661	320,279	Part A	0,1,2,3	31,719	81.5%	97.4%	70.3%	9.6%	9.9%	9.4%		0.8%	Pearson Spring 2017
5	868654	4134-M03661	320,279	Part B	0,1,2,3	31,719	83.2%	98.6%	30.0%	14.3%	21.9%	31.0%		2.7%	Pearson Spring 2017
5	914139	VH055372	46,453	Part A	0,1	4,627	93.4%	100.0%	84.6%	12.9%				2.4%	Pearson Spring 2017
5	914139	VH055372	46,453	Part B	0,1,2	4,627	89.9%	99.4%	46.3%	26.0%	24.1%			3.6%	Pearson Spring 2017
5	914195	0154-M00796	92,904	Part B	0,1,2	9,282	95.9%	99.8%	80.4%	8.3%	6.4%			4.8%	Pearson Spring 2017
5	868655	M03555	4,052	Comprehensive	0,1,2,3	408	84.4%	99.8%	74.5%	15.1%	3.3%	2.7%		4.5%	Pearson Spring 2017
5	868655	M03555	53,396	Comprehensive	0,1,2,3	11,338	84.0%	99.0%	61.0%	26.0%	6.0%	5.0%		1.0%	DRC Spring 2017
5	914148	M03888	72,736	Comprehensive	0,1,2,3	7,272	86.7%	98.9%	39.9%	27.5%	13.2%	18.7%		0.8%	Pearson Spring 2017

Math Grade 6

Grade	IDEAS ID	PARCC UIN	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 %	Report Source
6	868798	1298-M21432	2,653	Part A	0,1,2,3	266	96.2%	98.9%	81.2%	3.1%	2.4%	5.6%		7.7%	Pearson Spring 2017
6	868798	1298-M21432	2,653	Part B	0,1,2,3	266	96.6%	99.2%	82.6%	2.1%	1.9%	3.5%		9.8%	Pearson Spring 2017
6	868798	1298-M21432	52,395	Part A	0,1,2,3	10,800	91.0%	98.0%	71.0%	7.0%	7.0%	15.0%		0.0%	DRC Spring 2017
6	868798	1298-M21432	52,395	Part B	0,1,2,3	10,800	93.0%	98.0%	77.0%	3.0%	6.0%	13.0%		0.0%	DRC Spring 2017
6	914244	5012-M25073	112,484	Part A	0,1,2	11,249	88.4%	99.6%	41.5%	17.7%	39.4%			1.3%	Pearson Spring 2017
6	914244	5012-M25073	112,484	Part B	0,1,2	11,249	92.0%	99.6%	59.6%	17.1%	19.6%			3.7%	Pearson Spring 2017
6	914286	5047-M25308	38,543	Part A	0,1,2	3,854	87.6%	99.5%	77.4%	15.4%	1.2%			5.9%	Pearson Spring 2017
6	914286	5047-M25308	38,543	Part B	0,1,2	3,854	96.7%	99.4%	75.9%	14.2%	2.5%			7.4%	Pearson Spring 2017
6	914231	1740-M23030	89,916	Comprehensive	0,1,2,3	8,905	70.5%	96.2%	40.2%	18.0%	20.4%	19.0%		2.3%	Pearson Spring 2017
6	870825	M21477	38,264	Comprehensive	0,1,2,3	3,822	94.1%	99.6%	63.3%	25.4%	4.2%	2.3%		4.8%	Pearson Spring 2017
6	870825	M21477	52,418	Comprehensive	0,1,2,3	11,165	93.0%	100.0%	53.0%	36.0%	6.0%	3.0%		2.0%	DRC Spring 2017
6	868818	M22225	166,058	Comprehensive	0,1,2,3	16,588	88.3%	98.9%	61.7%	10.7%	10.1%	14.6%		3.0%	Pearson Spring 2017
6	868818	M22225	52,418	Comprehensive	0,1,2,3	11,080	90.0%	99.0%	68.0%	11.0%	9.0%	10.0%		1.0%	DRC Spring 2017
6	914281	M25152	112,484	Comprehensive	0,1,2,3	11,247	89.0%	99.0%	53.8%	14.1%	12.3%	17.2%		2.6%	Pearson Spring 2017

Math Grade 7

Grade	IDEAS ID	PARCC UIN	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 %	Report Source
7	914333	1300-M21438	111,058	Part A	0,1,2	10,988	92.0%	99.8%	52.4%	34.4%	7.1%			6.0%	Pearson Spring 2017
7	914333	1300-M21438	111,058	Part B	0,1,2	10,988	93.9%	99.7%	48.2%	35.3%	9.0%			7.5%	Pearson Spring 2017
7	914339	VH151385	88,725	Part A	0,1,2	8,838	95.4%	99.4%	66.9%	7.8%	20.9%			4.4%	Pearson Spring 2017
7	914339	VH151385	88,725	Part B	0,1,2	8,838	95.5%	99.7%	77.2%	5.7%	9.7%			7.4%	Pearson Spring 2017
7	868849	M22018	325,765	Comprehensive	0,1,2,3	32,435	91.9%	99.3%	54.2%	23.8%	7.8%	11.4%		2.8%	Pearson Spring 2017
7	868849	M22018	51,790	Comprehensive	0,1,2,3	10,988	97.0%	100.0%	65.0%	20.0%	6.0%	8.0%		1.0%	DRC Spring 2017
7	914329	M25580	111,124	Comprehensive	0,1,2,3	11,013	87.9%	99.1%	47.6%	19.8%	18.4%	10.0%		4.3%	Pearson Spring 2017
7	914336	M25731	88,721	Comprehensive	0,1,2,3	8,591	84.7%	99.2%	26.2%	31.4%	30.6%	8.8%		3.0%	Pearson Spring 2017
7	878705	VH095906	13,001	Comprehensive	0,1,2,3	1,279	90.5%	98.0%	70.8%	10.0%	7.7%	3.6%		7.9%	Pearson Spring 2017
7	878705	VH095906	51,790	Comprehensive	0,1,2,3	10,720	96.0%	99.0%	84.0%	9.0%	5.0%	1.0%		2.0%	DRC Spring 2017
7	914362	VH083535	100,577	Part A	0,1,2,3	19,892	90.4%	98.5%	74.7%	5.1%	4.9%	12.5%		2.8%	Pearson Spring 2016
7	914362	VH083535	100,577	Part B	0,1,2,3	19,892	90.3%	98.8%	70.8%	5.6%	5.6%	13.8%		4.3%	Pearson Spring 2016

Math Grade 8

Grade	IDEAS ID	PARCC UIN	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 %	Report Source
8	892626	VH095800	29,600	Part A	0,1,2,3	2,775	92.0%	99.3%	68.5%	8.4%	6.3%	10.7%		6.1%	Pearson Spring 2017
8	892626	VH095800	29,600	Part B	0,1,2,3	2,774	94.1%	99.1%	78.9%	4.2%	3.1%	5.4%		8.4%	Pearson Spring 2017
8	892626	VH095800	44,761	Part A	0,1,2,3	9,566	87.0%	99.0%	61.0%	12.0%	7.0%	18.0%		2.0%	DRC Spring 2017
8	892626	VH095800	44,761	Part B	0,1,2,3	9,566	92.0%	100.0%	76.0%	8.0%	5.0%	9.0%		2.0%	DRC Spring 2017
8	914400	M21215	87,197	Comprehensive	0,1,2,3	8,580	87.3%	98.9%	64.9%	14.1%	10.1%	3.5%		7.4%	Pearson Spring 2017
8	914381	M25425	69,637	Comprehensive	0,1,2,3,4	6,943	90.8%	99.1%	52.2%	12.5%	26.2%	1.9%	1.0%	6.2%	Pearson Spring 2017
8	914376	VH049778	285,447	Comprehensive	0,1,2,3	28,266	84.0%	98.9%	59.1%	19.5%	10.9%	6.2%		4.4%	Pearson Spring 2017
8	878156	VH057654	99,440	Comprehensive	0,1,2,3	9,865	89.6%	98.5%	74.6%	9.6%	9.0%	2.5%		4.2%	Pearson Spring 2017
8	878156	VH057654	44,810	Comprehensive	0,1,2,3	9,714	93.0%	100.0%	77.0%	9.0%	11.0%	2.0%		2.0%	DRC Spring 2017
8	868882	VH081259	29,600	Part B	0,1,2	2,954	94.5%	99.7%	79.5%	7.4%	1.4%			11.7%	Pearson Spring 2017
8	800273	M20195	44,810	Comprehensive	0,1,2,3,4	9,878	90.0%	100.0%	56.0%	26.0%	7.0%	3.0%	6.0%	2.0%	DRC Spring 2017
8	800273	M20195	47,273	Comprehensive	0,1,2,3,4	10,152	88.0%	99.0%	46.0%	27.0%	8.0%	4.0%	8.0%	6.0%	DRC Spring 2016

English I

Task	IDEAS ID	Spring 2018 Form	PARCC UIN	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 %	Report Source
LAT	902152	Op	VH017536_2T	126,939	RCWE	0,1,2,3,4	8,443	14,816	114,377	23,741	73.6%	99.3%	36.7%	36.2%	16.6%	5.3%	1.8%	3.4%	Pearson Spring 2017
LAT	902152	Op	VH017536_2T	126,939	Conv	0,1,2,3	8,443	14,816	114,377	23,741	73.4%	99.6%	31.5%	37.1%	20.4%	7.5%		3.4%	Pearson Spring 2017
NWT	906152	AE	VH084830	61,936	WE	0,1,2,3,4	3,125	7,776	53,955	10,498	73.3%	98.7%	30.3%	21.8%	27.3%	8.7%	4.2%	7.6%	Pearson Spring 2017
NWT	906152	AE	VH084830	61,936	Conv	0,1,2,3	3,125	7,776	53,955	10,498	74.4%	99.4%	28.1%	27.7%	25.4%	11.2%		7.6%	Pearson Spring 2017
NWT	906152*	AE	VH084830	4,663	WE	0,1,2,3,4	n/a	n/a	n/a	1,076	81.0%	99.0%	22.0%	34.0%	29.0%	10.0%	1.0%	2.0%	DRC Fall 2017, Op
NWT	906152*	AE	VH084830	4,663	Conv	0,1,2,3	n/a	n/a	n/a	1,076	78.0%	99.0%	25.0%	36.0%	26.0%	10.0%		2.0%	DRC Fall 2017, Op
RST	902161	Op, AE	VH017542_2T	123,860	RCWE	0,1,2,3,4	2,656	13,063	116,406	23,334	76.1%	99.5%	22.2%	33.3%	24.3%	12.4%	3.7%	4.1%	Pearson Spring 2017
RST	902161	Op, AE	VH017542_2T	123,860	Conv	0,1,2,3	2,656	13,063	116,407	23,334	76.1%	99.6%	23.1%	32.7%	23.5%	16.6%		4.1%	Pearson Spring 2017
RST	902161	Op, AE	VH017614_2T	3,179	RCWE	0,1,2,3,4	3,012	317	0	620	82.6%	99.4%	43.9%	33.3%	12.0%	2.5%	0.8%	7.6%	Pearson Spring 2017
RST	902161	Op, AE	VH017614_2T	3,179	Conv	0,1,2,3	3,012	317	0	620	79.4%	99.7%	47.8%	30.2%	11.7%	2.8%		7.6%	Pearson Spring 2017
RST	902161*	Op, AE	VH017542_2T	4,674	RCWE	0,1,2,3,4	n/a	n/a	n/a	982	78.0%	99.0%	12.0%	34.0%	40.0%	13.0%	0.0%	0.0%	DRC Fall 2017, Op
RST	902161*	Op, AE	VH017542_2T	4,674	Conv	0,1,2,3	n/a	n/a	n/a	982	78.0%	99.0%	14.0%	32.0%	38.0%	15.0%		0.0%	DRC Fall 2017, Op

*Handscored by DRC in Fall of 2017

English II

Task	IDEAS ID	Spring 2018 Form	PARCC UIN	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 %	Report Source
NWT	902354	AE	7064	4,409	WE	0,1,2,3,4	4,189	435	0	844	84.5%	100.0%	42.5%	19.5%	13.5%	6.0%	2.0%	16.5%	Pearson Spring 2017
NWT	902354	AE	7064	4,409	Conv	0,1,2,3	4,189	435	0	844	85.1%	100.0%	39.7%	21.7%	15.1%	7.0%		16.5%	Pearson Spring 2017
NWT	902354*	AE	7064	9,271	WE	0,1,2,3,4	n/a	n/a	n/a	2,098	81.0%	100.0%	46.0%	17.0%	19.0%	12.0%	2.0%	2.0%	DRC Fall 2017, Op
NWT	902354*	AE	7064	9,271	Conv	0,1,2,3	n/a	n/a	n/a	2,098	83.0%	100.0%	43.0%	18.0%	22.0%	14.0%		2.0%	DRC Fall 2017, Op
RST	902331	Op, AE	VH004490**	2,605	RCWE	0,1,2,3,4	1,915	263	646	827	81.9%	99.3%	51.5%	28.1%	7.2%	0.9%	0.1%	12.3%	Pearson Spring 2017
RST	902331	Op, AE	VH004490**	2,605	Conv	0,1,2,3	1,915	263	646	827	84.8%	99.8%	53.3%	26.3%	7.4%	0.7%		12.3%	Pearson Spring 2017
RST	902331*	Op, AE	VH004490**	9,305	RCWE	0,1,2,3,4	n/a	n/a	n/a	2,020	79.0%	100.0%	37.0%	24.0%	25.0%	11.0%	2.0%	2.0%	DRC Fall 2017, Op
RST	902331*	Op, AE	VH004490**	9,305	Conv	0,1,2,3	n/a	n/a	n/a	2,020	77.0%	99.0%	35.0%	23.0%	27.0%	14.0%		2.0%	DRC Fall 2017, Op
LAT	906197	Op	HH428127697	57,407	RCWE	0,1,2,3,4	28,405	6,463	27,217	14,673	75.4%	99.2%	27.0%	30.6%	27.3%	9.0%	1.3%	4.8%	Pearson Spring 2017
LAT	906197	Op	HH428127697	57,407	Conv	0,1,2,3	28,405	6,463	27,218	14,673	74.8%	99.1%	27.1%	30.0%	27.5%	10.7%		4.8%	Pearson Spring 2017
Task	IDEAS ID	Spring 2018 Form	PARCC UIN	Responses Available	Trait	Score Points	Score Count	n/a	n/a	Reliability Read Count	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	CC %	Report Source
RST	902331	Op, AE	VH004490**	126,270	RCWE	0,1,2,3,4	121,660	n/a	n/a	16,036	76.6%	99.7%	22.7%	34.8%	23.4%	8.3%	2.0%	8.8%	Pearson Spring 2016
RST	902331	Op, AE	VH004490**	126,270	Conv	0,1,2,3	121,507	n/a	n/a	16,003	76.9%	99.7%	25.1%	32.6%	24.0%	9.5%		8.8%	Pearson Spring 2016

*Handscored by DRC in Fall of 2017

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

English III (All Report Data Scored by DRC and MI [AI])

IDEAS ID	Spring 2018 Form	Total Reads	Trait	Score Points	Read 2x	Exact%	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	NS%	Source of Data
851368	Form U, Op	7,783	Content	1, 2, 3, 4	2,204	83%	17%	0%		17%	39%	38%	5%	2%	Op Fall 2016
		7,783	Style	1, 2, 3, 4	2,204	83%	17%	0%		12%	32%	44%	11%	2%	
		7,783	Formation (F)	0, 1	2,204	83%	17%		37%	61%				2%	
		7,783	Usage (U)	0, 1	2,204	78%	22%		21%	78%				2%	
		7,783	Mechanics (M)	0, 1	2,204	89%	11%		22%	77%				2%	
		7,783	Spelling (S)	0, 1	2,204	83%	17%		25%	73%				2%	
851368	Form U, Op	686	Content	1, 2, 3, 4	486	94%	6%	0%		56%	26%	2%	0%	15%	Op Summer 2017
		686	Style	1, 2, 3, 4	486	94%	6%	0%		48%	30%	6%	1%	15%	
		686	Formation (F)	0, 1	486	95%	5%		57%	27%				15%	
		686	Usage (U)	0, 1	486	96%	4%		71%	14%				15%	
		686	Mechanics (M)	0, 1	486	93%	7%		57%	28%				15%	
		686	Spelling (S)	0, 1	486	93%	7%		35%	50%				15%	
851370*	Form R, AE	8,382	Content	1, 2, 3, 4	3,672	91%	9%	0%		18%	34%	28%	11%	10%	Op Fall 2017
		8,382	Style	1, 2, 3, 4	3,672	91%	9%	0%		14%	34%	33%	9%	10%	
		8,382	Formation (F)	0, 1	3,672	94%	6%		27%	63%				10%	
		8,382	Usage (U)	0, 1	3,672	93%	7%		46%	44%				10%	
		8,382	Mechanics (M)	0, 1	3,672	93%	7%		28%	62%				10%	
		8,382	Spelling (S)	0, 1	3,672	96%	4%		21%	69%				10%	
851370*	Form R, AE	182	Content	1, 2, 3, 4	62	81%	19%	0%		5%	43%	46%	5%	1%	AE Fall 2016
		182	Style	1, 2, 3, 4	62	87%	13%	0%		5%	40%	51%	3%	1%	
		182	Formation (F)	0, 1	62	87%	13%		13%	86%				1%	
		182	Usage (U)	0, 1	62	90%	10%		30%	69%				1%	
		182	Mechanics (M)	0, 1	62	94%	6%		13%	86%				1%	
		182	Spelling (S)	0, 1	62	87%	13%		14%	85%				1%	
Form U (851368), will be AI scored by MI with human backreads by DRC.															
Form R (851370), the AE form, will be handscored by DRC supervisors in Spring 2018.															

ELA Grade 3

Task	IDEAS ID	PARCC UIN	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 %	Report Source
NWT	913497	AA431426588	118,416	WE	0,1,2,3	34,298	13,546	84,911	27,299	71.2%	98.6%	30.0%	56.0%	10.6%	1.6%		1.8%	Pearson Spring 2017
NWT	913497	AA431426588	118,416	Conventions	0,1,2,3	34,298	13,546	84,910	27,299	68.6%	98.6%	33.3%	46.8%	16.0%	2.1%		1.8%	Pearson Spring 2017
RST	868500	VF581237	66,149	RCWE	0,1,2,3	62,975	6,643	3,282	13,959	70.8%	98.5%	42.8%	41.0%	10.9%	1.3%		3.9%	Pearson Spring 2017
RST	868500	VF581237	66,149	Conventions	0,1,2,3	62,975	6,643	3,282	13,959	68.8%	98.3%	37.1%	41.2%	15.7%	2.1%		3.9%	Pearson Spring 2017
RST	868500	VF581237	57,098	RCWE	0,1,2,3	n/a	n/a	n/a	12,140	75.0%	99.0%	41.0%	43.0%	13.0%	1.0%		2.0%	DRC Spring 2017
RST	868500	VF581237	57,098	Conventions	0,1,2,3	n/a	n/a	n/a	12,140	76.0%	100.0%	25.0%	46.0%	24.0%	3.0%		2.0%	DRC Spring 2017

ELA Grade 4

Task	IDEAS ID	PARCC UIN	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 %	Report Source
RST	802342	1805	4,565	RCWE	0,1,2,3,4	n/a	n/a	n/a	878	75.9%	97.3%	41.1%	37.9%	10.8%	1.1%	0.1%	9.0%	Pearson Spring 2016
RST	802342	1805	4,565	Conventions	0,1,2,3	n/a	n/a	n/a	878	74.3%	97.7%	39.6%	40.0%	10.1%	1.2%		9.0%	Pearson Spring 2016
RST	802342	1805_AT	65	RCWE	0,1,2,3,4	n/a	n/a	n/a	42	100.0%	100.0%	46.3%	30.5%	14.6%	0.0%	0.0%	8.5%	Pearson Spring 2016
RST	802342	1805_AT	65	Conventions	0,1,2,3	n/a	n/a	n/a	42	81.0%	100.0%	43.9%	36.6%	11.0%	0.0%		8.5%	Pearson Spring 2016
RST	802342	1805_TTS	9,851	RCWE	0,1,2,3,4	n/a	n/a	n/a	1,951	80.6%	97.5%	46.4%	38.0%	6.7%	0.4%	0.1%	8.4%	Pearson Spring 2016
RST	802342	1805_TTS	9,851	Conventions	0,1,2,3	n/a	n/a	n/a	1,951	78.7%	97.7%	45.3%	39.9%	5.8%	0.6%		8.4%	Pearson Spring 2016
RST	802342	1805	54,392	RCWE	0,1,2,3,4	n/a	n/a	n/a	14,620	73%	99%	15%	45%	33%	4%	0%	2%	DRC Spring 2016
RST	802342	1805	54,392	Conventions	0,1,2,3	n/a	n/a	n/a	14,620	72%	98%	14%	44%	35%	4%		2%	DRC Spring 2016
RST	802342	1805	2,605	RCWE	0,1,2,3,4	n/a	n/a	n/a	525	75.6%	97.7%	46.7%	32.6%	6.0%	0.6%	0.1%	14.0%	Pearson Spring 2017
RST	802342	1805	2,605	Conventions	0,1,2,3	n/a	n/a	n/a	525	75.6%	97.3%	44.1%	35.9%	5.2%	0.7%		14.0%	Pearson Spring 2017
LAT	913567	VH170170	121,461	RCWE	0,1,2,3,4	35,658	13,901	87,168	28,425	67.3%	98.5%	33.5%	40.0%	21.2%	3.5%	0.5%	1.3%	Pearson Spring 2017
LAT	913567	VH170170	121,461	Conventions	0,1,2,3	35,659	13,893	87,168	28,418	69.3%	99.1%	28.1%	45.7%	21.0%	3.9%		1.3%	Pearson Spring 2017

ELA Grade 5

Task	IDEAS ID	PARCC UIN	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 %	Report Source
NWT	913663	CC434542756	130,346	WE	0,1,2,3	34,715	4,660	96,145	20,437	71.5%	98.8%	30.9%	29.1%	27.5%	11.5%		0.9%	Pearson Spring 2017
NWT	913663	CC434542756	130,346	Conventions	0,1,2,3	34,710	4,660	96,143	20,440	69.1%	98.7%	32.0%	31.2%	26.3%	9.6%		0.9%	Pearson Spring 2017
RST	868590	C1381	13,923	RCWE	0,1,2,3,4	13,559	1,295	0	2,584	66.8%	96.9%	29.2%	42.0%	21.9%	4.2%	0.3%	2.4%	Pearson Spring 2017
RST	868590	C1381	13,923	Conventions	0,1,2,3	13,559	1,295	0	2,584	67.1%	97.3%	27.9%	39.3%	25.2%	5.2%		2.4%	Pearson Spring 2017
RST	868590	C1381	53,560	RCWE	0,1,2,3,4	n/a	n/a	n/a	12,210	73.0%	99.0%	28.0%	47.0%	20.0%	4.0%	0.0%	0.0%	DRC Spring 2017
RST	868590	C1381	535,360	Conventions	0,1,2,3	n/a	n/a	n/a	12,210	72.0%	99.0%	23.0%	45.0%	25.0%	6.0%		0.0%	DRC Spring 2017

ELA Grade 6

Task	IDEAS ID	PARCC UIN	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 %	Report Source
NWT	913694	D1466	127,628	WE	0,1,2,3,4	34,718	14,034	93,800	29,433	75.9%	99.4%	40.3%	20.6%	22.9%	10.0%	4.3%	1.8%	Pearson Spring 2017
NWT	913694	D1466	127,628	Conventions	0,1,2,3	34,718	14,034	93,800	29,433	75.0%	99.6%	33.4%	30.2%	23.3%	11.3%		1.8%	Pearson Spring 2017
RST	870622	VH082653	52,594	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,928	72.0%	99.0%	24.0%	40.0%	29.0%	6.0%	1.0%	0.0%	DRC Spring 2017
RST	870622	VH082653	52,594	Conventions	0,1,2,3	n/a	n/a	n/a	10,928	73.0%	99.0%	22.0%	38.0%	33.0%	6.0%		0.0%	DRC Spring 2017

ELA Grade 7

Task	IDEAS ID	PARCC UIN	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 %	Report Source
NWT	913842	EE430133306	128,845	WE	0,1,2,3,4	37,606	14,582	91,555	30,289	72.7%	98.6%	34.2%	12.7%	20.4%	17.9%	12.6%	2.2%	Pearson Spring 2017
NWT	913842	EE430133306	128,845	Conventions	0,1,2,3	37,605	14,582	91,555	30,289	71.5%	99.0%	28.5%	20.9%	23.8%	24.7%		2.2%	Pearson Spring 2017
RST	868704	VH069361	158,086	RCWE	0,1,2,3,4	n/a	n/a	n/a	27,575	74.6%	99.5%	20.9%	33.7%	27.8%	12.2%	3.5%	1.9%	Pearson Spring 2016
RST	868704	VH069361	158,086	Conventions	0,1,2,3	n/a	n/a	n/a	27,471	75.7%	99.5%	21.3%	33.0%	28.2%	15.6%		1.9%	Pearson Spring 2016
RST	868704	VH069361	52,091	RCWE	0,1,2,3,4	n/a	n/a	n/a	10,916	72.0%	99.0%	16.0%	33.0%	32.0%	15.0%	3.0%	0.0%	DRC Spring 2017
RST	868704	VH069361	52,091	Conventions	0,1,2,3	n/a	n/a	n/a	10,916	73.0%	99.0%	17.0%	31.0%	33.0%	18.0%		0.0%	DRC Spring 2017

ELA Grade 8

Task	IDEAS ID	PARCC UIN	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 %	Report Source
LAT	913958	F1460	128,084	RCWE	0,1,2,3,4	36,606	4,234	89,633	19,154	69.6%	99.8%	26.4%	31.2%	25.9%	10.9%	3.1%	2.4%	Pearson Spring 2017
LAT	913958	F1460	128,084	Conventions	0,1,2,3	36,606	4,234	89,634	19,154	71.9%	99.7%	22.9%	30.7%	29.2%	14.7%		2.4%	Pearson Spring 2017
RST	868741	5013	7,826	RCWE	0,1,2,3,4	7,573	783	0	1,547	74.5%	99.4%	52.8%	26.4%	11.1%	2.4%	0.5%	6.8%	Pearson Spring 2017
RST	868741	5013	7,826	Conventions	0,1,2,3	7,573	783	0	1,547	77.5%	99.2%	53.1%	26.5%	10.5%	3.0%		6.8%	Pearson Spring 2017
RST	868741	5013	50,673	RCWE	0,1,2,3,4	n/a	n/a	n/a	11,002	72.0%	99.0%	12.0%	21.0%	29.0%	25.0%	12.0%	1.0%	DRC Spring 2017
RST	868741	5013	50,673	Conventions	0,1,2,3	n/a	n/a	n/a	11,002	78.0%	99.0%	12.0%	21.0%	29.0%	38.0%		1.0%	DRC Spring 2017

Biology

IDEAS ID	Spring 2018 Form	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	NS%	Report Source	Source of Data
845340	Form J, Op	n/a	0-4	This item has not previously been scored operationally by DRC. Was AE item in spring of 2017 with 9 total responses.											
845226	Form K, AE	12,560	0-4	3,680	91%	9%	0%	45%	24%	11%	6%	3%	12%	Op Fall 2016	DRC - handscore
845226	Form K, AE	2,952	0-4	1,128	98%	2%	0%	61%	16%	2%	0%	0%	20%	Op Sum 2017	DRC - handscore
These items will be handscored by DRC in Spring 2018.															

U.S. History ERs (EOC)

IDEAS ID	Spring 2018 Form	Total Reads	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	NS%	Source of Data
848632	Form E, Spring 2018 Op	9,847	0-4	3,086	89%	11%	0%	53%	30%	12%	4%	1%	1%	Fall 2016, Op
848632	Form E, Spring 2018 Op	1,609	0-4	1,158	95%	5%	1%	85%	12%	2%	0%	0%	1%	Summer 2017, Op
848291*	Form C, Spring 2018 Admin Error	216	0-4	86	79%	21%	0%	6%	13%	28%	30%	23%	0%	AE Fall 2016
All report data scored by DRC - AI [MI]														
Form E (848632) will be AI scored by MI with DRC human backreads.														
Form C (848291), the AE form, will be handscored by DRC supervisors.														

U.S. History ERs and CRs (LEAP 2025)

IDEAS ID	Item Type	Spring 2018 Form	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj%	Non-Adj%	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Source of Data
894104*	ER	Op, AE	5,000	Content	0-4	5,000	62%	33%	5%	31%	34%	22%	9%	4%	Spring 2017 FT
894104*	ER	Op, AE	5,000	Claims	0-4	5,000	61%	32%	7%	39%	28%	21%	9%	4%	Spring 2017 FT
894256*	ER	Op	5,000	Content	0-4	5,000	66%	32%	2%	28%	37%	24%	8%	3%	Spring 2017 FT
894256*	ER	Op	5,000	Claims	0-4	5,000	64%	33%	3%	34%	33%	21%	8%	3%	Spring 2017 FT
892955*	ER	Op	5,000	Content	0-4	5,000	65%	32%	3%	34%	29%	25%	9%	3%	Spring 2017 FT
892955*	ER	Op	5,000	Claims	0-4	5,000	64%	32%	4%	37%	26%	25%	10%	3%	Spring 2017 FT
894047*	ER	Op	5,000	Content	0-4	5,000	66%	31%	3%	39%	29%	19%	9%	4%	Spring 2017 FT
894047*	ER	Op	5,000	Claims	0-4	5,000	66%	30%	4%	40%	29%	19%	9%	4%	Spring 2017 FT
894225**	CR	Op	1,660		0-2	320	71%	29%	0%	44%	34%	22%			Spring 2017 FT
892994**	CR	Op	1,659		0-2	318	68%	31%	1%	13%	43%	44%			Spring 2017 FT
894188**	CR	Op, AE	6,150		0-2	1,190	85%	15%	0%	53%	33%	14%			Fall 2017 Op
894188**	CR	Op, AE	1,655		0-2	310	74%	25%	1%	33%	38%	28%			Spring 2017 FT
894149**	CR	Op, AE	6,056		0-2	1,132	87%	13%	0%	68%	18%	14%			Fall 2017 Op
894149**	CR	Op, AE	1,653		0-2	306	76%	21%	3%	44%	26%	30%			Spring 2017 FT

All data from DRC handscoring. Nonscores excluded.

*AI scoring models built by MI using 2500 FT responses, scored 2x and resolved. Op ER responses will be AI scored by MI with DRC human backreads. (See U.S. History 5-level model building data earlier in Appendix.)

**All other Spring 2018 USH items (operational 2-point CRs and all AE form items) will be handscored.

Social Studies Grades 3-8

Gr	IDEAS ID	Item Type	Total Reads	Trait	Score Points	Read 2x	Exact %	Adj %	Non-Adj %	SP 0 %	SP 1 %	SP 2 %	SP 3 %	SP 4 %	Source of Data
3	801223	ER	5,428	Content & Claims	0-4	5,428	74	24	2	61	30	5	1	0	Spring 2016 FT
3	801185	CR	63,014		0-2	11,542	82	17	1	23	35	37			Spring 2017 Op
3	801185	CR	1,651		0-2	302	79	19	2	25	48	22			Spring 2016 FT
3	890671	CR	1,650		0-2	300	73	26	1	47	34	14			Spring 2017 FT
4	801572	ER	4,048	Content	0-4	4,048	74	25	2	51	39	9	1	0	Spring 2016 FT [online]
4	801572	ER	4,048	Claims	0-4	4,048	74	25	1	51	39	9	1	0	Spring 2016 FT [online]
4	801572	ER	4,124	Content	0-4	4,124	69	28	3	46	39	11	2	0	Spring 2016 FT [paper]
4	801572	ER	4,124	Claims	0-4	4,124	69	28	2	46	39	11	2	0	Spring 2016 FT [paper]
4	890774	CR	1,657		0-2	314	81	19	0	13	49	36			Spring 2017 FT
4	801582	CR	62,365		0-2	11,578	83	17	0	41	48	6			Spring 2017 Op
4	801582	CR	1,657		0-2	314	78	21	2	42	41	9			Spring 2016 FT
5	807728*	ER	5,198	Content	0-4	5,198	76	23	1	61	23	13	3	0	Spring 2016 FT
5	807728*	ER	5,198	Claims	0-4	5,198	78	20	2	66	20	11	2	0	Spring 2016 FT
5	807709	CR	57,809		0-2	10,582	82	18	0	65	31	4			Spring 2017 Op
5	807709	CR	1,566		0-2	280	81	19	0	61	34	4			Spring 2016 FT
5	890929	CR	1,653		0-2	306	73	27	0	60	31	10			Spring 2017 FT
6	804944*	ER	5,949	Content	0-4	5,949	73	25	1	61	32	7	1	0	Spring 2016 FT
6	804944*	ER	5,949	Claims	0-4	5,949	75	24	1	64	31	5	1	0	Spring 2016 FT
6	891206	CR	1,651		0-2	302	79	20	1	56	26	18			Spring 2017 FT
6	891161	CR	1,654		0-2	308	81	19	0	37	47	16			Spring 2017 FT
7	805569*	ER	5,220	Content	0-4	5,220	68	31	1	32	42	21	4	0	Spring 2016 FT
7	805569*	ER	5,220	Claims	0-4	5,220	66	32	1	35	41	20	4	0	Spring 2016 FT
7	891311	CR	1,654		0-2	308	73	27	1	34	42	24			Spring 2017 FT
7	891331	CR	1,659		0-2	318	74	25	1	48	40	12			Spring 2017 FT
8	809006*	ER	5,086	Content	0-4	5,086	63	35	2	27	38	24	8	2	Spring 2016 FT
8	809006*	ER	5,086	Claims	0-4	5,086	62	36	2	33	33	24	8	2	Spring 2016 FT
8	892278	CR	1,656		0-2	312	79	20	1	43	41	15			Spring 2017 FT
8	892336	CR	1,668		0-2	336	70	29	1	23	46	30			Spring 2017 FT

All data from DRC handscoring. Nonscores excluded.
*AI scoring models built by MI using 2500 FT responses, scored 2x and resolved. Grades 5-8 Op ER responses will be AI scored by MI with DRC human backreads. (See Grades 5-8 SS model building data earlier in Appendix.)
**All other Spring 2018 SS items (operational 2-point CRs and grade 3 and 4 ER items) will be handscored.

Appendix E. Classical Item Statistics

Table E.1 Operational Item Statistics—English I Fall Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥4,590	≥4,590	0.58	0.23	0.00	36.18	12.28	51.54		
2	ESR	≥4,590	≥4,580	0.43	0.36	0.04	53.13	8.28	38.55		
3	ESR	≥4,590	≥4,580	0.53	0.32	0.09	29.91	33.26	36.75		
4	TE	≥4,590	≥4,590	0.39	0.43	0.02	60.40	1.87	37.70		
5	ESR	≥4,590	≥4,590	0.46	0.50	0.00	36.90	33.54	29.56		
6	ESR	≥4,590	≥4,580	0.56	0.43	0.04	42.15	3.59	54.21		
7	ESR	≥4,590	≥4,580	0.76	0.34	0.09	13.57	20.69	65.63		
8	TE	≥4,590	≥4,580	0.39	0.45	0.02	43.19	35.81	20.95		
9	CR	≥4,590	≥4,520	0.39	0.74	1.29	11.91	33.70	40.25	12.48	0.28
10	CR	≥4,590	≥4,520	0.39	0.74	1.29	11.91	33.70	40.25	12.48	0.28
11	CR	≥4,590	≥4,520	0.52	0.74	1.29	13.05	32.15	38.53	14.9	
12	ESR	≥4,590	≥4,570	0.50	0.48	0.09	38.40	23.83	37.68		
13	MS	≥4,590	≥4,570	0.52	0.53	0.13	34.13	28.52	37.22		
14	TE	≥4,590	≥4,570	0.33	0.29	0.15	48.92	36.72	14.20		
15	MS	≥4,590	≥4,570	0.53	0.28	0.09	30.92	32.14	36.85		
16	CR	≥4,590	≥4,410	0.33	0.70	3.51	22.52	33.78	29.43	9.45	1.11
17	CR	≥4,590	≥4,410	0.41	0.70	3.51	24.59	35.24	26.55	9.91	
18	ESR	≥4,590	≥4,560	0.38	0.20	0.31	44.88	33.38	21.43		
19	ESR	≥4,590	≥4,570	0.42	0.46	0.24	56.16	3.62	39.97		
20	ESR	≥4,590	≥4,560	0.52	0.33	0.31	40.78	13.35	45.56		
21	ESR	≥4,590	≥4,560	0.55	0.41	0.37	39.62	9.47	50.53		
22	TE	≥4,590	≥4,530	0.36	0.46	0.96	39.43	48.18	11.43		
23	TE	≥4,590	≥4,530	0.64	0.55	1.16	14.31	41.72	42.81		
24	ESR	≥4,590	≥4,570	0.32	0.36	0.07	62.17	11.38	26.39		
25	TE	≥4,590	≥4,560	0.41	0.29	0.33	42.73	31.94	25.01		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
26	MS	≥4,590	≥4,570	0.43	0.35	0.13	20.80	71.65	7.43		
27	ESR	≥4,590	≥4,570	0.45	0.30	0.09	45.85	18.09	35.98		
28	ESR	≥4,590	≥4,570	0.11	0.17	0.04	85.17	8.06	6.68		
29	ESR	≥4,590	≥4,570	0.45	0.35	0.04	53.45	3.67	42.79		
30	ESR	≥4,590	≥4,570	0.59	0.41	0.04	34.80	12.15	52.95		
31	ESR	≥4,590	≥4,570	0.56	0.53	0.07	33.01	21.84	45.00		
32	ESR	≥4,590	≥4,570	0.30	0.21	0.04	59.09	22.04	18.74		
33	ESR	≥4,590	≥4,570	0.41	0.23	0.00	50.87	15.14	33.88		

Copyright © 2019 by Louisiana Department of Education.

Table E.2 Operational Item Statistics—English II Fall Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥6,000	≥6,000	0.79	0.36	0.02	13.72	13.77	72.49		
2	ESR	≥6,000	≥5,990	0.31	0.31	0.13	51.6	35.48	12.79		
3	ESR	≥6,000	≥5,990	0.63	0.31	0.17	15.48	42.51	41.82		
4	TE	≥6,000	≥5,990	0.45	0.31	0.12	18.73	73.06	8.08		
5	MS	≥6,000	≥5,990	0.53	0.38	0.18	30.49	32.75	36.56		
6	ESR	≥6,000	≥6,000	0.79	0.46	0.05	14.10	12.95	72.88		
7	MS	≥6,000	≥5,990	0.40	0.46	0.18	32.20	54.60	13.00		
8	ESR	≥6,000	≥5,990	0.62	0.48	0.17	29.04	17.57	53.21		
9	CR	≥6,000	≥5,930	0.39	0.74	1.17	19.45	25.71	34.67	16.55	2.40
10	CR	≥6,000	≥5,930	0.39	0.74	1.17	19.45	25.71	34.67	16.55	2.40
11	CR	≥6,000	≥5,930	0.54	0.74	1.17	17.93	23.51	36.58	20.75	
12	ESR	≥6,000	≥5,980	0.42	0.29	0.07	47.78	20.67	31.49		
13	ESR	≥6,000	≥5,980	0.29	0.32	0.17	64.69	11.52	23.62		
14	MS	≥6,000	≥5,980	0.46	0.24	0.17	12.47	82.39	4.97		
15	MS	≥6,000	≥5,980	0.28	0.42	0.12	54.32	35.69	9.87		
16	CR	≥6,000	≥5,860	0.37	0.72	2.10	27.25	21.52	27.53	18.08	3.47
17	CR	≥6,000	≥5,860	0.50	0.73	2.10	24.72	20.22	31.87	21.04	
18	MS	≥6,000	≥5,960	0.26	0.32	0.45	57.23	32.15	10.15		
19	MS	≥6,000	≥5,970	0.25	0.44	0.18	70.77	7.80	21.24		
20	MS	≥6,000	≥5,980	0.23	0.23	0.10	66.26	21.64	11.99		
21	MS	≥6,000	≥5,970	0.31	0.28	0.18	50.72	36.91	12.17		
22	TE	≥6,000	≥5,960	0.63	0.44	0.37	14.32	45.34	39.95		
23	MS	≥6,000	≥5,970	0.62	0.57	0.25	22.12	32.07	45.54		
24	ESR	≥6,000	≥5,980	0.78	0.35	0.02	16.05	12.37	71.57		
25	ESR	≥6,000	≥5,970	0.49	0.37	0.10	33.66	34.38	31.86		
26	ESR	≥6,000	≥5,970	0.62	0.49	0.07	22.75	30.84	46.35		
27	ESR	≥6,000	≥5,970	0.63	0.33	0.12	23.98	25.89	50.01		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
28	MS	≥6,000	≥5,970	0.48	0.44	0.03	43.14	17.30	39.46		
29	TE	≥6,000	≥5,960	0.40	0.39	0.27	46.26	26.12	27.28		
30	MS	≥6,000	≥5,960	0.39	0.53	0.25	47.65	25.46	26.58		
31	ESR	≥6,000	≥5,960	0.62	0.53	0.15	34.18	7.30	58.25		
32	MS	≥6,000	≥5,970	0.24	0.33	0.08	61.24	28.66	9.89		
33	MS	≥6,000	≥5,970	0.41	0.52	0.00	32.43	52.16	15.28		

Copyright © 2019 by Louisiana Department of Education.

Table E.3 Operational Item Statistics—Algebra I Fall Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	MC	≥2,940	≥2,930	0.48	0.30	0.17		48.38	15.87	20.77	14.81		
2	MPS	≥2,940	≥2,930	0.54	0.62	0.03	21.85	49.08	29.04				
3	MC	≥2,940	≥2,920	0.35	0.11	0.20		8.69	27.03	28.9	35.17		
4	MC	≥2,940	≥2,920	0.39	0.14	0.27		17.66	38.48	38.68	4.91		
5	MPS	≥2,940	≥2,930	0.35	0.37	0.17	41.39	47.55	10.88				
6	MC	≥2,940	≥2,920	0.73	0.35	0.27		14.3	72.32	6.98	6.13		
7	MC	≥2,940	≥2,920	0.47	0.33	0.34		12.94	29.45	47.05	10.21		
8	MPS	≥2,940	≥2,930	0.27	0.16	0.00	53.34	38.48	8.18				
9	MC	≥2,940	≥2,930	0.48	0.38	0.14		11.18	47.68	25.22	15.78		
10	TE	≥2,940	≥2,930	0.10	0.21	0.31	81.9	15.07	2.72				
11	MC	≥2,940	≥2,930	0.56	0.29	0.03		21.46	55.68	10.88	11.94		
12	MC	≥2,940	≥2,920	0.25	0.38	0.44		33.64	27.89	12.72	25.31		
13	MC	≥2,940	≥2,930	0.63	0.32	0.10		18.2	62.72	9.12	9.86		
14	MC	≥2,940	≥2,920	0.38	0.09	0.17		8.38	37.97	26.21	27.27		
15	CR	≥2,940	≥2,700	0.17	0.56	5.40	56.3	16.81	13.17	3.77	1.8		
16	MC	≥2,940	≥2,930	0.52	0.49	0.20		19.86	10.58	17.38	51.97		
17	TE	≥2,940	≥2,910	0.14	0.25	0.41	85.89	13.66					
18	MS	≥2,940	≥2,930	0.44	0.52	0.20	56.33	43.47					
19	MS	≥2,940	≥2,920	0.17	0.38	0.20	83.03	16.77					
20	SA	≥2,940	≥2,910	0.74	0.44	0.24	26.3	73.22					
21	CR	≥2,940	≥2,770	0.28	0.59	3.85	47.97	20.05	18.73	7.59			
22	MC	≥2,940	≥2,930	0.31	0.22	0.14		23.18	31.39	15.20	30.10		
23	MPS	≥2,940	≥2,920	0.43	0.42	0.00	31.90	49.59	18.51				
24	MC	≥2,940	≥2,920	0.29	0.11	0.41		14.38	28.94	34.76	21.51		
25	TE	≥2,940	≥2,920	0.30	0.40	0.14	47.71	43.82	8.33				
26	CR	≥2,940	≥2,310	0.03	0.44	14.48	68.99	9.36	0.44	0.17	0.03		
27	CR	≥2,940	≥2,630	0.15	0.47	7.27	61.63	17.49	8.9	1.46			

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	CR	≥2,940	≥2,890	0.49	0.72	1.05	29.71	4.62	12.56	6.86	7.54	15.01	22.07
29	MC	≥2,940	≥2,910	0.42	0.43	0.31		36.20	42.11	11.85	9.53		
30	CR	≥2,940	≥2,850	0.24	0.49	0.98	38.34	50.70	3.7	4.18			
31	MPS	≥2,940	≥2,920	0.68	0.48	0.07	25.17	12.88	61.89				
32	MC	≥2,940	≥2,920	0.66	0.31	0.14		4.41	22.71	65.95	6.80		
33	MS	≥2,940	≥2,930	0.37	0.35	0.10	63.33	36.57					
34	MC	≥2,940	≥2,910	0.33	0.38	0.31		17.11	30.53	18.89	33.16		
35	CR	≥2,940	≥2,790	0.04	0.40	0.00	86.08	6.11	2.11	0.61			
36	MC	≥2,940	≥2,920	0.40	0.19	0.31		26.82	40.22	23.52	9.13		
37	SA	≥2,940	≥2,930	0.30	0.63	0.07	49.34	14.13	12.56	13.69	10.21		
38	TE	≥2,940	≥2,920	0.55	0.38	0.17	44.55	55.28					

Copyright © 2019 by Louisiana Department of Education.

Table E.4 Operational Item Statistics—Geometry Fall Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	ESR	≥5,460	≥5,460	0.56	0.51	0.00	21.32	45.53	33.15				
2	CR	≥5,460	≥5,200	0.16	0.58	2.97	68.73	11.21	12.71	2.64			
3	TE	≥5,460	≥5,420	0.60	0.43	0.24	40.04	59.72					
4	MC	≥5,460	≥5,440	0.68	0.44	0.11		68.32	13.62	10.08	7.87		
5	TE	≥5,460	≥5,390	0.45	0.54	1.28	27.47	53.72	17.53				
6	MC	≥5,460	≥5,440	0.56	0.41	0.13		28.38	11.34	55.60	4.55		
7	TE	≥5,460	≥5,440	0.79	0.41	0.04	5.76	31.06	63.14				
8	ESR	≥5,460	≥5,430	0.37	0.54	0.11	45.93	33.31	20.65				
9	TE	≥5,460	≥5,410	0.21	0.53	0.86	78.15	20.99					
10	MC	≥5,460	≥5,430	0.53	0.22	0.46		16.28	53.04	24.05	6.17		
11	MC	≥5,460	≥5,430	0.37	0.29	0.24		36.67	11.66	39.13	12.30		
12	TE	≥5,460	≥5,430	0.36	0.57	0.49	64.08	35.42					
13	MC	≥5,460	≥5,420	0.22	0.19	0.18		30.26	31.55	22.04	15.96		
14	CR	≥5,460	≥4,930	0.21	0.70	5.42	52.96	4.08	10.44	12.21	3.20	5.29	2.12
15	MC	≥5,460	≥5,450	0.82	0.28	0.05		7.77	2.84	6.91	82.43		
16	CR	≥5,460	≥4,940	0.20	0.70	6.13	57.60	12.13	9.32	4.75	6.94		
17	MC	≥5,460	≥5,440	0.58	0.36	0.07		4.35	28.32	58.08	9.18		
18	SA	≥5,460	≥5,420	0.29	0.48	0.42	50.68	39.94	8.96				
19	CR	≥5,460	≥5,300	0.22	0.69	0.70	53.20	7.16	14.85	5.31	12.12	2.11	2.34
20	SA	≥5,460	≥5,440	0.35	0.60	0.35	64.87	34.79					
21	TE	≥5,460	≥5,430	0.29	0.51	0.06	50.52	40.87	8.55				
22	MC	≥5,460	≥5,430	0.63	0.27	0.17		62.93	13.38	17.26	6.25		
23	CR	≥5,460	≥4,770	0.07	0.50	8.46	73.55	10.08	3.39	0.48			
24	CR	≥5,460	≥5,090	0.14	0.66	0.00	62.30	18.42	7.36	3.33	1.89		
25	TE	≥5,460	≥5,430	0.72	0.41	0.04	28.06	71.91					
26	MC	≥5,460	≥5,430	0.45	0.39	0.17		44.53	17.17	25.68	12.45		
27	MC	≥5,460	≥5,430	0.29	0.39	0.22		28.56	51.51	12.79	6.92		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	MC	≥5,460	≥5,430	0.68	0.38	0.18		7.38	3.01	21.55	67.88		
29	MC	≥5,460	≥5,420	0.26	0.35	0.18		25.74	51.72	16.71	5.64		
30	SA	≥5,460	≥5,390	0.65	0.37	0.79	34.51	64.7					
31	TE	≥5,460	≥5,450	0.40	0.63	0.11	20.28	33.45	22.95	13.26	9.95		
32	MS	≥5,460	≥5,420	0.19	0.46	0.20	80.46	19.34					
33	MC	≥5,460	≥5,430	0.39	0.39	0.18		38.75	11.64	10.17	39.26		
34	MC	≥5,460	≥5,440	0.49	0.39	0.35		11.67	26.15	13.35	48.48		
35	SA	≥5,460	≥5,440	0.17	0.54	0.20	83.13	16.67					
36	TE	≥5,460	≥5,450	0.49	0.43	0.09	50.73	49.18					
37	MC	≥5,460	≥5,420	0.14	0.15	0.24		24.80	19.12	42.07	13.77		
38	SA	≥5,460	≥5,370	0.31	0.38	1.30	48.38	40.12	10.19				

Copyright © 2019 by Louisiana Department of Education.

Table E.5 Operational Item Statistics—English I Spring Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥46,180	≥46,170	0.41	0.35	0.02	47.82	22.36	29.80		
2	TE	≥46,180	≥46,080	0.36	0.48	0.23	51.95	22.91	24.91		
3	ESR	≥46,180	≥46,130	0.60	0.30	0.12	30.49	18.74	50.64		
4	ESR	≥46,180	≥46,130	0.39	0.30	0.12	51.01	20.77	28.10		
5	ESR	≥46,180	≥46,120	0.74	0.37	0.13	20.60	10.16	69.12		
6	ESR	≥46,180	≥46,150	0.57	0.39	0.08	36.13	12.89	50.89		
7	CR	≥46,180	≥44,880	0.24	0.77	1.42	30.53	44.32	19.11	2.92	0.30
8	CR	≥46,180	≥44,880	0.24	0.77	1.42	30.53	44.32	19.11	2.92	0.30
9	CR	≥46,180	≥44,880	0.36	0.76	1.42	24.39	45.32	24.06	3.41	
10	ESR	≥46,180	≥46,090	0.35	0.39	0.21	57.79	13.29	28.70		
11	TE	≥46,180	≥46,010	0.71	0.45	0.38	11.20	34.51	53.91		
12	ESR	≥46,180	≥46,030	0.69	0.44	0.32	18.83	24.37	56.48		
13	ESR	≥46,180	≥46,010	0.45	0.36	0.37	49.31	11.50	38.82		
14	ESR	≥46,180	≥46,170	0.53	0.27	0.04	39.80	13.51	46.65		
15	ESR	≥46,180	≥46,130	0.41	0.39	0.12	54.28	8.77	36.83		
16	ESR	≥46,180	≥46,130	0.54	0.33	0.11	29.01	34.53	36.36		
17	TE	≥46,180	≥46,140	0.33	0.44	0.09	66.14	2.06	31.72		
18	ESR	≥46,180	≥46,140	0.43	0.51	0.10	40.09	32.88	26.94		
19	ESR	≥46,180	≥46,120	0.53	0.46	0.14	44.60	4.10	51.16		
20	ESR	≥46,180	≥46,130	0.72	0.45	0.13	17.47	21.60	60.80		
21	TE	≥46,180	≥46,120	0.38	0.45	0.13	44.78	34.72	20.37		
22	CR	≥46,180	≥44,690	0.36	0.81	1.77	15.69	36.59	31.64	11.46	1.39
23	CR	≥46,180	≥44,690	0.36	0.81	1.77	15.69	36.59	31.64	11.46	1.39
24	CR	≥46,180	≥44,690	0.48	0.80	1.77	15.60	35.68	32.15	13.34	
25	MS	≥46,180	≥46,160	0.52	0.35	0.06	26.98	42.20	30.76		
26	ESR	≥46,180	≥46,110	0.63	0.40	0.15	29.95	14.77	55.12		
27	ESR	≥46,180	≥46,120	0.48	0.35	0.13	40.30	23.80	35.76		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
28	TE	≥46,180	≥46,130	0.52	0.30	0.12	44.14	7.96	47.78		
29	ESR	≥46,180	≥46,120	0.47	0.44	0.14	41.98	21.88	35.99		
30	TE	≥46,180	≥46,070	0.45	0.43	0.25	34.58	39.67	25.50		
31	ESR	≥46,180	≥46,030	0.21	0.22	0.33	63.25	30.38	6.04		
32	ESR	≥46,180	≥46,050	0.59	0.34	0.29	20.68	41.01	38.02		
33	ESR	≥46,180	≥46,030	0.35	0.47	0.33	56.71	16.23	26.73		
34	TE	≥46,180	≥45,900	0.40	0.50	0.62	44.73	29.90	24.76		

Copyright © 2019 by Louisiana Department of Education.

Table E.6 Operational Item Statistics—English II Spring Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥41,880	≥41,870	0.42	0.33	0.03	52.86	10.35	36.76		
2	TE	≥41,880	≥41,840	0.45	0.30	0.11	41.97	26.83	31.09		
3	ESR	≥41,880	≥41,790	0.24	0.15	0.21	66.36	19.62	13.82		
4	MS	≥41,880	≥41,800	0.39	0.47	0.19	52.81	16.88	30.12		
5	MS	≥41,880	≥41,800	0.21	0.29	0.19	67.90	22.40	9.51		
6	TE	≥41,880	≥41,770	0.39	0.33	0.27	27.02	67.79	4.93		
7	CR	≥41,880	≥40,350	0.32	0.79	2.12	18.77	37.79	33.73	5.63	0.42
8	CR	≥41,880	≥40,350	0.32	0.79	2.12	18.77	37.79	33.73	5.63	0.42
9	CR	≥41,880	≥40,350	0.43	0.79	2.12	19.23	36.26	34.38	6.47	
10	ESR	≥41,880	≥41,610	0.50	0.39	0.65	41.77	15.59	41.99		
11	ESR	≥41,880	≥41,560	0.55	0.34	0.78	39.48	9.81	49.94		
12	MS	≥41,880	≥41,490	0.29	0.32	0.92	47.00	46.17	5.91		
13	TE	≥41,880	≥41,170	0.25	0.42	1.70	59.78	27.25	11.28		
14	ESR	≥41,880	≥41,870	0.79	0.38	0.03	14.30	13.09	72.57		
15	ESR	≥41,880	≥41,830	0.33	0.28	0.12	49.68	35.45	14.75		
16	ESR	≥41,880	≥41,830	0.63	0.32	0.11	14.92	43.37	41.60		
17	TE	≥41,880	≥41,830	0.44	0.30	0.13	19.88	72.46	7.53		
18	MS	≥41,880	≥41,840	0.54	0.42	0.09	30.09	32.58	37.24		
19	ESR	≥41,880	≥41,840	0.81	0.45	0.09	13.11	11.30	75.50		
20	MS	≥41,880	≥41,790	0.40	0.48	0.21	32.77	53.38	13.63		
21	ESR	≥41,880	≥41,820	0.61	0.49	0.14	30.02	18.40	51.44		
22	CR	≥41,880	≥40,350	0.39	0.80	2.12	11.81	35.36	36.00	11.58	1.59
23	CR	≥41,880	≥40,350	0.39	0.80	2.12	11.81	35.36	36.00	11.58	1.59
24	CR	≥41,880	≥40,350	0.49	0.80	2.12	13.99	35.56	35.66	11.13	
25	ESR	≥41,880	≥41,870	0.77	0.35	0.04	15.88	13.43	70.66		
26	ESR	≥41,880	≥41,830	0.52	0.36	0.11	31.85	33.04	34.99		
27	ESR	≥41,880	≥41,830	0.60	0.49	0.13	23.06	33.41	43.41		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
28	ESR	≥41,880	≥41,820	0.64	0.33	0.15	22.54	26.54	50.77		
29	MS	≥41,880	≥41,820	0.51	0.43	0.13	38.38	22.02	39.47		
30	MS	≥41,880	≥41,820	0.50	0.50	0.15	35.73	29.20	34.92		
31	MS	≥41,880	≥41,800	0.44	0.40	0.19	19.88	72.42	7.51		
32	TE	≥41,880	≥41,700	0.44	0.48	0.42	34.47	43.41	21.70		
33	MS	≥41,880	≥41,770	0.28	0.39	0.26	58.69	25.37	15.68		
34	TE	≥41,880	≥41,640	0.34	0.51	0.58	45.67	39.19	14.57		

Copyright © 2019 by Louisiana Department of Education.

Table E.7 Operational Item Statistics—Algebra I Spring Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	MC	≥46,010	≥45,950	0.31	0.16	0.14		18.71	9.12	41.11	30.92		
2	MC	≥46,010	≥45,920	0.53	0.31	0.21		53.14	21.49	14.96	10.19		
3	MC	≥46,010	≥45,890	0.38	0.17	0.26		18.43	38.64	37.61	5.05		
4	MC	≥46,010	≥45,980	0.75	0.36	0.07		5.80	8.01	74.88	11.24		
5	CR	≥46,010	≥42,880	0.13	0.61	5.02	67.77	4.36	8.96	4.39	3.97	2.20	1.54
6	MC	≥46,010	≥45,920	0.36	0.31	0.21		36.15	27.78	29.02	6.83		
7	MC	≥46,010	≥45,900	0.46	0.16	0.24		45.49	31.80	11.53	10.93		
8	MC	≥46,010	≥45,960	0.48	0.39	0.12		13.83	29.39	47.82	8.86		
9	CR	≥46,010	≥42,760	0.06	0.46	4.91	78.19	10.23	2.05	1.23	1.23		
10	MPSR	≥46,010	≥46,000	0.29	0.21	0.04	50.77	41.13	8.06				
11	MC	≥46,010	≥45,900	0.43	0.20	0.24		13.53	43.12	15.72	27.40		
12	MC	≥46,010	≥45,900	0.37	0.08	0.25		14.82	37.20	17.58	30.15		
13	MPSR	≥46,010	≥45,970	0.27	0.21	0.10	54.13	37.27	8.50				
14	SA	≥46,010	≥46,000	0.69	0.60	0.03	9.30	11.44	14.07	24.51	40.65		
15	MC	≥46,010	≥45,930	0.48	0.39	0.17		10.15	47.58	26.92	15.17		
16	TE	≥46,010	≥45,950	0.19	0.22	0.14	66.01	30.31	3.54				
17	TE	≥46,010	≥45,420	0.34	0.53	1.29	48.54	33.95	16.22				
18	MC	≥46,010	≥45,950	0.55	0.30	0.15		18.85	54.83	11.99	14.18		
19	MC	≥46,010	≥45,960	0.24	0.28	0.12		23.72	9.84	44.58	21.73		
20	SA	≥46,010	≥45,350	0.29	0.53	1.44	70.34	28.22					
21	MC	≥46,010	≥45,870	0.38	0.09	0.32		8.74	38.14	27.30	25.51		
22	MC	≥46,010	≥45,930	0.52	0.47	0.18		18.65	11.44	17.92	51.82		
23	SA	≥46,010	≥45,950	0.52	0.59	0.14	32.01	31.69	36.16				
24	MPSR	≥46,010	≥45,720	0.32	0.40	0.64	43.15	49.66	6.56				
25	CR	≥46,010	≥46,000	0.59	0.60	0.03	11.81	21.51	44.54	22.10			
26	MC	≥46,010	≥45,900	0.35	0.24	0.25		21.74	35.09	14.98	27.94		
27	MC	≥46,010	≥45,830	0.32	0.25	0.40		32.01	30.80	26.45	10.34		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	MC	≥46,010	≥45,900	0.35	0.23	0.26		12.01	35.27	31.26	21.20		
29	MC	≥46,010	≥45,850	0.44	0.11	0.35		10.97	26.98	17.43	44.27		
30	CR	≥46,010	≥41,820	0.26	0.62	6.63	53.20	14.26	15.00	8.43			
31	CR	≥46,010	≥40,810	0.20	0.57	8.54	50.61	24.79	10.24	3.06			
32	MPSR	≥46,010	≥45,930	0.42	0.25	0.19	32.99	49.17	17.65				
33	MC	≥46,010	≥45,930	0.34	0.28	0.18		13.21	17.03	35.57	34.01		
34	MS	≥46,010	≥45,950	0.36	0.43	0.15	63.83	36.02					
35	MC	≥46,010	≥45,910	0.36	0.42	0.24		15.95	29.57	17.85	36.40		
36	CR	≥46,010	≥44,440	0.05	0.49	2.58	85.85	7.25	2.33	1.15			
37	MC	≥46,010	≥45,870	0.39	0.17	0.32		23.86	38.97	27.99	8.86		
38	SA	≥46,010	≥44,830	0.16	0.24	2.57	81.82	15.61					
39	CR	≥46,010	≥40,460	0.19	0.52	9.08	38.78	39.46	4.21	4.08	1.39		

Copyright © 2019 by Louisiana Department of Education.

Table E.8 Operational Item Statistics—Geometry Spring Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	TE	≥36,800	≥36,720	0.61	0.42	0.20	38.96	60.84					
2	MC	≥36,800	≥36,690	0.10	0.28	0.29		17.28	33.76	38.54	10.13		
3	MPSR	≥36,800	≥36,760	0.53	0.55	0.10	28.91	36.89	34.10				
4	MPSR	≥36,800	≥36,790	0.37	0.26	0.03	39.90	45.44	14.64				
5	MC	≥36,800	≥36,740	0.45	0.45	0.17		14.33	16.47	24.35	44.69		
6	MPSR	≥36,800	≥36,650	0.31	0.37	0.40	50.85	36.07	12.68				
7	TE	≥36,800	≥36,780	0.79	0.40	0.06	5.71	31.47	62.76				
8	TE	≥36,800	≥36,650	0.23	0.54	0.41	76.34	23.26					
9	TE	≥36,800	≥36,640	0.22	0.43	0.43	78.05	21.52					
10	MS	≥36,800	≥36,770	0.61	0.58	0.09	38.96	60.95					
11	MS	≥36,800	≥36,780	0.34	0.36	0.05	65.83	34.12					
12	TE	≥36,800	≥36,780	0.46	0.57	0.04	9.09	30.44	36.22	17.14	7.07		
13	SA	≥36,800	≥36,390	0.37	0.43	1.11	62.57	36.32					
14	TE	≥36,800	≥36,650	0.37	0.54	0.41	62.37	37.22					
15	CR	≥36,800	≥34,220	0.34	0.60	4.66	43.32	14.92	25.68	9.07			
16	CR	≥36,800	≥32,900	0.23	0.70	7.18	50.46	4.22	9.14	13.48	3.55	5.84	2.72
17	MS	≥36,800	≥36,760	0.20	0.42	0.11	79.72	20.18					
18	MC	≥36,800	≥36,790	0.85	0.27	0.03		6.05	2.46	6.15	85.32		
19	CR	≥36,800	≥33,040	0.20	0.70	6.83	56.81	12.31	9.88	4.83	5.96		
20	SA	≥36,800	≥36,640	0.29	0.47	0.44	51.47	38.69	9.40				
21	MPSR	≥36,800	≥36,780	0.47	0.42	0.05	19.30	68.32	12.32				
22	SA	≥36,800	≥36,650	0.35	0.60	0.39	64.78	34.83					
23	CR	≥36,800	≥34,530	0.18	0.72	4.45	46.22	24.17	9.24	4.94	4.61	3.18	1.47
24	MC	≥36,800	≥36,730	0.60	0.29	0.18		60.09	14.39	20.12	5.22		
25	CR	≥36,800	≥32,380	0.09	0.53	8.37	71.91	9.33	6.04	0.73			
26	TE	≥36,800	≥36,790	0.72	0.40	0.02	28.06	71.91					
27	MC	≥36,800	≥36,740	0.42	0.35	0.16		16.87	42.18	18.00	22.79		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	MC	≥36,800	≥36,770	0.28	0.43	0.08		28.33	52.75	11.64	7.21		
29	CR	≥36,800	≥33,560	0.06	0.57	5.53	76.51	8.88	3.57	1.36	0.89		
30	SA	≥36,800	≥36,700	0.46	0.58	0.27	53.96	45.76					
31	SA	≥36,800	≥36,270	0.38	0.57	1.43	61.18	37.39					
32	MC	≥36,800	≥36,740	0.37	0.38	0.16		36.47	10.44	9.78	43.15		
33	MC	≥36,800	≥36,750	0.29	0.35	0.13		28.80	25.94	23.49	21.63		
34	MC	≥36,800	≥36,760	0.42	0.14	0.11		41.44	41.73	11.97	4.75		
35	SA	≥36,800	≥36,680	0.17	0.55	0.32	82.49	17.19					
36	MC	≥36,800	≥36,770	0.71	0.37	0.07		12.31	5.85	71.27	10.50		
37	SA	≥36,800	≥36,520	0.24	0.51	0.75	75.40	23.85					
38	TE	≥36,800	≥36,780	0.23	0.36	0.07	59.47	35.37	5.09				

Copyright © 2019 by Louisiana Department of Education.

Table E.9 Operational Item Statistics—English I Summer Administration

Item	Item Type	Total N	Adj. N	p-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥60	≥60	0.44	0.01	0.00	44.62	23.08	32.31		
2	MS	≥60	≥60	0.42	0.50	0.00	33.85	49.23	16.92		
3	ESR	≥60	≥60	0.58	0.51	0.00	32.31	20.00	47.69		
4	ESR	≥60	≥60	0.63	0.51	0.00	29.23	15.38	55.38		
5	ESR	≥60	≥60	0.64	0.50	0.00	32.31	7.69	60.00		
6	TE	≥60	≥60	0.28	0.70	1.54	55.38	30.77	12.31		
7	ESR	≥60	≥60	0.32	0.48	0.00	61.54	12.31	26.15		
8	ESR	≥60	≥60	0.31	0.54	0.00	58.46	21.54	20.00		
9	CR	≥60	≥50	0.20	0.87	6.15	41.54	29.23	16.92	3.08	0.00
10	CR	≥60	≥50	0.20	0.87	6.15	41.54	29.23	16.92	3.08	0.00
11	CR	≥60	≥50	0.28	0.84	6.15	38.46	32.31	15.38	4.62	
12	MS	≥60	≥60	0.25	0.37	0.00	60.00	29.23	10.77		
13	MS	≥60	≥60	0.55	0.54	0.00	15.38	58.46	26.15		
14	ESR	≥60	≥60	0.22	0.45	0.00	70.77	13.85	15.38		
15	ESR	≥60	≥60	0.48	0.60	0.00	47.69	9.23	43.08		
16	CR	≥60	≥50	0.14	0.58	7.69	50.77	30.77	6.15	0.00	1.54
17	CR	≥60	≥50	0.18	0.70	7.69	52.31	27.69	7.69	1.54	
18	MS	≥60	≥60	0.42	0.55	0.00	41.54	32.31	26.15		
19	ESR	≥60	≥60	0.35	0.46	0.00	56.92	16.92	26.15		
20	ESR	≥60	≥60	0.33	0.28	0.00	56.92	20.00	23.08		
21	TE	≥60	≥60	0.46	0.17	0.00	49.23	9.23	41.54		
22	ESR	≥60	≥60	0.37	0.13	0.00	55.38	15.38	29.23		
23	TE	≥60	≥60	0.28	0.33	1.54	52.31	36.92	9.23		
24	MS	≥60	≥60	0.48	0.51	0.00	27.69	47.69	24.62		
25	MS	≥60	≥60	0.32	0.50	0.00	53.85	27.69	18.46		
26	MS	≥60	≥60	0.28	0.41	0.00	61.54	20.00	18.46		
27	TE	≥60	≥60	0.06	0.39	0.00	90.77	6.15	3.08		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	p-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
28	ESR	≥60	≥60	0.09	0.00	0.00	86.15	9.23	4.62		
29	ESR	≥60	≥60	0.31	0.50	0.00	66.15	6.15	27.69		
30	ESR	≥60	≥60	0.39	0.58	0.00	53.85	13.85	32.31		
31	ESR	≥60	≥60	0.36	0.61	1.54	55.38	15.38	27.69		
32	ESR	≥60	≥60	0.25	0.29	0.00	63.08	24.62	12.31		
33	ESR	≥60	≥60	0.20	0.45	0.00	72.31	15.38	12.31		

Copyright © 2019 by Louisiana Department of Education.

Table E.10 Operational Item Statistics—English II Summer Administration

Item	Item Type	Total N	Adj. N	p-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
1	ESR	≥70	≥70	0.30	0.34	0.00	64.86	10.81	24.32		
2	ESR	≥70	≥70	0.56	0.48	1.35	33.78	18.92	45.95		
3	ESR	≥70	≥70	0.38	0.46	0.00	56.76	10.81	32.43		
4	ESR	≥70	≥70	0.50	0.33	0.00	45.95	8.11	45.95		
5	TE	≥70	≥70	0.20	0.49	4.05	70.27	13.51	12.16		
6	TE	≥70	≥70	0.21	0.46	5.41	55.41	39.19	0.00		
7	CR	≥70	≥60	0.08	0.74	9.46	58.11	20.27	2.70	0.00	0.00
8	CR	≥70	≥60	0.08	0.74	9.46	58.11	20.27	2.70	0.00	0.00
9	CR	≥70	≥60	0.12	0.74	9.46	54.05	24.32	2.70	0.00	
10	ESR	≥70	≥70	0.54	0.38	1.35	39.19	12.16	47.30		
11	ESR	≥70	≥70	0.39	0.50	1.35	44.59	31.08	22.97		
12	ESR	≥70	≥70	0.44	0.51	1.35	40.54	29.73	28.38		
13	ESR	≥70	≥70	0.55	0.39	1.35	32.43	24.32	41.89		
14	ESR	≥70	≥70	0.47	0.46	0.00	51.35	4.05	44.59		
15	TE	≥70	≥70	0.08	0.27	1.35	82.43	16.22	0.00		
16	ESR	≥70	≥70	0.46	0.39	1.35	37.84	31.08	29.73		
17	ESR	≥70	≥70	0.35	0.26	1.35	58.11	12.16	28.38		
18	ESR	≥70	≥70	0.26	0.29	0.00	64.86	18.92	16.22		
19	ESR	≥70	≥70	0.39	0.63	0.00	52.70	16.22	31.08		
20	MS	≥70	≥70	0.32	0.46	1.35	54.05	27.03	17.57		
21	TE	≥70	≥70	0.27	0.57	0.00	59.46	27.03	13.51		
22	CR	≥70	≥50	0.11	0.77	14.86	47.30	24.32	4.05	0.00	0.00
23	CR	≥70	≥50	0.11	0.77	14.86	47.30	24.32	4.05	0.00	0.00
24	CR	≥70	≥50	0.20	0.70	14.86	36.49	32.43	6.76	0.00	
25	ESR	≥70	≥70	0.32	0.16	0.00	50.00	35.14	14.86		
26	TE	≥70	≥70	0.35	0.72	1.35	56.76	14.86	27.03		
27	ESR	≥70	≥70	0.35	0.62	0.00	51.35	27.03	21.62		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	p-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D
28	TE	≥70	≥70	0.25	0.58	1.35	63.51	21.62	13.51		
29	MS	≥70	≥70	0.28	0.38	0.00	58.11	27.03	14.86		
30	MS	≥70	≥70	0.33	0.56	0.00	55.41	22.97	21.62		
31	MS	≥70	≥70	0.34	0.30	0.00	33.78	63.51	2.70		
32	TE	≥70	≥70	0.25	0.37	1.35	55.41	37.84	5.41		
33	MS	≥70	≥70	0.18	0.47	0.00	68.92	25.68	5.41		
34	TE	≥70	≥70	0.23	0.61	2.70	60.81	28.38	8.11		

Copyright © 2019 by Louisiana Department of Education.

Table E.11 Operational Item Statistics—Algebra I Summer Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	MC	≥80	≥80	0.37	-0.11	0.00		36.59	28.05	20.73	14.63		
2	MPSR	≥80	≥80	0.38	0.51	0.00	40.24	42.68	17.07				
3	MC	≥80	≥80	0.21	0.10	0.00		10.98	28.05	40.24	20.73		
4	MC	≥80	≥80	0.45	0.20	0.00		20.73	30.49	45.12	3.66		
5	MPSR	≥80	≥80	0.29	0.47	0.00	47.56	46.34	6.10				
6	MC	≥80	≥80	0.57	0.23	0.00		21.95	57.32	10.98	9.76		
7	MC	≥80	≥80	0.32	0.42	0.00		12.20	41.46	31.71	14.63		
8	MPSR	≥80	≥80	0.28	0.40	0.00	51.22	41.46	7.32				
9	MC	≥80	≥80	0.32	0.36	0.00		20.73	31.71	35.37	12.20		
10	TE	≥80	≥80	0.12	0.29	0.00	78.05	19.51	2.44				
11	MC	≥80	≥80	0.45	0.12	0.00		19.51	45.12	17.07	18.29		
12	MC	≥80	≥80	0.16	0.36	0.00		21.95	37.80	24.39	15.85		
13	MC	≥80	≥80	0.51	0.32	0.00		13.41	51.22	23.17	12.20		
14	MC	≥80	≥80	0.28	-0.01	1.22		8.54	28.05	41.46	20.73		
15	CR	≥80	≥60	0.07	0.54	13.41	69.51	6.10	6.10	1.22	0.00		
16	MC	≥80	≥80	0.26	0.56	0.00		25.61	17.07	31.71	25.61		
17	TE	≥80	≥80	0.09	0.16	1.22	90.24	8.54					
18	MS	≥80	≥80	0.22	0.47	0.00	78.05	21.95					
19	MS	≥80	≥80	0.12	0.40	1.22	86.59	12.20					
20	SA	≥80	≥70	0.53	0.53	3.66	45.12	51.22					
21	CR	≥80	≥70	0.13	0.66	8.54	69.51	6.10	6.10	4.88			
22	MC	≥80	≥80	0.28	0.22	0.00		13.41	28.05	23.17	35.37		
23	MPSR	≥80	≥80	0.35	0.47	1.22	46.34	35.37	17.07				
24	MC	≥80	≥80	0.28	0.04	0.00		12.20	28.05	37.80	21.95		
25	TE	≥80	≥80	0.20	0.47	1.22	63.41	31.71	3.66				
26	CR	≥80	≥50	0.02	0.62	21.95	60.98	6.10	0.00	0.00	0.00		
27	CR	≥80	≥60	0.14	0.70	17.07	54.88	15.85	6.10	1.22			

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	CR	≥80	≥70	0.25	0.78	7.32	51.22	4.88	12.20	8.54	2.44	3.66	9.76
29	MC	≥80	≥80	0.25	0.57	1.22		31.71	24.39	26.83	15.85		
30	CR	≥80	≥70	0.16	0.57	8.54	58.54	23.17	0.00	6.10			
31	MPSR	≥80	≥80	0.43	0.62	0.00	51.22	12.20	36.59				
32	MC	≥80	≥80	0.60	0.24	0.00		7.32	25.61	59.76	7.32		
33	MS	≥80	≥80	0.21	0.34	0.00	79.27	20.73					
34	MC	≥80	≥80	0.28	0.43	2.44		23.17	25.61	21.95	26.83		
35	CR	≥80	≥70	0.03	0.50	3.66	90.24	1.22	1.22	1.22			
36	MC	≥80	≥80	0.32	0.08	0.00		19.51	31.71	37.80	10.98		
37	SA	≥80	≥80	0.20	0.57	0.00	51.22	28.05	10.98	8.54	1.22		
38	TE	≥80	≥80	0.48	0.35	1.22	51.22	47.56					

Copyright © 2019 by Louisiana Department of Education.

Table E.12 Operational Item Statistics—Geometry Summer Administration

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
1	MPSR	≥70	≥70	0.48	0.70	0.00	33.80	36.62	29.58				
2	CR	≥70	≥50	0.36	0.87	14.08	43.66	1.41	7.04	21.13			
3	TE	≥70	≥70	0.45	0.65	0.00	54.93	45.07					
4	MC	≥70	≥70	0.62	0.52	0.00		61.97	14.08	14.08	9.86		
5	TE	≥70	≥70	0.41	0.88	1.41	46.48	22.54	29.58				
6	MC	≥70	≥70	0.61	0.42	0.00		21.13	14.08	60.56	4.23		
7	TE	≥70	≥70	0.61	0.67	0.00	19.72	39.44	40.85				
8	MPSR	≥70	≥70	0.42	0.73	0.00	45.07	26.76	28.17				
9	TE	≥70	≥70	0.29	0.89	1.41	70.42	28.17					
10	MC	≥70	≥70	0.49	0.68	0.00		14.08	49.30	26.76	9.86		
11	MC	≥70	≥70	0.42	0.42	0.00		42.25	11.27	28.17	18.31		
12	TE	≥70	≥70	0.31	0.85	1.41	67.61	30.99					
13	MC	≥70	≥70	0.37	0.61	0.00		18.31	21.13	36.62	23.94		
14	CR	≥70	≥50	0.34	0.90	21.13	38.03	2.82	2.82	2.82	8.45	9.86	8.45
15	MC	≥70	≥70	0.68	0.50	0.00		5.63	7.04	19.72	67.61		
16	CR	≥70	≥50	0.29	0.85	21.13	39.44	8.45	5.63	7.04	9.86		
17	MC	≥70	≥70	0.52	0.61	0.00		11.27	25.35	52.11	11.27		
18	SA	≥70	≥70	0.37	0.68	1.41	45.07	33.80	19.72				
19	CR	≥70	≥60	0.27	0.83	5.63	53.52	4.23	8.45	4.23	11.27	1.41	11.27
20	SA	≥70	≥70	0.34	0.76	0.00	66.20	33.80					
21	TE	≥70	≥70	0.35	0.75	0.00	47.89	33.80	18.31				
22	MC	≥70	≥70	0.55	0.37	0.00		54.93	23.94	14.08	7.04		
23	CR	≥70	≥50	0.22	0.79	19.72	42.25	14.08	15.49	1.41			
24	CR	≥70	≥50	0.29	0.93	14.08	46.48	5.63	8.45	7.04	12.68		
25	TE	≥70	≥70	0.61	0.48	0.00	39.44	60.56					
26	MC	≥70	≥70	0.38	0.51	0.00		38.03	22.54	28.17	11.27		
27	MC	≥70	≥70	0.39	0.72	0.00		39.44	29.58	26.76	4.23		

Copyright © 2019 by Louisiana Department of Education.

Item	Item Type	Total N	Adj. N	*p*-Value	Pbis	% Omit	% at 0	% at 1/A	% at 2/B	% at 3/C	% at 4/D	% at 5	% at 6
28	MC	≥70	≥70	0.49	0.67	0.00		12.68	9.86	28.17	49.30		
29	MC	≥70	≥70	0.38	0.69	0.00		38.03	22.54	32.39	7.04		
30	SA	≥70	≥60	0.54	0.52	2.82	45.07	52.11					
31	TE	≥70	≥70	0.41	0.87	0.00	26.76	29.58	14.08	11.27	18.31		
32	MS	≥70	≥70	0.24	0.65	0.00	76.06	23.94					
33	MC	≥70	≥70	0.31	0.50	0.00		30.99	12.68	14.08	42.25		
34	MC	≥70	≥70	0.32	0.69	0.00		22.54	33.80	11.27	32.39		
35	SA	≥70	≥70	0.30	0.90	1.41	69.01	29.58					
36	TE	≥70	≥70	0.38	0.68	0.00	61.97	38.03					
37	MC	≥70	≥70	0.27	0.62	0.00		14.08	23.94	35.21	26.76		
38	SA	≥70	≥70	0.35	0.71	0.00	46.48	38.03	15.49				

Copyright © 2019 by Louisiana Department of Education.

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

Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publication.

Center for Assessment. (June 2017). *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.

Dorans, N. J., & Schmitt, M. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational Testing Service.

Data Recognition Corporation. (2016). *Interpretive guide: Grades 3–8 ELA and math* Maple Grove, MN.

Educational Testing Service, Pearson, & Measured Progress. (2016). *Final technical report for 2015 administration*. PARCC. Retrieved from https://parcc-assessment.org/wp-content/uploads/2018/02/PARCC-2015-Tech-Report.pdf

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.

Copyright © 2019 by Louisiana Department of Education.

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.Wainerand H. I. Braun (Eds.), Test Validity, pp. 129-145. Hillsdale, NJ: Erlbaum.

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*, 179–197.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 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://parcc-assessment.org/parcc_pls_techreport_011316_toparcc-final/.

Pearson. (2017). *PARCC: Final technical report for 2016 administration.* PARCC. Retrieved from https://parcc-assessment.org/wp-content/uploads/2018/02/PARCC-2016-Tech-Report.pdf.

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, 479.

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

Copyright © 2019 by Louisiana Department of Education.

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*, 233–251.

Copyright © 2019 by Louisiana Department of Education.