2019 LEAP 2025
Operational Technical Summary
English Language Arts and Mathematics

Submitted to the
Louisiana Department of Education

April 2020

DATA RECOGNITION
DRC
CORPORATION

2019 Technical Summary

The tests used in Louisiana are carefully constructed to fairly assess the progress of Louisiana students. This document provides an overview of the process and summarizes some of the key psychometric information.

INTRODUCTION

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

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

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

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

In March 2016, BESE approved the Louisiana Student Standards in ELA and mathematics. In the 2016–2017, 2017–2018, and 2018–2019 school years, students in grades 3–8, except those qualifying for an alternate assessment for students with the most significant cognitive disabilities (the LAA 1 in 2016–2017 or LEAP Connect in 2017–2018 and 2018–2019), were administered forms for ELA and mathematics that consisted of PARCC assessment items while developing some Louisiana-owned items to enhance the PARCC item bank. This allowed for the continued comparability to forms administered in the 2014–2015 and 2015–2016 school years. The grade 3 and 4 forms were administered with a paper-based test (PBT) and a computer-based test (CBT), and grades 5 through 8 were a CBT administration.

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.

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.

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 make up the available item pool for the 2019 LEAP 2025 forms construction. Please refer to the PARCC Model Content Frameworks for ELA/Literacy (Grades 3-11) and PARCC Model Content Frameworks for Mathematics (Grades 3-11) for additional information about the development of item specifications and blueprints for the PARCC flagship assessments. These resources can be accessed via the Research page of New Meridian's website. The LDOE and DRC confirmed that all items selected for use on the LEAP 2025 forms were appropriate for use on Louisiana assessments by convening committees of Louisiana educators who reviewed and approved items from the item banks prior to form selection.

The ELA and mathematics LEAP 2025 assessments for grades 3–8 were developed based on the requirements of RFP #678PUR-LEAP 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 and, for grades 3 through 8 assessments, to Louisiana’s 2015–2017 assessments;

- constructed to use Louisiana's grades 3 through 8 ELA and mathematics assessments as the baseline scale[1] to report test results for grades 3 through 8 students;
- developed to limit the amount of testing time required and to be in compliance with state law regarding testing time;
- developed and reviewed with Louisiana 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 in both paper- and computer-based formats.

The item selection process for forms construction was a content-focused, collaborative process between the LDOE and DRC content specialists, followed by a psychometric evaluation of each form selection. The item selection process consisted of the following:

- Content-area assessment specialists verified that each item aligned to the content standards specified in the Louisiana Student Standards for ELA and mathematics and was not offensive, inappropriate, or biased.
- Using a PARCC pool of items, content-area assessment specialists selected tasks and/or passage sets to match the blueprint. The items covered a range of standards and addressed the appropriate claims, subclaims, and subcategories.
- Content-area assessment specialists verified that each item met psychometric guidelines for excellence given available item-performance data.
- Content-area assessment specialists verified that each item met 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.

After items were reviewed by Louisiana educator committees, LDOE, and DRC content-area assessment specialists, selected items were then placed in forms. In constructing the forms, DRC and LDOE content-area assessment specialist used these 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

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

- 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;

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.

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 are reported in Table 1. These reliability coefficients were computed using the census data. The reliability statistics ranged from 0.86 to 0.92 for all ELA forms. For mathematics, the reliabilities ranged from 0.92 to 0.94. These results indicate acceptable reliability coefficients for the LEAP 2025 tests.

Table 1 Reliability in English Language Arts and Mathematics

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

*Number of Items for ELA does not include the RI/RL component

CONSTRUCT-RELATED VALIDITY

In addition to content validity addressed in the Test Content Development and Reliability sections, additional evidence of validity, especially construct-related validity, is demonstrated through studies of convergent and divergent validity.

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

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.

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.

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 the content and the 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.

To present evidence of construct-related validity, the 2019 Louisiana LEAP 2025 Technical Report describes in detail the following validity studies:

- Decision Accuracy,
- Decision Consistency,
- Principal Components Analysis,
- Correlations among Claims, Subclaims, and Subcategories
- Reliability of Claims, Subclaims, and Subcategories, and
- Divergent (Discriminant) Validity.

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

Test-Level Scores

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

Two types of test-level scores are reported to indicate a student's achievement on the LEAP 2025: (1) the scale score and (2) its associated level of achievement.

Scale Scores

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

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

Levels of Achievement

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

Use of Test-Level Scores

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

Reporting Category and Subcategory Subscores

A student's performance on the ELA reporting categories (i.e., Reading and Writing) and mathematics categories (i.e., Major Content, Additional & Supporting Content, Expressing Mathematical Reasoning, and Modeling & Application) is reported in one of three ratings: *Weak*, *Moderate*, or *Strong*. Additionally, subcategory ratings are reported at the student level for ELA and mathematics. ELA has three subcategories for reading (i.e., literary text, informational text, and vocabulary) and two subcategories for writing (i.e., written expression and knowledge and use of language conventions). Mathematics has four subcategories and they differ by grade. Subcategory performance is reported in one of three ratings of achievement: *Strong*, *Moderate*, or *Weak*.

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

Use of the Reporting Category and Subcategory-Level Subscores

The purpose of reporting category- or subcategory-level performance ratings on LEAP 2025 assessments is to show, for each student, the relationship between the overall achievement being measured and the skills in each of the areas defined by the categories and subcategories in ELA and 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. School system and school administrators may compare their aggregate results with the state means to better understand their strengths and weaknesses within a content area. Caution should be exercised when comparing category- or subcategory-level ratings between students or across years, because different items will comprise these ratings and these items may vary in difficulty.

EQUATING OF TEST FORMS

A statistical process called equating is needed to convert the scale of the form administered in the current administration to the scale of the forms in previous administrations. This is to ensure that scores from different administrations have the same meaning. Detailed technical information describing this process can be found in the full technical report. This process places the form scores on the same scale, such that students performing on an assessment at the same level of (underlying) achievement should receive the same scaled score.

All forms for a given grade and content area should provide comparable scores, and the passing standards across different administrations should be equivalent. Therefore, a form-equating procedure is conducted every year to establish score equivalency across forms. The form-equating process ensures that students are not given an unfair advantage or disadvantage, despite whether a particular form students take is “easier” or “harder” than a form taken by other students.

Measurement Model

All 2019 LEAP 2025 item calibration and equating were performed based on item response theory (IRT)

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

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

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

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

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

where djx is the relative difficulty of score category x of item j.

IRTPRO (Cai, Thissen, & du Toit, 2011) was used for the IRT calibrations. IRTPRO is a multipurpose program that implements a variety of IRT models associated with mixed-item formats and associated statistics. IRTPRO has been used to calibrate large data sets such as those of Partnership for Assessment of Readiness for College and Careers (PARCC). The program implements marginal maximum likelihood (MML) estimation techniques for items and MLE estimation of theta.

Methodology

Calibration and linking methodology used for the Spring 2019 LEAP 2025 administration closely followed most of the PARCC methods referenced in the PARCC document "Final Technical Report for the 2015 Administration." To maintain comparability to PARCC, the 2PL/GPC IRT model was applied to item calibration using the software IRTPRO (Cai et al., 2011). To avoid local independence between the two Writing traits, Written Expression (WE) and Written Knowledge and Use of Language Conventions (WKL) were separately calibrated using sparse matrix.

The Stocking and Lord (1983) procedure was the applied linking method using the software STUIRT (Kim, S. and Kolen, M., 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.

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

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

RESULTS SUMMARY

Table 2 and Table 3 present the mean scale-score data based on all test scores valid for reporting purposes of the 2019 Spring LEAP 2025 ELA and Mathematics operational administration.

Table 2 State-Level Scale Score Statistics: English Language Arts

Grade	N	Mean SS	S.D. SS	Percentile				
				10th	25th	50th	75th	90th
3	≥52,940	745.85	42.01	690	715	746	776	802
4	≥54,800	744.11	34.37	699	721	744	768	789
5	≥54,910	742.46	32.08	701	718	742	765	784
6	≥54,800	739.70	30.18	701	718	740	761	778
7	≥52,350	744.57	35.94	695	720	747	769	789
8	≥50,720	745.65	36.32	696	721	746	771	793

Table 3 State-Level Scale Score Statistics: Mathematics

Grade	N	Mean SS	S.D. SS	Percentile				
				10th	25th	50th	75th	90th
3	≥52,820	742.28	32.66	700	720	744	764	784
4	≥54,690	740.74	31.75	699	719	742	763	781
5	≥54,730	737.10	29.68	699	714	737	757	778
6	≥54,710	733.92	29.53	696	714	733	754	773
7	≥52,090	733.40	26.05	701	714	732	752	766
8*	≥44,520	729.58	34.03	689	707	730	752	775

*Eighth grade students who enrolled in Algebra I had the option of taking the Algebra EOC test instead of the LEAP 2025 Mathematics Grade 8 test.

Table 4 and Table 5 summarize the percentages of students in each achievement level for the 2019 LEAP 2025 operational administration. Percentages may not add up to 100 due to rounding.

Table 4 Percentage of Students in Achievement Level: English Language Arts

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

Table 5 Percentage of Students in Achievement Level: Mathematics

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

*Eighth grade students who enrolled in Algebra I had the option of taking the Algebra EOC test instead of the LEAP 2025 Mathematics Grade 8 test.

REFERENCES

Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012, January). Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments. Austin, TX: Pearson. Retrieved from http://researchnetwork.pearson.com/wpcontent/uploads/Bulletin21_Evidence_Based_Standard_Setting.pdf

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

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.

Kim, S. & Kolen, M. (2004). STUIRT: A computer program for scale transformation under unidimensional item repose theory models [Computer program]. University of Iowa. Version 1.0.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201–210.