LEAP 2025 Biology
Technical Report: 2018–2019

Prepared by DRC, Pearson, and WestEd

FOREWORD

Improving student achievement is a primary goal of any educational assessment program such as the Louisiana Educational Assessment Program 2025 (LEAP 2025). This technical report and its associated materials have been produced in a way that can help educators understand the technical characteristics of the assessment used to measure student achievement.

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) and in the new edition, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Table of Contents

FOREWORD 2

1. Introduction 7

Summary of the 2018–2019 Activities 7

2. Assessment Framework 9

3. Overview of the Test Development Process 11

Item Development Plan 11

Proposal and Review of Topics and Sources 16

Performance Expectation Bundling 16

Phenomena Selection and Outline Development 17

Matching Phenomena to Item Sets 18

Outline and Stimuli Development 19

Item Writing and Review Process 20

4. Construction of Embedded Test Forms 26

Test Design 26

Initial Construction 28

Operational Form 28

Revision and Review 31

Psychometric Approval of Operational Forms 31

LDOE Review 32

Version of Test Forms 33
Online and Accommodated Print Forms 33
Accommodated Forms 33
Braille Forms 33

5. Test Administration 35

Training of School Systems 35
Ancillary Materials 36
Time 41
Online Forms Administration 42
Accessibility and Accommodations 42
Testing Windows 44
Test Security Procedures 44

6. Scoring Activities 45

Answer Key Verification 45

7. Data Analysis 58

Classical Item Statistics 58
Differential Item Functioning 59
Item Calibration and Scaling 63
Measurement Models 63
Operational Item Parameters 64
Item Fit 64
Dimensionality and Local Item Independence 66
Unidimensionality and Principal Component Analysis 67

Scaling 68

8. Reporting for Biology 70

Biology Standard Setting 71

Standard Setting Executive Report 72

LEAP 2025 Biology Standard Setting Process and Results 72

General Method 74

Results for LEAP 2025 Biology 76

9. Data Review Process and Results 77

10. Reliability and Validity 79

Internal Consistency Reliability Estimation 79

Student Classification Accuracy and Consistency 80

Validity 82

11. Statistical Summaries 85

References 87

Appendix A: Training Agendas 91

Appendix B: Test Summary 102

Appendix C: Item Analysis Summary Report 110

Appendix D: Dimensionality 123

Appendix E: Scale Distribution and Statistical Report 127

Appendix F: Reliability and Classification Accuracy 133

1. Introduction

The Louisiana Department of Education (LDOE) has a long and distinguished history in the development and administration of assessments that support its state accountability system and are aligned to the Louisiana Student Standards. Per state law, the LDOE is to administer statewide summative science assessments in grades 3–8 and in Biology. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. Further, the LDOE and the BESE are committed to the development of rigorous assessments as one component of their comprehensive plan—Louisiana Believes—designed to ensure that every Louisiana student is on track to be successful in postsecondary education and the workforce.

The purpose of this technical report is to describe the process for the embedded field test (EFT) and operational test administration of the statewide summative science assessment for high school Biology. This report outlines the testing procedures, forms construction, administration, calibration, analyses, standard-setting, and reporting of scores.

Summary of the 2018–2019 Activities

WestEd and Pearson, in partnership with the LDOE and Data Recognition Corporation (DRC), the administration vendor, developed a timeline to capture the major activities necessary to produce the spring 2019 Biology operational forms with EFT. Table 1.1 summarizes those key activities along with the months during which the activities were completed.

Table 1.1

Key Activities from October 2017 to May 2019

Date	Activity
October–December 2017	• Started item development planning for spring 2019 test • Item development plans and outlines approved by LDOE staff
December 2017–February 2018	• WestEd updated content development specifications and style guide • WestEd began item writing and development
April 2018	• WestEd updated 2018–2019 Framework and Test Construction Document based on feedback from LDOE
April–May 2018	• LDOE staff reviewed proposed item sets, tasks, and standalones • LDOE staff reviewed 2018–2019 Framework and Test Construction Document and proposed changes
May 2018	• WestEd started development of achievement-level descriptors
June 2018	• WestEd and LDOE convened Item Content/Bias Review Committee • LDOE and WestEd staff held Reconciliation meeting
July–August 2018	• Virtual planning meeting held to discuss early data results in science • Pearson, WestEd, and LDOE convened Data Review meeting for spring 2018 results • Pearson, WestEd, and LDOE reconciled results of data review • Test construction activities began
August–September 2018	• Planning meeting held • LDOE staff reviewed the fall and fall AE forms
October 2018	• Achievement-level descriptors format finalized with LDOE • 2018–2019 Framework and Test Construction Document finalized • Remaining spring 2019 materials delivered to administration vendor
November 2018	• Technical Advisory Committee Meeting convened • LDOE staff reviewed proposed spring 2019 EFT selections
November–December 2018	• Fall 2018 test administered
August 2018	• Online content delivered to administration vendor
December 2018–March 2019	• Biology achievement-level descriptors uploaded to LDOE • LDOE reviewed achievement-level descriptors
January 2019	• LDOE/WestEd/DRC met for planning meeting
April–May 2019	• Spring 2019 test administered, including EFT • Pearson initiated Standard Setting

2. Assessment Framework

An assessment framework addresses the test design, test blueprint, range of standards covered, reporting categories, percentages of assessment items and score points by reporting category, projected testing times, numbers of forms to be administered, and select psychometric analysis activities.

Measuring student proficiency of the full depth and breadth of the Louisiana Student Standards for Science (LSSS) requires assessments built from a range of item types. As a general rule, the choice of a specific item type is a function of efficient and effective measurement of the target content. Multiple-choice (MC) and multiple-select (MS) item types provide students an opportunity to select the correct answer or answers from a set of answer choices. MS items can elicit a greater depth of understanding than traditional MC items by requiring the selection of more than one correct response, efficiently scored by an automated scoring engine. Constructed-response (CR) and extended-response (ER) items allow students to develop an explanation, describe a model, design a solution, and/or otherwise apply and communicate scientific understanding as required by the Science and Engineering Practices (SEPs) and Crosscutting Concepts (CCCs). These types of student-produced responses are scored by teams of trained readers. Technology-enhanced (TE) items allow students to apply and communicate scientific knowledge and understanding as required by the SEPs and CCCs in ways that may not be addressed by MC or MS item types, but in a manner more cost-effective and less time-consuming than CR and ER item types with automated engine scoring. TE items may ask students to develop models or to sort processes by dragging components into a valid order, construct viable explanations by selecting words or phrases from several drop-down menus, or complete other tasks. The complexity of the TE items reduces the probability of randomly guessing the correct answer. Two-part items involve the application of understanding different but related knowledge to a concept or to support assertions with evidence.

For two-part items, students may construct an explanation and support the explanation with evidence or make a claim and evaluate evidence to support that claim. Another application of two-part items is to develop a model in part A and to evaluate the model in

part B. A range of item types and applications allows greater test taker engagement and provides a more authentic assessment experience.
The test design includes item sets, a task, and standalone items. A stimulus that describes a scientific phenomenon anchors each item set or task. A focus that details some aspects of a phenomenon provides the common anchor for standalone items. Item sets are composed of four items associated with a common stimulus. The item sets may include 1-point selected-response items (single-select and/or MS formats), 1- and 2-point TE items, and 2-point two-part items (two-part independent [TPI] and/or two-part dependent [TPD] formats). Three of the item sets also include a 2-point CR item. In addition to the item sets, the assessment contains one task. Tasks are made up of five items tied to a common stimulus. Tasks may include 1-point selected-response items (single-select and/or MS formats), 1- and 2-point TE items, 2-point two-part items (TPI and/or TPD formats), and a 9-point ER item. Standalone items may be either 1-point selected-response items (both single-select and MS formats), 1- and 2-point TE items, or 2-point two-part items (TPI and/or TPD formats). The standalone items provide flexibility to meet the test blueprint and afford greater coverage of the standards while still requiring students to make connections among the three dimensions of the LSSS. All points associated with the task set contribute to a student's overall score, but the 9-point ER item is not a component of the current blueprint and therefore not included in the proportional representation of content assessed by other parts of the test.

The assessment is administered primarily online. However, an accommodated paper version of the assessment is available for students who are unable to test online. For accommodated paper forms, TE items are adapted to a paper format to assess the same content.

The Assessment Framework was reviewed by LDOE content and psychometric staff to ensure that the test designs, blueprints, and form designs met the necessary content, reporting, and psychometric requirements.

3. Overview of the Test Development Process

Item Development Plan

A table of acronyms used in item and test development is presented below.

Table 3.1a

Acronyms Used in Biology Item and Test Development

Acronym	Meaning
ARG	Engaging in Argument from Evidence
CCC	Crosscutting Concepts
C/E	Cause and Effect
DATA	Analyzing and Interpreting Data
DCI	Disciplinary Core Ideas
E/M	Energy and Matter
E/S	Constructing Explanations and Designing Solutions
INFO	Obtaining, Evaluating, and Communicating Information
INV	Planning and Carrying Out Investigations
LEAP	Louisiana Educational Assessment Program
LS	Life Science
LSSS	Louisiana Student Standards for Science
MCT	Using Mathematics and Computational Thinking
MOD	Developing and Using Models
PAT	Patterns
PE	Performance Expectation
Q/P	Asking Questions and Defining Problems
S/C	Stability and Change
SEP	Science and Engineering Practices
S/F	Structure and Function
SPQ	Scale, Proportion, and Quantity
SYS	Systems and System Models

The blueprint components that guided item development projections for biology are presented in the following tables.

Table 3.1b

Test Blueprint for LEAP 2025 Biology: DCI Domain Coverage

Biology: DCI Domain Coverage			
	# of PEs in LSSS	Relative % in LSSS	% by points of all items
LS1	8	40%	35%–45%
LS2	4	20%	15%–25%
LS3	3	15%	10%–20%
LS4	5	25%	20%–35%
Total	20	100%	

LS1 From Molecules to Organisms: Structures and Processes
LS2 Ecosystems: Interactions, Energy, and Dynamics
LS3 Heredity: Inheritance and Variation of Traits
LS4 Biological Evolution: Unity and Diversity

Table 3.1c

Test Blueprint for LEAP 2025 Biology: Minimal PE Coverage

Biology: Minimal PE Coverage Every PE will be included at least one time in a test			
	SEP	CCC	Min items
HS-LS1-1	6E/S	S/F	1
HS-LS1-2	2MOD	SYS	1
HS-LS1-3	3INV	S/C	1
HS-LS1-4	2MOD	SYS	1
HS-LS1-5	2MOD	E/M	1
HS-LS1-6	6E/S	E/M	1
HS-LS1-7	2MOD	E/M	1
HS-LS1-8	8INFO	SPQ	1
HS-LS2-1	5MCT	SPQ	1
HS-LS2-4	5MCT	E/M	1
HS-LS2-6	7ARG	S/C	1
HS-LS2-7	6E/S	S/C	1
HS-LS3-1	1Q/P	C/E	1
HS-LS3-2	7ARG	C/E	1
HS-LS3-3	4DATA	SPQ	1
HS-LS4-1	4DATA	PAT	1
HS-LS4-2	6E/S	C/E	1
HS-LS4-3	4DATA	PAT	1
HS-LS4-4	6E/S	C/E	1
HS-LS4-5	7ARG	C/E	1

Table 3.1d

Test Blueprint for LEAP 2025 Biology: CCC Coverage

CCC Overall	# of PEs in LSSS	Relative % in LSSS	% by Points of CCC Items
CCC 1 – PAT	2	10%	5%–15%
CCC 2 – C/E	5	25%	20%–30%
CCC 3 – SPQ	3	15%	10%–20%
CCC 4 – SYS	2	10%	5%–15%
CCC 5 – E/M	4	20%	15%–25%
CCC 6 – S/F	1	5%	5%–15%
CCC 7 – S/C	3	15%	10%–20%
Total	20	100%	

Table 3.1e

Test Blueprint for LEAP 2025 Biology: SEP Coverage

SEP Overall	# in PEs in LSSS	Relative % in LSSS	% by Points of SEP Items
SEP 1 – Q/P	1	5%	5%–15%
SEP 2 – MOD	4	20%	15%–25%
SEP 3 – INV	1	5%	5%–15%
SEP 4 – DATA	3	15%	10%–20%
SEP 5 – MCT	2	10%	5%–15%
SEP 6 – E/S	5	25%	20%–30%
SEP 7 – ARG	3	15%	10%–20%
SEP 8 – INFO	1	5%	5%–15%
Total	20	100%	

Table 3.1f

Test Blueprint for LEAP 2025 Biology: SEP Reporting Category Coverage

SEP Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (1 & 3)	2	11%	6%–16%	7
Reporting Category 2 (4, 5, 7)	8	42%	37%–47%	7
Reporting Category 3 (2 & 6)	9	47%	42%–52%	7
Total	19	100%		

Note that for SEP reporting category coverage, SEP 8 (Obtaining, evaluating, and communicating information) is assumed to be embedded within each reporting category (1–3), so SEP 8 is not being repeated across the reporting categories.

Table 3.1g

Test Blueprint for LEAP 2025 Biology: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	# SR	# CR, TE, Two-part	# ER	Total Items	Total Points
4-Item set	5	4	6	2	2	0	20	30
Standalone items	1	16	22	10	6	0	16	22
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	14	10	1	41	67

The Biology assessment item development plan was created in conjunction with LDOE content staff. The development plan allowed for item attrition throughout the item development process, including reviews by LDOE assessment staff and by a content and bias review committee consisting of Louisiana educators. In addition, the number of items to be field tested also allowed for item loss due to deviations from psychometric criteria for item statistics based on student performance.

The development plan and the content distribution determined the focus of the item and task sets and standalone items to be developed. This section describes the processes used to develop the item sets, task sets, and standalone items. Table 3.2 shows the item

development plan for the number of items developed by WestEd by reporting category. Note that the test design specified that the test alternates by year between field testing item sets and tasks. Spring 2019 was designated as an "item set" year for field testing, therefore no tasks were proposed for development.

Table 3.2

Number of Items Developed for Biology Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/TPI	ER	CR	Total Number of Items (non-ER/CR)
Item sets	3	11	3	6	7	0	3	27
Tasks	0	0	0	0	0	0	0	0
Standalone items	n/a	6	1	3	6	0	0	16

Proposal and Review of Topics and Sources

Performance Expectation Bundling

In the previous item development cycle, WestEd used the 2017 LSSS to recommend how performance expectations could be bundled in a task or item set to ensure that the breadth of all dimensions of constituent PEs are assessed in a meaningful way. Key to this bundling was the need to ensure that bundles and phenomena achieved a "natural fit" that supported the assessment of each phenomenon. Therefore, not all PEs were bundled, and some PEs were bundled in multiple groupings. Based on the specific nature of the performance expectations comprising each bundle, the LDOE and WestEd determined that some item sets and tasks would allow a "mix and match" approach in which the disciplinary core idea (DCI) and crosscutting concept (CCC) for one of the PEs in a bundle could be used to develop items aligned to the other PE in the bundle. Within each task or item set, each item was given a primary assignment to a single PE in the bundle, and to two or three of the dimensions comprising the three-dimensional structure

of the performance expectation. However, the items in each item set or task work together to assess the multidimensional nature of the performance expectations bundle. At the end of this process, LDOE approved 28 bundles for the 2017–2018 Biology assessment

An additional two bundles were proposed for the 2018–2019 cycle. Of the total of 30 bundles, 3 were targeted for development in the 2018–2019 cycle. One bundle continued to be kept on hold for use in other contexts.

Phenomena Selection and Outline Development

Phenomena describe observable events in nature and include relevant data, images, and text that provide students with the information they need to engage in the scientific practices described in the LSSS. The stimuli for the LEAP 2025 Biology assessment center around scientific phenomena and text, images, tables, graphs, models, and graphic organizers created by WestEd's Design Team.

Phenomena and bundles were chosen to represent the breadth of assessable science content. As part of the item development plan, all PEs were aligned to at least one standalone item or an item in an item set.

After studying the LSSS, the content lead generated lists of bundled and associated phenomena for item sets.

When identifying a phenomenon, the content lead considered:

- the emphasis of each performance expectation, as described in the clarification statements for each performance expectation;
- whether a proposed phenomenon was rich enough to support the required number of items, including overage;
- whether the phenomenon fit with the "PE bundles" developed earlier to provide meaningful, three-dimensional assessment of performance expectations; and
- whether the phenomenon was well suited for an item set (rather than a task).

Phenomena were chosen to represent the breadth of content described in the LSSS. The process of determining phenomena and associated bundles was iterative and included the identification of phenomena that could be assessed with a particular bundle, as well as understanding the need to assess PEs that had not been assessed in the previous field test.

Matching Phenomena to Item Sets

As the test design called for item sets and tasks to be field tested in alternate years, item sets were targeted for development for the 2018–2019 development cycle. The narrowing of set types to item sets influenced the selection of phenomena. Like the tasks, the item sets are phenomena-based, but unlike the tasks, they are made up of independent items that do not build upon each other. Also, unlike the tasks, the items in the item sets do not scaffold to help discriminate student performance levels, do not require a specific order, and do not contain a three-dimensional extended-response (ER) item. Although an item set does not need to contain a constructed-response (CR) item, for the 2018–2019 development cycle, WestEd developed CRs for all item sets. In total, 3 CRs were developed for 3 item sets.

For the item sets, WestEd offered a document containing descriptions of 8 phenomena associated with bundles to the LDOE for review prior to item development. Based on the list, the LDOE identified 2 phenomena to be developed into stimuli for the item sets. Additionally, one phenomenon submitted during the 2017–2018 development cycle was also identified for development. Upon approval of the phenomena, WestEd submitted item outlines containing stimuli and item descriptions to the LDOE. Once the item outlines were approved, item development for the item sets began.

In contrast to item sets and tasks, standalone items reflected independent content and are supported by a focus. A focus differs from a phenomenon in that it explores only certain key aspects of an event and is typically supported by less data. As stated previously, the standalone items were included within the blueprints to provide greater coverage of the standards assessed and to provide flexibility in meeting the blueprints and test characteristic curve targets across test administrations. The WestEd content lead

developed the foci for standalone items, based on standards that lacked coverage across the item sets and tasks. Consequently, these items were developed last. For standalone items, WestEd submitted the items and corresponding foci simultaneously; there was no separate focus approval phase for these items.

Outline and Stimuli Development

WestEd used both experienced internal and external science assessment editors to develop the phenomena-based stimuli for the item sets. Before the editors began the process, the WestEd content lead trained them on the process of conducting an effective literature search, on the LDOE's objectives, and on best practices for accessibility, as well as bias and sensitivity issues. For an outline of the training, see Appendix A for the LEAP 2025 Biology Training Agenda (2018–2019).

To support the outline development process, writers were given the Louisiana Student Standards for Science. They were also provided specific item set templates that described the PE bundle to be written to, as well as the point value, item types, dimensional alignment of each of the items in the set, and whether the dimensions of the bundled PEs could be mixed or matched. The outline contained space for writers to enter the primary sources they used in researching their phenomenon and writing their stimulus, space for the writers to include a draft of the stimulus and its supporting data, as well as space to describe each item and its metadata. Writers submitted their item outlines to the editors, who finalized the item set outlines before they were submitted to the content lead and manager for senior review. After this review, the outlines were submitted to the LDOE.

Evaluating the Reading Level of Stimuli. WestEd performed Lexile and ATOS analyses on each stimulus to obtain quantitative measures of the readability of the texts. The Lexile Analyzer, developed by MetaMetrics, analyzes the semantic and syntactic features of a text and assigns it a Lexile measure. MetaMetrics also provides grade-level ranges corresponding to Lexile ranges. It should be noted that the grade-level ranges include overlap across grade levels. The ATOS text analysis tool, developed by Renaissance Learning, takes into account the most important predictors of text complexity, including average sentence length and average word length, and uses a graded vocabulary list of more than 100,000 words to analyze word difficulty level. It reports on a grade-level scale.

In addition to the Lexile and ATOS measures, the LSSS were used as an additional measure of grade-level appropriateness. WestEd and the LDOE also drew on the professional experience of educators, during Content and Bias Committee review, to verify that sources would be accessible to students, and made changes based on their feedback. Most of the stimuli developed for the assessments were found to be below or at grade level; however, some of the science vocabulary was evaluated as above grade level. In those cases, additional support such as parenthetical definitions (glossing) was added for words that were above grade level and for words or phrases that were thought to be sources of potential confusion for students. The appropriateness of the stimuli for both content and readability was an explicit part of the content review process with Louisiana teachers.

Item Writing and Review Process

WestEd employed a cadre of item writers for the Biology assessment. All writers' resumes were reviewed and approved by the LDOE before engaging in any item development activities. As the first step in the item writing process, the WestEd content lead provided a webinar training to all writers in January 2018. For an outline of the information covered, see Appendix A for the LEAP 2025 Biology Item Training Agenda (2018–2019). In the training, writers were provided context for the assessment, including LDOE expectations, the LSSS, and a review of best practices for item development. The item writers were provided the approved item topics and drafts of the stimuli, as well as item outlines that provided explanations of the phenomena underlying the item sets. Item writers were also provided with alignment to the Science and Engineering Practices, Crosscutting Concepts, and Disciplinary Core Ideas of the LSSS, and guidance on how each item set should be developed. The use of item set overviews allowed WestEd to provide direction to the items developed during the development cycle. For standalone development, item writers were provided with assignments that indicated the number of items to write to each performance expectation, as well as the specific dimensions to align to for each item.

The item writing assignments for each item set also specified the set type, the item types (e.g., SR, MS, TE, TPI, TPD, CR), the number of items to be written, as well as potential item stems to be used for each item. Significant attention was devoted to understanding how to write TE items as well as scoring guides for CR items. Although all the writers were

science writers with experience in writing three-dimensional items, WestEd also gave instructions in basic assessment item writing principles. Writers were instructed to make certain that the vocabulary and context of the items were grade-level appropriate, to ensure that the distractors were incorrect but plausible, and to avoid cueing and outliers in the items. Writers were also provided training in universal design and bias/sensitivity. A variety of items were presented and reviewed using universal design and bias/sensitivity lenses. This training also included an overview of these topics (see Appendix A for the LEAP 2025 Biology Item Writer Training Agenda). WestEd provided training and feedback to the writers throughout the development cycle, as the LDOE and WestEd gained a clearer understanding of how the stimuli, items, and item sets worked together.

WestEd provided additional training to a subset of editors outlining the specific responsibilities for those who served as editors for the Biology assessment. For an outline of the information covered, see Appendix A for the LEAP 2025 Biology Training Agenda (2018–2019). Items went through two rounds of content editing that examined characteristics of items including alignment to the dimensions of the performance expectations of the LSSS, content accuracy, cognitive complexity, and quality of distractors. Items then went through one round of proofreading, which focused on grammar, usage, and consistent style of graphics, and a final round of review before being submitted to the LDOE for their first round of review.

Item Development Platform. Items were developed in Assessment Banking and Building solutions for Interoperable assessment (ABBI), Pearson's proprietary item development platform. In addition to the items and stimuli, the platform captured item metadata and allowed viewers to preview items using Pearson's format viewer (TestNav 8). In this view, items appeared together with all the associated stimuli in the set. The ability to examine the items and stimuli as a set was critical in the item review and in the evaluation of the sets' content and cognitive demands on students.

Style Guidelines. Style guidelines continued to be based on documentation established with the LEAP 2025 Social Studies and Biology assessments. This documentation was amended and updated as the development cycle progressed. When questions of style arose that were unanswered by existing documentation, WestEd consulted the LDOE, and approved changes were added to the project style guide.

LDOE Content Review. As writing and editing for batches of item sets and standalone items were completed, these batches were sent to the LDOE for review by the LDOE Science Assessment Coordinators; Director of Assessment Development for Math, Science, and Special Populations; Elementary Assessment Coordinator; Special Populations Assessment Coordinator; and Science Program Coordinator. Feedback from the LDOE review was implemented before the content and bias review meetings.

Content and Bias Review. After the completion of item development, WestEd coordinated face-to-face content and bias review meetings, convened in Baton Rouge. The meetings were led by facilitators from the LDOE and from WestEd. Participants included current classroom teachers, retired teachers, content specialists, and school administrators. For both content and bias review meetings, participants completed nondisclosure agreements as part of the activities. The recruitment process, conducted by LDOE staff, included participants from regions across the state. Participants represent the population of Louisiana students served—including special education, English Learners, and students with disabilities—as well as the diverse geographic and demographic composition of the state. Table 3.3 provides the demographic characteristics of the review committee.

Table 3.3

Representation of Educators Participating in 2018–2019 Content and Bias Reviews

Characteristic	Number of Participants
Classroom Teacher	9
Content/Curriculum Specialist	0
School Administrator	0
Other Staff	2
ELL Teacher	0
Special Education Teacher	0
Special Ed Teacher – Gifted	0
Visually- or Hearing-Impaired Teacher	1
Black or African American	2
Asian	0
Hispanic/Latino	1
White	7
Male	3
Female	7
Total Participants	10

Note: As teachers may fulfill multiple roles, representation of roles exceeds number of total participants.

Before the committee members began the item review process, they received an orientation from the LDOE about the LEAP 2025 Biology assessment, and the WestEd content lead provided training on the criteria for evaluating items for content and bias considerations and the use of ABBI for item review. The committee members individually reviewed PE, SEP, DCI, and CCC alignment for each item and recorded the degree of alignment for each dimension and overall alignment on a worksheet on a scale of 0 (not aligned) to 3 (well aligned), referring to LSSS Appendix A (Learning Progressions). An item was considered to have a high degree of alignment if it aligned to the particular bullet listed in the PE. An item was considered to have a lower degree of alignment if it aligned

to another bullet listed in the learning progression for that SEP or CCC. Committee members also recorded whether the science for each item was accurate and whether each item was free of bias. Areas of concern considered included opportunity and access, portrayal of groups represented, and protecting privacy and avoiding offensive content.

After the review of each item, each member voted in ABBI on whether to accept, accept with edits, or reject each item, recording comments for any item where they noted issues with science accuracy or bias. (If participants skipped an item or chose not to record a decision for a given item, the system registered the response as “No Vote” for that individual review. “No Vote” was recorded as the consensus rating when an initial group decision on an item was not reached, and the committee failed to return to that item and register a final vote to accept, revise, or reject the item.) Participants used personal laptops or laptops provided by WestEd to access ABBI. At the end of each day, WestEd made certain that the participants cleared their computer caches and deleted their download histories for the day. WestEd monitored participants to be sure that they did not use their cell phones at the table. WestEd also collected all materials at the end of each day, including notepads provided to the participants to write notes on as they reviewed the items.

Following the individual reviewers’ votes, the group came together to view and discuss each stimulus and item as it was projected on-screen with the goal of achieving consensus. The WestEd facilitators compiled detailed notes about committee decisions for implementation after the review. Because of the limited time available, there was not a review and discussion of every set as a full committee. In those cases, the LDOE facilitator reviewed the individual comments of the participants and provided a final decision for those items and stimuli.

Results of Content Review. The results of the reviewers’ individual judgments were captured in ABBI. Table 3.4 provides these results, based on the participants’ individual votes on each item following their initial review.

Table 3.4

Vote Totals Based on Individual Votes Following Initial Review

Item Type	Number of Items	Votes to Accept	Votes to Accept with Edits	No Vote	Votes to Reject	Total Votes
CR	3	27	1	0	0	28
ER	0	0	0	0	0	0
MC	16	151	7	0	1	159
MS	1	8	1	0	0	9
TE	13	122	4	0	0	126
TPD	9	76	9	0	0	85
TPI	4	36	3	0	0	39
Stimulus	3	17	1	1	0	19
All Biology	**49**	**437**	**26**	**1**	**1**	**465**

After the committee members voted individually on each item, items were discussed as a whole group and a determination was made to accept, revise, or reject each item. At the end of the meeting, no items were rejected by the group. The others were either accepted as is or accepted with edits. None of the item sets were rejected by the committee.

Post-Review Finalization. After the content and bias review, the WestEd staff implemented the committee's feedback and then met virtually with LDOE staff for reconciliation. WestEd provided records of all implemented changes to the LDOE prior to the virtual reconciliation meetings. During the reconciliation meeting, content leads from the LDOE and WestEd reviewed items to ensure that the items reflected the content, clarity, and style appropriate for inclusion in the field test. Following the reconciliation meetings, which focused on the finalization of item content, the LDOE and WestEd content leads worked together to finalize the scoring guides for CR and ER items through a separate series of communications. Once all content considerations were resolved, all items and stimuli went through a final formal fact-checking round and two additional rounds of proofreading. Any changes resulting from these reviews were submitted to the LDOE for approval.

4. Construction of Embedded Test Forms

Test Design

To assess the integrated nature of the content, practices, and crosscutting concepts of the LSSS, the LEAP 2025 Biology Assessment involved a set-based design. The test included item sets and a task, each anchored by a common stimulus or stimuli. Additionally, standalone items were included to support meeting the specific targets of the test blueprint. Table 4.1 shows the Test Design for Biology.

Table 4.1
Test Design for Biology

Test Session	Number of Items
Session 1: OP Item set	1–3 OP item set SR item(s) 0–3 OP item set TE item(s) 0–2 OP item set TPI/TPD item(s) 0–1 OP item set CR item(s)
OP Item set	1–3 OP item set SR item 0–3 OP item set TE item(s) 0–2 OP item set TPI/TPD item(s) 0–1 OP item set CR item(s)
OP Item set	1–3 OP item set SR item(s) 0–3 OP item set TE item(s) 0–2 OP item set TPI/TPD item(s) 0–1 OP item set CR item(s)
OP Standalone items	1 OP standalone SR item(s) 0–2 OP standalone TE item(s) 0–2 OP standalone TPI/TPD item(s)
FT standalone item	0–1 FT standalone SR item(s) 0–1 FT standalone TE item(s) 0–1 FT standalone TPI/TPD item(s)
Session 2: OP Task	1–4 FT task set SR item(s) 0–3 FT task set TE item(s) 1 FT task set ER item

Test Session	Number of Items
FT Item set	1–3 FT item set SR item(s) 0–3 FT item set TE item(s) 0–2 FT item set TPI/TPD item(s) 0–1 FT item set CR item(s)
OP Standalone items	1 OP standalone SR item(s) 0–2 OP standalone TE item(s) 0–2 OP standalone TPI/TPD item(s)
FT Standalone item	0–1 FT standalone SR item(s) 0–1 FT standalone TE item(s) 0–1 FT standalone TPI/TPD item(s)
Session 3: OP Item set	1–3 OP item set SR item(s) 0–3 OP item set TE item(s) 0–2 OP item set TPI/TPD item(s) 0–1 OP item set CR item(s)
OP Item set	1–3 OP item set SR item(s) 0–3 OP item set TE item(s) 0–2 OP item set TPI/TPD item(s) 0–1 OP item set CR item(s)
Operational standalone item	8 OP standalone SR items 0–2 OP standalone TE item(s) 0–2 OP standalone TPI/TPD item(s)
FT Standalone items	0–2 FT standalone SR item(s) 0–2 FT standalone TE item(s) 0-2 FT standalone TPI/TPD item(s)
Total Operational Items Tested for Biology Fall 2018	16 OP standalone SR items 1 OP task set SR item 2 OP task set TE items 1 OP task set TPD item 1 OP task set ER item 10 OP item set SR items 7 OP item set TE items 3 OP item set CR items

Test Session	Number of Items
Total Operational Items Tested Across Forms for Biology Spring 2019	9 OP standalone SR items 3 OP standalone TE items 4 OP standalone TPD/TPI items 2 OP task set SR items 4 OP task set TE items 2 OP task set TPD item 2 OP task set ER items 9 OP item set SR items 3 OP item set TE items 5 OP item set TPD/TPI items 3 OP item set CR items
Total Items Field Tested Across Forms for Biology Spring 2019 (includes re-embedded operational items)	37 FT standalone SR items 22 FT standalone TE items 13 FT standalone TPD/TPI items 33 OP item set SR items 30 OP item set TE items 12 OP item set TPD/TPI items 10 OP item set CR items

Initial Construction

The purpose of the fall 2018, spring 2019, and summer 2019 forms construction activities was to create operational forms using the spring 2018 embedded field test and to embed field test items in the spring 2019 form for potential use in future operational assessments. This section describes the process used to create operational and field test forms.

Operational Form

Data review-approved items from the spring 2018 embedded field test were available for use on the fall 2018 and spring 2019 operational assessments. (See the *LEAP 2025 Biology Technical Report: 2017–2018 Field Test* for results from the data review and reconciliation of the spring 2018 field test items.)

WestEd completed item selection for one operational (OP) form and one administrative error (AE) form for the fall 2018 administration. The designation of these forms was reversed for the spring 2019 administration so that the fall operational assessment became the spring 2019 administrative error form and the fall administrative error form became the basis for creating the spring operational form. Two operational forms were created for the spring administration. The difference between the two forms was the task that was selected. Otherwise, the item sets and standalone items were the same across the two forms.

WestEd worked with the LDOE content staff to select items for the forms following the data review meeting in August and submitted these forms to Pearson psychometricians for consideration before formal submission to the LDOE for approval. The operational and administrative error forms were designed to adhere to the blueprint for Biology and exhibit the broadest possible balance of breadth of PE coverage. Based on these considerations, the WestEd content lead selected the task first and followed with a combination of item sets and standalone items that would ensure that the relative distribution of score points by reporting category would meet the blueprint for the operational assessment and administrative error forms for Biology while avoiding similar content and topics across the balance of items and item types. Placeholder items were included on the fall operational and administrative error forms to match the location and item types of the field test items that would appear on the spring 2019 forms. The spring 2019 administrative error form included placeholder items. Table 4.2 provides the operational test composition for Biology for fall 2018 and spring 2019.

Table 4.2
LEAP 2025 Biology: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	# SR	# CR, TE, Two-part	# ER	Total Items	Total Points
4-Item set	5	4	6	2	2	0	20	30
Standalone items	1	16	22	10	6	0	16	22
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	14	10	1	41	67

Field Test Versions

Twenty-four embedded field test forms were administered in spring 2019 for Biology. This number is greater than the number of item sets available for field testing. Because standards for the Biology assessment were to be set in summer 2019 following the spring 2019 administration, there was a need to re-embed as many items as possible so that those items would appear on the same scale for future operational assessments. All the items selected for the fall 2018 operational administration were embedded in field test positions. Additional item sets that had been field tested in spring 2018 were also embedded and re-field tested. One or two versions of each item set were field tested as needed.

Items to be field tested were embedded within the three sessions of the operational form. The field test items included an item set in session 2, one standalone item in session 1, one standalone item in session 2, and two standalone items in session 3. Thus, the field test design included a subset of item types (item sets and standalone items) that appear within the operational portion of the form.

Because fewer standalone items were developed than positions were available across the 24 field test forms, standalone items were repeated as necessary across the forms.

In addition to content balance, the WestEd content lead was careful to avoid cueing and clanging between items. Cueing occurs when content in one item provides clues to the answer of another item. Clanging refers to overlap or similarity of content. Because content was purposefully distributed across the forms, cueing and clanging were intended to have been avoided; however, developers also conducted a separate review of the forms to check for inadvertent cueing or clanging.

Following the final item placement by the WestEd content lead, test maps containing each item's unique identification number (UIN) were created. The test maps captured details about each proposed form, including test session, item sequence, unique item number, and associated item metadata. Item descriptions were also included for each item, to aid in the review of the selection and placement of individual items.

Revision and Review

Psychometric Approval of Operational Forms

Prior to submitting the forms to LDOE staff for review, Pearson psychometricians and WestEd content specialists participated in an iterative process of reviewing and revising the forms. The psychometric review consisted of comparisons of the expected representation and the actual representation of reporting categories, science and engineering practices, disciplinary core ideas, crosscutting concepts, performance expectations, and item types—SR, CR, TE, TPI, TPD, and ER—on the operational forms.

The answer keys for MC items also were examined, to determine whether any forms had significantly non-uniform distributions of correct responses (A, B, C, and D). Spreadsheets were used to generate frequency tables of reporting categories, science and engineering practices, disciplinary core ideas, crosscutting concepts, performance expectations, item types, and MC answer keys for each form and across forms. Deviations from the blueprint were identified and addressed. Test characteristic curves (TCC) based on item response theoretic models were applied to data, and conditional standard errors of measurement were computed for each iteration during the test construction process to evaluate how well a proposed test form matched psychometric targets. Psychometric approval from

Pearson was provided for all forms prior to submission to the LDOE for their review. Please refer to the following table for criteria to flag items based on scoring point.

Table 4.3
Summary of Flagging Criteria to Select/Flag Items: Classical Analysis and IRT

Point	*P*-value		P-B	DIF	IRT		
	Low Bound	Upper Bound	Lower Bound	Exclude	a	b	C
1	0.25	0.90	0.20	C	0.35 – 3.50	-3.00 – 3.00	< 0.35
2 and higher	0.25	0.90	0.20		0.35 – 3.50	-3.00 – 3.00	N/A

Note: Detailed information can be found from 2018–2019 Framework and Test Construction Document. It should be noted that these values are psychometric recommendations. Actual item decision occurs by content staff based on these recommendation criteria.

LDOE Review

Following the psychometric reviews, the test maps and constructed sets were delivered to the LDOE for approval. Forms were reviewed by both LDOE content and psychometric staff. Based on the LDOE review, sets or items were replaced and the sequence of answer choices (for field test items) and the sequence of items within sets were revised as requested. Following these changes, the overall balance of answer choices and key runs was re-evaluated, and final adjustments were made to achieve the appropriate balance.

Finalized test maps were used to create PDF versions of paper forms, which were reviewed by WestEd's proofreaders before the items were transferred from ABBI to DRC.

Version of Test Forms

Online and Accommodated Print Forms

The LEAP 2025 Biology assessment is administered as Computer Based Tests (CBT) with an accommodated print form only for students who cannot complete the assessment online. For fall 2018, Form A was the operational base form and Form B was used as the administrative error form. Both forms contained item set and standalone placeholders. For spring 2019, Form B and Form C were administered as the operational base form. Twelve field test versions of Form B and twelve field test versions of form C were administered. Form A was used as the administrative error form. For summer 2019, Form A was used as the operational base form, with item set and standalone placeholders. Form B (with item set and standalone placeholders) was used as the administrative error form.

Accommodated Forms

For each administration, the accommodated print form was selected based on the field test version that contained fewest and least complex technology-enhanced items. This version was identified as Version 1. The technology-enhanced items in this version were converted to a paper and pencil format that allowed students to record their responses, or have their responses transcribed into the test booklet. In addition, alternate text was written for all stimuli and items containing graphics.

Braille Forms

Braille forms were constructed to enable students with visual impairments to participate in the LEAP 2025 assessments. The operational items in Version 1 of the accommodated print forms for fall 2018 and spring 2019 were used to construct the fall 2018 and spring 2019 braille forms. There are not large-print versions of the Biology accommodated print forms. Instead, students needing a large-print version in Biology use larger-sized monitors and/or the magnification features of the online testing system. All online test content has been developed to scale in relation to the available area on larger monitors while

maintaining the correct aspect ratio. Specific recommendations on how to transcribe items into braille were provided by the braille publisher to produce the braille version of the LEAP 2025 assessments and the test administrator's notes that accompany the braille forms. The goal was to maximize the number of items on the braille forms that could be transcribed into braille.

For students who were administered a large-print or braille test form, examiners are instructed to transcribe students' responses from the large-print test or braille test form into the online testing system (INSIGHT), exactly as the responses appear in the original form.

5. Test Administration

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

Training of School Systems

To ensure that LEAP 2025 HS assessments are administered and scored in accordance with the Department's mandates, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE offers monthly webinars, and weekly office hours to school system testing coordinators to communicate with and train school systems. The LDOE provides train-the-trainer opportunities for the school system test coordinators, who in turn convey test administration training to schools within their 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 systems. They disseminate information to each school, offer assistance with test administration, and serve as liaisons between the LDOE and their school system. The LDOE also provides assistance with and interpretation of assessment data and test results.

Ancillary Materials

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

For the spring 2019 test administration, DRC produced an administration manual, the *LEAP 2025 High School/EOC Test Administration Manual* (TAM), which serves for the LEAP 2025 administrations.

DRC also produced a test coordinator manual. LDOE assessment staff review, provide feedback, and give final approval for the test administration and test coordinator manuals. The manuals are inclusive of all LEAP 2025 HS assessments in ELA, mathematics, social studies, and science. The manuals provide detailed instructions for school systems and school test coordinators' responsibilities to distribute and collect test materials and to return test materials to DRC when appropriate as outlined in its table of contents.

Test Coordinators Manual Table of Contents

1. Key Dates
2. Resources Available in eDIRECT
3. LEAP 2025 and EOC 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. LEAP 2025/EOC 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. Testing Conditions
   8.5. Testing Schedule
   8.6. Extended Time for Testing
   8.7. Extended Breaks

8.8. Makeup Testing
8.9. LEAP 2025 High School and End-of-Course Testing Times
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
11.2. eDIRECT User Guide
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 LEAP 2025/EOC Tests
13. Request for Rescoring
14. Void Notification

The TAM provides detailed instructions for administering the LEAP 2025 HS assessments. The manual includes instructions for test security, test administrator responsibilities, test preparation, administration of online tests, and post-test procedures. Information included in the TAM 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
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 and Geometry
13.3. LEAP 2025 Biology

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
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 relevant to test administration. Information in the test administration manuals addresses these in the following manner.

Directions for test administration found in the manual address Standard 4.15 from the *Standards*, 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 TAM provides instructions for before-, during-, and after-testing activities 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.

Furthermore, the standardized procedures addressed in the TAM 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). To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the EOC was administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The test administration manuals included instructions for scheduling the test within the state testing window. The test administration manual also contained the schedule for timing each test session.

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

Department staff release annual test security reports about testing concerns observed during monitoring visits. These reports describe a wide range of improper activities that may occur during testing, including copying and reviewing test questions with students or using a calculator on parts of the test where it is not allowed.

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

The TAM outlines the steps that teachers should take to prepare classroom environment testing for administering the LEAP 2025 online test. These include the following:

- Determine the layout of the classroom environment.
- Plan seating arrangements. Allow enough space between students to prevent the sharing of answers.
- Eliminate distractions such as bells or telephones.
- Use a Do Not Disturb sign on the door of the testing room.
- Make sure classroom maps, charts, and any other materials that relate to the content and processes of the test are covered or removed or are 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 version of the LEAP 2025 assessment, examiners are instructed to transcribe students' responses from the braille test book into the online testing system (INSIGHT) exactly as they responded in the braille test book.

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 is presented under "Test Security" in the test administration manuals.

Time

Each session of each content area test was timed to provide sufficient time for students to attempt all items. The test administration manuals provided examiners with timing guidelines for the assessments.

Online Forms Administration

The online forms were administered via DRC's INSIGHT online assessment system. School system and school personnel set up test sessions via DRC's online testing portal, eDIRECT, and printed test tickets. Students entered their ticket information to access the test in INSIGHT. In addition, students had access to Online Tools Training, which allowed them to practice using tools and features within INSIGHT.

Students were required to experience the Online Tools Training (OTT) before the computer-based test administration. The OTT allows students to observe and practice features of the Online Assessment Software prior to an actual test administration. Students were also required to view the Student Tutorials, which present visual and verbal descriptions of the properties and features of the DRC INSIGHT Online Assessment Software.

Accessibility and Accommodations

Accessibility features and accommodations include Access for All, Accessibility Features, and Accommodations.

- Access for All features are available to all students taking an assessment.
- Accessibility Features are available to students when deemed appropriate by a team of educators.
- Accommodations must appear in a student's IEP/504/EL plan.

Accommodations may be used with students who qualify under the Individuals with Disabilities Education Act (IDEA) and have an IEP or Section 504 of the Americans with Disabilities Act and have a Section 504 plan, or who are identified 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, and 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 Test Administration Manual contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP assessments. The following accommodations were offered for this administration:

- Braille
- Answers Recorded
- Extended Time
- Transferred Answers
- Tests Read Aloud
- English/Native Language Word-to-Word Dictionary
- Directions Read Aloud/Clarified in Native Language
- Text-to-Speech
- Human Read Aloud
- Directions in Native Language

For more details about these accommodations, please refer to the LEAP Accessibility and Accommodations Manual.

Testing Windows

The 2018–2019 assessments were administered to students within the state testing windows of November 28 to December 14, 2018; April 15 to May 17, 2019; and June 17–21, 2019.

Test Security Procedures

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 EOC/LEAP 2025 HS assessments. Test security procedures are discussed throughout the TCM and TAM.

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 (e.g., test administrators and the school test coordinator). During the testing sessions, test administrators are directly responsible for the security of the EOC/LEAP 2025 HS and must account for all test materials and supervise the test administrations at all times.

The LDOE routinely conducts comprehensive data forensics with the administration vendor. Incidents that warrant further investigation with prospective voided test results include plagiarism, excessive wrong-to-right response changes, web-monitoring, and patterns of unusual school-level gains. In addition, 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.

6. Scoring Activities

Answer Key Verification

After a targeted number of tests were administered, DRC conducted an answer key verification. The purpose of this verification was to verify that the correct answers were being properly applied during the scoring process.

Directory of Test Specifications (DOTS) process. DRC created a DOTS file, based on the approved test selection. The DOTS is a document containing information about each item on a test form, such as item identifier, item sequence, answer key, score points, subtest, session, content standard, and prior use of item. WestEd reviewed and confirmed the contents of the DOTS file as part of test review rounds. The DOTS file was then provided to the LDOE for multiple rounds of review, then final approval. Once approved, the information contained in the DOTS was used in scoring the test and in reporting.

Selected-Response Item Keycheck. TRIAN, a standardized Pearson program that calculates MC item statistics, was used to verify that MC field test items were keyed correctly (i.e., that the true correct response was applied during scoring). Items were flagged if their item statistics fell outside expected ranges. For example, items were flagged if few students selected the correct response (*p*-value less than 0.15), if the item did not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) selected a certain incorrect response. Lists of flagged items, with the reasons for flagging, were provided to WestEd content staff for key verification. Scoring of MS items was evaluated at data review.

Scoring of TEs and Adjudication. All TE and MS items were processed through DRC's autoscoring engine and scored according to the assigned scoring rules as established during content creation by WestEd in conjunction with the LDOE. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring any TE items. DRC established an adjudication process for technology-enhanced items to verify that correct

answers were identified. DRC's technology-enhanced scoring process included the following procedures:

- A scoring rubric was created for each TE 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.

TE items and other eligible items identified in the test map were automatically scored as tests were processed. TE items were scored according to scoring rules in the Directory of Test Specifications (DOTS), which includes scoring information for all item types.

The adjudication process focuses on detecting possible errors in scoring TE and MS items. DRC provides a report listing the frequency distributions of TE item responses and MS items. Members of the LDOE and WestEd content staff examine the TE and MS response distributions and the auto-frequency reports to evaluate whether the items were scored appropriately. In the event that scoring issues are identified, WestEd content staff and the

LDOE review recommend changes to the scoring algorithm. Any changes to the scoring algorithm are based on the LDOE's decisions. DRC, in turn, applies the approved scoring changes to any affected items.

Constructed- and Extended-Response Item Scoring Process. The constructed- and extended-response items were scored by human raters trained by DRC. Human scorers provided second reads to 10% of these responses as well as handscoring supervisory reviews.

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 Biology 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 2018–2019 LEAP 2025 high school Biology test responses had at least a four-year college degree.

DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, or recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to demonstrate their proficiency in writing by responding to a DRC writing topic and their

proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and also demonstrates their ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

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 LDOE-approved training materials. These materials were developed by DRC and LDOE staff from a selection scored by Louisiana educators at rangefinding and include the following:

- 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.
The following table details the composition of the training materials for Biology.

Table 6.1
Biology Training Set Composition

Set Type*	Biology Training Materials	Annotated
Anchor set (2-point CRs)	Item-specific anchor sets containing three responses per score point	Yes
Anchor set (9-point ERs)	Item-specific anchor sets containing two responses per score point	
Training sets	Two training sets for each CR item and three training sets for each ER item ● 10 responses per training set ● All numeric score points represented*	No
Qualifying sets	Two qualifying sets for each CR item and two qualifying sets for each ER item ● 10 responses per qualifying set ● All numeric score points represented*	No

*Examples of responses at the top score points or for all score-point combinations were not present in some anchor, training, and qualifying sets as there were few or no examples found during rangefinding or subsequent field test scoring. DRC scoring directors identified examples of these scores during live scoring to supplement reader training.

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

Any scorer who did not qualify by the end of the qualifying process for an item was not allowed to score live student responses. The qualifying standards for the Biology constructed- and extended-response items are shown in Table 6.2.

Table 6.2
Biology Qualifying Standards

Course and Item Type	Qualifying Standard	
Biology 0–2 point CR	0–2 Rubric	Scorers must qualify with 80% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
Biology 0–9 point multi-part ER*	0–3 Rubric	Scorers must qualify with 70% exact agreement or higher on one or more of the qualifying sets in order to score student responses.
	0–6 Rubric	Scorers must qualify with 60% exact agreement or higher on one or more of the qualifying sets in order to score student responses.

*Qualifying sets are made up of 10 responses comparable to the anchor set responses. For multi-part Biology ERs, the appropriate qualifying standard should be achieved on each part of the item. For example, if an item has Part A with a top score of 6 and Part B with a top score of 3, a scorer would need to achieve 60% perfect agreement on Part A and 70% perfect agreement on Part B on one or more of the qualifying sets. A scorer may qualify on one part in the first qualifying set and the other part in the second qualifying set.

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)

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- and extended-response items.

Reader Monitoring Procedures. Throughout the handscoring process, DRC project managers, scoring directors, and team leaders reviewed the statistics that were generated

daily. DRC used one team leader for every 10 to 12 readers. 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, adjusting to that ratio as needed. If a supervisor disagreed with a reader's scores during monitoring, the supervisor 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 validity responses. Validity responses were inserted among the live student responses.

The validity responses were added to DRC's image handscoring system prior to the beginning of scoring. Validity reports compared readers' scores to predetermined 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 handscored 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 also removed all unreported scores that were assigned by the scorer during the period in question. The responses were then reassigned and rescored.

To monitor inter-rater reliability, DRC produced scoring summary reports daily. 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

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

Table 6.3
Agreement Rate Requirements for Validity and Inter-Rater Reliability

Subject	Score Point Range	Perfect Agreement	Perfect Agreement + Adjacent
Biology CR	0–2	80%	95%
Biology (multi-part) ER	0–3	70%	95%
	0–6	60%	93%

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. DRC used these calibration sets to perform calibration across the entire scorer population for an item if trends were detected (e.g., low agreement between certain score points or if a certain type of response was missing from initial training). These calibrations were designed to help refocus scorers on how to properly use the scoring guidelines. They were selected to help illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After readers scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and 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 regarding reliability, score point distribution, and validity performance. In addition to the reader feedback logs, DRC provides the LDOE with handscoring quality control reports for review throughout the scoring window.

Inter-Rater Reliability. A minimum of 10% of the responses in Biology were scored independently by a second reader. 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.

Tables 6.4–6.9 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the 2018–2019 forms.

Table 6.4

Operational Constructed-Response Inter-Rater Reliability

Admin.	Item	Read 2X	Inter-Rater Reliability %		
			EX	AD	NA
Fall 2018	Item 1	≥1,510	92	8	0
	Item 2	≥1,480	86	14	0
	Item 3	≥1,440	93	7	0
Spring 2019	Item 1	≥11,440	89	10	1
	Item 2	≥9,750	94	5	0
	Item 3	≥9,700	86	13	1
	Item 4	≥550	94	6	0
	Item 5	≥290	91	9	0
	Item 6	≥410	87	11	2
	Item 7	≥450	91	7	2
	Item 8	≥430	93	7	0
	Item 9	≥490	95	4	1
	Item 10	≥440	88	12	0
Summer 2019	Item 1	≥150	99	1	0
	Item 2	≥170	100	0	0
	Item 3	≥140	100	0	0

Table 6.5

Operational Constructed-Response Score Point Distributions

Administration	Item	Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
Fall 2018	Item 1	≥7,270	71	25	2	0
	Item 2	≥7,540	57	31	11	0
	Item 3	≥7,350	78	13	7	0
Spring 2019	Item 1	≥43,690	60	21	11	0
	Item 2	≥42,920	82	10	4	0
	Item 3	≥41,210	59	30	5	0
	Item 4	≥1,810	25	25	43	0
	Item 5	≥1,730	49	33	14	0
	Item 6	≥1,670	66	18	12	0
	Item 7	≥1,680	42	28	23	0
	Item 8	≥1,640	60	29	2	1
	Item 9	≥1,720	41	32	24	0
	Item 10	≥1,700	53	28	14	0
Summer 2019	Item 1	≥310	50	15	1	0
	Item 2	≥310	49	9	0	0
	Item 3	≥290	61	2	0	0

Table 6.6

Field Test Constructed-Response Inter-Rater Reliability

Admin.	Item	Read 2X	Inter-Rater Reliability %		
			EX	AD	NA
Spring 2019	Item 1	≥490	98	2	0
	Item 2	≥480	98	2	0
	Item 3	≥460	98	2	0

Table 6.7

Field Test Constructed-Response Score Point Distributions

Administration	Item	Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
Spring 2019	Item 1	≥1,820	76	17	1	0
	Item 2	≥1,690	78	10	6	0
	Item 3	≥1,740	53	38	5	0

Table 6.8

Operational Extended-Response Inter-Rater Reliability

Admin.	Item	Part	Read 2X	Inter-Rater Reliability %		
				EX	AD	NA
Fall 2018	Item 1	Part A (0–6)	≥1,580	87	10	3
		Part B (0–3)		85	14	1
Spring 2019	Item 1	Part A (0–3)	≥6,050	82	17	1
		Part B (0–6)		80	15	6
	Item 2	Part A (0–6)	≥4,680	88	10	2
		Part B (0–3)		88	10	2
Summer 2019	Item 1	Part A (0–6)	≥120	95	5	0
		Part B (0–3)		100	0	0

Table 6.9

Operational Extended-Response Score Point Distributions

Admin	Item	Total	Score Point Distribution								
			Part	% "0" Rating	% "1" Rating	% "2" Rating	% "3" Rating	% "4" Rating	% "5" Rating	% "6" Rating	% Blank
Fall 2018	Item 1	≥7,440	Part A (0–6)	55	13	13	14	2	1	1	0
			Part B (0–3)	41	35	11	11				0
Spring 2019	Item 1	≥23,240	Part A (0–3)	9	40	29	15				0
			Part B (0–6)	31	20	19	10	8	2	3	0
	Item 2	≥19,850	Part A (0–6)	44	14	14	19	2	1	2	0
			Part B (0–3)	34	34	11	16				0
Summer 2019	Item 1	≥290	Part A (0–6)	57	13	3	3	0	0	0	0
			Part B (0–3)	51	21	2	3				0

7. Data Analysis

Classical Item Statistics

This section shows the results of the classical item analysis for data obtained from the LEAP operational tests. These item analysis results serve two purposes: 1) to inform item performance; and 2) to provide item statistics for the item bank. LEAP classical item analysis consists of the following types of items: key/multiple option-based items, rule-based machine-scored items such as technology-embedded items, and hand-scored constructed response items. For each operational item, the analysis produces item difficulty (i.e., *p*-value) and item discrimination (p-b serial).

Appendix C: Item Analysis Summary Report includes tables and figures that provide the information on classical item statistics for operational items. Tables C.1–C.4 show summaries of classical item statistics. A measure of item difficulty, *p* (or "the *p*-value"), indicates the average proportion of total points earned on an item. For example, if *p* = 0.50 on an MC item, then half of the examinees earned a score of 1. If *p* = 0.50 on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). The corrected point-biserial correlation is a measure of item discrimination. Items with higher item-total correlations provide better information about how well items discriminate between lower- and higher-performing students.

The difficulty of an item is commonly expressed as a *p*-value, which is the mean score on an item. For the desirable ranges of *p*-values for any item type at the time of test construction is set to 0.25 MC, TE, CR, and ER items. Please note that these recommendations should be considered as a "rule of thumb" rather than strict cut-off values. The point biserial correlation of any MC item should be greater than 0.20. Any item with negative point-biserial correlation should not be selected. However, there may be cases in which items required to meet content guidelines do not meet the point-biserial correlation guideline. The following flagging criteria was also used to review any field test items for data review:

- Correct Response *p*-value < 0.25
- Correct Response point-biserial < 0.20
- Distractor *p*-value > 0.40

It should be noted that statistical results of FT items can be found at Pearson ABBI.

Differential Item Functioning

Differential Item Functioning (DIF) analyses are intended to statistically signal potential item bias. DIF is defined as a difference between similar ability groups' (e.g., males or females that attain the same total test score) probability of getting an item correct. Because test scores can reflect many sources of variation, the test developers' task is to create assessments that measure the intended knowledge and skills without introducing construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores may reflect those extraneous elements in addition to what the test is purported to measure. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). Different cultural and socioeconomic experiences are among some factors that can confound test scores intended to reflect the measured construct.

One DIF methodology applied to dichotomous items was the Mantel–Haenszel (*MH*) DIF statistic (Holland & Thayer, 1988; Mantel & Haenszel, 1959). The MH method is a frequently used method that offers efficient statistical power (Clauser & Mazor, 1998). The *MH* chi-square statistic is

$$MH_{\chi^2} = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k Var(F_k)},$$

where F_k is the sum of scores for the focal group at the k^{th} level of the matching variable (Zwick, Donoghue, & Grima, 1993). Note that the *MH* statistic is sensitive to *N* such that larger sample sizes increase the value of the chi-square.

In addition to the *MH* chi-square statistic, the *MH* delta statistic (*ΔMH*), first developed by the Educational Testing Service (ETS), was computed. To compute the *ΔMH DIF*, the *MH* alpha (the odds ratio) is calculated:

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k},$$

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*. The *MH DIF* statistic is based on a 2×2×*M* (2 groups × 2 item scores × *M* strata) frequency table, in which students in the reference (male or white) and focal (female or black) groups are matched on their total raw scores.

The *ΔMH DIF* is then computed as

$$\Delta MH\ DIF = -2.35\ln(\alpha_{MH}).$$

Positive values of *ΔMH DIF* indicate items that favor the focal group (i.e., positive DIF items are differentially easier for the focal group); negative values of *ΔMH DIF* indicate items that favor the reference group (i.e., negative DIF items are differentially easier for the reference group). Ninety-five percent confidence intervals for *ΔMH DIF* are used to conduct statistical tests.

The *MH* chi-square statistic and the *ΔMH DIF* were used in combination to identify operational test items exhibiting strong, weak, or no DIF (Zieky, 1993). Table 7.1 defines the DIF categories for dichotomous items.

Table 7.1
DIF Categories for Dichotomous Items

DIF Category	Criteria
A (negligible)	| *ΔMH DIF* | is not significantly different from 0.0 or is less than 1.0.
B (slight to moderate)	1. | *ΔMH DIF* | is significantly different from 0.0 but not from 1.0, and is at least 1.0; OR 2. | *ΔMH DIF* | is significantly different from 1.0 but is less than 1.5. Positive values are classified as "B+" and negative values as "B–."
C (moderate to large)	| *ΔMH DIF* | is significantly greater than 1.0 and is at least 1.5. Positive values are classified as "C+" and negative values as "C–."

For polytomous items, the standardized mean difference (*SMD*) (Dorans & Schmitt, 1991; Zwick, Thayer, & Mazzeo, 1997) and the Mantel χ^2 statistic (Mantel, 1963) are used to identify items with DIF. *SMD* estimates the average difference in performance between the

reference group and the focal group while controlling for student ability. To calculate the *SMD*, let *M* represent the matching variable (total test score). For all $M = m$, identify the students with raw score *m* and calculate the expected item score for the reference group (E_{rm}) and the focal group (E_{fm}). DIF is defined as $D_m = E_{fm} - E_{rm}$, and *SMD* is a weighted average of D_m using the weights $w_m = N_{fm}$ (the number of students in the focal group with raw score *m*), which gives the greatest weight at score levels most frequently attained by students in the focal group.

$$\text{SMD} = \frac{\sum_m w_m(E_{fm} - E_{rm})}{\sum_m w_m} = \frac{\sum_m w_m D_m}{\sum_m w_m}$$

The *SMD* is converted to an effect-size metric by dividing it by the standard deviation of item scores for the total group. A negative *SMD* value indicates an item on which the focal group has a lower mean than the reference group, conditioned on the matching variable. On the other hand, a positive *SMD* value indicates an item on which the reference group has a lower mean than the focal group, conditioned on the matching variable.

The *MH DIF* statistic is based on a $2\times(T+1)\times M$ (2 groups × T+1 item scores × M strata) frequency table, where students in the reference and focal groups are matched on their total raw scores (T = maximum score for the item). The Mantel χ^2 statistic is defined by the following equation:

$$\text{Mantel } \chi^2 = \frac{\left(\sum_m \sum_t N_{rtm} Y_t - \sum_m \frac{N_{r+m}}{N_{++m}} \sum_t N_{+tm} Y_t\right)^2}{\sum_m Var\left(\sum_t N_{rtm} Y_t\right)}.$$

The *p*-value associated with the Mantel χ^2 statistic and the *SMD* (on an effect-size metric) are used to determine DIF classifications. Table 7.2 defines the DIF categories for polytomous items.

Table 7.2

DIF Categories for Polytomous Items

DIF Category	Criteria
A (negligible)	Mantel χ^2 *p*-value > 0.05 or $\lvert SMD/SD \rvert \leq 0.17$
B (slight to moderate)	Mantel χ^2 *p*-value < 0.05 and $0.17 < \lvert SMD/SD \rvert < 0.25$
C (moderate to large)	Mantel χ^2 *p*-value < 0.05 and $\lvert SMD/SD \rvert \geq 0.25$

Three DIF analyses were conducted for field test items: female/male, black/white, and Hispanic/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The same methods were used to detect items on which black or white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Table 7.2 provide the percentages of items flagged for DIF. Items flagged with B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Very few field test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the Mantel–Haenszel statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 and *SMD* statistics. Table 7.3 summarizes DIF statistics for the 2019 spring operational items. It should be noted that all DIF results can be found in Pearson ABBI.

All items exhibiting statistical DIF were reviewed by the LDOE and WestEd content staff. Per the LDOE's standard practice, if multiple items exhibiting statistical DIF must be used on a test, the items to be used are purposefully reviewed and selected to ensure that the DIF flags do not consistently favor or disfavor the same comparison group. At the 2019 data review, no items were found to exhibit bias, and no items were rejected from the prospective item pool strictly based on DIF analysis results and content reviews.

Table 7.3
Summary of DIF Flags for Biology Operational Items

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	46	[0],[0]	[0],[0]
African American – White	42	[0],[3]	[0],[1]
Hispanic – White	44	[0],[2]	[0],[0]

The results of classical test theoretic data analyses—item *p*-values, item discrimination indices, and *MH DIF* indices—and analyses based on item theoretic methods are reviewed by committees of Louisiana educators for potential bias. It should be also noted that for data review on field test item analysis results, particularly, any statistically flagged items

evaluated for and determined to present potential bias are rejected from inclusion in the item pool.

Item Calibration and Scaling

LEAP 2025 Biology assessments are standards-based assessments that have been constructed to align to the LSSS, as defined by the LDOE and Louisiana educators. For each course, the content standards specify the subject matter students should know and the skills they should be able to perform. In addition, performance standards specify how much of the content standards students need to master in order to achieve proficiency. Constructing tests to content standards enables the tests to assess the same constructs from one year to the next.

Item Response Theory (IRT) models were used in the item calibration for the LEAP 2025 Biology test. All calibration activities were independently replicated by Pearson staff as an added quality-control check.

Scaling is the process whereby student performance is associated with an ordered value, typically a number. The most common and straightforward way to score a test is to simply use the sum of points a student earned on the test, namely, the raw score. Although the raw score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Typically, a scaled metric is used, on which test forms from different years are equated.

Measurement Models

IRTPRO, a software application for item calibration and test scoring, was used to estimate IRT parameters from LEAP 2025 data. MC, MS, and some TE items were scored dichotomously (0/1), so the three-parameter logistic model (3PL) was applied to those data:

$$p_i(\theta_j) = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta_j - b_i)}}.$$

In that model, $p_i(\theta_j)$ is the probability that student *j* would earn a score of 1 on item *i*, b_i is the difficulty parameter for item *i*, a_i is the slope (or discrimination) parameter for item *i*, c_i is the pseudo-chance (or guessing) parameter for item *i*, and *D* is the constant 1.7. This test also included five types of polytomous items: TE items scored 0–2, CR items scored 0–2, TPI items scored 0–2, TPD items scored 0–2, and ER items scored 0–9. Data from polytomous items were used to estimate parameters for the generalized partial credit model (GPCM) (Muraki, 1992):

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$, $p_{im}(\theta_j)$ is the probability of an examinee with θ_j getting score *m* on item *i*, and *Mi* is the number of score categories of item *i* with possible item scores as consecutive integers from 0 to *Mi* – 1. In the GPCM, the *d* parameters define the "category intersections" (i.e., the θ value at which examinees have the same probability of scoring 0 and 1, 1 and 2, etc.).

Operational Item Parameters

The distributions of item parameters are summarized in Table C.5. Figures in Appendix C provide graphical displays of the distributions of IRT parameter estimates for each grade. TPI, TPD, CR, and ER items have no *c* parameters because they are polytomous items and are therefore modeled using the GPCM. The number of item parameters associated with the ER items reflect item parameter estimates associated with particular "part scores" that comprise the total ER item. It should be noted that statistical results of FT items can be found at Pearson ABBI.

Item Fit

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q_1 statistic (Yen, 1981) is used as an index for how well theoretical item curves match observed item responses. Q_1 is computed by first conducting an IRT item parameter estimation, then estimating students'

achievement using the estimated item parameters, and, finally, using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at 10 selected equal intervals across the range of student achievement. Q_1 is computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-square (χ^2) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis (the expected item performance). Q_1 for each item type has varying degrees of freedom because the different item types have different numbers of IRT parameters. Therefore, Q_1 is not directly comparable across item types. An adjustment or linear transformation (translation to a *Z*-score, Z_{Q_1}) is made for different numbers of item parameters and sample size to create a more comparable statistic.

Yen's Q_1 statistic (Yen, 1981) was calculated to evaluate item fit for field test items by comparing observed and expected item performance. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. For dichotomous items, Q_1 is computed as

$$Q_{1i} = \sum_{j=1}^{j} \frac{N_{ij}(O_{ij}-E_{ij})^2}{E_{ij}(1-E_{ij})},$$

where N_{ij} is the number of examinees in interval (or group) *j* for item *i*, O_{ij} is the observed proportion of the examinees in the same interval, and E_{ij} is the expected proportion of the examinees for that interval. The expected proportion is computed as

$$E_{ij} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_i(\hat{\theta}_a),$$

where $P_i(\hat{\theta}_a)$ is the item characteristic function for item *i* and examinee *a*. The summation is taken over examinees in interval *j*.

The generalization of Q_1 for items with multiple response categories is

$$Gen\ Q_{1i} = \sum_{j=1}^{10} \sum_{k=1}^{m_i} \frac{N_{ij}(O_{ikj}-E_{ikj})^2}{E_{ikj}},$$

where

$$E_{ikj} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_{ik}\,(\hat{\theta}_a).$$

Both Q_1 and generalized Q_1 results are transformed to ZQ_1 and are compared to a criterion $ZQ_{1,\text{crit}}$ to determine whether fit is acceptable. The conversion formulas are

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}}$$

and

$$ZQ_{1,crit} = \frac{N}{1500} * 4,$$

where *df* is the degrees of freedom (the number of intervals minus the number of independent item parameters). Items are categorized as exhibiting either fit or misfit.

A summary of IRT item parameter statistics and item fit is displayed in Appendix D: Dimensionality.

Dimensionality and Local Item Independence

By fitting all items simultaneously to the same achievement scale, IRT is operating under the assumption that there is a single predominant construct that underlies the performance of all items. Under this assumption, item performance should be related to achievement and, additionally, any relationship of performance between pairs of items should be explained or accounted for by variance in students' levels of achievement. This is the "local item independence" assumption of unidimensional IRT and is associated with a test for unidimensionality called the Q_3 statistic (Yen, 1984).

Computation of the Q_3 statistic starts with expected student performance on each item, which is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference is the remainder in performance after accounting for underlying achievement. If performance on an item is driven by a predominant achievement construct, then the residual will be small (as tested by the Q_1

statistic), and the correlation between residuals of the item pairs will also be small. These correlations are analogous to partial correlations or the relationship between two variables (items) after accounting for the effects of a third variable (underlying achievement). The correlation among IRT residuals is the Q_3 statistic.

When calculating the level of local item dependence for two items (*i* and *j*), the Q_3 statistic is

$$Q_3 = r_{d_i d_j}.$$

The correlation between d_i and d_j values is the correlation of the residuals—that is, the difference between expected and observed scores for each item. For test taker *k*,

$$d_{ik} = u_{ik} - P_i(\theta_k),$$

where u_{ik} is the score of the *k*th test taker on item *i* and $P_i(\theta_k)$ represents the probability of test taker *k* responding correctly to item *i*.

With *n* items, there are $n(n - 1)/2$ Q_3 statistics. If an assessment consists of 48 items, for example, there are 1,128 Q_3 values. The Q_3 values should all be small. Summaries of the distributions of Q_3 are provided in Appendix D: Dimensionality. Specifically, Q_3 data are summarized by minimum, 5th percentile, median, 95th percentile, and maximum values for LEAP 2025 Science grades 3 through 8. To add perspective to the meaning of Q_3 distributions, the average zero-order correlation (simple intercorrelation) among item responses is also shown. If the achievement construct accounts for the relationships between items, Q_3 values should be much smaller than the zero-order correlations. The Q_3 summary tables in the dimensionality reports in Appendix D show for all grades and subjects that at least 90% (between the 5th and 95th percentiles) of the items are expectedly small. These data, coupled with the Q_1 data, indicate that the unidimensional IRT model provides a reasonable solution to capture the essence of student science achievement defined by the selected set of items for each grade level.

Unidimensionality and Principal Component Analysis

It should be noted that Appendix D provides information about principal component analysis of Biology. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, in the case of scholastic achievement, one-dimensional

scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 2004; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983). To check the unidimensionality of the 2019 LEAP assessments, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the SAS program. The first and the second principal component eigenvalues were compared *without rotation*. Table D.4 and Figure D.4 summarize the results of the first and second principal component eigenvalues of the assessments.

A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the table and figures, the first component is substantially larger than the second eigenvalue across the assessments: the first eigenvalue was at least 5 times as big as the second eigenvalue. In addition, the figure indicates that the second component sharply drops from the first and gets flat. As a result, we could conclude that the unidimensionality assumption of 2019 assessment was met.

Scaling

Based on the panelist recommendations and LDOE approval, the scale is set using two cut scores, Basic and Mastery, with fixed scale score points of 725 and 750, respectively. The scale scores for Approaching Basic and Advanced vary by grade level. The highest obtainable scale score (HOSS) and lowest obtainable scale score (LOSS) for the scale determined by the LDOE are 650 and 850.

IRT ability estimates (θs) are transformed to the reporting scale with a linear transformation equation of the form

$$SS = A\theta + B,$$

where *SS* is scale score, θ is IRT ability, *A* is a slope coefficient, and *B* is an intercept. The slope can be calculated as

$$A = \frac{SS_{Mastery} - SS_{Basic}}{\theta_{Mastery} - \theta_{Basic}},$$

where $\theta_{Mastery}$ is the Mastery cut score on the theta scale, and θ_{Basic} is the Basic cut score on the theta scale. $SS_{Mastery}$ and SS_{Basic} are the Mastery and Basic scale score cuts, respectively. With *A* calculated, *B* are derived from the equation

$$SS_{Mastery} = A\theta_{Mastery} + B,$$

which are rearranged as

$$B = SS_{Mastery} - A\theta_{Mastery} \text{ or } B = SS_{Mastery} - \frac{SS_{Mastery} - SS_{Basic}}{\theta_{Mastery} - \theta_{Basic}}\theta_{Mastery}.$$

Thus, the general equation for converting θs to scale scores is

$$SS = \left(\frac{SS_{Mastery} - SS_{Basic}}{\theta_{Mastery} - \theta_{Basic}}\right)\theta + \left(SS_{Mastery} - \frac{SS_{Mastery} - SS_{Basic}}{\theta_{Mastery} - \theta_{Basic}}\theta_{Mastery}\right).$$

The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same scaling constants *A* and *B* are used to convert student ability estimates to the reporting scale until new achievement-level standards are set. Descriptive Statistics and Frequency Distribution of LEAP 2025 Biology Scale Scores can be found in Appendix E: Scale Distribution and Statistical Report.

8. Reporting for Biology

Additional information regarding score reporting can be found in the *Interpretive Guide English I, English II, Algebra I, Geometry, U.S. History, and Biology 2018–2019* document. Detailed Information can be found at the following link: https://www.louisianabelieves.com/resources/library/assessment.
The elements of the table of contents are provided below.

- Introduction to the Interpretive Guide
    - Overview
        - Purpose of the Interpretive Guide
    - Test Design
    - Scoring
        - Item Types and Scoring
    - Interpreting Scores and Achievement Levels
        - Scale Score
        - Achievement Level Definitions
        - Student Rating by Reporting Category and Subcategory
- Student-Level Reports
    - Sample Student Report: Explanation of Results and Terms
    - Sample Student Report A
    - Sample Student Report B
    - Parent Guide to the LEAP 2025 High School Student Reports
- School Roster Report
    - Sample School Roster Report: Explanation of Results and Terms
    - Sample School Roster Report

Biology Standard Setting

Ancillary documentation of the standard setting process and results can be found in the *LEAP 2025 Standard Setting Meeting* document. The full report summarizes the processes and results of the standard setting. Excerpts from the Executive Report are provided in the following pages. The elements of the table of contents are listed below.

- Executive Report
- Chapter 1 – Overview of the Standard Setting Process
    - Goals of the Standard Setting Meeting
    - LEAP 2025 Achievement Levels
    - The LEAP 2025 Standard Setting Process
- Chapter 2 – Pre-Meeting Development
    - LEAP 2025 Achievement Level Descriptors
    - Development of the Participant Materials
    - Preparation of the Ordered Item Book
    - Development of the Presentation Materials
    - Facilitator Training
    - Preparation for Data Analysis during the Meetings
- Chapter 3 – Standard Setting Meetings
    - Purpose of the Standard Setting Meetings
    - Committee Participant Composition
    - Standard Setting Meeting Facilitators and Staff
    - Materials
    - Procedure
    - Standard Setting Meetings and Proceedings
    - Recommended LEAP 2025 Cut Scores from Standard Setting Committees
- Chapter 4 – Post-Standard Setting
    - Vertical Articulation Meeting
    - Standards Policy Review Committee
    - Scaling Process
- Chapter 5 – Evidence of Procedural Validity of the Standard Setting Process
    - Internal Procedures
    - Committee Representation
    - Committee Training
    - Perceived Participant Validity of the Workshop
- References
- Appendices

Standard Setting Executive Report

10 July 2018

This report summarizes the process and results of setting achievement levels for the Louisiana Educational Assessment Program (LEAP) 2025 Biology assessment. The Louisiana Department of Education (LDOE) and WestEd with Pearson (LEAP 2025 Biology assessment contractors) recommend the achievement levels shown in Table 2 of the Standard Setting Report for adoption by the Board of Elementary and Secondary Education (BESE).

LEAP 2025 Biology Standard Setting Process and Results

Achievement levels are used to classify student achievement on an assessment. In order to classify student achievement into the different achievement levels, the following components are required: 1) policy definitions; 2) Achievement Level Descriptors (ALDs); and 3) cut scores. Policy definitions describe the achievement levels in general terms that apply to all courses or subject areas. ALDs illustrate the achievement levels in terms that are specific to a course or subject area. Cut scores represent the lowest boundary of each achievement level on the scale.

The process of recommending achievement standards for the LEAP 2025 Biology test was similar to the processes followed for previous assessments in Louisiana and in line with national best practice. Results and details of the process are presented in the following sections.

Policy Definitions

Achievement level policy definitions for the LEAP 2025 Biology assessment are shown in Table 1. These policy definitions are also used for the social studies grades 3–8 assessments, English language arts (ELA) assessments, and mathematics assessments. The titles and descriptions of the achievement levels were defined to be part of a cohesive

assessment system, and the achievement levels indicate a student's ability to demonstrate proficiency on the LSSS defined for a specific course.

Table 8.1
Achievement Level Policy Definitions for LEAP 2025

Achievement Level	Achievement Level Policy Definition
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 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.

Achievement Level Descriptors (ALDs)

ALDs for the Biology test are shown in the appendix of the Standard Setting Executive Report. A multi-step iterative process was used in developing, reviewing, and approving the ALDs. Prior to the standard setting committee, a draft set of ALDs representing a gradual increase in expectations across the achievement levels was created by LDOE content staff in cooperation with WestEd content specialists. Panelists who participated in the standard setting committees had the opportunity to provide suggestions and edits to the draft set of ALDs based on the recommended cut score for each achievement level and the items in the ordered item book. To produce the final set of ALDs, the LDOE edited

the set of draft ALDs based on suggestions generated by the panelists in the standard setting meeting.

Cut Scores

The cut scores recommended for adoption by BESE are shown in Table 2. This table shows the scale score ranges corresponding to each achievement level. The cut scores for the achievement levels are the lowest cut score within each range. There is no cut score for *Unsatisfactory*, since 650 is the lowest obtainable scale score a student can earn.

Table 8.2
Scale Score Ranges for LEAP 2025 Achievement Levels for Biology

Achievement Level	Scale Score Ranges Biology
Advanced	774 to 850
Mastery	750 to 773
Basic	725 to 749
Approaching Basic	711 to 724
Unsatisfactory	650 to 710

Details pertaining to the general method for obtaining the recommended cut scores are provided below.

General Method

Prior to the standard setting committee, on April 26, 2018, a policy committee was convened of teachers, school and school system leaders, and LDOE staff. The purpose of the meeting was to review information that would be useful in considering the policy implications of the cut scores for the LEAP 2025 Biology assessment and to provide a set of recommended ranges for the cut scores that would be presented to panelists during the standard setting meeting. The information that was shared with the committee included the impact data from the spring 2017 administration of the LEAP 2025 Science assessments for grades 3–8 and the high school assessment for Biology, the Louisiana

high school graduation rates in 2016, and the results of a contrasting groups teacher study performed for the Biology assessment during spring 2018. After a review and discussion of the data presented during the meeting, the policy committee members approved recommended ranges for the cut scores. The ranges, shown in Table 3, represent the maximum and minimum percentage of students that could be reasonably expected to be classified into each achievement level or higher based on the policy considerations. These ranges helped guide the standard setting committee in understanding policy considerations as part of the standard setting process.

Table 8.3
Recommended Ranges from the Policy Committee

Achievement Levels	Cumulative Impact Data	
	Minimum	Maximum
Advanced	5%	15%
Mastery	25%	40%
Basic	50%	65%
Approaching Basic	70%	85%

From July 9 to July 10, 2018, after the first year of operational administration, a standard setting committee meeting was conducted to provide cut score recommendations for the LEAP 2025 Biology assessment. The committee was composed of 13 individuals, including teachers and non-teacher educators, who were selected for the standard setting committee to provide content expertise during the committee meeting and to be representative of the state's educators. The evidence-based bookmark method was used for the standard setting meeting (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001; Schultz & Mitzel, 2009). The key material used by the committee was a book of test items arranged in order of difficulty. Participants identified and discussed the knowledge, skills, and abilities required to respond to the test items and divided the items into two groups—items that a student who is minimally qualified for an achievement level would likely answer correctly and items too difficult for students at that same achievement level. Additionally, the participants were provided the recommended ranges from the policy committee to review and consider as part of the judgment process.

In order to create a common point of reference across the science assessments, cut scores and measures of student achievement on all LEAP 2025 assessments are translated to a scale that ranges from 650 to 850 points, a *Basic* cut of 725 and a *Mastery* cut of 750. The common values of 725 for the *Basic* cut score and 750 for the *Mastery* cut score across assessments do not mean that they reflect that same difficulty, or that achievement levels can be compared in difficulty through the scale values of their cut scores across grades and subjects. Similarly, the percentage of students in an achievement level is not directly comparable across grades and subjects. The population of students tested is different for each assessment. Achievement levels from different tests are not comparable because the cut scores for these tests are criterion referenced—they are based on content-specific expectations of what students should know and be able to do.

Results for LEAP 2025 Biology

Table 4 shows the percent of students who took the LEAP 2025 Biology assessment during the spring 2018 administration that would be classified into achievement levels based on the cut score recommendations from the standard setting committee.

Table 8.4
Percent of Students in Achievement Levels

Achievement Level	Assessment
	Biology
Advanced	9%
Mastery	19%
Basic	32%
Approaching Basic	15%
Unsatisfactory	24%

9. Data Review Process and Results

During data review of the spring 2018 FT items, content experts and psychometric support staff reviewed field-tested items with accompanying data to make judgments about the appropriateness of items for use on future operational test forms. Statistically flagged items were not rejected on the sole basis of statistics; only items with identifiable flaws based on content were rejected.

The data review meeting began with a refresher presentation to data review. The presentation included a review of item statistics (difficulty, discrimination, DIF, score distributions), appropriate interpretations and inferences, what would be considered reasonable values, and how the values might differ across item types.

Facilitators from Pearson and WestEd led the data review. Statistical information was evaluated for each item to determine whether the item functioned as intended. Each item's suitability for future operational tests was then evaluated in the context of the field-test statistics. Judgments to accept, accept with edits (or "revise/re-field test"), or reject were then recorded for each item. Table 9.1 summarizes the disposition of field-tested items from data review. If the decision was to edit or to reject an item, additional information was captured to document the reason for the decision.

Table 9.1

Summary of Data Review Votes

Item Type		Number of Items			
	Accept	Accept with Edits	Reject	Total	% of Total
CR	1	1	1	3	6.52
ER	0	0	0	0	0.00
MC	13	2	1	16	34.78
MS	0	1	0	1	2.17
TE	13	0	0	13	28.26
TPI	3	1	0	4	8.70
TPD	6	3	0	9	19.57
Total	36	8	2	46	100.00

Following the data review meeting, LDOE content specialists reviewed items and the data review judgments with a focus on items that were rejected or accepted with edits. This reconciliation process provided the LDOE with an additional opportunity to review item content and consider possible revisions that would allow items to be field tested again for future operational use. Final item dispositions were determined by outcomes from the reconciliation process.

10. Reliability and Validity

Internal Consistency Reliability Estimation

Internal consistency methods use data from a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures that require multiple test administrations. One of the most frequently used internal consistency reliability estimates is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the assumption that inter-item covariances constitute true-score variance and the fact that the average true-score variance of items is greater than or equal to the average inter-item covariance. The formula for coefficient alpha is

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} s_{Y_i}^2}{s_X^2}\right),$$

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the ith item or component, and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 Biology tests is evidenced through a dimensionality analysis. Dimensionality analyses results are discussed in "Chapter 7. Data Analysis."

The reliability and classification accuracy reports in Appendix F: Reliability and Classification Accuracy provide coefficient alpha and IRT model-based or "marginal reliability" (Thissen, Chen, & Bock, 2003) for the total test. Coefficient alpha values range from 0.85 to 0.86, and the marginal alpha value was 0.97. Marginal reliability is described as "an average reliability over levels of θ or theta" (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 "posterior standard deviations" (SEMs) in the IRTPRO output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31 θs. Marginal reliability is the

average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals. The formula for marginal reliability is

$$\overline{\rho} = \frac{s_\theta^2 - E(SEM_\theta^2)}{s_\theta^2},$$

where s_θ^2 is the variance of a given θ (is 1 for standardized θ) and $E(SEM_\theta^2)$ is the average error variance or the mean of the squared posterior standard deviations by weighting population density. Marginal reliability can be interpreted in the same way as traditional internal consistency reliability estimates such as coefficient alpha.

Additional reliabilities were calculated on various demographic subgroups[1] using the population of students (Appendix F: Reliability and Classification Accuracy). Included with coefficient alpha in the tables is the number of students responding to the test, the mean score obtained by this group of students, and the standard deviation of the scores obtained for this group.

Coefficient alpha estimates are computed for the entire test and each subscale by reporting category. Subscore reliability will generally be lower than total score reliability because reliability is influenced by the number of items as well as their covariation. In some cases, the number of items associated with a subscore is small (10 or fewer). Subscore results must be interpreted carefully when these measures reflect the limited number of items associated with the score.

Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their scale scores. It is important to know the reliability of student scores in any examination; but, assessing the reliability of the classification decisions based on these scores is of even greater importance. Classification decision reliability is estimated by the probabilities of correct

[1] The subgroups are male/female, white/Black/Hispanic/Asian/American Indian or Alaska Native/Native Hawaiian or Other Pacific Islander/multi-racial, and English Learners.

and consistent classification of students. Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive accuracy and consistency classification measures.

Accuracy of Classification. According to Livingston and Lewis (1995, p. 180), the classification accuracy is "the extent to which the actual classifications of the test takers . . . agree with those that would be made on the basis of their true scores, if their true scores could somehow be known." Accuracy estimates are calculated from cross-tabulations between "classifications based on an observable variable (scores on a test) and classifications based on an unobservable variable (the test takers' true scores)." True score is also referred to as a hypothetical mean of scores from all possible forms of the test if they could be somehow obtained (Young & Yoon, 1998).

Consistency of Classification. Classification consistency is "the agreement between classifications based on two non-overlapping, equally difficult forms of the test" (Livingston & Lewis, 1995, p. 180). Consistency is estimated using actual response data from a test and the test's reliability to statistically model two parallel forms of the test and compare the classifications on those alternate forms.

Accuracy and Consistency Indices. Three types of accuracy and consistency indices were generated: *overall*, *conditional-on-level*, and *cut point*, provided in Appendix F: Reliability and Classification Accuracy. The *overall accuracy* of performance-level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels. It is a proportion (or percentage) of correct classification across all the levels. The overall accuracy index ranges from 0.669 to 0.672 for LEAP 2025 Biology.

Another way to express overall consistency is to use Cohen's Kappa (κ) coefficient (Cohen, 1960). The overall coefficient Kappa when applying all cutoff scores together is

$$\kappa = \frac{P - P_c}{1 - P_c},$$

where P is the probability of consistent classification, and P_c is the probability of consistent classification by chance (Lee, Hanson, & Brennan, 2000). P is the sum of the

diagonal elements, and P_c is the sum of the squared row totals. The PChance index ranges from 0.245 to 0.250 for LEAP 2025 Biology.

Kappa is a measure of "how much agreement exists beyond chance alone" (Fleiss, 1973), which means that it provides the proportion of consistent classifications between two forms after removing the proportion of consistent classifications expected by chance alone. The Kappa index ranges from 0.413 to 0.415 across forms.

Consistency conditional-on-level is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all the students classified into that level (marginal entry).

Accuracy conditional-on-level is analogously computed. The only difference is that in the consistency table both row and column marginal sums are the same, whereas in the accuracy table, the sum that is based on true status is used as a total for computing accuracy conditional on level.

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.

Validity

"Validity is defined as … the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests" (AERA/APA/NCME, 2014). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process.

The 2018–2019 LEAP 2025 Biology test was designed and developed to provide fair and accurate scores that support appropriate, meaningful information for educational decisions. Validity evidence may be found in the following portions: Chapter 2

(Assessment Framework), Chapter 3 (Overview of the Test Development Process), Chapter 4 (Construction of Test Forms), Chapter 5 (Test Administration), Chapter 6 (Scoring Activities), Chapter 7 (Data Analysis), Chapter 8 (Reporting for Biology), Chapter 9 (Data Review Process and Results), Chapter 10 (Reliability and Validity), and Chapter 11 (Statistical Summaries). As the technical report has evolved, chapter by chapter, it reflects phases of the testing cycle. Each part of the technical report details the procedures and processes applied in the creation of LEAP 2025 and their results.

The knowledge, expertise, and professional judgment offered by Louisiana educators ultimately ensure that the content of the LEAP 2025 Biology assessment is an adequate and representative sample of appropriate content, and that the content is a legitimate basis upon which to derive valid conclusions about student achievement.

Chapters 3 and 4 of the technical report address test-form development. Chapter 3 presents a general discussion of test book creation and the editing process, describing the selection of operational test items, the content distribution of embedded field test items, and the process to obtain approvals from the LDOE. The test design process and participation by Louisiana educators throughout the process—from item development, content review, and bias review to test selection—reinforce confidence in the content and design of LEAP 2025 to derive valid inferences about Louisiana student performance.

Chapter 5 of the technical report describes the process, procedures, and policies that guide the administration of the LEAP 2025 assessments, including accommodations, test security, and detailed written procedures provided to test administrators and school personnel.

Chapter 6 describes scoring processes and activities for the LEAP 2025 Biology assessment.

Chapter 7 describes classical data analysis and item response theoretic calibration, scaling, and equating methods, as well as processes and procedures to clean data to ensure replicable, iterative calibrations and scaling of the 2018–2019 LEAP 2025 Biology test to derive scale scores from students' raw scores. Some references to introductory and advanced discussions of IRT are provided. Chapter 7 also describes an analysis of DIF.

Complete tables of gender and ethnicity DIF results for all 2018–2019 LEAP 2025 Biology operational items are presented in Appendix C.

Chapter 8 of the technical report summarizes the test results, score distributions, and achievement-level information.

Chapter 9 describes the data review process and results.

Chapter 10 addresses Cronbach's alpha and marginal alpha as measures of internal consistency and describes analysis procedures for classification consistency and classification accuracy.

Chapter 11 reports the statistical summaries of the LEAP 2025 Biology assessment for 2018–2019.

Additional, corroborating evidence consistent with the validity, reliability, and consistency of the LEAP 2025 Biology assessment has been documented in the LEAP Biology framework, test development plans, and the 2019 Biology standard setting technical report.

11. Statistical Summaries

The LEAP 2025 test results for Biology are not on a vertical scale, and therefore the scale scores across grades cannot be compared. While the lowest obtainable scale score on the Science tests is 650, the highest obtainable scale score is 850. Test results are presented in Table 11.1. Scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state and disaggregated into various demographic groups. In addition to the descriptive statistics presented in Table 11.1, scale score frequency distributions are presented in Appendix E.

The current years' unidimensionality results can be found in Appendix D. We continue to conduct a principal component analysis. Measurement implies order and magnitude along a single dimension (Andrich, 1989). In the case of scholastic achievement, a one-dimensional scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 1988; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983). To check the unidimensionality, the relative sizes of the eigenvalues associated with a principal component analysis of the item set will be examined using the SAS program.

Table 11.1

Spring 2019 LEAP 2025 State Test Results Biology

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥34,140	733.94	25.90	14	20	36	23	7
Gender								
Female	≥17,390	734.86	24.76	12	20	38	23	7
Male	≥16,750	732.98	27.00	16	20	35	23	7
Ethnicity								
Hispanic/Latino	≥2,100	729.89	28.30	20	19	34	20	6
American Indian or Alaska Native	≥240	739.27	25.45	7	17	42	24	9
Asian	≥550	750.21	25.12	5	12	27	34	21
Black	≥14,170	722.91	24.05	23	29	35	12	2
Native Hawaiian or Other Pacific Islander	≥30	744.21	31.24	12	6	39	18	24
White	≥16,430	743.16	23.11	6	14	38	32	10
Multi-Racial	≥610	738.23	25.34	10	17	40	24	9
Economically Disadvantaged (Economic Status)								
No	≥13,440	743.15	24.41	7	13	36	32	11
Yes	≥20,700	727.95	25.08	18	25	37	17	4
LEP Status								
Fully English Proficient	≥33,240	734.61	25.59	13	20	37	23	7
English Learner	≥900	709.32	25.49	45	29	20	5	1

References

AERA/APA/NCME. (1999/2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Andrich, A. (1988). Rasch models for measurement. Newbury Park, CA: SAGE Publications, Inc.

Andrich, A. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & H. H. Lovibond (Eds.), *Mathematical and theoretical systems*. North-Holland: Elsevier Science Publisher B.V.

Andrich, A. (2004). *Modern measurement and analysis in social science*. Murdoch University, Perth, Western Australia.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Warner (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, 63(4), 602–614.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31–44.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–47.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Educational Testing Service.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Green, D. R. (1975, December). Procedures for assessing bias in achievement tests. Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000, October). Procedures for computing classification consistency and accuracy indices with multiple categories (ACT Research Report Series 2000–10). Iowa City: ACT, Inc.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179–197.

Loehlin, J. C. (1987). *Latent variable models*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8, 452-461.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

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

Orlando, M. (2004, June). Critical issues to address when applying item response theory (IRT) models. Paper presented at the Drug Information Association, Bethesda, MD.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. In W. E. Hathaway (Ed.), *Testing in the schools. New directions for testing and measurement* (p. 19). San Francisco: Jossey-Bass.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E.P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies: A revised core vocabulary*. Austin, TX: Steck-Vaughn.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161–186). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). MULTILOG (version 7) [Computer software]. In Mathilda du Toit (Ed.), *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. Chicago: Scientific Software International.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.

Young, M. J., & Yoon, B. (1998, April). Estimating the consistency and accuracy of classifications in a standards-referenced assessment (CSE Technical Report 475). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: University of California, Los Angeles.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 26, 44–66.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, 10(4), 321–344.

Appendix A: Training Agendas

LEAP 2025 Biology Item Outline Development Training Agenda
Item Development Cycle for the 2018–2019 LEAP 2025 Science Assessment

I. **Item Development Process**
   a. Overview
   b. Steps in process

II. **Outlines**
   a. What outlines are
      i. Definition and purpose
      ii. Components
   b. What outlines are not
      i. Characteristics
      ii. Non-examples
   c. Outline assignments
      i. Tasks
         Components
            a. Stimulus
               i. Purpose of graphics, data tables, and graphs
               ii. Reading level
            b. Item types (G3,4 vs 5-EOC/Bio)
            c. Bundling of PEs
      ii. Item sets
         Components
            a. Stimulus
            b. Item types (G3,4 vs 5-EOC/Bio)
            c. Bundling of PEs
      iii. Standalone
            a. Purpose
            b. Use of graphics, data tables, and graphs
            c. Item Types
            d. Single PEs
      iv. Template

III. **Considerations**

a. Tasks
    i. Needed number of items and ERs
    ii. Dimensionality
    iii. Number of items seen by students vs. number of items developed
    iv. Use of PEs
    v. Use of scaffolding within the task

b. Item sets
    i. Needed number of items and ERs
    ii. Dimensionality
    iii. Interchangeability
    iv. Use of PEs (mix and match)
    v. Number of items seen by students vs. number of items developed

c. Phenomena list (topics to avoid)

d. Bias and Sensitivity
    i. Definitions
        1. Bias
        2. Sensitivity
        3. Stereotyping
        4. Fairness
    ii. Rationale for Removing Bias and Sensitivity
        1. Portrayal of groups within Louisiana's diverse population
        2. Protection of privacy and avoidance of offensive content
    iii. Potential Sources of Bias
        1. Ethnicity
        2. Culture
        3. Religion
        4. Disability
        5. Gender/age stereotypes
        6. Geography
        7. Socioeconomic status
        8. Controversial issues or contexts
        9. English language proficiency
    iv. Strategies to Avoid Bias
        1. Include non-DCI related information needed to understand stimulus/make stimulus accessible to students regardless of background.
        2. Use familiar language and contexts to avoid accessibility bias.
        3. Avoid issues and themes that demean, offend, or inaccurately portray any religion, ethnicity, culture, gender, social group, disability
        4. Avoid topics that will offend the privacy of values and beliefs of students, parents, or public

LEAP 2025 Biology Item Writer Training Agenda
Item Development Cycle for the 2018–2019 LEAP 2025 Science Assessment

I. **Project Overview**
   a. Purpose of LEAP project in science
   b. Characteristics of assessment
      i. Grade specific, ending the current practice of grade span assessments in grades 4 and 8;
      ii. Designed to be accessible for use by the widest possible range of students, including but not limited to students with disabilities and English Learners (ELs);
      iii. Constructed to yield valid and reliable test results while reporting student performance to five achievement levels;
      iv. Developed and/or reviewed with Louisiana educator and student involvement;
      v. Non-computer-adaptive; and
      vi. Administered online.

II. **Louisiana Student Standards for Science (LSSS)**
   a. New science standards were approved in early March 2017.
      i. The LSSS represent the knowledge and skills needed for students to successfully transition to postsecondary education and the workplace. The standards call for students to:
         1. Apply content knowledge to real-world phenomena and to design solutions;
         2. Demonstrate the practices of scientists and engineers;
         3. Connect scientific learning to all disciplines of science; and
         4. Express ideas grounded in scientific evidence.
   b. The Louisiana Student Standards are not the NGSS!

III. **Anatomy of the LSSS**
   a. Descriptor
   b. Grade level
   c. Standard
   d. Domain
   e. Topic number
   f. Performance Expectation
      i. Science and Engineering Practices
      ii. Disciplinary Core Ideas
      iii. Crosscutting Concepts

IV. **More Acronyms**
   a. SEP key
      i. 1. Q/P = Asking Questions and Defining Problems
      ii. 2. MOD = Developing and Using Models
      iii. 3. INV = Planning and Carrying Out Investigations

iv. 4. DATA = Analyzing and Interpreting Data
v. 5. MCT = Using Mathematics and Computational Thinking
vi. 6. E/S = Constructing Explanations and Designing Solutions
vii. 7. ARG = Engaging in Argument from Evidence
viii. 8. INFO = Obtaining, Evaluating, and Communicating Information

b. CCC key
   i. PAT = Patterns
   ii. C/E = Cause and Effect
   iii. SPQ = Scale, Proportion, and Quantity
   iv. SYS = Systems and System Models
   v. E/M = Energy and Matter
   vi. S/F = Structure and Function
   vii. S/C = Stability and Change
c. "Acronyms Cheat Sheet"

Multidimensional Standards → Multidimensional Assessment

d. Dimensions are never to be taught in isolation, and therefore are never tested in isolation.
e. The goal of a multidimensional assessment is to gather evidence that a student has proficiency in each of the three dimensions.
    i. Every item must align to at least two of the three dimensions (with one exception for ERs—"mix and match").
    ii. Assessment must reflect the different dimensional combinations.
        1. SEP and DCI
        2. DCI and CCC
        3. SEP and CCC (not content)
        4. SEP, DCI, CCC

V. **Aligning to Multiple Dimensions**

a. SEP:
    i. Develop and model; Analyze data; Construct an explanation
b. DCI:
c. CCC:
    i. Energy and Matter; Patterns; Scale, Proportion, and Quantity

VI. **Phenomena: Keystone of 3-D Assessments**

a. Phenomena: Observable events that students can use the three dimensions to explain or make sense of.
    i. Links to phenomena websites are available in the "LEAP Phenomena and Context" document.

VII. **Context: How Phenomena Are Presented**

a. Contexts are the setting in which phenomena are presented (stimuli).
b. A single phenomenon can be presented in many different contexts.
c. Phenomena ≠ context; context ≠ phenomena

VIII. **Contexts and Stimuli**

a. Stimuli contain contexts in which phenomena are presented.
b. Contexts and stimuli should be unique and novel.
    i. Non-textbook
    ii. Think outside the box
c. Stimuli must be student friendly and grade appropriate.
    i. Engaging to students
    ii. Free of bias and sensitivity issues
        1. Definitions
            a. Bias
            b. Sensitivity
            c. Stereotyping
            d. Fairness
        2. Rationale for Removing Bias and Sensitivity
            a. Portrayal of groups within Louisiana's diverse population
            b. Protection of privacy and avoidance of offensive content

3. Potential Sources of Bias
    a. Ethnicity
    b. Culture
    c. Religion
    d. Disability
    e. Gender/age stereotypes
    f. Geography
    g. Socioeconomic status
    h. Controversial issues or contexts
    i. English language proficiency
4. Strategies to Avoid Bias
    a. Include non-DCI related information needed to understand stimulus/make stimulus accessible to students regardless of background.
    b. Use familiar language and contexts to avoid accessibility bias.
    c. Avoid issues and themes that demean, offend, or inaccurately portray any religion, ethnicity, culture, gender, social group, disability
    d. Avoid topics that will offend the privacy of values and beliefs of students, parents, or public

d. Phenomena, contexts, and stimuli need to be the right grain size.
e. Goldilocks—provide only the information that is needed.

IX. **Phenomena and PE Bundles**

a. *PE bundle* is usually 2 PEs, but 1-PE and 3-PE bundles are acceptable.
b. PE bundling is used in two of the three "item groupings" on LSSS assessment.
c. See "Phenomena and Context Overview" and "Contexts and Stimuli" documents for more information.

X. **Assessment Design: Item Components**

a. The LSSS assessment will consist of three distinct "components."
    i. Tasks (PE bundles; phenomena)
    ii. Item sets (PE bundles; phenomena)
    iii. Standalone items (single PE only; foci)

XI. **Component: Task**

a. Tasks (stimulus; four items + ER; dependency OK; phenomenon/PE bundle)
b. Tasks include a stimulus and a dependent set of four 1- or 2-point SRs and/or TE items, culminating with one 3-dimensional extended response.
c. Items in tasks may require a specific order.
d. Information in one item may be used in another item (but NOT cue!).
e. Items may be scaffolded to help discriminate student performance levels.
f. All items help make sense of or explain a phenomenon.
g. No CRs
h. For ER: Can "mix and match" within dimensions from PE bundle as long as the ER aligns with one SEP, one DCI, and one CCC

XII. **Component: Item Set**

a. Item set (stimulus; four items total; CR possible; no inter-item dependency)
    i. Item sets are composed of a stimulus and four 1- or 2-point SR, TE, and/or CR items.
    ii. Some item sets will contain one 2-point CR.
    iii. Item sets without a CR will contain one 2-point TE item (likely an evidence-based selected response [EBSR]).
    iv. Items are independent of one another, but all items must depend on the common stimulus.
    v. Like tasks, the item set makes sense of or explains a phenomenon using a PE bundle. No ERs are included in item sets.

XIII. **Component: Standalone Items**

a. Standalone items (single PE; no parts)
    i. Standalone items will have a "focus" rather than a phenomenon upon which a stimulus is built. This is because a phenomenon is too large to explain or make sense of with one item.
    ii. Item types include 1- and 2-point formats: no CRs or ERs.

XIV. **Item Types: Selected Response (SR) Formats**

a. Multiple choice (MC) (1 point)
    i. Four answer options with one and only one correct answer
b. Multiple select (MS) (1 point)
    i. Five or six answer options with two or three correct answers

XV. **Item Types: Open-Response Formats**

a. Constructed response (CR) (2 points)
    i. Students enter text into a response space
    ii. Can be two parts
    iii. Aligns to PE bundle
    iv. 2-D or 3-D
    v. Used in item sets ONLY (not all)
b. Extended response (ER) (grades 3 and 4: 6 points; grades 5–EOC: 9 points)
    i. Students enter text into a response space
    ii. Can be up to three parts
    iii. 3-D: Aligns to one SEP, one DCI, and one CCC (mix and match from PE bundle)
    iv. Can include additional stimulus
    v. Can reference or depend on previous item in task
    vi. Role of scaffolding
    vii. Used in tasks ONLY

XVI. **Item Types:**

a. Technology-enhanced items (TEIs)
    i. TEIs are worth 1 or 2 points
    ii. Used in tasks, item sets, and standalone items
    iii. TEI types (NO TEIs in grades 3 and 4!)
        1. Graphic Gap Match
            - Graphic Gap Match Response Interactions allow graphic gaps and graphic choices. This item type can also be used to create regular gap matches by creating the background in art.
        2. Order Interaction
            - An Order Interaction Response Interaction consists of choices that may be placed in order or sequence and is a drag-and-drop interaction type. Typically, this interaction type will have three or more choices. The test taker drags the options to the desired order.
        3. Hot Spot
            - A Hot Spot Response Interaction includes an art image or graphic. The initial state of this item type has no choices selected. This interaction type has a specific set of choices or hot spots that are defined within areas of the art image. One or more choices may be selected in this interaction.
        4. Hot Text
            - Hot Text Response Interactions include only text. The initial state of this item type has no choices selected. This interaction type has a specific set of hot text selections that are defined within areas of the text. One or more choices may be selected in this interaction.

5. Fill in the Blank (FIB)
    - A Text Entry (FIB) Response Interaction includes a free-form field where the test taker enters text, without the ability to use the return or enter key. This interaction will not support multi-line responses.

b. Evidence-based selected response (EBSR): Combination of two questions; second question asks students to identify evidence used from the text to support their response to the first question

XVII. **Development Process Overview**

XVIII. **Universal Design**

a. Ensures that a fair test is developed that provides an accurate measure of what all assessed students know and can do without compromising reliability or validity
    i. Use consistent naming and graphics conventions;
    ii. Ensure reading level suitable for the grade level being tested;
    iii. Replace low-frequency words with simple, common words;
    iv. Avoid irregularly spelled words, words with ambiguous or multiple meanings, technical terms unless defined and integral to meaning, and concepts with multiple names, symbols, or representations;
    v. Ensure clarity of noun-pronoun relationships (eliminate pronouns wherever possible);
    vi. Simplify keys and legends;
    vii. Use grade-appropriate content; and
    viii. Avoid differential familiarity for any group, based on language, socioeconomic status, regional/geographic area, or prior knowledge or experience unrelated to the subject matter being tested (bias/sensitivity).
b. See “Universal Design” for more information.

XIX. **Item Difficulty**

a. Item difficulty allows students to be placed along a learning progression and assigned to one of the FIVE proficiency levels (to be set at a future date).
    i. Want a range of difficulty items among each item grouping
    ii. Cognitive complexity is not difficulty.
b. See “Item Difficulty Overview” for more information.

XX. **Cognitive Complexity***

a. Need for a range of items of varied cognitive complexity
b. Existing models of cognitive complexity (e.g., DOK)
c. Development of a model to address three-dimensional items of LEAP assessment*
d. (*As the TAGS-M model was in development during the early portion of the 2018-19 development cycle, item writers used their understanding of cognitive complexity to develop two- and three-dimensional items aligned to the PEs of the LSSS, targeting a broad range of cognitive complexities. These items were then coded by WestEd staff after the TAGS-M model was complete.)

XXI. **Sourcing**

a. Sources are required for specific information, such as species, planets, stars, elements, or designs of existing solutions.
    i. Sources are not needed for commonly known facts.
        1. Formula for photosynthesis
        2. The definition of speed
    ii. If in doubt, source!
    iii. Use reputable sources.
    iv. See “Sources” for more information.

XXII. **Graphics**

a. Graphics are used to convey ideas, data, and/or concepts in a simplified visual form.
    i. Graphics are essential components of science and include:
        1. Tables, diagrams, models, graphs, images
    ii. All graphics must be introduced appropriately with an introductory statement. Some graphics require only a brief introduction; some require a bit more, e.g.:
        1. The students’ results are shown in the table below.
        2. Students made a scale drawing of their prototype. The scale drawing is shown below.
    iii. Be aware that some graphics may be changed during production to control for colorblindness.
    iv. See “General Guidelines for Graphics” document for more information.
    v. Style guide

XXIII. **Development Process Overview**

XXIV. **Information Security**

a. Do NOT email!
b. We will send/receive items and assignments using a secure system.
c. General questions about processes OK

LEAP 2025 Biology Editor Training Agenda
Item Development Cycle for the LEAP 2025 Science Assessment

I. **Item Set/Task/Standalone Item Overview**
   a. Criteria for review
II. **Item Development Process**
   a. One round of items slated for development in 2018-19
   b. All batches will go through four rounds of LDOE review at different stages of development before committee:
      i. Outline review (item descriptions; graphic roughs)
      ii. Item development
         1. R1 (fully fleshed-out items; functional TE items; graphics; sources)
         2. R2 (implementation of LDOE feedback; rewrites possible; revisions expected)
         3. R3 (final look before committee review—no editing, all comments are for committee review)
   c. Committee review
III. **Process Overview for Intake/E1**
IV. **Intake/E1 Rules for Returning Item Sets/Tasks/Standalone Item Submissions to Writers**
V. **Feedback to Writers**
VI. **Process Overview for Intake/E2**
VII. **Intake/E1 Rules for Returning Item Sets/Tasks/Standalone Item Submissions to E1 Writer**
VIII. **Use of the Style Guides**
   a. Social Studies/Science Content Style Guide
   b. TEI Guide
   c. Graphics Style Guide

Appendix B: Test Summary

Biology

Contents
Table B.1 Test Blueprint Distribution by Reporting Category for Spring 2019 Operational Biology: Percentage of Points by Reporting Category (includes Task Items)
Tables B.2.1–B.2.2 Standard Coverage by Form: Spring 2019 Operational Biology
Table B.3 Summary of Spring 2019 EFT Item Development (Field-Tested Items by Item Type)
Table B.4 Spring 2019 Operational Item Summary for Biology
Table B.5 Raw Score Summary: Spring 2019 Operational Biology
Table B.6 Raw Score Summary by Reporting Category: Spring 2019 Operational Biology
Tables B.7.1–B.7.2 Scale Score and Raw Score Summary: Spring 2019 Operational Biology

Table B.1

Test Blueprint Distribution by Reporting Category for Spring 2019 Operational Biology: Percentage of Points by Reporting Category (includes Task Items)

Reporting Category	Form B	Form C
Investigate	22.0%	26.8%
Evaluate	14.6%	17.1%
Reason Scientifically	31.7%	22.0%

Table B.2
Standard Coverage by Form: Spring 2019 Operational Biology

Table B.2.1
Form B

Reporting Categories and GLEs		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
Investigate	HS-LS2-1			1					3.57
	HS-LS2-4							1	3.57
	HS-LS2-6		1			2			10.71
	HS-LS3-2		1						3.57
	HS-LS3-3			1					3.57
	HS-LS4-5			1		1			7.14
	Subtotal		**2**	**3**		**3**		**1**	**32.14**
Evaluate	HS-LS1-3		1		1				7.14
	HS-LS3-1		1						3.57
	HS-LS3-3		1			2			10.71
	Subtotal		**3**		**1**	**2**			**21.43**
Reason Scientifically	HS-LS1-2	1			1				7.14
	HS-LS1-4							1	3.57
	HS-LS1-5	1							3.57
	HS-LS1-6		1						3.57
	HS-LS1-7				1	1			7.14
	HS-LS2-7		1	1		1	1		14.29
	HS-LS4-2					1			3.57
	HS-LS4-4				1				3.57
	Subtotal	**2**	**2**	**1**	**3**	**3**	**1**	**1**	**46.43**
Total		**2**	**7**	**4**	**4**	**8**	**1**	**2**	**100.00**

Table B.2.2

Form C

Reporting Categories and GLEs		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
Investigate	HS-LS2-1			1					3.70
	HS-LS2-4							1	3.70
	HS-LS2-6		1			2			11.11
	HS-LS3-2		2				1		11.11
	HS-LS3-3			1					3.70
	HS-LS4-5			1		1			7.41
	Subtotal		**3**	**3**		**3**	**1**	**1**	**40.74**
Evaluate	HS-LS1-3		1		1				7.41
	HS-LS3-1		1			1			7.41
	HS-LS3-3	1				2			11.11
	Subtotal	**1**	**2**		**1**	**3**			**25.93**
Reason Scientifically	HS-LS1-2	1			1				7.41
	HS-LS1-4							1	3.70
	HS-LS1-5	1							3.70
	HS-LS1-6		1						3.70
	HS-LS1-7				1	1			7.41
	HS-LS4-2					1			3.70
	HS-LS4-4				1				3.70
	Subtotal	**2**	**1**		**3**	**2**		**1**	**33.33**
Total		**3**	**6**	**3**	**4**	**8**	**1**	**2**	**100.00**

Table B.3

Summary of Spring 2019 EFT Item Development (Field-Tested Items by Item Type)

Item Type	Item Count	Percent
CR	10	6%
MC	62	39%
MS	8	5%
TE	52	33%
TPD	15	10%
TPI	10	6%

Table B.4

Spring 2019 Operational Item Summary for Biology

Form	MC	MS	TE	CR	ER	TPD	TPI
B	13	6	8	3	1	8	2
C	13	6	8	3	1	7	3

Table B.5

Raw Score Summary: Spring 2019 Operational Biology

Form	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
B	≥17,700	25	10	0	59	0.33	0.39	0.86	3.91
C	≥16,440	25	10	0	57	0.33	0.39	0.85	3.89

Note: Reliability is coefficient alpha.

Table B.6

Raw Score Summary by Reporting Category: Spring 2019 Operational Biology

Core Test Form	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
B	Investigate	4.84	2.19	0	14	0.41	0.30	0.24	1.91
	Evaluate	8.26	3.86	0	20	0.37	0.45	0.64	2.32
	Reason Scientifically	15.21	6.79	0	38	0.32	0.45	0.75	3.40
C	Investigate	7.17	2.76	0	17	0.42	0.30	0.26	2.37
	Evaluate	11.19	5.46	0	29	0.33	0.46	0.72	2.89
	Reason Scientifically	9.84	4.29	0	25	0.29	0.40	0.57	2.81

Table B.7
Scale Score and Raw Score Summary: Spring 2019 Operational Biology

Table B.7.1
Form B

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥17,700	100.00	734.60	26.42	25	10
Female	≥8,910	50.34	735.57	25.14	26	10
Male	≥8,790	49.66	733.61	27.62	25	11
African American	≥7,340	41.49	723.62	24.39	21	9
American Indian or Alaska Native	≥130	0.75	741.62	26.46	28	11
Asian	≥280	1.60	747.33	26.49	31	11
Hispanic/Latino	≥1,210	6.84	727.78	29.29	23	11
Multi-Racial	≥300	1.74	737.13	25.76	26	10
Native Hawaiian or Other Pacific Islander	≥20	0.12	747.00	34.29	31	14
White	≥8,400	47.46	744.51	23.41	29	10
Economically Disadvantaged	≥10,920	61.71	728.61	25.72	23	10
English Language Learners	≥610	3.45	708.13	25.42	16	8

Note: These tables report the number of students, scaled-score means, and standard deviations for subgroups.

Table B.7.2
Form C

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥16,440	100.00	733.23	25.32	25	10
Female	≥8,480	51.57	734.12	24.34	25	10
Male	≥7,960	48.43	732.29	26.28	24	10
African American	≥6,820	41.51	722.14	23.66	20	9
American Indian or Alaska Native	≥110	0.67	736.48	24.01	26	10
Asian	≥270	1.64	753.25	23.26	33	11
Hispanic/Latino	≥880	5.41	732.76	26.65	25	10
Multi-Racial	≥300	1.85	739.33	24.91	27	10
Native Hawaiian or Other Pacific Islander	≥10	0.07	739.33	25.70	27	11
White	≥8,030	48.84	741.76	22.70	28	10
Economically Disadvantaged	≥9770	59.44	727.22	24.31	22	9
English Language Learners	≥290	1.78	711.80	25.50	17	8

Note: These tables report the number of students, scaled-score means, and standard deviations for subgroups.

Appendix C: Item Analysis Summary Report

Summary Statistics Reports

Biology

Contents
Table C.1 *P*-Value by Item Type: Spring 2019 Operational Biology
Plot C.1 *P*-Value by Item Type: Spring 2019 Operational Biology
Table C.2 Item-Total Correlation, Point-Biserial Correlation: Spring 2019 Operational Biology
Plot C.2 Item-Total Correlation, Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology
Table C.3 Corrected* Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology
Plot C.3 Corrected* Point-Biserial Correlation: Spring 2019 Operational Biology
Table C.4 Item-Total Correlation Summary by Reporting Category: Spring 2019 Operational Biology
Table C.5 Statistically Flagged Operational Items: Spring 2019 Operational Biology
Table C.6 IRT Item Parameters: Spring 2019 Operational Biology
Plot C.4 IRT a-Parameter: Spring 2019 Operational Biology
Plot C.5 IRT b-Parameter: Spring 2019 Operational Biology
Plot C.6 IRT c-Parameter: Spring 2019 Operational Biology

Table C.1

P-Value Summary by Item Type: Spring 2019 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.030	0.030	0.044	0.092	0.092
ER	2	0.119	0.121	0.169	0.232	0.248
MC	14	0.199	0.342	0.422	0.600	0.713
MS	6	0.071	0.081	0.139	0.276	0.384
TE	10	0.088	0.274	0.397	0.515	0.700
TPD	9	0.178	0.249	0.302	0.537	0.657
TPI	2	0.158	0.158	0.364	0.569	0.569

Plot C.1

P-Value by Item Type: Spring 2019 Operational Biology

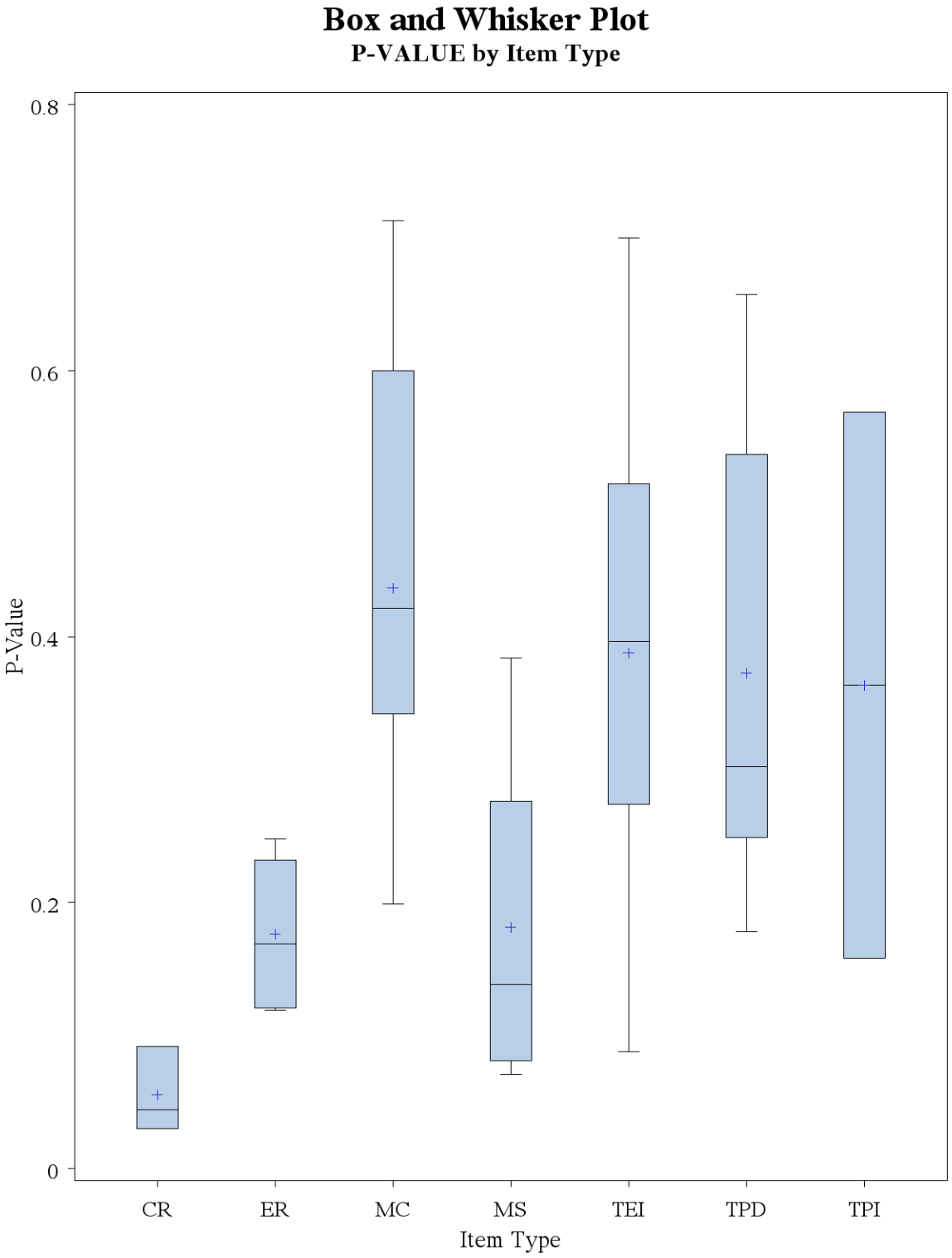


Table C.2

Item-Total Correlation, Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.249	0.249	0.350	0.479	0.479
ER	2	0.568	0.569	0.577	0.609	0.631
MC	14	0.129	0.203	0.355	0.519	0.566
MS	6	0.106	0.290	0.318	0.393	0.490
TE	10	0.239	0.329	0.402	0.504	0.593
TPD	9	0.175	0.337	0.408	0.503	0.623
TPI	2	0.367	0.367	0.512	0.658	0.658

Plot C.2

Item-Total Correlation, Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology

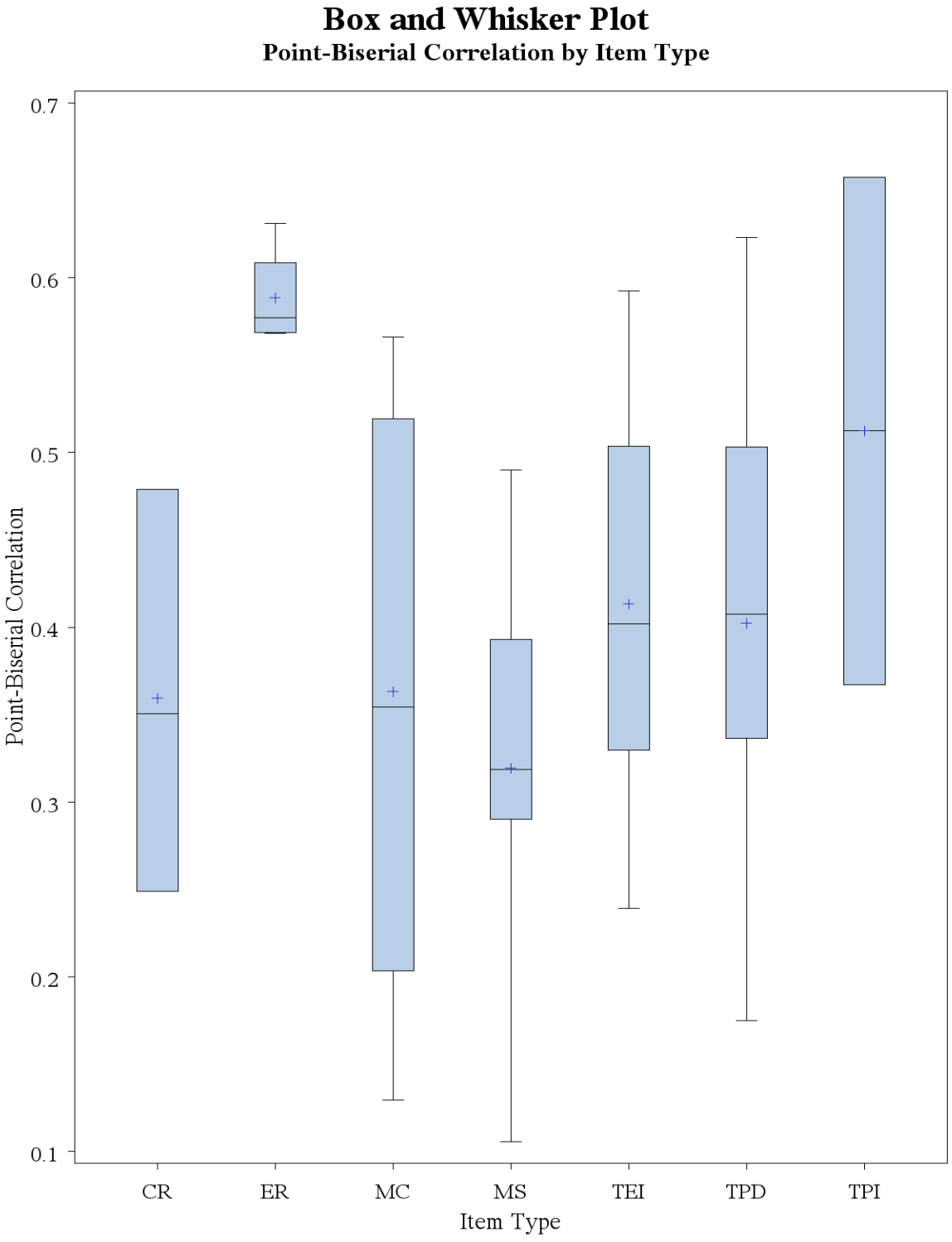


Table C.3

Corrected Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.218	0.218	0.325	0.440	0.440
ER	2	0.491	0.494	0.498	0.522	0.546
MC	14	0.089	0.155	0.312	0.482	0.531
MS	6	0.080	0.265	0.281	0.362	0.455
TE	10	0.212	0.264	0.356	0.437	0.536
TPD	9	0.097	0.272	0.326	0.437	0.564
TPI	2	0.320	0.320	0.464	0.607	0.607

Note: *Corrected point-biserial correlation which is slightly more robust than point-biserial correlation, calculates the relationship between the item score and the total test score after removing the item score from the total test score.

Plot C.3

Corrected* Point-Biserial Correlation by Item Type: Spring 2019 Operational Biology

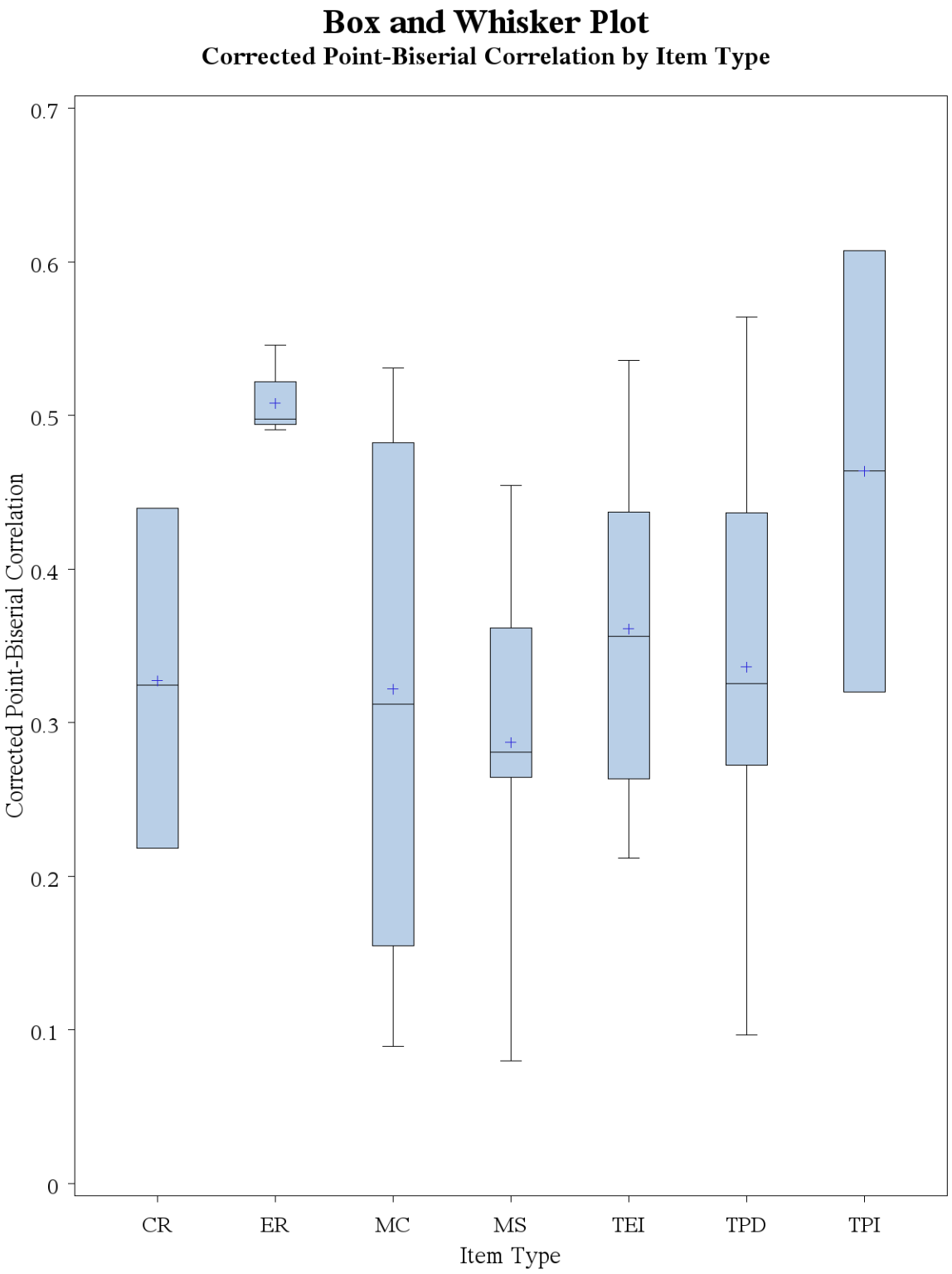


Table C.4

Item-Total Correlation by Reporting Category: Spring 2019 Operational Biology

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Evaluate	1	0.25	0.25	0.25	0.25	0.25
CR	Reason Scientifically	1	0.35	0.35	0.35	0.35	0.35
ER	Evaluate	1	0.57	0.57	0.60	0.63	0.63
ER	Reason Scientifically	1	0.57	0.57	0.58	0.59	0.59
MC	Evaluate	3	0.36	0.36	0.53	0.57	0.57
MC	Investigate	3	0.30	0.30	0.33	0.35	0.35
MC	Reason Scientifically	3	0.20	0.20	0.40	0.48	0.48
MS	Investigate	1	0.33	0.33	0.33	0.33	0.33
MS	Reason Scientifically	3	0.31	0.31	0.40	0.49	0.49
TE	Evaluate	3	0.36	0.36	0.46	0.57	0.57
TE	Reason Scientifically	1	0.50	0.50	0.50	0.50	0.50
TPD	Evaluate	3	0.34	0.34	0.41	0.53	0.53
TPD	Investigate	3	0.18	0.18	0.30	0.35	0.35
TPD	Reason Scientifically	2	0.41	0.41	0.52	0.62	0.62
TPI	Reason Scientifically	2	0.37	0.37	0.51	0.66	0.66

Table C.5

Statistically Flagged Operational Items: Spring 2019 Operational Biology

Item Type	*N* OP Items	*N* Items Flagged for *P*-Value	*N* Items Flagged for Mean	*N* Items Flagged for Point-Biserial Correlation	*N* Items Flagged for DIF	*N* Items Flagged for Omitting
CR	3	3	3	0	0	3
MC	14	2	0	3	1	9
MS	6	4	0	1	0	5
TEI	10	2	0	0	0	6
TPD	9	3	3	1	0	6
TPI	2	1	1	0	0	2

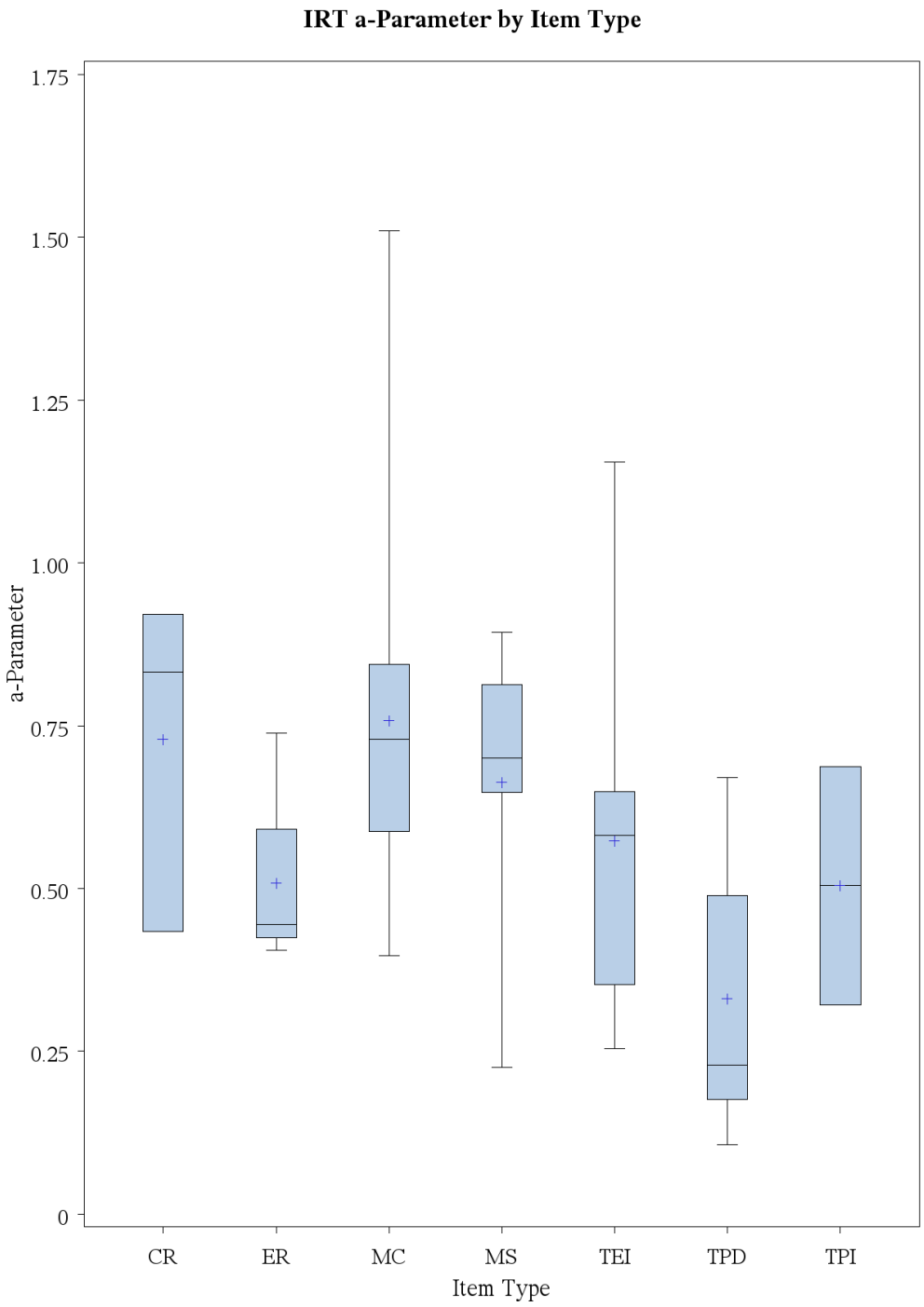
Table C.6

IRT Item Parameters by Item Type: Spring 2019 Operational Biology

Item Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.434	0.434	0.833	0.922	0.922
CR	b	3	1.206	1.206	1.892	2.120	2.120
ER	a	2	0.405	0.424	0.444	0.592	0.738
ER	b	2	-0.033	0.364	0.967	1.396	1.619
MC	a	14	0.396	0.587	0.729	0.844	1.510
MC	b	14	-1.095	-0.270	0.547	1.624	2.800
MC	c	14	0.018	0.073	0.180	0.248	0.331
MS	a	6	0.226	0.648	0.701	0.814	0.894
MS	b	6	0.739	0.875	1.851	2.517	3.511
MS	c	6	0.007	0.016	0.043	0.059	0.064
TEI	a	10	0.254	0.352	0.582	0.648	1.155
TEI	b	10	-1.032	-0.193	0.659	1.105	2.768
TEI	c	10	0.010	0.020	0.042	0.161	0.266
TPD	a	9	0.106	0.176	0.229	0.490	0.670
TPD	b	9	-0.965	-0.226	1.339	2.534	4.737
TPI	a	2	0.322	0.322	0.505	0.688	0.688
TPI	b	2	-0.805	-0.805	1.090	2.985	2.985

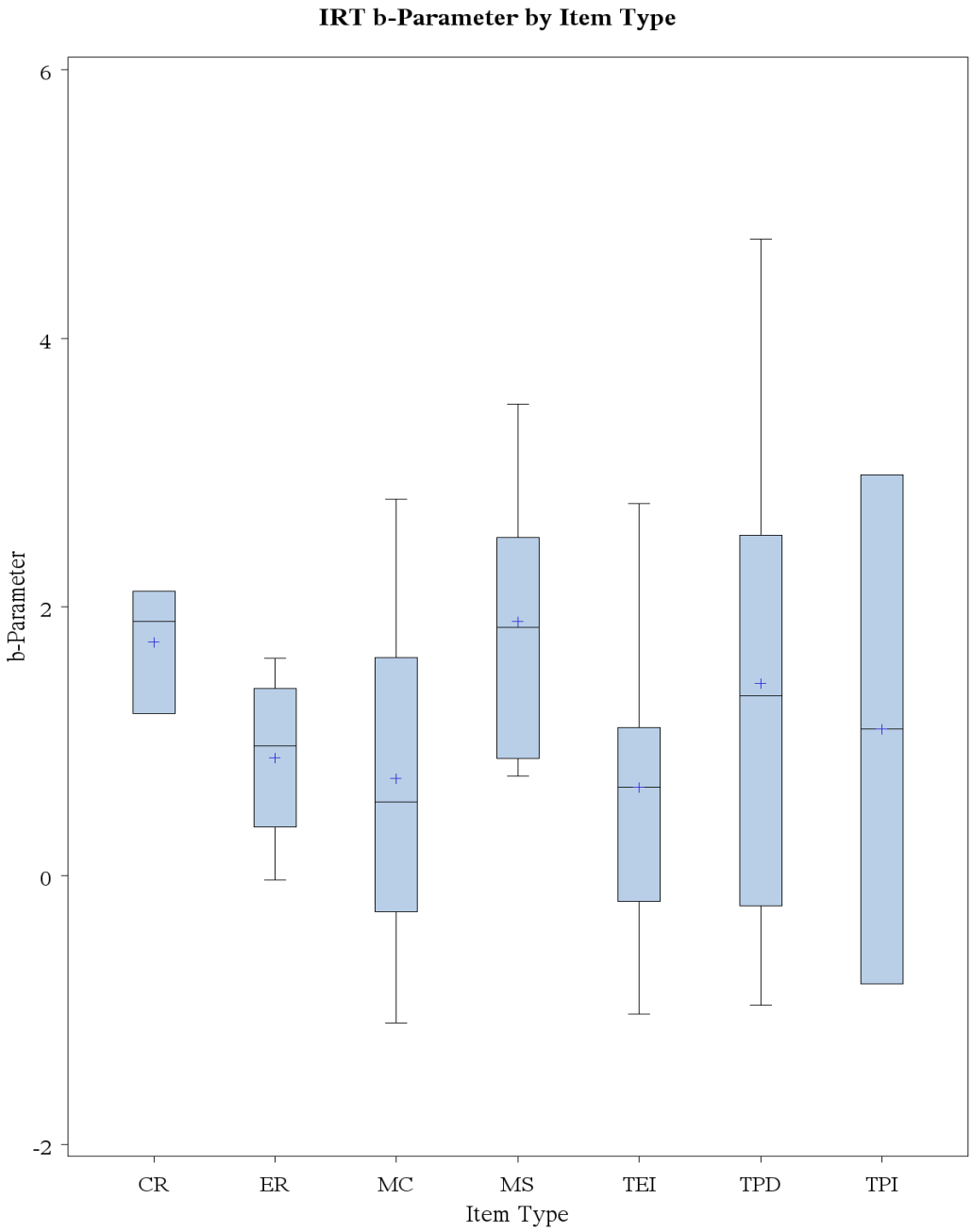
Plot C.4

IRT a-Parameter: 2019 Spring Operational Biology

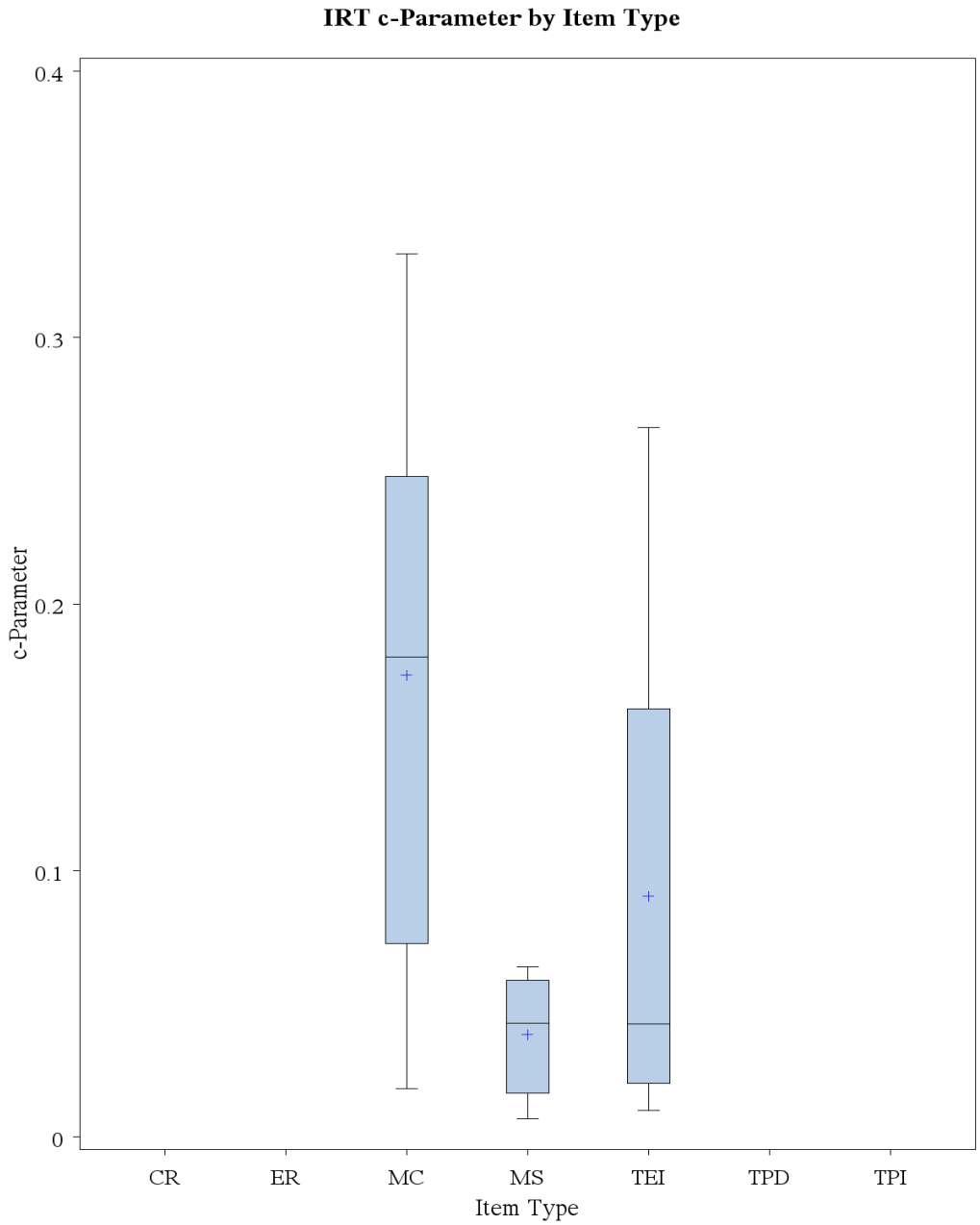


Plot C.5

IRT b-Parameter: 2019 Spring Operational Biology



Plot C.6
IRT c-Parameter: 2019 Spring Operational Biology



Note: Only dichotomous items (scored 0 or 1) have c parameters.

Appendix D: Dimensionality

Dimensionality Reports
Biology

Contents
Table D.1 Zq1 Statistics and Summary Data: Spring 2019 Operational Biology
Table D.2 Q3 Statistics and Summary Data: Spring 2019 Operational Biology
Table D.3.1–D.3.2 Reporting Category Intercorrelation Coefficients: Spring 2019 Operational Biology
Table D.4 First and Second Eigenvalue: Spring 2019 Operational Biology
Figure D.4 Principal Component Analysis Plot: Spring 2019 Operational Biology

Table D.1

Zq1 Statistics and Summary Data: Spring 2019 Operational Biology

Form	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
B	CR	28	28.0	50.0	56.0	56	0
	ER	105	105.0	174.5	244.0	244	2
	MC	1	16.0	23.0	27.0	77	0
	MS	4	5.0	13.0	21.0	49	0
	TEI	1	9.5	35.0	59.0	821	1
	TPD	15	22.0	52.0	79.0	271	1
	TPI	13	13.0	32.0	51.0	51	0
C	CR	28	28.0	50.0	56.0	56	0
	ER	56	56.0	145.0	234.0	234	2
	MC	1	16.0	23.0	27.0	77	0
	MS	4	5.0	13.0	21.0	49	0
	TEI	6	13.0	39.5	63.5	821	2
	TPD	15	17.0	36.0	86.0	271	1
	TPI	13	13.0	51.0	58.0	58	0

Table D.2

Q3 Statistics and Summary Data: Spring 2019 Operational Biology

Form	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
B	0.125	-0.080	-0.047	-0.015	0.076	0.341
C	0.113	-0.109	-0.055	-0.016	0.086	0.336

Table D.3

Reporting Category Intercorrelation Coefficients for Spring 2019 Operational Biology

Table D.3.1

Form B

Reporting Category	Investigate	Evaluate	Reason Scientifically
Investigate	1.00		
Evaluate	0.47	1.00	
Reason Scientifically	0.53	0.83	1.00

Table D.3.2

Form C

Reporting Category	Investigate	Evaluate	Reason Scientifically
Investigate	1.00		
Evaluate	0.60	1.00	
Reason Scientifically	0.62	0.81	1.00

Table D.4

First and Second Eigenvalue: Spring 2019 Operational Biology

Grade	Form	First Eigenvalue	Second Eigenvalue
High School	B	7.180	1.301
	C	6.772	1.254

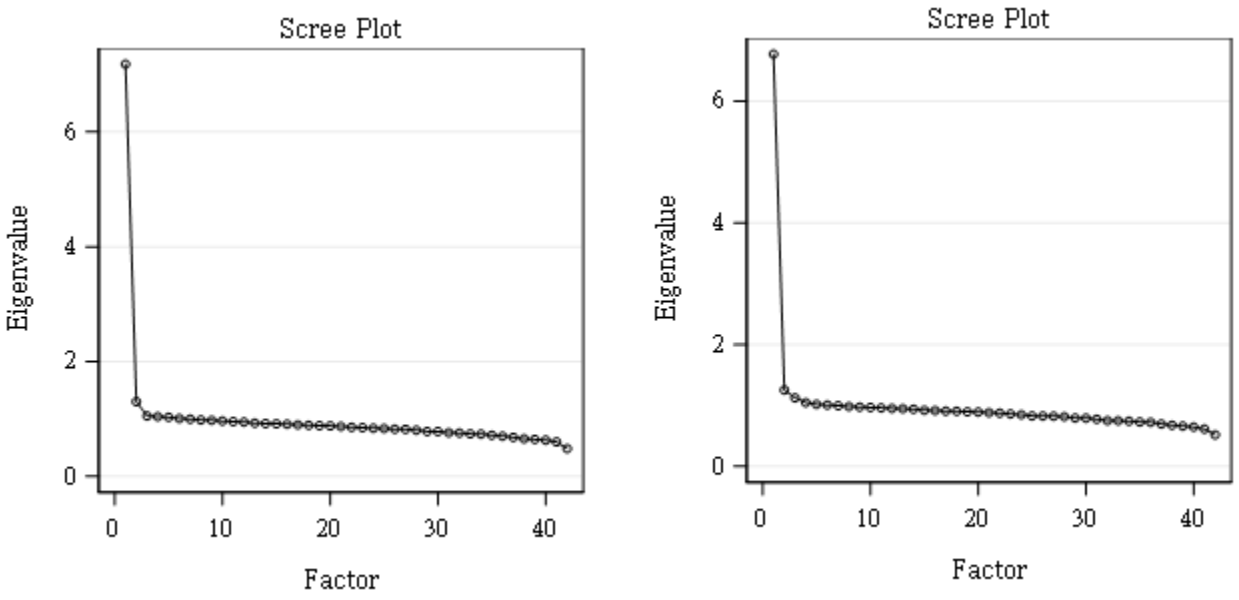


Figure D.4

Principal Component Analysis Plot: Spring 2019 Operational Biology

Appendix E: Scale Distribution and Statistical Report

Table E.1 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
BIOLOGY
ALL STUDENTS
Form ALL

N	≥34140		
Mean	733.94	Median	735.00
Std deviation	25.90	Variance	670.92
Skewness	-0.2486	Kurtosis	0.0262
Mode	720.00	Std Error Mean	0.1402
Range	173.00	Interquartile Range	35.00

Quantile	Estimate
100% Max	823
99%	787
95%	774
90%	766
75% Q3	752
50% Median	735
25% Q1	717
10%	700
5%	691
1%	672
0% Min	650

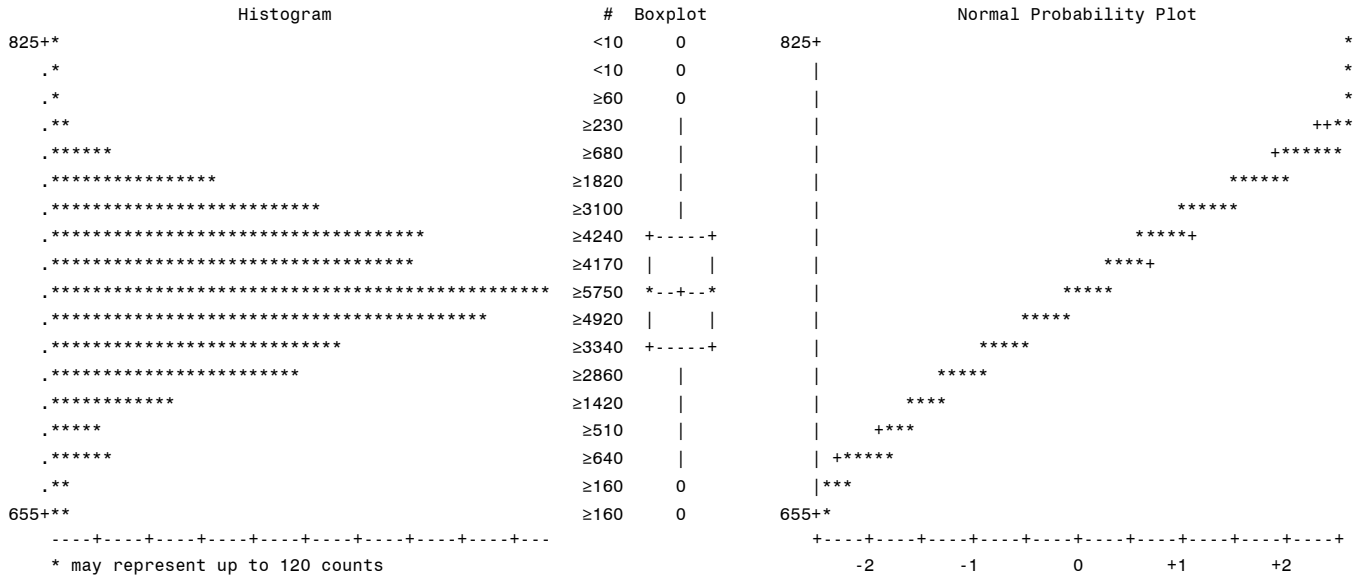


Table E.2 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
BIOLOGY
ALL STUDENTS
Form ALL

SCALE_SCORE		Freq	Freq	Cum. Percent	Cum. Percent
650	********	≥150	≥150	0.46	0.46
661	********	≥160	≥320	0.47	0.94
672	**********	≥250	≥570	0.74	1.68
679	**********	≥380	≥960	1.14	2.83
686	**********	≥510	≥1480	1.51	4.34
691	**********	≥620	≥2110	1.84	6.18
696	**********	≥790	≥2900	2.33	8.52
700	**********	≥880	≥3790	2.59	11.11
704	**********	≥910	≥4710	2.68	13.80
707	**********	≥1050	≥5770	3.10	16.90
711	**********	≥1010	≥6780	2.97	19.87
714	**********	≥1110	≥7890	3.25	23.12
717	**********	≥1220	≥9120	3.58	26.71
720	**********	≥1240	≥10370	3.66	30.37
722	**********	≥1240	≥11610	3.66	34.02
725	**********	≥1210	≥12830	3.54	37.57
727	**********	≥1210	≥14040	3.56	41.14
730	**********	≥1200	≥15250	3.51	44.66
732	**********	≥1170	≥16420	3.43	48.10
735	**********	≥1140	≥17570	3.36	51.46
737	**********	≥1100	≥18670	3.22	54.68
739	**********	≥1130	≥19800	3.32	58.01
741	**********	≥1130	≥20940	3.31	61.32
743	**********	≥1020	≥21960	3.00	64.32
746	**********	≥990	≥22950	2.90	67.22
748	**********	≥1020	≥23980	3.00	70.23
750	**********	≥940	≥24930	2.78	73.01
752	**********	≥880	≥25810	2.58	75.60
754	**********	≥880	≥26700	2.60	78.20
756	**********	≥810	≥27520	2.39	80.59
758	**********	≥700	≥28230	2.07	82.67
760	**********	≥730	≥28960	2.16	84.83
762	**********	≥670	≥29640	1.97	86.80
764	**********	≥660	≥30300	1.95	88.75
766	**********	≥520	≥30830	1.54	90.30
768	**********	≥490	≥31330	1.46	91.75
770	**********	≥480	≥31820	1.43	93.18
772	**********	≥410	≥32230	1.22	94.40
774	**********	≥370	≥32610	1.09	95.49
776	**********	≥300	≥32910	0.90	96.40
779	**********	≥230	≥33150	0.69	97.09
781	**********	≥230	≥33380	0.67	97.76
783	**********	≥160	≥33550	0.49	98.26
785	**********	≥160	≥33710	0.47	98.74
787	**********	≥110	≥33830	0.34	99.08
790	**********	≥70	≥33910	0.23	99.32
792	**********	≥50	≥33970	0.16	99.48
795	**********	≥60	≥34030	0.18	99.66
798	**********	≥30	≥34070	0.11	99.77
801	**********	≥30	≥34100	0.10	99.87
804	**********	≥10	≥34120	0.04	99.92
814	**********	<10	≥34140	0.00	99.99
816	**********	<10	≥34140	0.00	99.99
817	**********	<10	≥34140	0.00	99.99
819	**********	<10	≥34140	0.00	99.99

100 200 300 400 500 600 700 800 900 1000 1100 1200
Frequency

Table E.3 Scale Score Descriptive Statistics and Plots

Biology
ALL STUDENTS
Form B

N	≥17700		
Mean	734.60	Median	737.00
Std deviation	26.42	Variance	697.93
Skewness	-0.2910	Kurtosis	0.0452
Mode	722.00	Std Error Mean	0.1985
Range	173.00	Interquartile Range	37.00

Quantile	Estimate
100% Max	823
99%	790
95%	776
90%	768
75% Q3	754
50% Median	737
25% Q1	717
10%	700
5%	691
1%	661
0% Min	650

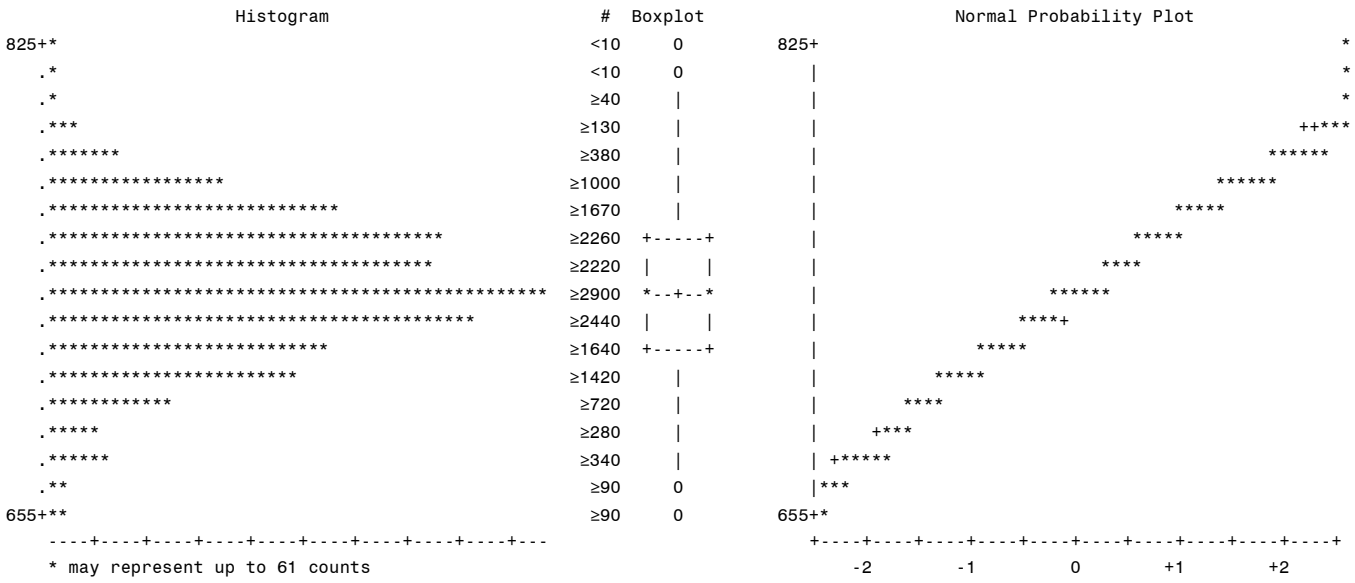


Table E.4 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
BIOLOGY
Form B

SCALE_SCORE		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥80	≥80	0.50	0.50
661	**********	≥90	≥180	0.55	1.06
672	**********	≥140	≥330	0.81	1.88
679	**********	≥200	≥530	1.15	3.04
686	**********	≥280	≥820	1.62	4.67
691	**********	≥310	≥1130	1.75	6.43
696	**********	≥400	≥1540	2.30	8.74
700	**********	≥440	≥1980	2.49	11.23
704	**********	≥460	≥2440	2.60	13.83
707	**********	≥520	≥2970	2.95	16.79
711	**********	≥500	≥3480	2.87	19.66
714	**********	≥550	≥4030	3.13	22.80
717	**********	≥570	≥4610	3.26	26.06
720	**********	≥600	≥5220	3.44	29.51
722	**********	≥620	≥5840	3.51	33.02
725	**********	≥610	≥6460	3.50	36.52
727	**********	≥590	≥7050	3.33	39.86
730	**********	≥580	≥7630	3.28	43.14
732	**********	≥590	≥8230	3.35	46.49
735	**********	≥580	≥8810	3.28	49.77
737	**********	≥570	≥9380	3.26	53.03
739	**********	≥560	≥9950	3.21	56.25
741	**********	≥610	≥10570	3.46	59.71
743	**********	≥530	≥11100	3.03	62.75
746	**********	≥540	≥11650	3.10	65.85
748	**********	≥520	≥12180	2.98	68.83
750	**********	≥490	≥12670	2.77	71.61
752	**********	≥460	≥13140	2.65	74.27
754	**********	≥470	≥13620	2.69	76.97
756	**********	≥430	≥14050	2.43	79.40
758	**********	≥390	≥14440	2.21	81.61
760	**********	≥410	≥14850	2.32	83.93
762	**********	≥340	≥15200	1.95	85.89
764	**********	≥360	≥15560	2.04	87.93
766	**********	≥280	≥15850	1.61	89.55
768	**********	≥260	≥16120	1.51	91.06
770	**********	≥260	≥16380	1.49	92.56
772	**********	≥240	≥16620	1.36	93.91
774	**********	≥190	≥16810	1.07	94.99
776	**********	≥170	≥16990	1.00	96.00
779	**********	≥130	≥17130	0.76	96.76
781	**********	≥120	≥17250	0.69	97.45
783	**********	≥100	≥17350	0.57	98.02
785	**********	≥90	≥17440	0.52	98.55
787	**********	≥60	≥17510	0.37	98.93
790	**********	≥40	≥17560	0.28	99.21
792	**********	≥30	≥17590	0.18	99.40
795	**********	≥30	≥17630	0.20	99.61
798	**********	≥10	≥17650	0.09	99.71
801	**********	≥20	≥17670	0.14	99.84
804	**********	<10	≥17680	0.05	99.90
808	**********	<10	≥17690	0.05	99.94

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570 600
Frequency

Table E.5 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
BIOLOGY
Form C

N	≥16440		
Mean	733.23	Median	735.00
Std deviation	25.32	Variance	640.91
Skewness	-0.2043	Kurtosis	0.0042
Mode	717.00	Std Error Mean	0.1974
Range	162.00	Interquartile Range	35.00

Quantile	Estimate
100% Max	812
99%	787
95%	774
90%	766
75% Q3	752
50% Median	735
25% Q1	717
10%	700
5%	691
1%	672
0% Min	650

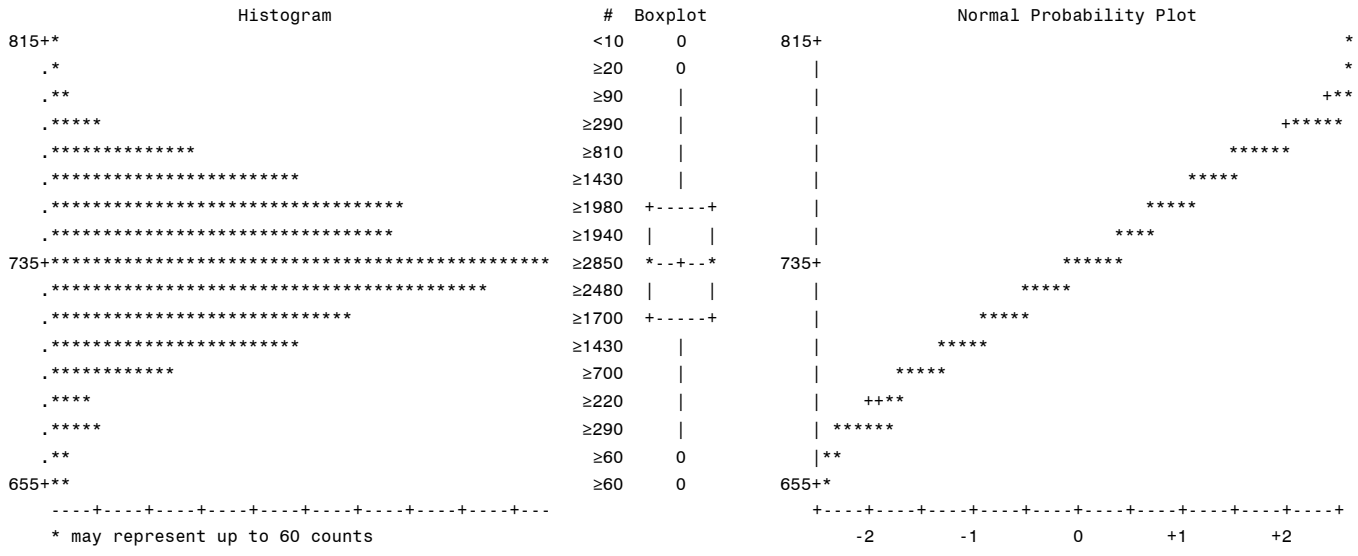


Table E.6 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
BIOLOGY
Form C

SCALE_SCORE		Freq	Cum. Freq	Cum. Percent	Cum. Percent
650	*********	≥60	≥60	0.42	0.42
661	*********	≥60	≥130	0.40	0.81
672	**********	≥100	≥240	0.66	1.47
679	**********	≥180	≥420	1.12	2.60
686	**********	≥220	≥650	1.39	3.99
691	**********	≥310	≥970	1.93	5.92
696	**********	≥380	≥1360	2.35	8.27
700	**********	≥440	≥1800	2.71	10.98
704	**********	≥450	≥2260	2.77	13.76
707	**********	≥530	≥2790	3.25	17.01
711	**********	≥500	≥3300	3.07	20.09
714	**********	≥550	≥3860	3.38	23.47
717	**********	≥640	≥4500	3.93	27.40
720	**********	≥640	≥5140	3.89	31.29
722	**********	≥620	≥5770	3.81	35.10
725	**********	≥590	≥6360	3.59	38.71
727	**********	≥620	≥6990	3.81	42.52
730	**********	≥610	≥7610	3.76	46.29
732	**********	≥580	≥8190	3.53	49.82
735	**********	≥560	≥8760	3.45	53.28
737	**********	≥520	≥9280	3.18	56.46
739	**********	≥560	≥9850	3.44	59.90
741	**********	≥510	≥10360	3.15	63.05
743	**********	≥480	≥10850	2.96	66.01
746	**********	≥440	≥11290	2.69	68.70
748	**********	≥490	≥11790	3.03	71.74
750	**********	≥450	≥12250	2.78	74.52
752	**********	≥410	≥12660	2.51	77.03
756	**********	≥380	≥13460	2.34	81.88
758	**********	≥310	≥13780	1.92	83.80
760	**********	≥320	≥14100	1.98	85.79
762	**********	≥320	≥14430	1.98	87.77
764	**********	≥300	≥14740	1.85	89.63
766	**********	≥240	≥14980	1.47	91.10
768	**********	≥220	≥15210	1.39	92.50
770	**********	≥220	≥15430	1.36	93.86
772	**********	≥170	≥15610	1.07	94.93
774	**********	≥180	≥15790	1.10	96.04
776	**********	≥120	≥15920	0.78	96.83
779	**********	≥100	≥16020	0.61	97.44
781	**********	≥100	≥16130	0.65	98.10
783	*********	≥60	≥16200	0.41	98.52
785	*********	≥60	≥16270	0.42	98.94
787	*******	≥50	≥16320	0.31	99.25
790	****	≥20	≥16350	0.18	99.43
792	***	≥20	≥16370	0.13	99.56
795	***	≥20	≥16390	0.15	99.71
801	*	≥10	≥16430	0.07	99.91
804	*	<10	≥16430	0.04	99.95
811		<10	≥16440	0.01	99.99

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570 600 630
Frequency

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

Biology

Contents
Table F.1 Reliability for Overall and Subgroups: Spring 2019 Operational Biology
Table F.2 Cronbach Alpha and Marginal Reliability: Spring 2019 Operational Biology
Table F.3.1–F.3.7 Classification Accuracy and Decision Consistency: Spring 2019 Operational Biology

Table F.1

Reliability for Overall and Subgroups: Spring 2019 Operational Biology

Subgroup	Form B	Form C
All Students	0.863	0.854
Female	0.855	0.846
Male	0.872	0.864
African American	0.825	0.818
American Indian or Alaska Native	0.871	0.847
Asian	0.878	0.853
Hispanic/Latino	0.878	0.861
Multi-Racial	0.853	0.861
Native Hawaiian or Other Pacific Islander	0.914	0.849
White	0.845	0.835
English Learners	0.818	0.814

Table F.2

Cronbach Alpha and Marginal Reliability: Spring 2019 Operational Biology

Form	Cronbach Alpha	Marginal Reliability
B	0.86	0.97
C	0.85	0.97

Table F.3
Classification Accuracy and Decision Consistency: Spring 2019 Operational Biology

Table F.3.1
Estimates of Accuracy and Consistency of Achievement-Level Classification by Form

Form	Accuracy	Consistency	PChance	Kappa
B	0.669	0.558	0.245	0.415
C	0.672	0.560	0.250	0.413

Table F.3.2
Accuracy of Classification at Each Achievement Level for Each Form

Form	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
B	0.808	0.592	0.690	0.625	0.682
C	0.788	0.599	0.685	0.645	0.717

Table F.3.3
Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.942	0.896	0.886	0.941
C	0.936	0.886	0.891	0.954

Table F.3.4
Consistency of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.915	0.854	0.840	0.919
C	0.907	0.841	0.847	0.936

Table F.3.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.663	0.672	0.630	0.377
C	0.632	0.652	0.624	0.413

Table F.3.6

Accuracy of Dichotomous Categorizations: False Positive Rates (PAC Metric)

Form	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.025	0.047	0.058	0.045
C	0.027	0.051	0.061	0.035

Table F.3.7

Accuracy of Dichotomous Categorizations: False Negative Rates (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.033	0.058	0.056	0.014
C	0.037	0.063	0.048	0.010