LEAP 2025 Biology
Technical Report: 2021–2022

Prepared by DRC, Pearson, and WestEd

EXECUTIVE SUMMARY

The Louisiana Educational Assessment Program 2025 (LEAP 2025) is composed of tests that are carefully constructed to fairly assess the achievement of Louisiana students. This technical report provides information on the operational test administrations, scoring activities, analyses, and results of the spring 2022 administration of the LEAP 2025 Biology test, which included both operational and field test items.

While 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 information is primarily 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, 2014).

The chapters of this technical report outline general information about the assessment framework, test development process, embedded test form construction, content and data review, administration and scoring activities of the LEAP 2025 test, CTT (Classical Test Theory) and IRT (Item Response Theory) analysis results, test results, demographic characteristics of students, interpretation of the scores on the tests, and reliability and validity. Additionally, because of conditions related to COVID-19, please use caution when making any inferences from the statistical results of the spring 2022 administration.

Table of Contents

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, statistical analyses, IRT (Item Response Theory) calibration, test results, reliability and validity, and reporting of scores.

Summary of the 2019–2022 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 2022 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 August 2019 to May 2022

Date	Activity
August–December 2019	• Started item development planning for spring 2021 test • Item development plans and outlines approved by LDOE • WestEd updated content development specifications and style guide • Technical Advisory Committee meeting convened
December 2019–July 2020	• WestEd began item writing and development • LDOE staff reviewed proposed item sets, tasks, and standalones
January–March 2020	WestEd updated 2020–2022 Framework and Test Construction Document based on LDOE comments and LDOE reviewed and approved • Technical Advisory Committee meeting convened
July 2020	• Item development put on hold due to the pandemic
August 2020	• Planning meeting held
February 2021	• Planning meeting held
March 2021	• WestEd and LDOE convened Item Content/Bias Review Committee • LDOE and WestEd staff held reconciliation meeting
April–July 2021	• Content finalized and LDOE approved • Online content delivered to administration vendor • Field test forms selected using operational base form previously selected for spring 2020 but never administered and field test selection is approved by LDOE
August 2021	• Virtual planning meeting held
October 2021	• LDOE staff reviewed proposed spring 2019 EFT selections in administration platform
November–December 2021	• Fall 2021 test administered
February 2022	• LDOE/WestEd/DRC met for planning meeting
April–May 2022	• Spring 2022 test administered, including EFT

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 sets, tasks, and standalone items to be developed and to be revised and refield tested. This section describes the processes used to develop new item sets, tasks, and standalone

items and used to revise existing item sets or tasks. Note that the test design specified that the test alternates by year between field testing item sets and tasks. Spring 2022 was designated as a "task" year for field testing, so only task and standalone development that was used on the spring 2022 field test is included in the table. Table 3.2 shows the item development plan for the number of items developed by WestEd.

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	0	0	0	0	0	0	0	0
Tasks	3	10	4	9	6	6	0	29
Standalone items	33	9	6	8	10	0	0	33

Table 3.3 shows the item development plan for the revised and refield test tasks that were used on the spring 2022 field test.

Table 3.3
Number of Items Revised for Biology Assessment for Item Sets and Tasks

	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	0	0	0	0	0	0	0	0
Tasks	4	11	4	8	9	8	0	32

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. In previous development, the LDOE and WestEd determined that some item sets and tasks would allow a "mix and match" approach in which the Science and Engineering Practice (SEP) for one of the PEs in a bundle could be used to develop items aligned to the disciplinary core idea (DCI) and crosscutting concept (CCC) of the other PE in the bundle. This approach was discontinued beginning with the current cycle because it generated some items with a SEP alignment outside the reporting category for the PE the item aligned to and therefore did not fit the reporting category. 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.

One additional bundle was proposed for the 2019–2022 cycle. Of the total of 31 bundles, 3 were targeted for task development in the 2019–2022 cycle.

Revise and Refield Test Tasks and Sets

In addition to new development, tasks and item sets that had items that did not perform well were flagged for revision and refield-testing. During the 2019–2022 cycle, four tasks were designated for revision and refield-testing.

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 Tasks and Foci to Standalone Items

As the test design called for item sets and tasks to be field tested in alternate years, tasks were targeted for development for the 2019–2022 development cycle. The narrowing of set types to tasks influenced the selection of phenomena. Like the item sets, the tasks are phenomena-based, but unlike the item sets, the items build upon each other, require a specific order, and contain a three-dimensional extended-response (ER) item.

For the tasks, WestEd offered a document containing descriptions of 7 phenomena associated with bundles to the LDOE for review prior to item development. Based on the list, the LDOE identified 3 phenomena to be developed into stimuli for the tasks. 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 tasks 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 (2019–2022).

To support the outline development process, writers were given the Louisiana Student Standards for Science (LSSS). 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 determined to be above grade level by the software, but on grade level according to the LSSS, 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 2019 and January 2020. For an outline of the information covered, see Appendix A for the LEAP 2025 Biology Item Training Agenda (2019–2022). 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 (2019–2022). 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 virtual content and bias review meetings, held using Zoom. 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 2021–2022 Content and Bias Reviews

Characteristic	Number of Participants
Classroom Teacher	6
Content/Curriculum Specialist	0
School Administrator	0
Instructional Lead or Supervisor	1
ELL Teacher	0
Special Education Teacher	0
Special Ed Teacher – Gifted	0
Visually or Hearing-Impaired Teacher	1
Black or African American	1
Asian	0
Hispanic/Latino	0
White	7
Male	1
Female	7
Total Participants	8

Prior to joining the virtual review, committee members were required to watch a prerecorded training, including content training on how to evaluate the items as part of the committee review process. Participants were also provided meeting materials in advance of the review. At the start of the virtual committee, 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 to access ABBI and only had access to ABBI during meeting times. Participants were locked out of ABBI when the meeting was not in progress. 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 required cameras to be on at all times during the meeting in order to monitor participants to be sure that participants were in a secure space and that they did not use their cell phones. Content security was stressed in the prerecorded training, during the meeting introduction, throughout the meetings, at the end of each day, and at the conclusion of each meeting.

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	Votes to Reject	No Vote	Total Votes
CR	6	43	2	0	0	45
ER	6	34	12	0	0	46
MC	29	186	34	0	0	220
MS	6	40	6	0	0	46
TE	54	337	72	0	1	410
TPD	11	76	9	0	0	85
TPI	9	62	7	0	0	69
Stimulus	9	48	18	0	1	67
All Biology	**130**	**826**	**160**	**0**	**2**	**988**

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.

Data Review Process and Results

During data review of the spring 2019 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/refield test"), or reject were then recorded for each item. Table 3.5 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. Appendix A has comprehensive information on data review training.

Table 3.5

Summary of Data Review Votes

Item Type		Number of Items		
	Accept	Accept with Edits	Reject	% of Total
CR	0	0	0	0%
ER	8	3	3*	13%
MC	18	4	7	28%
MS	0	0	0	0%
TE	27	9	2	36%
TPI	6	1	3	10%
TPD	7	4	3	13%
Total	66	21	18	0%

* These items were rejected at rangefinding.

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.

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
Total Operational Items Tested Across Forms for Biology Spring 2022	7 OP standalone SR items 6 OP standalone TE items 3 OP standalone TPD/TPI items 1 OP task set SR items 2 OP task set TE items 1 OP task set TPD item 1 OP task set ER items 6 OP item set SR items 9 OP item set TE items 1 OP item set TPD/TPI items 3 OP item set CR items
Total Items Field Tested Across Forms for Biology Spring 2022	8 FT standalone SR items 14 FT standalone TE items 9 FT standalone TPD/TPI items

Initial Construction

The purpose of the spring 2022 forms construction activities was to create an operational form using the spring 2018 and spring 2019 items that were approved for operational use and to embed field test items in the spring 2022 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 2019 embedded field test were available for use on the spring 2022 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 for the spring 2022 administration.

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 2022 forms. The spring 2022 administrative error form included placeholder items. Table 4.2 provides the operational test composition for Biology for spring 2022.

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

Sixteen embedded field test forms were administered in spring 2022 for Biology. This number is greater than the number of tasks available for field testing.

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

Because fewer standalone items were developed than positions were available across the 16 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 in the 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 2021 window 1, Form A was the operational base form and Form B was used as the administrative error form. For fall 2021 window 2, Form B was the operational base form and Form A was used as the administrative error form. Both forms contained item set and standalone placeholders. For spring 2022, Form D was administered as the operational base form. Sixteen field test versions of Form D were administered. Form B was used as the administrative error form. For summer 2022, 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 the 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. Detailed information can be found at Appendix G, Accommodated Print and Braille Creation.

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 spring 2022 were used to construct the spring 2022 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. Detailed information can be found at Appendix G, Accommodated Print and Braille Creation.

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 policies, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE provides train-the-trainer opportunities for district test coordinators, who in turn convey test administration training to schools within their school systems. The LDOE conducts quality-assurance visits during testing to ensure school system adherence to the standardized administration of the tests.

The district test coordinators are responsible for the schools within their school system. 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 each test administration, Data Recognition Corporation (DRC) produces an administration manual, the LEAP 2025 High School Test Administration Manual (TAM). The TAM provides detailed instructions for administering the LEAP 2025 HS assessments. The manual includes information on test security, test administrator responsibilities, test preparation, administration of online tests, and post-test procedures.

Test Administration Manual

Table of Contents

14.2. Returning Test Materials to the School Test Coordinator

15. Index

DRC also produces a Test Coordinator Manual (TCM). The TCM provides detailed instructions for district and school test coordinators' responsibilities for distributing, collecting, and returning test materials.

Test Coordinator Manual

Table of Contents

10. Managing Test Sessions and 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. DRC INSIGHT Portal User Guide
    11.3. LEAP 2025 Accommodations and Accessibility Manual
    11.4. DRC INSIGHT Technology User Guide
    11.5. Student Tutorials
    11.6. Online Tools Training (OTT)
12. Post-Administration Rescoring Process for LEAP 2025 HS Assessments
13. Request for Rescoring
14. Void Notification

LDOE assessment staff review, provide feedback, and give final approval for the manuals. The manuals are inclusive of LEAP 2025 HS assessments in English Language Arts (ELA), Mathematics, Social Studies, and Science.

The Standards contain multiple references relevant to test administration. Information in the TAM addresses these in the following manner.

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

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

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

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 LEAP 2025 tests were administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The TCM included instructions for scheduling the test within the state testing window. The TAM and TCM 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 that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; allowing students to correct or complete answers after tests have been submitted; splitting sessions into two parts; ignoring the standardized directions for the assessment; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the Individualized Education Program (IEP), Individual Accommodation Plan (IAP), or EL Checklist; allowing accommodations for students who do not have an IEP, IAP, or EL Checklist; or defining terms on the test.

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

The TAM outlines the steps that teachers should take to prepare the classroom testing environment 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 thesharing 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 thecontent 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 the student 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 manuals.

Time

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

Online Forms Administration

The online forms are administered via DRC’s INSIGHT online assessment system. School system and school personnel set up test sessions via DRC’s online testing portal, DRC INSIGHT Portal (eDIRECT), and print test tickets. Students enter their ticket information to access the test in INSIGHT. In addition, students have access to Online Tools Training before the testing window, which allows them to practice using tools and features within INSIGHT. Tutorials with online video clips that demonstrate features of the system are also available to students before testing.

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 ateam 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 English Learners (ELs).

Accommodations must be specified in the qualifying student's individual plan and must beconsistent 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 TAM contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP 2025 assessments. The following accommodations were provided by DRC for this administration:

- Braille
- Text-to-Speech
- Directions in Native Language

The following additional access and accommodation features were also available:

- Answers Recorded
- Extended Time
- Transferred Answers
- Individual/Small Group Administration
- 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 Accommodationsand Accessibility Features User Guide*.

Testing Windows

The 2021–2022 assessments were administered to students within the state testing windows of November 30–December 17, 2021, or January 5–24, 2022; April 14–May 13, 2022; and June 20–24, 2022.

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

Data Forensic Analyses

Due to the importance of the LEAP 2025 HS assessments, it is prudent to confirm that the results from the assessments are based on true student achievement. To help ensure that scores are related to actual learning and that results are valid, data forensic analyses take place to assist in separating meaningful gains from spurious gains. It is important to note that although the results of the analyses may be used to identify potential problems within a school, the identification of a problem is not an accusation of misconduct.

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

- Response Change Analysis
- Score Fluctuation Analysis

- Web Monitoring
- Plagiarism Detection

Response Change Analysis. Students make changes to answer choices when taking the LEAP 2025 HS assessments, and this behavior is expected. Unfortunately, changes to student answers are sometimes influenced by school personnel who want to improve performance. Therefore, the response change analysis is conducted to identify school- and test administrator-level response change patterns that are statistically improbable when compared to the expected pattern at the state level.

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

Web Monitoring. The content of the LEAP 2025 assessments should not appear outside the boundaries of the forms administered. To protect Louisiana test content, the internet is monitored for postings that contain, or appear to contain, potentially exposed and/or copied test content. When test content is verified, steps are taken to quickly remove the infringing content.

Plagiarism Detection. The LDOE monitors for two different plagiarism situations: copying from student to student and copying from an outside source, such as Wikipedia or other internet sources. Instances of possible plagiarism are identified by human scorers. Alerts are set to identify responses that indicate the possibility of teacher interference or plagiarism. Alerted responses are given additional review so that the appropriate action can be taken.

Alerts for Disturbing Content

Scorers for the LEAP 2025 HS assessments also have the ability to apply an alert flag to student responses that may indicate disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). All alerted responses are automatically routed to the scoring director, who reviews and forwards appropriate responses to senior project staff for review. If it is concluded that a response warrants an alert, project management will contact the LDOE to take the necessary action. At no point during this process do scorers or staff have access to demographic information for any students participating in the assessment.

6. Scoring Activities

Directory of Test Specifications (DOTS) process. DRC creates 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 reviews and confirms the contents of the DOTS file as part of test review rounds. The DOTS file is then provided to the LDOE for multiple rounds of review, then final approval. Once approved, the information contained in the DOTS is used in scoring the test and in reporting.

Selected-Response (SR) Item Keycheck. SR items for Biology include multiple-choice (MC) and multiple-select (MS) questions. Pearson calculates MC and MS item statistics and flags items if item statistics fall outside expected ranges. For example, items are flagged if few students select the correct response (p-value less than 0.15), if the item does not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) select a certain incorrect response.

Lists of flagged MC and MS items, with the reasons for flagging, are provided to LDOE and WestEd content staff for key verification. The staff reviews the list of flagged MC and MS items to confirm that the answer keys are accurate. Scoring of MC and MS items is also evaluated at data review.

Scoring of Technology-Enhanced (TE) Items. All TE items are processed through DRC's autoscoring engine and scored according to the assigned scoring rules established duringcontent creation by WestEd in conjunction with the LDOE. DRC ensures that all rubrics and scoring rules are verified for accuracy before scoring any TE items. DRC has an established adjudication process for TE items to verify that correct answers are identified. DRC's TE scoring process includes the following procedures:

- A scoring rubric is created for each TE item. The rubric describes the one and only correct answer for dichotomously scored items (i.e.,

items scored as either right or wrong). If partial credit is possible, the rubric describes in detail the type of response that could receive credit for each score point.

- The information from the scoring rubric is entered into the scoring system within the item banking system so that the truth resides in one place along with the item image and other metadata. This scoring information designates specific information that varies by item type. For example, for a drag-and-drop item, the information includes which objects are to be placed in each drop region to receive credit.
- The information is then verified by another autoscoring expert.
- After testing starts, reports are generated that show every response, how many students gave that response, and the score the scoring system provided for that response.
- The scoring is then checked against the scoring rubric using two levels of verification.
- If any discrepancies are found, the scoring information is modified and verified again. The scoring process is then rerun. This checking and modification processcontinues until no other issues are found.
- As a final check, a final report is generated that shows all student responses,their frequencies, and their received scores.

In the case of braille test forms, student responses to items are transcribed into the online system by a test administrator.

Adjudication. TE items and other eligible items identified in the test map are automatically scored as tests are processed. TE items are 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 are scored

appropriately. In the event that scoring issues are identified, WestEd content staff and the LDOE committee review and 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-Response and Extended-Response Scoring

Constructed- and extended-response items are scored by human raters trained by DRC. Ten percent of the responses are scored twice to monitor and maintain inter-rater reliability. Scoring supervisors also conduct read-behinds and review all nonscores and alerts. Handscoring processing rules are detailed in the LEAP 2025 Spring 2022 Handscoring/AI Documentation document.

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 are selected and trained for the LEAP 2025 Biology assessment and monitored throughout the handscoring process.

Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2021–2022 LEAP 2025 HS Biology test responses have 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, and 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 does not demonstrate the ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

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

Readers only have access to student responses they are qualified to score. Each scorer is assigned a unique username and password to access DRC's imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

Each DRC scoring center is a secure facility. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials may leave the scoring center. To prevent the unauthorized duplication of secure materials, cell phone/camera use within the scoring rooms is strictly forbidden. Readers only have access to student responses they are qualified to score.

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

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

- Prompts and associated sources
- 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. The LDOE reviews training materials and oversees the training process.

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 perscore point	Yes
Anchor set (9-point ERs)	Item-specific anchor sets containing two responses perscore point	Yes
Training sets	Two training sets for each CR item and three training setsfor 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 qualifyingsets 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 demonstrate their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with true scores on qualifying sets). After each qualifying set is scored, the DRC scoring director responsible for training leads the scorers in a discussion of the set.

Any scorer who does not qualify by the end of the qualifying process for an item is 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 thatinvolves 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 group 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 arethe same, divided by the number of responses that were scored twice
- Percentage Adjacent (%AD)—total number of responses by reader where scoresare 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 point apart, divided by the number of responses thatwere 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 with the entire group, 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 scores 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 2021–2022 forms.

Table 6.4

Inter-Rater Reliability for Operational Constructed-Response Items

Admin.	Item	Inter-Rater Reliability*				
		2x	Total	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2021 Window 1	Item 1	≥4,830	≥12,970	96	4	0
	Item 2	≥5,170	≥12,660	98	2	0
	Item 3	≥4,480	≥12,440	97	3	0
Fall 2021 Window 2	Item 1	≥1,260	≥4,130	98	2	0
	Item 2	≥1,810	≥4,390	98	2	0
	Item 3	≥1,640	≥4,320	98	2	0
Spring 2022	Item 1	≥13,240	≥48,020	91	8	1
	Item 2	≥11,830	≥47,480	95	5	0
	Item 3	≥12,420	≥47,610	93	6	0
Spring 2022 (Seniors)	Item 1	≥1,030	≥2,710	97	2	1
	Item 2	≥1,370	≥2,820	98	2	0
	Item 3	≥1,140	≥2,830	97	3	0
Summer 2022	Item 1	≥1,820	≥4,300	98	2	0
	Item 2	≥2,140	≥4,290	98	2	0
	Item 3	≥1,810	≥4,150	100	0	0

***** The percent may not add up to 100% due to rounding.

Table 6.5

Score Point Distributions for Operational Constructed-Response Items

Administration	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
Fall 2021 Window 1	Item 1	≥12,970	53	17	8	0	22
	Item 2	≥12,660	59	13	1	0	26
	Item 3	≥12,440	64	9	6	0	20
Fall 2021 Window 2	Item 1	≥4,130	79	5	2	0	13
	Item 2	≥4,390	47	20	10	0	23
	Item 3	≥4,320	54	23	4	0	18
Spring 2022	Item 1	≥48,020	57	21	9	0	12
	Item 2	≥47,480	67	17	6	0	9
	Item 3	≥47,610	65	20	4	0	10
Spring 2022 (Seniors)	Item 1	≥2,710	67	5	2	0	25
	Item 2	≥2,820	52	9	0	1	38
	Item 3	≥2,830	54	15	3	0	27
Summer 2022	Item 1	≥4,300	58	10	1	0	30
	Item 2	≥4,290	51	7	0	1	41
	Item 3	≥4,150	63	3	1	0	32

* The percent may not add up to 100% due to rounding.

** Nonscore codes include Foreign language (F), Insufficient (I), Don't Understand (N), Refusal (R), Off Topic (T), and Unintelligible (U). Responses that cannot be assigned a score based on the rubric are assigned a nonscore code and count as zero points toward student scores.

Table 6.6

Inter-Rater Reliability for Operational Extended-Response Items

Admin.	Item	Inter-Rater Reliability*					
		2X	Total	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2021 Window 1	Item 1	≥3,990	≥12,230	Part A (0–6)	94	6	1
				Part B (0–3)	94	5	1
Fall 2021 Window 2	Item 1	≥1,500	≥4,250	Part A (0–3)	94	6	0
				Part B (0–6)	94	5	1
Spring 2022	Item 1	≥12,110	≥47,540	Part A (0–3)	88	11	1
				Part B (0–6)	82	14	4
Spring 2022 (Seniors)	Item 1	≥850	≥2,620	Part A (0–6)	96	4	1
				Part B (0–3)	95	4	1
Summer 2022	Item 1	≥1,520	≥4,040	Part A (0–6)	98	1	0
				Part B (0–3)	98	2	0

* The percent may not add up to 100% due to rounding.

Table 6.7

Score Point Distributions for Operational Extended-Response Items

Admin.	Item	Total	Score Point Distribution*								
			Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	"6" (%)	Blank (%)	Nonscore Codes (%)**
Fall 2021 Window 1	Item 1	≥12,230	Part A (0–6)	51	12	9	11	1	0	1	0	14
			Part B (0–3)	36	34	8	7	N/A	N/A	N/A	0	14
Fall 2021 Window 2	Item 1	≥4,250	Part A (0–3)	11	50	17	7	N/A	N/A	N/A	0	14
			Part B (0–6)	31	23	16	7	5	2	2	0	14
Spring 2022	Item 1	≥47,540	Part A (0–3)	9	53	21	8	N/A	N/A	N/A	0	9
			Part B (0–6)	29	24	19	9	6	2	2	0	9
Spring 2022 (Seniors)	Item 1	≥2,620	Part A (0–6)	57	13	8	4	0	0	0	0	17
			Part B (0–3)	41	33	6	2	N/A	N/A	N/A	0	17
Summer 2022	Item 1	≥4,040	Part A (0–6)	58	11	5	3	0	0	0	0	23
			Part B (0–3)	39	34	3	1	N/A	N/A	N/A	0	23

* The percent may not add up to 100% due to rounding.

* * Nonscore codes include Foreign language (F), Insufficient (I), Don't Understand (N), Refusal (R), Off Topic (T), and Unintelligible (U). Responses that cannot be assigned a score based on the rubric are assigned a nonscore code and count as zero points toward student scores.

Table 6.8

Inter-Rater Reliability for Spring 2022Field Test Extended-Response Items

Item	Inter-Rater Reliability*					
	2X	Total	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Item 1	≥500	≥2,280	Part A (0–3)	80	16	4
			Part B (0–3)	73	19	8
			Part C (0–3)	89	8	4
Item 2	≥500	≥2,310	Part A (0–4)	85	9	6
			Part B (0–5)	81	13	6
Item 3	≥540	≥2,290	Part A (0–4)	89	5	6
			Part B (0–3)	89	6	5
			Part C (0–2)	94	6	0
Item 4	≥580	≥2,290	Part A (0–2)	86	13	1
			Part B (0–3)	78	13	8
			Part C (0–4)	83	12	4
Item 5	≥540	≥2,280	Part A (0–3)	82	11	7
			Part B (0–3)	86	11	3
			Part C (0–3)	92	5	2
Item 6	≥540	≥2,280	Part A (0–3)	93	7	0
			Part B (0–3)	92	6	2
			Part C (0–3)	93	6	1
Item 7	≥580	≥2,290	Part A (0–2)	89	10	1
			Part B (0–3)	88	10	2
			Part C (0–4)	90	8	3

Table 6.8 (continued)

Item	Inter-Rater Reliability*					
	2X	Total	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Item 8	≥660	≥2,280	Part A (0–3)	94	4	3
			Part B (0–3)	97	2	1
			Part C (0–3)	95	2	3
Item 9	≥590	≥2,310	Part A (0–2)	94	5	1
			Part B (0–3)	96	2	1
			Part C (0–4)	97	2	0
Item 10	≥560	≥2,360	Part A (0–2)	86	13	2
			Part B (0–5)	89	7	4
			Part C (0–2)	90	9	1
Item 11	≥600	≥2,370	Part A (0–2)	86	13	1
			Part B (0–3)	84	12	4
			Part C (0–4)	92	5	3

* The percent may not add up to 100% due to rounding.

Table 6.9

Score Point Distributions for Spring 2022 Field Test Extended-Response Items

Item	Total	Score Point Distribution*								
		Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	Blank (%)	Nonscore Codes (%)**
Item 1	≥2,280	Part A (0–3)	57	15	21	3	N/A	N/A	0	4
		Part B (0–3)	60	18	12	5	N/A	N/A	0	4
		Part C (0–3)	81	11	3	1	N/A	N/A	0	4
Item 2	≥2,310	Part A (0–4)	33	27	9	18	9	N/A	0	5
		Part B (0–5)	44	22	20	6	2	2	0	5
Item 3	≥2,290	Part A (0–4)	48	6	27	5	5	N/A	0	8
		Part B (0–3)	67	10	12	2	N/A	N/A	0	8
		Part C (0–2)	53	32	6	N/A	N/A	N/A	0	8
Item 4	≥2,290	Part A (0–2)	37	40	15	N/A	N/A	N/A	0	7
		Part B (0–3)	53	19	13	8	N/A	N/A	0	7
		Part C (0–4)	67	12	10	2	2	N/A	0	7
Item 5	≥2,280	Part A (0–3)	48	11	27	8	N/A	N/A	0	6
		Part B (0–3)	59	14	15	6	N/A	N/A	0	6
		Part C (0–3)	81	9	3	1	N/A	N/A	0	6
Item 6	≥2,280	Part A (0–3)	23	43	25	4	N/A	N/A	0	5
		Part B (0–3)	85	8	2	0	N/A	N/A	0	5
		Part C (0–3)	83	10	1	0	N/A	N/A	0	5

Table 6.9 (continued)

Item	Total	Score Point Distribution*								
		Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	Blank (%)	Nonscore Codes (%)**
Item 7	≥2,290	Part A (0–2)	64	18	11	N/A	N/A	N/A	0	8
		Part B (0–3)	61	18	11	2	N/A	N/A	0	8
		Part C (0–4)	77	7	5	1	2	N/A	0	8
Item 8	≥2,280	Part A (0–3)	72	6	6	6	N/A	N/A	1	12
		Part B (0–3)	80	3	3	1	N/A	N/A	1	12
		Part C (0–4)	80	3	2	2	N/A	N/A	1	12
Item 9	≥2,310	Part A (0–2)	64	17	10	N/A	N/A	N/A	0	9
		Part B (0–3)	86	3	2	1	N/A	N/A	0	9
		Part C (0–4)	87	2	1	0	1	N/A	0	9
Item 10	≥2,360	Part A (0–2)	67	19	6	N/A	N/A	N/A	1	7
		Part B (0–5)	74	10	5	2	0	1	1	7
		Part C (0–2)	60	22	11	N/A	N/A	N/A	1	7
Item 11	≥2,370	Part A (0–2)	57	27	11	N/A	N/A	N/A	0	4
		Part B (0–3)	56	27	9	3	N/A	N/A	0	4
		Part C (0–4)	86	2	4	1	1	N/A	0	4

* The percent may not add up to 100% due to rounding.

** Nonscore codes include Foreign language (F), Insufficient (I), Don't Understand (N), Refusal (R),Off Topic (T), and Unintelligible (U). Responses that cannot be assigned a score based on the rubricare assigned a nonscore code and count as zero points toward student scores.

7. Data Analysis

Classical Item Statistics

This section describes the classical item analysis for data obtained from the operational LEAP 2025 HS Biology. The classical analysis includes statistical analysis based on the following types of items: multiple-choice/multiple-select items, rule-based machine-scored items such as technology-enhanced items, and handscored items such as constructed- and extended-response items. For each operational item, the statistical analysis produces item difficulty (p-value) and item discrimination (point-biserial).

Tables and figures that provide the additional information on classical item statistics for the spring 2022 test can be found in Appendix C: Item Analysis Summary Report. Tables C.1–C.5 show the summaries of classical item statistics. As 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). A measure of point-biserial correlation indicates 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. It should be noted that statistical analysis results for field-test (FT) items are stored in Pearson's Assessment Banking and Building solutions for Interoperable assessment (ABBI) system.

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 different 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×(*T*+1)×*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 | *SMD/SD* | ≤ 0.17
B (slight to moderate)	Mantel χ^2 *p*-value < 0.05 and 0.17< | *SMD/SD* | < 0.25
C (moderate to large)	Mantel χ^2 *p*-value < 0.05 and | *SMD/SD* | ≥ 0.25

Three DIF analyses were conducted for the operational test items only: 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 both black/white and Hispanic/white students performed unexpectedly well or unexpectedly poorly, given their performance on the full

assessment. The last two columns of Table 7.3 provide the number of items flagged for DIF. Items flagged with A-DIF show negligible 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 operational test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the *MH* statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 *p*-value and *SMD* statistics. Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table 7.3
Summary of DIF Flags: Spring 2022 Biology Operational Items

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	39	[1],[0]	[0],[1]
African American – White	41	[0],[0]	[0],[0]
Hispanic – White	40	[0],[0]	[0],[1]

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 (i.e., one-point) 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. Since the Biology test also included polytomous items scored higher than 1 point, the generalized partial credit model (GPCM) (Muraki, 1992) was used to estimate the parameters of these items:

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

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.).

Calibration and Linking

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.

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.

Thus, the primary purpose of form equating is to establish score equivalency between two (or more) forms. Equivalency is established by first building the forms to be equated according to content specifications. Then the form scores are placed on the same scale (by equating), such that students performing on two scaled assessments at the same level of underlying achievement should receive the same scale score on both forms, although they may not receive the same number-correct score (or raw score). LDOE and Pearson

strive to maintain equivalent samples or use near-census samples over the years, minimizing the potential differences caused by the different samples.

It should be noted that the spring 2021 is the first operational administration for Biology, and in the spring of 2021, the forms used were intact and when originally administered in 2019, they were post- equated and linked to the LEAP 2025 scale.

Table 7.4 provides scale scores at selected percentiles that can be used to compare the distributional characteristics of the spring 2022 test form to previous administrations. Although these scale scores are rounded values, there were differences in the scale score values for a given percentile across the forms. These variations could arise for several reasons: (1) differences in the proficiency (i.e., achievement) of the students in the samples or growth in student achievement across years; (2) unevenness in the respective distributions that combine with the number-correct-to-scale- score scoring method, leaving “gaps” in the scale; or (3) other sources of equating error. In general, however, the test characteristic function equating techniques will “level” the equated forms through the raw-to-scale- score adjustment.

Table 7.4

Comparisons of Scale Scores at Selected Percentiles: Biology Operational Forms

Percentile	2019 Spring Form B	2019 Spring Form C	2021 Spring Form B	2022 Spring Form D
99	790	787	787	787
95	776	774	773	772
90	768	766	765	765
85	762	760	759	760
80	758	756	753	755
75	754	752	748	751
70	750	748	744	747
65	746	743	740	744
60	743	741	738	740
55	739	737	733	736
50	737	735	729	734
45	732	730	726	730
40	730	727	722	726
35	725	722	719	723
30	722	720	714	719
25	717	717	711	714
20	714	711	706	709
15	707	707	702	706
10	700	700	695	700
5	691	691	687	688
1	661	672	666	670

Operational Item Parameters

The distributions of IRT item parameters are summarized in Appendix C. Appendix C also provides graphical displays of the distributions of IRT parameter estimates. 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. By the way, 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.

It should be noted that Yen's Q_1 statistic (Yen, 1981) was calculated to evaluate item fit for both operational and 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 $ZQ_{1,crit}$ 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 for operational items 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 kth 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 Biology. 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 the 2022 Biology test 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.

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

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

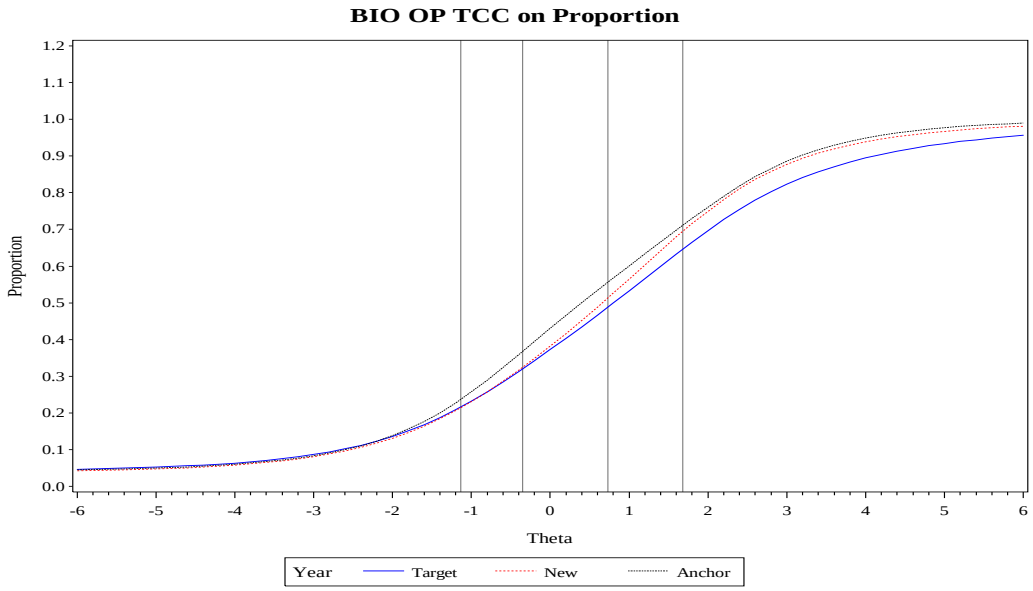
The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same 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.

Test Characteristic Curve

Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) across administrations of the LEAP 2025 assessments, as can be seen in the following figure. As seen from Plot 7.1, the TCCs between two years were similar across ability ranges. By the way, Plot 9.1 also indicates that the SEMs between two years are similar across ability ranges, especially in the middle ability ranges; each theta cut matches the scale score of each performance-level cut (i.e., 707, 725, 750, and 772).

Plot 7.1

Test Characteristic Curve: SPR 2022 Operational Biology



Note. The scale is on theta; Each theta cut matches the scale score of each performance cut: 707, 725, 750, and 772; Target = 2019 OP form; New = 2022 OP form; Anchor = Anchor Pool.

Test Information Curve, Score Distribution, and IRT Difficulty Distribution

In this section, student's biology score distribution, IRT item difficulty (i.e., b-parameter) distribution, and item information curve are presented. Compared to the base year (i.e., 2019 Biology test), the 2022 Biology test provides more test information around the middle range of theta than other rages, as can be observed from Table 7.5 and Plot 7.2.

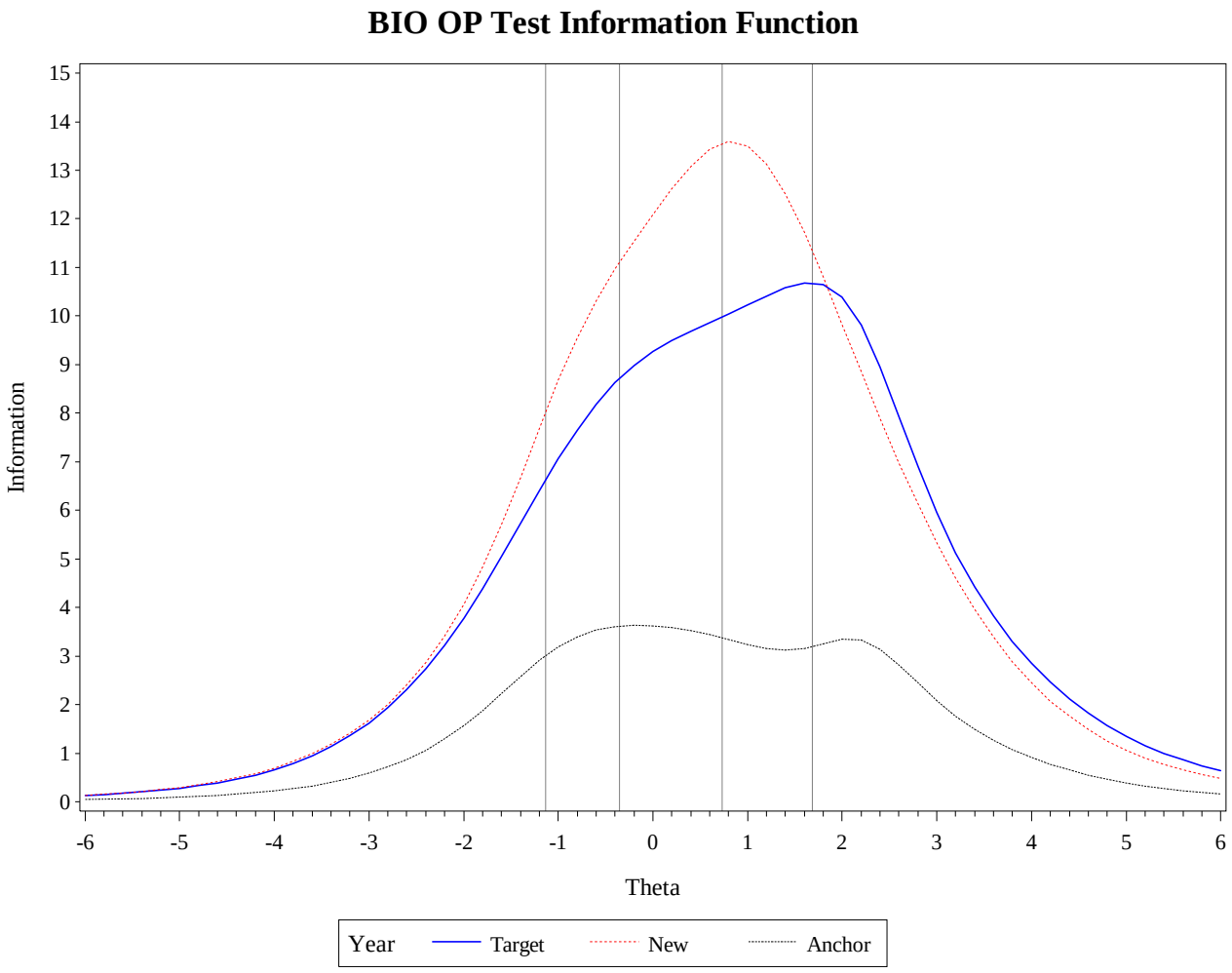
Table 7.5

Student's Score and IRT B-Parameter Distribution: Spring 2022 Operational Biology

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.19	theta < -3.5	0
0.27	-3.5 <= theta < -3.0	0
0.66	-3.0 <= theta < -2.5	0
2.23	-2.5 <= theta < -2.0	0
6.14	-2.0 <= theta < -1.5	1
10.68	-1.5 <= theta < -1.0	1
14.34	-1.0 <= theta < -0.5	4
14.86	-0.5 <= theta < 0.0	3
16.86	0.0 <= theta < 0.5	8
13.88	0.5 <= theta < 1.0	7
11.63	1.0 <= theta < 1.5	5
5.72	1.5 <= theta < 2.0	4
1.95	2.0 <= theta < 2.5	5
0.44	2.5 <= theta < 3.0	1
0.13	3.0 <= theta < 3.5	2
0.03	3.5 <= theta	0
-6.00	**Minimum**	-1.77
5.14	**Maximum**	3.47
-0.02	**Mean**	0.82
1.12	**SD**	1.18
≥38,820	**Number of Examinees**	41

Plot 7.2

Test Information Curve; SPR 2022 Operational Biology



Note. The scale is on theta; Each theta cut matches the scale score of each performance cut: 707, 725, 750, and 772; Target = 2019 OP form; New = 2022 OP form; Anchor = Anchor Pool.

Field Test Data Review

The process used to complete the field test item equating is an anchored item equating process. In this process the item parameters from the operational items from the 2019 administration were fixed as constant (i.e., to calculate Stocking-Lord equating constant) and the item parameters for the field test items were freely calibrated, placing the item parameters for the field test items on the same scale as the operational items.

As mentioned previously, field test items are reviewed at the data review meeting for all the same criteria as outlined previously. 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) based on CTT and IRT, appropriate interpretations and inferences, what would be considered reasonable values, and how the values might differ across item types. The result of such reviews is to determine if items are eligible to be placed in the item bank for future test construction or if items need to be updated and field tested again. It should be noted that all the results of SPR 2022 data review are saved in Pearson ABBI. It should be noted that the training presentation agenda for data evaluation is included in Appendix A: Training Agendas.

8. Test Results and Score Reports

This chapter provides information on the results of the spring LEAP Biology test. The scale score results and achievement level information are also presented here. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative description of student achievement. The levels are Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory. The results in Table 8.1 are presented as evidence of the reliability and validity of the scores from the LEAP 2025 Biology assessment.

Demographic Characteristics of Students

The operational Biology assessment was administered to all eligible students in the appropriate grade level during the spring 2022 first time administration window. Spring 2022 operational score results were reviewed based on the following student characteristics:

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

Test Results

For the spring 2022 Biology test, the lowest obtainable scale score (LOSS) on the tests is 650 and the highest obtainable scale score (HOSS) is 850. 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 8.1, scale score frequency distributions are presented in Appendix E: Scale Distribution and Statistical Report. Finally, because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table 8.1

LEAP 2025 State Test Results: Spring 2022 Operational Biology

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥38,820	732.65	25.64	18	20	36	20	6
Gender								
Female	≥19,470	733.27	24.54	15	21	38	20	5
Male	≥19,350	732.04	26.70	20	19	34	20	6
Ethnicity								
African American	≥16,000	721.56	23.21	28	27	33	10	2
American Indian or Alaska Native	≥260	735.63	22.33	10	22	41	23	5
Asian	≥690	752.48	25.47	6	7	27	38	22
Hispanic/Latino	≥2,940	727.36	27.41	25	20	33	17	4
Multi-Racial	≥1,060	736.98	24.32	13	17	39	24	7
Native Hawaiian or Other Pacific Islander	≥30	744.54	24.54	6	6	54	23	11
White	≥17,800	742.40	22.99	7	14	40	29	9
Economically Disadvantaged*								
No	≥14,090	744.24	23.42	7	13	38	31	11
Yes	≥23,320	726.39	24.46	23	24	35	15	3

Table 8.1 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥37,670	733.45	25.33	16	20	37	21	6
Yes	≥1,140	706.64	21.99	54	25	19	2	NR
Education Classification								
Gifted/Talented	≥2,290	758.75	22.08	2	5	24	40	29
Regular	≥33,230	733.18	24.07	15	21	39	21	5
Special	≥3,290	709.21	23.21	52	24	18	5	1
Section 504								
No	≥35,240	733.45	25.58	17	19	37	21	6
Yes	≥3,570	724.78	24.89	25	26	32	12	4
Migrant								
No	≥38,760	732.67	25.64	18	20	36	20	6
Yes	≥60	722.81	24.33	28	19	38	16	
Homeless Status								
No	≥38,140	732.80	25.65	17	20	36	21	6
Yes	≥680	724.33	24.16	25	27	32	14	2
Military Affiliation								
No	≥38,300	732.51	25.64	18	20	36	20	6
Yes	≥520	743.19	23.93	7	15	37	30	11
Foster Care Status								
No	≥38,750	732.68	25.64	17	20	36	20	6
Yes	≥70	719.81	23.72	33	22	38	4	3

* Economic status was not available for all students.

Effect Size

One way to evaluate the magnitude of the standardized mean difference (SMD) is to calculate the ES. Cohen's *d* was used to calculate the ES and is given by the following formula:

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

where is x_a is the mean score of group A, x_b is the mean score of group B, $s_a{}^2$ is the variance of group A, $s_b{}^2$ is the variance of group B, n_a is the number of students in group A, and n_b is the number of students in group B.

Cohen's *d*, then, expresses the difference in group means in terms of the standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the *d* statistic: *d* = 0.20 is a small ES, *d* = 0.50 is a medium ES, and *d* = 0.80 is a large ES. Based on Cohen's (1988) guidelines, certain trends are observable in Table B.6. Although no big difference in Biology test scores was seen between females and males, mean raw scores and ESs show that Asian and White students tend to outperform other ethnicity groups. There were clear performance differences among regular education, gifted/talented education, and special education students in Education Classification and Non-English Learner and English Learner in EL status. Performance differences were also observed from Economically Disadvantaged status, Homeless status, Foster Care status, and Military Affiliation status.

Score Reports

Score reports are the primary means of communicating test scores to appropriate school system personnel (e.g., testing coordinators or superintendents), teachers, and parents. Interpretations of test scores from each administration are disseminated in two ways: the individual score report and the LEAP Interpretive Guide. The LDOE and DRC strive to create documents that will be accessible to parents, teachers, and all other stakeholders. The Individual Student-Level Report (ISR) is the primary means for sharing student test results with parents. As such, it is a standalone document from which parents can glean information that is relevant to understanding their children's test scores. For more information about the test, parents are provided Parent Guide to the LEAP 2025 Student Reports. In the 2021–2022 administration year, student reports for each school were posted by subject, then downloaded and printed from eDIRECT by the school systems and schools. eDIRECT is DRC's secure online system that provides schools and districts access to student tests and reports.

School Roster Report. A School Roster Report, which provides summary information about student performance on the LEAP 2025 high school Biology assessment, is available to school systems and schools through eDIRECT. Total test scores and achievement level indicators are shown for the test of interest. Category and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by category and subcategory are summarized. More details can be found in the LEAP 2025 High School Interpretive Guide (iGUIDE) 2021-2022.

Individual Student-Level Report. The ISR is another type of report available through the eDIRECT system. ISRs may be downloaded and printed by schools to be sent home to parents. At the top of the page, overall student performance is reported by scale score and achievement level. In the middle of the page, category and subcategory performance indicators are reported. When a student does not receive a scale score, their achievement level will be left blank. ISRs for students whose scores were invalidated will display a blank scale score for a given course.

LEAP 2025 High School Interpretive Guide (iGUIDE) 2021-22. The LEAP 2025 High School Interpretive Guide (iGUIDE) 2021-2022 was written to help Louisiana school system and school administrators, teachers, parents, and the general public understand the LEAP 2025 Biology test. The LEAP 2025 High School Interpretive Guide (iGUIDE) 2021-2022 was developed collaboratively by DRC and LDOE staff. LDOE staff had opportunities to review the guide, provide feedback, and give final approval. 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

Achievement Level Policy Definitions and Cut Scores

Achievement level policy definitions for the LEAP 2025 Biology assessment are shown in Table 8.2. 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. The standard-setting section of the LEAP 2025 Biology 2018-2019 technical report contains comprehensive information.

Table 8.2

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.

It should be noted that the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information address multiple best practices of the testing industry. Table 8.3 shows the cut of each performance level, and the CSEM for each performance level can be found at Table 9.1 in Section, Reliability. The standard-setting section of the LEAP 2025 Biology 2018-2019 technical report contains comprehensive information.

Table 8.3

Performance Level Cuts at the Approaching Basic, Basic, Mastery, and Advanced: Operational 2022 LEAP Biology

Form	Approaching Basic	Basic	Mastery	Advanced
	Cut Score	Cut Score	Cut Score	Cut Score
D	707	725	750	772

9. Reliability

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 i_{th} 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.

While coefficient alpha value was 0.90, the marginal alpha value was 0.91 for the 2022 Biology test. 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

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

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 using the population of students. (Please refer to Table F.1.) Included with coefficient alpha in the tables are 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.

Classical Standard Error of Measurement

The classical standard error of measurement (SEM) represents the amount of variance in a score that results from random factors other than what the assessment is intended to measure. Because underlying traits such as academic achievement cannot be measured with perfect precision, the SEM is used to quantify the margin of uncertainty in test scores. For example, factors such as chance error and differential testing conditions can cause a student's observed score (the score achieved on a test) to fluctuate above or below his or her true score (the student's expected score). The SEM is calculated using both the standard deviation and the reliability of test scores, as follows:

$$SEM = \sigma_x\sqrt{(1 - P'_{xx})},$$

where P'_{xx} is the reliability estimate and σ_x is the standard deviation of raw scores on the test. A standard error provides some sense of the uncertainty or error in the estimate of the true score using the observed score. For example, suppose a student achieves a raw score of 50 on a test with an SEM of 3. Placing a one-SEM band around this student's score would result in a raw score range of 47 to 53. If the student took the test 100 times and 100 similar raw score ranges were computed, about 68 of those score ranges would include the student's true score.

It is important to note that the SEM provides an estimate of the average test score error for all students regardless of their individual proficiency levels. It is generally accepted that the SEM varies across the range of student proficiencies (Peterson, Kolen, & Hoover, 1989). For this reason, it is useful to report test-level SEM, and SEM for 2022 Biology was 3.72, as seen from Table B.4.

Conditional Standard Error of Measurement

It is important to note that the SEM index provides only an estimate of the average test score error for all students regardless of their individual levels of proficiency. By comparison, conditional standard error of measurement (CSEM) provides a reliability estimate at each score point on a test. Like the SEM, the CSEM reflects the amount of variance in a score resulting from random factors other than what the assessment is designed to measure, but it provides an estimate conditional on proficiency. The CSEM is usually smallest, and thus scores are most reliable, near the middle of the score distribution. Typically, achievement tests included relatively large numbers of moderately difficult items. Because these items are usually well-matched to a majority of students' ability, they provide the most reliable estimates of ability. It is desirable, for an achievement test where students are classified into pass/fail categories, that the CSEM be lowest at the cut score for passing. The CSEMs at the four cut scores that define the performance levels are presented in Table 9.1.

Table 9.1

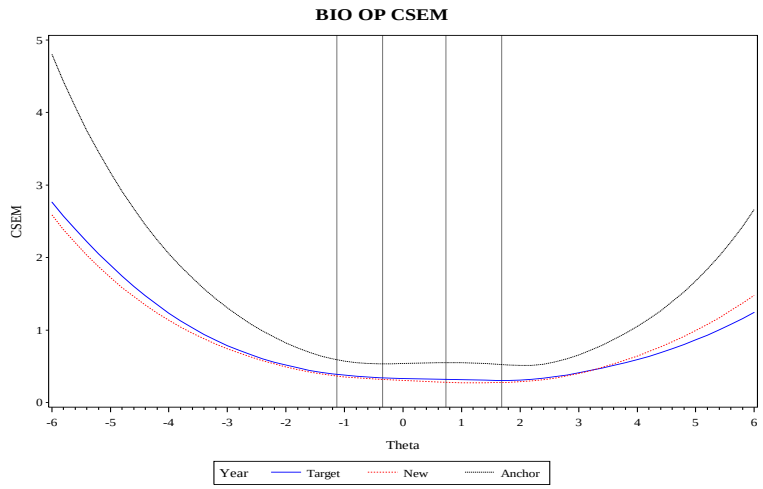
Conditional Standard Errors of Performance Level Cuts: SPR 2022 Operational Biology

Form	Approaching Basic		Basic		Mastery		Advanced	
	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
D	707	8	725	7	750	6	772	6

IRT methods are used for estimating CSEM and are presented in the following graph. With fixed-form assessments, the estimates of measurement error tend to be higher at the low and high ends of the scale-score range, where few items measure the ability levels. Generally, there are few students with extreme scores, and these score levels cannot be estimated as accurately as levels toward the middle of the ability range. The middle of the ability range, where cut scores are located, shows lower measurement error than the low and high ends of the ability ranges. Plot 9.1 demonstrates that the tests are designed so that measurement error is minimized in the middle of the scale range, where most students are located.

Plot 9.1

CSEM Curves: SPR 2022 Operational Biology



Note. The scale is on theta; Each theta cut matches the scale score of each performance level: 707, 725, 750, and 772; Target = 2019 test; New = 2022 OP form; Anchor = anchor items.

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; 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 and consistent classification of students. Procedures from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) were used 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. While the overall accuracy index was 0.723, the overall consistency index was 0.62 for the 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 was 0.245 for the 2022 Biology test.

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 was 0.497 for the 2022 Biology test.
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.

10. 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 2021–2022 LEAP 2025 Biology test was designed and developed to provide fair and accurate scores that support appropriate, meaningful information for educational decisions. 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 2, 3, and 4 provide 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. The data review process and results are also discussed. 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 2022 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 2022 Biology operational items are presented in Appendix C. Chapter 8 of the technical report summarizes the test results, score distributions, score reports, and achievement level information. Chapter 9 addresses Cronbach's alpha and marginal alpha as measures of internal consistency and describes analysis procedures for classification consistency and classification accuracy. In addition, test validity is addressed in this chapter.

Evidence for Construct-Related Validity

Evidence for construct-related validity—the meaning of test scores and the inferences they support—is the central concept underlying the LEAP 2025 validation process. Validity evidence, from the design of the test to item development and scoring, is created throughout the entire assessment process. Therefore, evidence of validity is described throughout the LEAP 2025 technical report.

Internal Structure of Reporting Categories

The 2022 Biology test contains three reporting categories: *Investigate, Evaluate, and Reason Scientifically.* Table D.1 shows correlations among the reporting categories, and the moderate correlations were observed among the reporting categories; since we used distinct items for each reporting category, a moderate correlation was anticipated.

Content-Related Evidence

Content validity is frequently defined in terms of the sampling adequacy of test items. That is, content validity is the extent to which the items in a test adequately represent the domain of items or the construct of interest (Suen, 1990). Consequently, content validity provides judgmental evidence in support of the domain relevance and representativeness of the content in the test (Messick, 1989). It should be noted that the 2022 Biology operational test forms were built exclusively using an ABBI bank program which contained both content and statistical information about both operational and field-tested items.

Dimensionality and Principal Component Analysis

Appendix D: Dimensionality provides information about principal component analysis of the Biology tests. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, 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, 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 spring 2022 assessment, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the Statistical Analysis System (SAS) program. The first and second principal component eigenvalues were compared without rotation. Table D.4 and Plot D.1 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 figure, the first component is substantially larger than the second eigenvalue for the spring 2022 test. Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Evidence Based on Relations to Other Variables

Evidence based on *relations to other variables* is typical utility of criterion-related validity evidence to measure concurrent or predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). Thus, external variables can be used to evaluate hypothesized relationships between test scores and other measures of

student achievement (e.g., test scores on other tests) to evaluate the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades).

A significant number of students who took the LEAP Biology test also took the ACT Science, ACT Mathematics, ACT English, LEAP USH, LEAP Algebra, LEAP Geometry, LEAP English 1, and LEAP English II tests. For the total student group, in general, moderate correlation was observed between LEAP Biology and ACT Science exams. In general, however, English Learner, Special Education group, and Migrant groups slightly lower correlations than other groups. A separate report, External Validity Study: SPR 2022, that was submitted to LDOE has more specific information.

Item Development and Field-Test Analysis

Test development for LEAP Biology is ongoing and continuous. Content specialists, teachers from across Louisiana, WestEd/Pearson, and the LDOE were greatly involved in developing and reviewing test items. Committees such as content review and bias review reviewed all of the items, which were finally stored in the item bank. Specifically, an internal review by LDOE and WestEd/Pearson staff for alignment and quality required a great deal of time and energy. More specific information on item (test) development and review can be obtained in Chapter 3, Overview of the Test Development Process.

Field test items were embedded and administered in one of 10 test forms. Once these items were scored, the LDOE and WestEd/Pearson conducted additional item analysis and content review. Any field test items that exhibited statistical results that suggested potential problems were carefully reviewed by both LDOE and WestEd/Pearson content specialists. A determination was then made as to whether an item should be accepted, rejected, and revised/refield-tested. Information on statistical analyses for field test items can be obtained in Chapter 7, Data Analysis.

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. Finally, Table 10.1 summarizes the sources of validity evidence and indicates where the evidence can be found in the technical report.

Table 10.1

Evidence of Validity and the Corresponding Technical Report Chapter

Source of Validity	Related Information	Related Chapter/Source
Evidence Based on Test Content	Item Development Process	Chapter 3 LEAP 2025 High School Biology Assessment Frameworks
	Test Blueprint and Item Alignment to Curriculum and Standards	Chapters 2 & 3 Appendix A LEAP 2025 High School Biology Assessment Frameworks
	Item Bias, Sensitivity, and Content Appropriateness	Chapter 3
	Accommodations	Chapter 4
Evidence Based on Response Processes	Field Test Analysis Data Review	Chapters 3, 7, & 9 LEAP 2025 High School Biology Assessment Frameworks
	Classical Item analysis IRT Analysis	Chapter 7
Evidence Based on Internal Structure	Differential Item Functioning	Chapter 7
	Reliability and Standard Errors of Measurement	Chapter 9
	Correlation among Reporting Categories	Chapter 9
	Dimensionality Analysis	Chapter 9
Evidence Based on Relations to Other Variables	Correlation Analysis between ACT and LEAP Biology Tests	Chapter 9
Evidence Based on the Consequences of Testing	Scale Score and Performance Level Information	Chapter 8
	Test Interpretive Guide	Chapter 8

References

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

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

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.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

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.

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.

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.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5–11.

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.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

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 Researchand 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 inEducation*, 10(4), 321–344.

Appendix A: Training Agendas

LEAP 2025 Biology Item Outline Development Training Agenda
Item Development Cycle for the 2019–2022 LEAP 2025 Science Assessment

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

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. **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

V. **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, or disability.
    4. Avoid topics that will offend the privacy of values and beliefs of students, parents, or the public.

LEAP 2025 Biology Item Writer Training Agenda
Item Development Cycle for the 2019–2022 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, or disability.
    d. Avoid topics that will offend the privacy of values and beliefs of students, parents, or the 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 2018–2019 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 2018–2019
   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

LEAP 2025 Biology and Grades 3-8 Content and Bias Item Review Committee Training Agenda
Item Development Cycle for the 2022-2023 LEAP Science Assessment

I. Welcome from LDOE
II. Introductions
III. Non-Disclosure Agreement
    a. Test security and student confidentiality are of utmost importance to WestEd and the Louisiana Department of Education.
    b. As a participant in the Science Content/Bias Item Review Meetings, you will have access to materials that must be regarded as secure.
    c. All materials must be treated as confidential. You are not to disclose the content of these materials or copy or reproduce any of the materials, directly or indirectly.
    d. By signing and submitting the form, you confirmed that you agree to adhere to these guidelines.
IV. LEAP Test Development Process
V. Purpose of Content and Bias Item Review
    a. To ensure high-quality science tests that:
        i. Reflect instructionally relevant content
        ii. Provide valid information to students, parents, teachers, administrators, policymakers, and the public
        iii. Are fair and appropriate for all students
VI. What to Consider
    a. Louisiana Student Standards for Science
    b. Performance Expectation and the Phenomenon
    c. Science Shifts
    d. Components
        i. Tasks
            a) Based on a common stimulus
            b) Items follow a prescribed order; items build on one another
            c) For field testing, different versions of items included culminating with an extended-response (ER) item
        ii. Item Sets
            a) Based on a common stimulus
            b) Items are not in a prescribed order
            c) 4 items on operational test; may have a constructed-response (CR) item
            d) For field testing, extra items included (12 items developed to get 4)
        iii. Standalone Items
VII. Item Types
VIII. Content alignment
    a. Alignment is the key element of content review.
        i. Is the item providing an appropriate measure of the PE and its related dimensions?

ii. Item content alignment is the degree to which an item measures the intended PE and its related dimensions.
iii. Put another way: An item is determined to be aligned if the item allows the student to provide evidence of his or her understanding of the specified PE and its related dimensions.

b. Additional considerations include:
i. Scoring/key accuracy
ii. Scientific accuracy

IX. Principles of LSSS for Science Alignment
a. Items must be aligned to at least two of the three dimensions.
b. Multiple aspects of the item and the item's alignment need to be considered.
c. Relative degrees of alignment need to be evaluated.
d. Holistic (not analytic) judgments are used to determine acceptable alignment.

X. Bias and Sensitivity Review
a. Items and stimuli should be free of bias and sensitivity concerns.
b. This helps to provide students with a fair opportunity to demonstrate their knowledge or skills, regardless of their backgrounds.
c. Bias is the presence of some language or content that prevents some members of a group from showing us their knowledge or skills in a particular content area.
i. Result: Two individuals of the same ability but from different groups perform differently.
d. What is sensitivity?
e. Any reference in a stimulus or item that might cause a student to have an emotional reaction and prevent the student from showing us their knowledge and skills for a particular content area.
i. Result: Two individuals of the same ability but from different groups perform differently.
f. If there are bias or sensitivity concerns for an item, the reviewer should be able to point to one of these areas as an area of concern.
i. Opportunity and Access
a) Problems:
i.) Not all Louisiana students have had the opportunity to visit different regions of the world, the US, or Louisiana.
ii.) Some students have stronger science skills than English skills.
b) Possible solutions:
i.) Include non-DCI information that makes a stimulus accessible to students from all backgrounds.
ii.) Avoid regional language or words with different meanings in different groups.
iii.) Avoid idioms and figurative language.
ii. Portrayal of Groups Represented

a) Problem:
    i.) A group is stereotyped (portrayed consistently in a particular way, which may be offensive to members of that group).
b) Possible solution:
    i.) Avoid issues and themes that demean, offend, or inaccurately portray a group, culture, ethnicity, disability.

iii. Protecting Privacy and Avoiding Offensive Content
a) Problem:
    i.) Some issues and contexts are controversial to particular groups.
b) Possible solution:
    i.) Avoid topics that will offend the privacy, values, and/or beliefs of students, parents, and the public.

XI. Cognitive Complexity and Difficulty
a. Cognitive complexity ≠ difficulty
b. Cognitive complexity refers to the type and level of thinking and reasoning required of students to answer a test question.
c. Difficulty refers to the amount of time and/or effort needed to answer a test question (easy or hard) and can be measured in percentage answering question correctly.
d. Task Analysis Guide in Science (Tekkumru-Kisa, Stein & Schunn, 2014)—focused on instruction
e. Modified TAGS model is a tool for coding 2- and 3-dimensional items
f. Cognitive Complexity in TAGS model

XII. Content Review Decisions
a. Yes ("Accept")
    i. Item is acceptable as is
    ii. Aligned
    iii. Scientifically accurate
    iv. Scoring information correct
    v. Free of bias concerns
b. No ("Accept with Edits" or "Reject")
    i. Due to content concerns
    ii. Metadata alignment with explanation
    iii. Science accuracy concern with explanation
    iv. Due to bias concerns
    v. With explanation
c. Reject when:
    i. Complete alignment mismatch
    ii. Unfixable context flaws
d. Revise when:
    i. Fixes can be made
    ii. Item Alignment Information

XIII. Reviewing Items
    a. Review items in ABBI online
    b. Your facilitator will walk you through a few items to help you learn how to use this tool.
    c. Use the Review Tool for alignment decisions
    d. Vote in ABBI
    e. You will select from:
        i. Accept
        ii. Accept with Edits
        iii. Reject
    f. “Accept with Edits” or “Reject” require comments/justification

XIV. Logistics
    a. Breaks will be announced by the facilitator
    b. ABBI access will be locked during non-meeting times
    c. Room will be locked over lunch
    d. At the conclusion of the meeting, you will receive email communications about:
        i. Stipend
        ii. Substitute Reimbursement Form
        iii. Evaluation survey

LEAP 2025 Biology and Grades 3–8 Data Review Training Agenda

I. What is a Data Review?

a. Statistical Definition: Classical Test Theory

1. P-value
2. Point-Biserial
3. Option/Distribution Analysis
4. Differential Item Function (DIF)
5. Flagging Value

Statistics	**Flagging Value**
P-value	≤ 0.25 or > 0.9
Omit Percentage	> 4%
Point-biserial Correlation	< 0.20
Distractor Percentage (MC only)	> 40%
Distractor Point-biserial Correlation (MC only)	> 0.00
DIF	B, C

b. Statistical Definition: Item Response Theory (IRT)

1. IRT Discrimination (a-parameter)
2. IRT Difficulty (b-parameter)
3. IRT Guessing (c-parameter)
4. Q1 (Zq1)
5. Item Fit Plot
6. Flagging Value

Flagging Value for IRT Item Parameters		
a (Discrimination)	b (Difficulty)	c (Guessing)
< 0.34	Lowe than -3.0 or Higher than 3.0	> 0.35

II. Judgement Task in ABBI

a. Accept
b. Accept with Edits
c. Reject

Appendix B: Test Summary

Biology

Contents
Table B.1 Percentage of Points by Reporting Category (includes Task Items): Spring 2022 Operational Biology
Table B.2 Standard Coverage: Spring 2022 Operational Biology
Table B.3 Item Type Summary: Spring 2022 Operational Biology
Table B.4 Raw Score Summary: Spring 2022 Operational Biology
Table B.5 Raw Score Summary by Reporting Category: Spring 2022 Operational Biology
Table B.6 Scale Score and Raw Score Summary: Spring 2022 Operational Biology

- Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table B.1

Percentage of Points by Reporting Category (includes Task Items): Spring 2022 Operational Biology

Reporting Category	Form D
N/A*	7.5%
Investigate	9.0%
Evaluate	35.8%
Reason Scientifically	47.8%

* N/A indicates no reporting category.

Table B.2

Standard Coverage: Spring 2022 Operational Biology

Reporting Categories		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
N/A*	HS-LS1-3			1					2.50
	HS-LS1-8			1				1	5.00
	Sub-Total			**2**				**1**	**7.50**
1 Investigate	HS-LS1-3			1	1				5.00
	HS-LS3-1	1		1					5.00
	Sub-Total	**1**		**2**	**1**				**10.00**
2 Evaluate	HS-LS2-1			1		1			5.00
	HS-LS2-4			1				1	5.00
	HS-LS2-6		1			1			5.00
	HS-LS3-2			2		1			7.50
	HS-LS3-3			1					2.50
	HS-LS4-1			3					7.50
	HS-LS4-3		1			1			5.00
	HS-LS4-5					1			2.50
	Sub-Total		**2**	**8**		**5**		**1**	**40.00**
3 Reason Scientifically	HS-LS1-1			1		1			5.00
	HS-LS1-2	1			1	1		1	10.00
	HS-LS1-5			1		1			5.00
	HS-LS1-6	1							2.50
	HS-LS2-7		1	2		1	1		12.50
	HS-LS4-2			1		2			7.50
	Sub-Total	**2**	**1**	**5**	**1**	**6**	**1**	**1**	**42.50**
Total		**3**	**3**	**17**	**2**	**11**	**1**	**3**	**100.00**

* N/A indicates no reporting category.

Table B.3

Item Type Summary: Spring 2022 Operational Biology

Admin.	MC	MS	TE*	CR	ER**	TPD	TPI
Spring 2022	11	2	17	3	1	3	3

* One of the TE items is a multiple-part, selected-response (MPSR) item.

** Classical analyses are calculated and estimated separately for each dimension of the ER item, and the result summarizes both dimensions.

Table B.4

Raw Score Summary: Spring 2022 Operational Biology

Admin.	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
Spring 2022	≥38,820	26.66	11.76	0	65	0.40	0.44	0.90	3.72

* Reliability is Cronbach's alpha.

Table B.5

Raw Score Summary by Reporting Category: Spring 2022 Operational Biology

Admin	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
Spring 2022	Investigate	1.66	1.39	0	6	0.30	0.41	0.42	1.06
	Evaluate	9.31	4.61	0	24	0.39	0.43	0.77	2.21
	Reason Scientifically	14.05	5.91	0	32	0.44	0.44	0.80	2.64

Table B.6

Scale Score and Raw Score Summary: Spring 2022 Operational Biology

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥38,820	100.00	732.65	25.64	26.66	11.76	-
Female	≥19,470	50.15	733.27	24.54	26.84	11.35	**-0.03**
Male	≥19,350	49.85	732.04	26.70	26.48	12.15	-
African American	≥16,000	41.23	721.56	23.21	21.51	9.91	**0.91**
American Indian or Alaska Native	≥260	0.68	735.63	22.33	27.79	10.65	**0.30**
Asian	≥690	1.79	752.48	25.47	36.36	12.30	**-0.47**
Hispanic/Latino	≥2,940	7.59	727.36	27.41	24.51	11.96	**0.59**
Multi-Racial	≥1,060	2.75	736.98	24.32	28.58	11.46	**0.23**
Native Hawaiian or Other PacificIslander	≥30	0.09	744.54	24.54	31.89	11.68	**-0.07**
White	≥17,800	45.86	742.40	22.99	31.13	11.19	-
Economically Disadvantaged:	≥14,090	36.30	744.24	23.42	32.05	11.44	**-0.75**
Economically Disadvantaged:	≥23,320	60.07	726.39	24.46	23.73	10.81	-
EL: No	≥37,670	97.04	733.45	25.33	26.99	11.69	**-0.97**
EL: Yes	≥1,140	2.96	706.64	21.99	15.69	7.90	-
Gifted or Talented	≥2,290	5.90	758.75	22.08	39.40	11.05	**-2.29**
Regular Education	≥33,230	85.60	733.18	24.07	26.77	11.13	**-0.92**
Special Education	≥3,290	8.49	709.21	23.21	16.69	9.03	-
Section 504: No	≥35,240	90.79	733.45	25.58	27.03	11.76	**-0.34**
Section 504: Yes	≥3,570	9.21	724.78	24.89	23.03	11.02	-
Migrant: No	≥38,760	99.84	732.67	25.64	26.67	11.76	**-0.38**
Migrant: Yes	≥60	0.16	722.81	24.33	22.17	10.26	-
Homeless: No	≥38,140	98.25	732.80	25.65	26.73	11.76	**-0.34**
Homeless: Yes	≥680	1.75	724.33	24.16	22.75	10.65	-
Military Affiliation: No	≥38,300	98.66	732.51	25.64	26.59	11.74	**0.43**
Military Affiliation: Yes	≥520	1.34	743.19	23.93	31.59	11.59	-
Foster Care: No	≥38,750	99.81	732.68	25.64	26.67	11.76	**-0.50**
Foster Care: Yes	≥70	0.19	719.81	23.72	20.85	9.80	-

Appendix C: Item Analysis Summary Report

Contents
Table C.1 P-Value Summary: Spring 2022 Operational Biology Table C.1.1 P-Value Summary by Item Type: Spring 2022 Operational Biology Plot C.1 P-Value Summary by Item Type: Spring 2022 Operational Biology
Table C.2. Item-Total Correlation Summary: Spring 2022 Operational Biology Table C.2.1 Item-Total Correlation Summary by Item Type: Spring 2022 Operational Biology Plot C.2 Item-Total Correlation Summary by Item Type: Spring 2022 Operational Biology
Table C.3. Corrected Point-Biserial Correlation Summary: Spring 2022Operational Biology Table C.3.1 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022Operational Biology Plot C.3 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational Biology
Table C.4 Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2022 Operational Biology
Table C.5.1 IRT-A Parameter Summary by Reporting Category: Spring 2022 Operational Biology Table C.5.2 IRT-B Parameter Summary by Reporting Category: Spring 2022 Operational Biology Table C.5.3 IRT Parameter Summary by Item Type: Spring 2022 Operational Biology
Plot C.5.1 IRT Parameter Summary: Spring 2022 Operational Biology: A-Parameter Plot C.5.2 IRT Parameter Summary: Spring 2022 Operational Biology: B-Parameter Plot C.5.3 IRT Parameter Summary: Spring 2022 Operational Biology: C-Parameter
Table C.6 Statistically Flagged Items by Item Type: Spring 2022 Operational Biology

- Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table C.1

P-Value Summary: Spring 2022 Operational Biology

Form	No. of Items	0 le p lt 0.2	0.2 le p lt 0.4	0.4 le p lt 0.6	0.6 le p lt 0.8	0.8 le p le 1.0
D	41	6	16	13	5	1

Table C.1.1

P-Value Summary by Item Type: Spring 2022 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.148	0.148	0.162	0.216	0.216
ER*	1	0.242	0.242	0.329	0.415	0.415
MC	11	0.276	0.317	0.407	0.541	0.804
MS	2	0.183	0.183	0.203	0.222	0.222
TE	17	0.097	0.292	0.393	0.533	0.752
TPD	3	0.352	0.352	0.485	0.719	0.719
TPI	3	0.189	0.189	0.637	0.685	0.685

* Classical analyses are calculated and estimated separately for each dimension of the ER item,and the result summarizes both dimensions.

Plot C.1

P-Value Summary by Item Type: Spring 2022 Operational Biology

Box and Whisker Plot

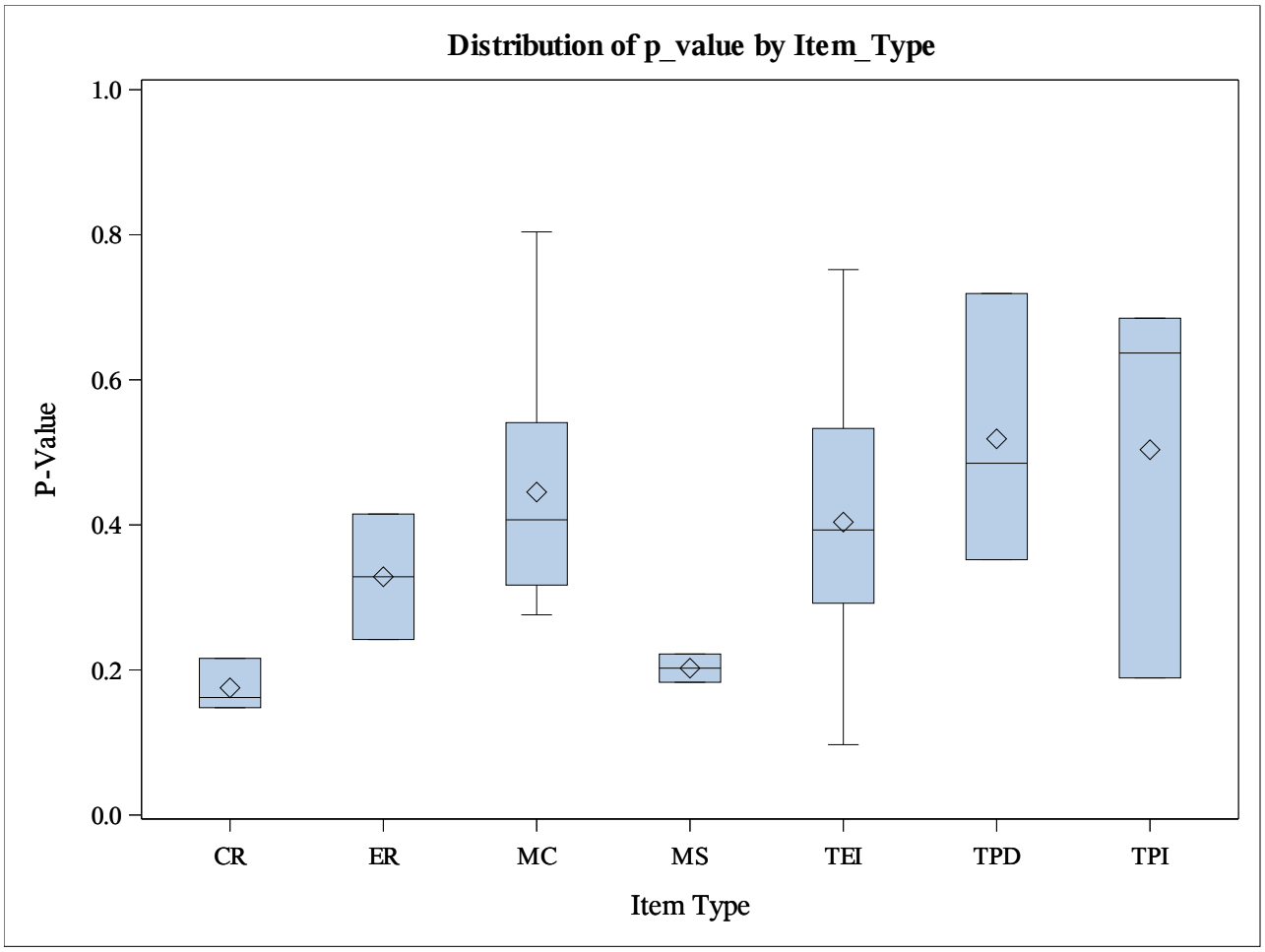


Table C.2

Item-Total Correlation Summary: Spring 2022 Operational Biology

No. of Items	r lt 0	0.0 le r lt 0.2	0.2 le r lt 0.3	0.3 le r lt 0.4	0.4 le r lt 0.5	r ge 0.5
41	0	3	6	3	13	16

Table C.2.1

Item-Total Correlation Summary by Item Type: Spring 2022 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.269	0.269	0.518	0.620	0.620
ER*	1	0.658	0.658	0.676	0.694	0.694
MC	11	0.121	0.293	0.407	0.460	0.497
MS	2	0.170	0.170	0.289	0.408	0.408
TE	17	0.168	0.400	0.471	0.534	0.621
TPD	3	0.471	0.471	0.584	0.591	0.591
TPI	3	0.370	0.370	0.549	0.573	0.573

* Classical analyses are calculated and estimated separately for each dimension of the ER item,and the result summarizes both dimensions.

Plot C.2
Item-Total Correlation Summary by Item Type: Spring 2022 Operational Biology

Box and Whisker Plot

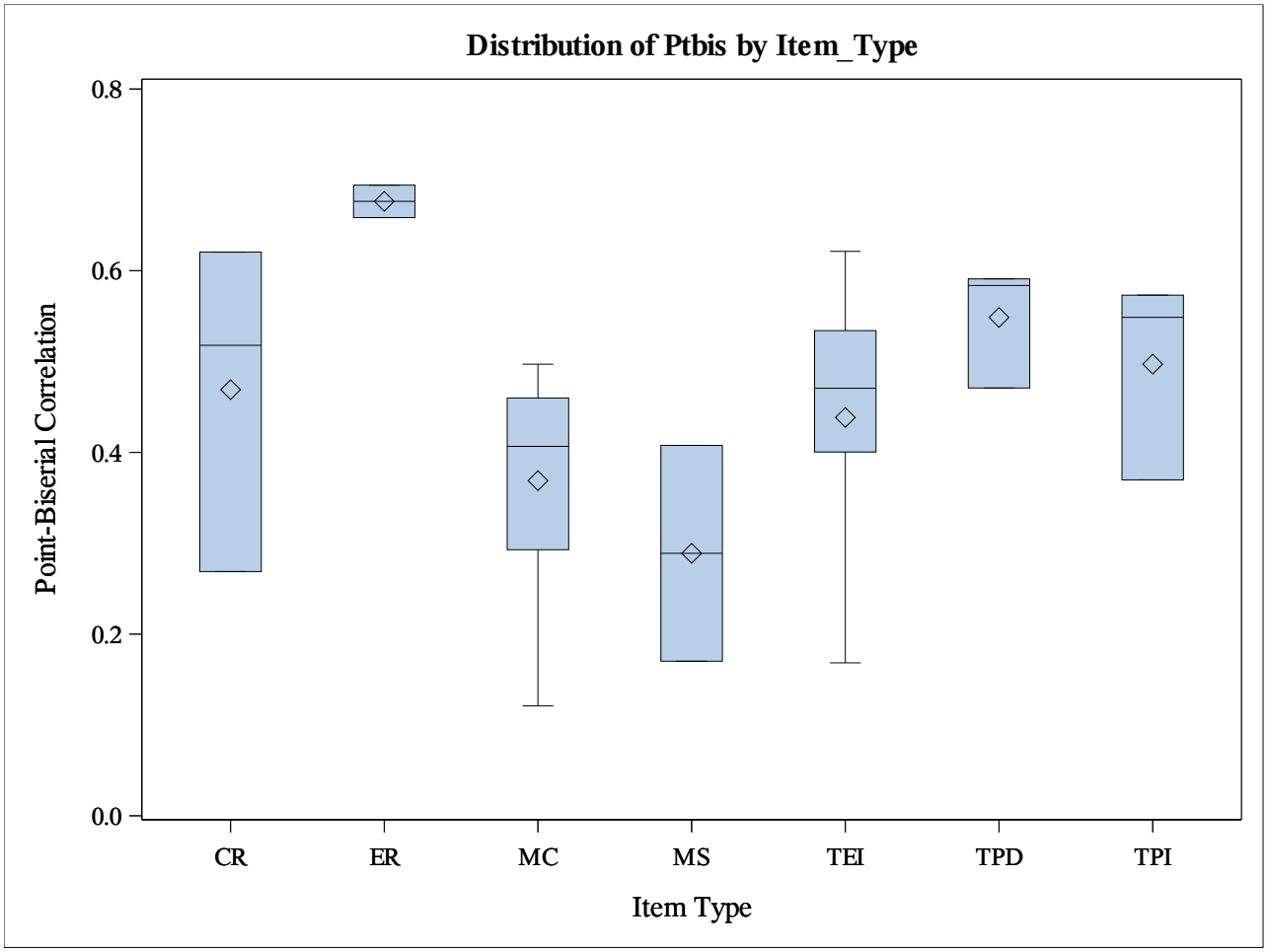


Table C.3

Corrected Point-Biserial Correlation Summary: Spring 2022 Operational Biology*

No. of Items	r lt 0	0.0 le r lt 0.2	0.2 le r lt 0.3	0.3 le r lt 0.4	0.4 le r lt 0.5	r ge 0.5
41	0	4	6	8	14	9

Table C.3.1

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational Biology*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.226	0.226	0.479	0.584	0.584
ER**	1	0.617	0.617	0.618	0.619	0.619
MC	11	0.083	0.253	0.371	0.428	0.464
MS	2	0.135	0.135	0.258	0.380	0.380
TE	17	0.130	0.352	0.412	0.493	0.578
TPD	3	0.415	0.415	0.535	0.542	0.542
TPI	3	0.325	0.325	0.498	0.530	0.530

* 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 itemscore from the total test score.

** Classical analyses are calculated and estimated separately for each dimension of the ER item, and the result summarizes both dimensions.

Plot C.3

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational Biology

Box and Whisker Plot

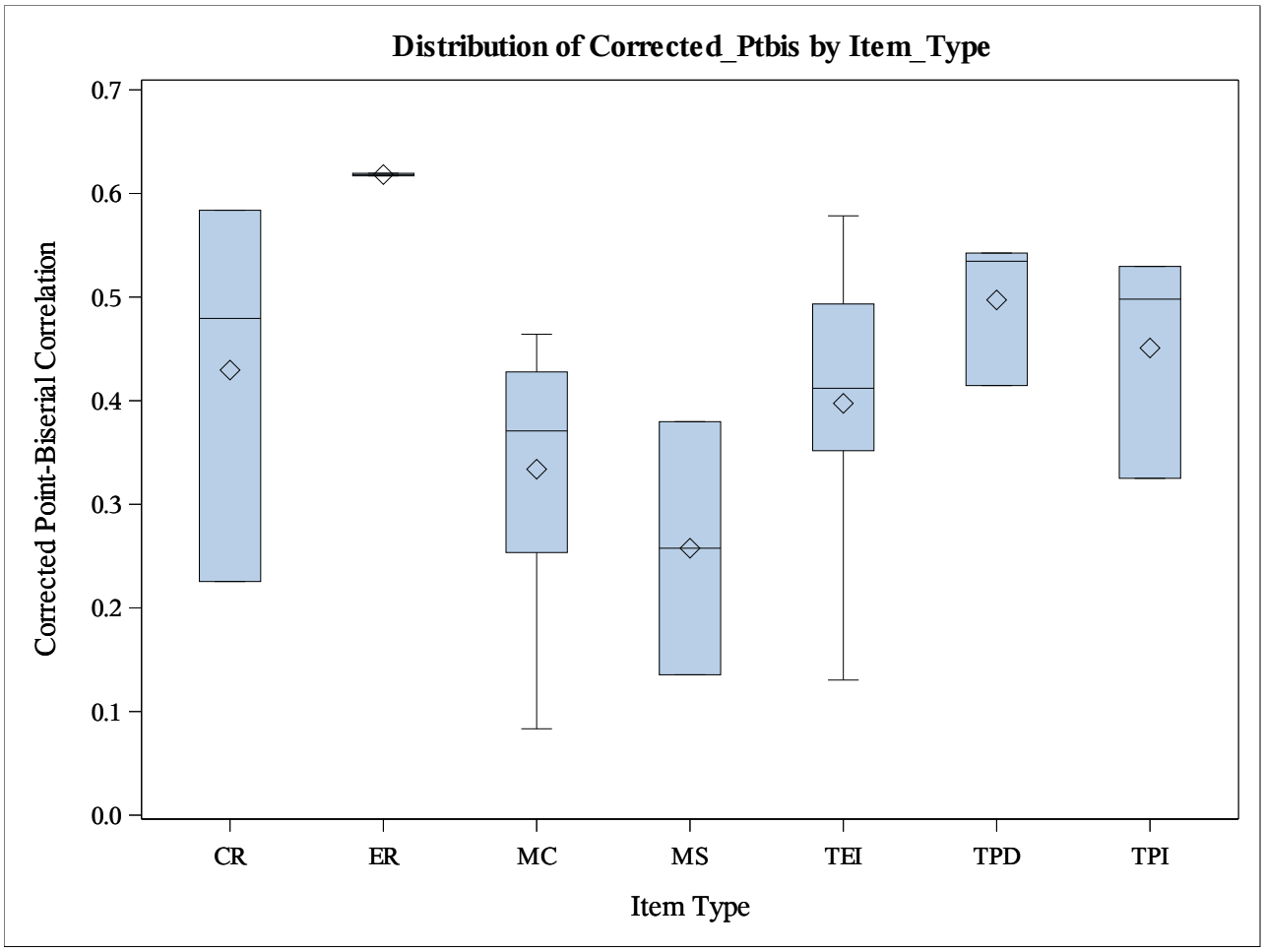


Table C.4

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2022 Operational Biology

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Evaluate	1	0.269	0.269	0.269	0.269	0.269
	Reason Scientifically	1	0.518	0.518	0.518	0.518	0.518
ER*	Reason Scientifically	1	0.658	0.658	0.676	0.694	0.694
MC	Evaluate	5	0.306	0.344	0.460	0.473	0.497
	Reason Scientifically	6	0.121	0.270	0.350	0.437	0.451
MS	Investigate	1	0.170	0.170	0.170	0.170	0.170
	Reason Scientifically	1	0.408	0.408	0.408	0.408	0.408
TEI	Investigate	2	0.543	0.543	0.557	0.571	0.571
	Evaluate	8	0.168	0.304	0.515	0.532	0.621
	Reason Scientifically	5	0.252	0.299	0.400	0.471	0.565
TPD	Evaluate	2	0.471	0.471	0.527	0.584	0.584
	Reason Scientifically	1	0.591	0.591	0.591	0.591	0.591
TPI	Investigate	1	0.370	0.370	0.370	0.370	0.370
	Reason Scientifically	2	0.549	0.549	0.561	0.573	0.573

* Classical analyses are calculated and estimated separately for each dimension of the ER item,and the result summarizes both dimensions.

Table C.5.1

IRT-A Parameter Summary by Reporting Category: Spring 2022 Operational Biology

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	1	0	1
0.2 <= a < 0.4	1	2	3	7
0.4 <= a < 0.6	1	4	3	9
0.6 <= a < 0.8	0	5	7	12
0.8 <= a < 1.0	0	2	3	6
1.0 <= a < 1.2	1	1	1	3
1.2 <= a < 1.4	0	1	0	1
1.4 <= a < 1.6	1	0	1	2
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.35	0.19	0.34	0.19
Maximum	1.47	1.31	1.53	1.53
Mean	0.86	0.64	0.71	0.69
SD	0.51	0.30	0.29	0.31
Number of Items	4	16	18	41

Table C.5.2

IRT-B Parameter Summary by Reporting Category: Spring 2022 Operational Biology

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	1	1
-1.5 <= b < -1.0	0	0	1	1
-1.0 <= b < -0.5	0	0	4	4
-0.5 <= b < 0.0	1	2	0	3
0.0 <= b < 0.5	0	4	4	8
0.5 <= b < 1.0	0	4	1	7
1.0 <= b < 1.5	1	2	1	5
1.5 <= b < 2.0	0	1	3	4
2.0 <= b < 2.5	2	1	2	5
2.5 <= b < 3.0	0	0	1	1
3.0 <= b < 3.5	0	2	0	2
3.5 <= b	0	0	0	0
Minimum	-0.01	-0.30	-1.77	-1.77
Maximum	2.44	3.47	2.65	3.47
Mean	1.40	1.02	0.52	0.82
SD	1.08	1.14	1.30	1.18
Number of Items	4	16	18	41

Table C.5.3

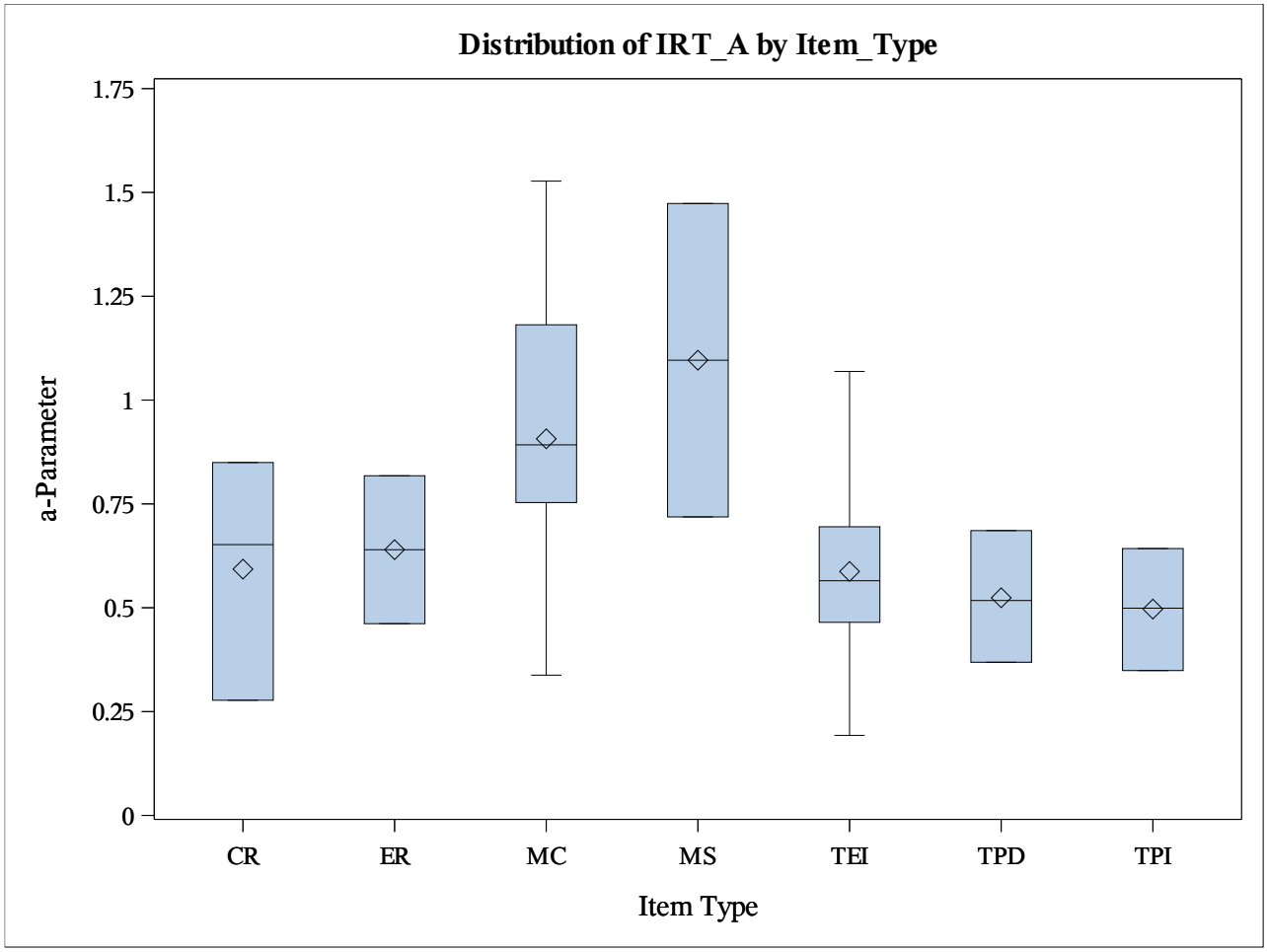
IRT Parameter Summary by Item Type: Spring 2022 Operational Biology

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.277	0.277	0.652	0.85	0.85
	b	3	1.193	1.193	1.661	3.26	3.26
ER*	a	1	0.462	0.462	0.64	0.818	0.818
	b	1	0.294	0.294	0.761	1.228	1.228
MC	a	11	0.338	0.753	0.892	1.181	1.527
	b	11	-1.04	0.005	0.886	1.591	2.141
	c	11	0.023	0.063	0.158	0.207	0.241
MS	a	2	0.719	0.719	1.096	1.473	1.473
	b	2	1.645	1.645	1.824	2.003	2.003
	c	2	0.009	0.009	0.093	0.177	0.177
TEI	a	17	0.193	0.465	0.565	0.695	1.069
	b	17	-1.771	0.056	0.667	1.313	3.466
	c	7	0.017	0.026	0.068	0.163	0.165
TPD	a	3	0.369	0.369	0.517	0.686	0.686
	b	3	-0.855	-0.855	0.098	0.857	0.857
TPI	a	3	0.349	0.349	0.499	0.643	0.643
	b	3	-0.902	-0.902	-0.646	2.442	2.442

* Classical analyses are calculated and estimated separately for each dimension of the ER item,and the result summarizes both dimensions.

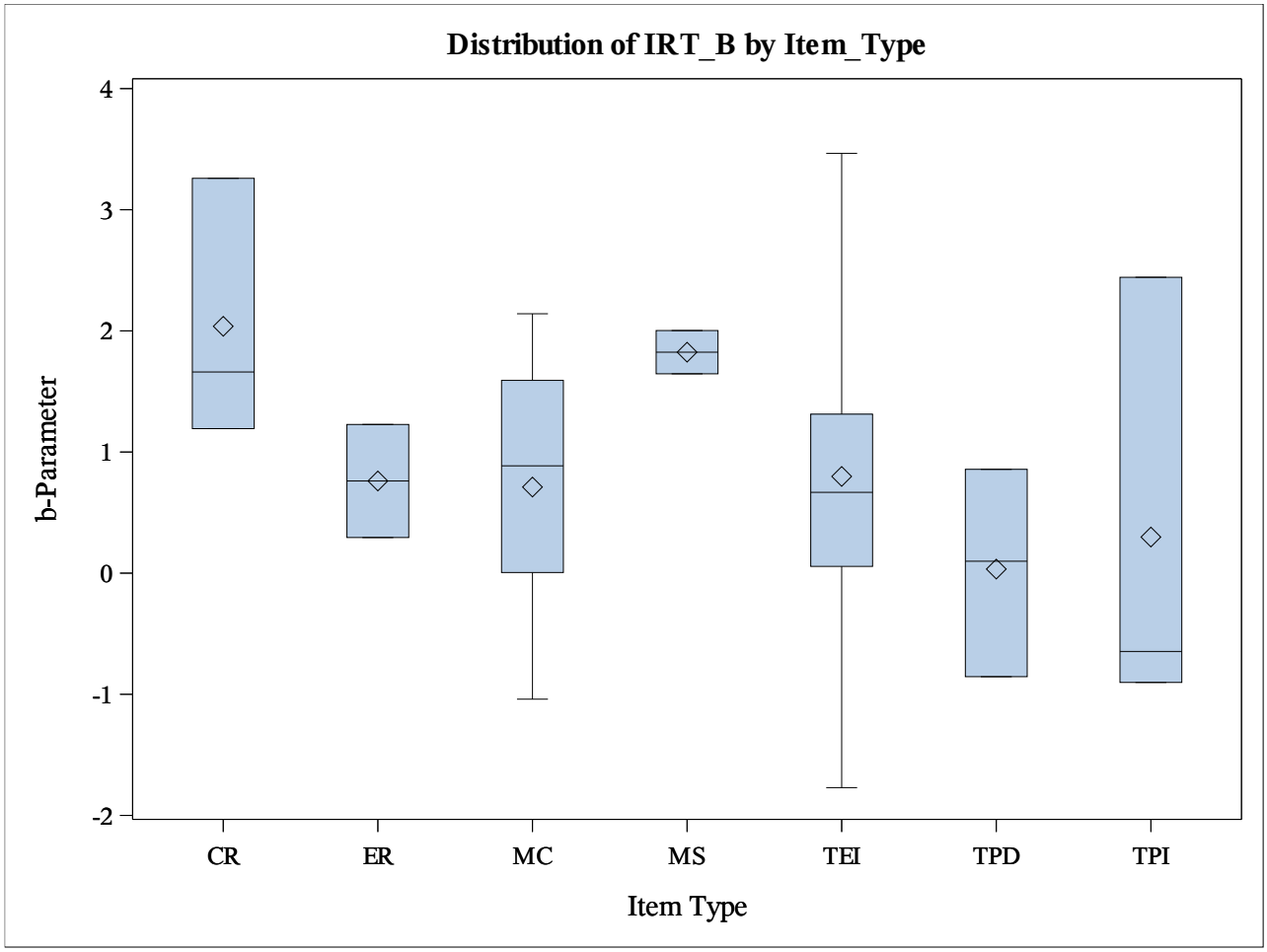
Plot C.5.1

IRT Item Parameter Summary for Spring 2022 Operational Biology: A-Parameter



Plot C.5.2

IRT Item Parameter Summary for Spring 2022 Operational Biology: B-Parameter



Plot C.5.3

IRT Item Parameter Summary for Spring 2022 Operational Biology: C-Parameter

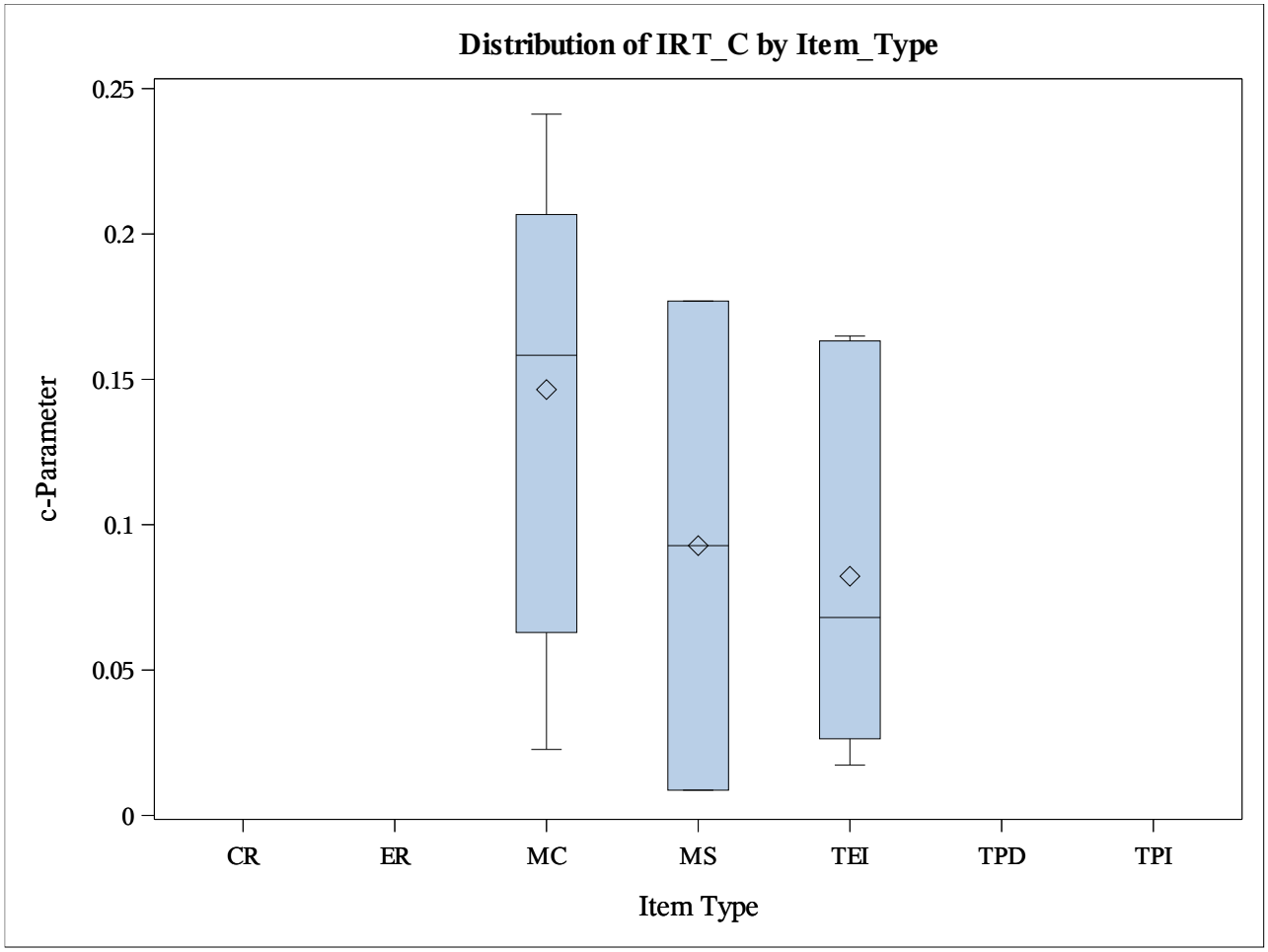


Table C.6

Statistically Flagged Operational Items: Spring 2022 Operational Biology

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	3	0	1	0
ER**	1	1	0	0	0
MC	11	0	1	1	0
MS	2	2	1	0	0
TEI	17	3	1	0	0
TPD	3	0	0	0	0
TPI	3	1	0	0	0

* The number of flagged DIF items includes both B and C DIF items.

** Classical analyses are calculated and estimated separately for each dimension of the ER item, and the result summarizes both dimensions.

Appendix D: Dimensionality

Dimensionality Reports
Biology

Contents
Table D.1 Zq1 Statistics and Summary Data: Spring 2022 Operational Biology
Table D.2 Q3 Statistics and Summary Data: Spring 2022 Operational Biology
Table D.3 Reporting Category Intercorrelation Coefficients: Spring 2022 Operational Biology
Table D.4 First and Second Eigenvalues: Spring 2022 Operational Biology Plot D.1 Principal Component Analysis: Spring 2022 Operational Biology

- Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table D.1

Zq1 Statistics and Summary Data: Spring 2022 Operational Biology

Form	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
D	CR	14.95	14.95	16.90	22.82	22.82	0
	ER	35.04	35.04	52.45	69.85	69.85	1
	MC	1.88	5.23	16.99	23.68	89.91	1
	MS	5.94	5.94	31.75	57.55	57.55	1
	TEI	2.34	6.14	14.85	15.89	145.78	3
	TPD	23.24	23.24	38.57	143.72	143.72	1
	TPI	25.34	25.34	36.73	74.05	74.05	1

Table D.2

Q3 Statistics and Summary Data: Spring 2022 Operational Biology

Form	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
D	0.174	-0.28	-0.126	-.002	0.155	0.378

Table D.3

Reporting Category Intercorrelation Coefficients: Spring 2022 Operational Biology

Reporting Category	Investigate	Evaluate	Reason Scientifically
Investigate	1.00		
Evaluate	0.60	1.00	
Reason Scientifically	0.59	0.78	1.00

Table D.4

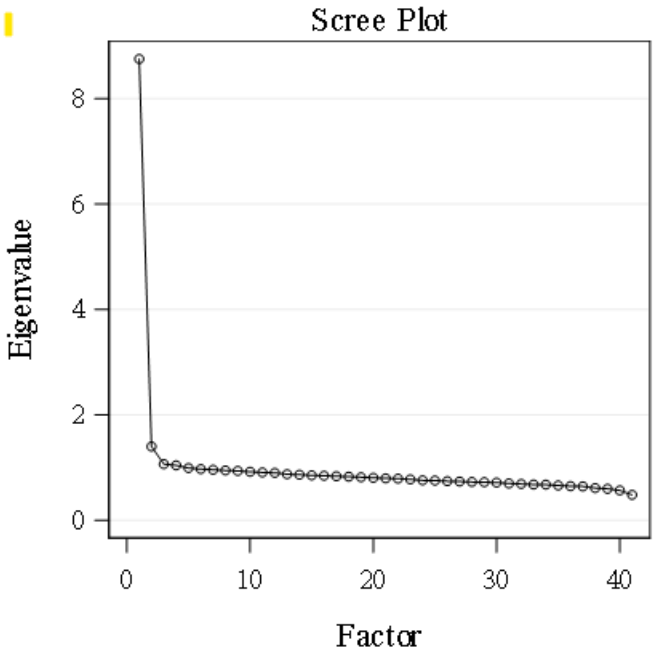
First and Second Eigenvalue: Spring 2022 Operational Biology*

Form	First Eigenvalue	Second Eigenvalue
D	8.753	1.399

* The ratio of first and second eigenvalues is about 5.565.

Plot D.1

Principal Component Analysis Plot: Spring 2022 Operational Biology



Appendix E: Scale Distribution and Statistical Report

Biology

Contents
Table E.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Biology
Table E.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Biology

- Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table E.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Biology

DESCRIPTIVE STATISTICS - SCALE SCORES
BIOLOGY
ALL STUDENTS

N	≥38820		
Mean	732.65	Median	734.00
Std deviation	25.64	Variance	657.60
Skewness	-0.1231	Kurtosis	-0.2347
Mode	734.00	Std Error Mean	0.1301
Range	200.00	Interquartile Range	37.00

Quantile	Estimate
100% Max	850
99%	787
95%	772
90%	765
75% Q3	751
50% Median	734
25% Q1	714
10%	700
5%	688
1%	670
0% Min	650

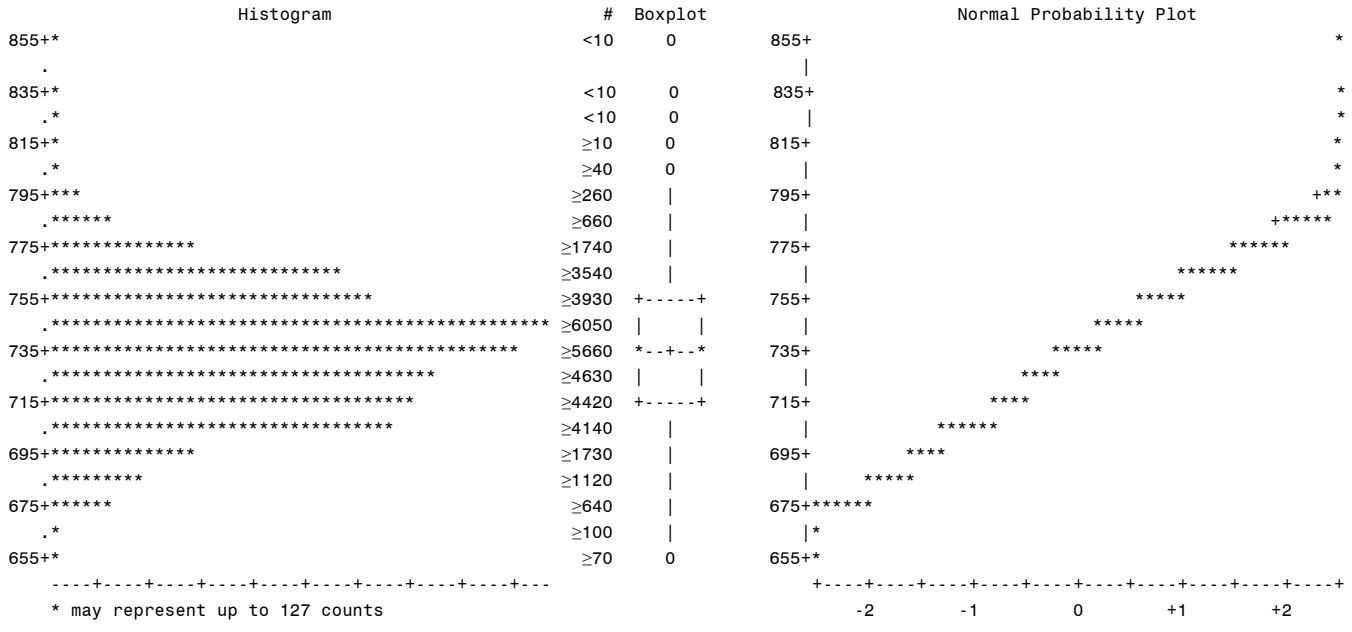
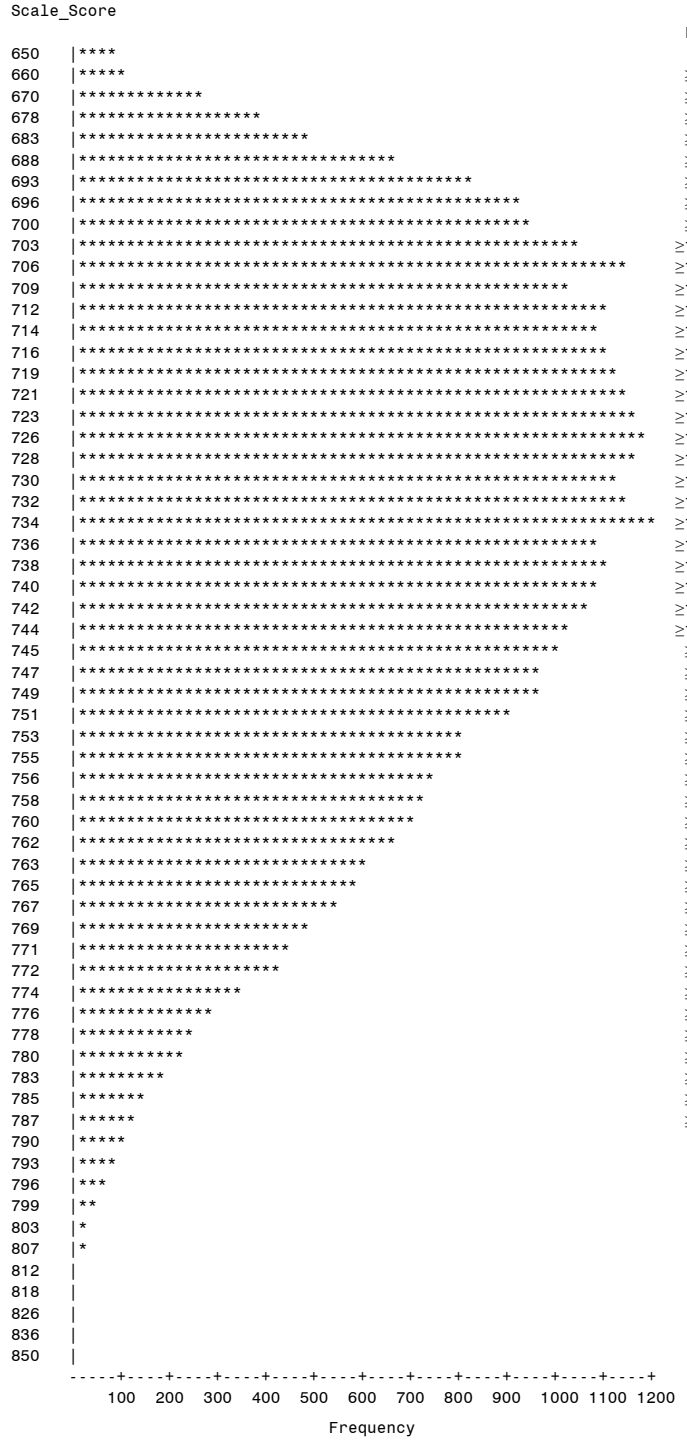


Table E. 2

Frequency Distribution of Scale Scores: Spring 2022 Operational Biology

Scale_Score	Freq	Cum. Freq	Percent	Cum. Percent
650	≥70	≥70	0.19	0.19
660	≥100	≥170	0.27	0.46
670	≥250	≥430	0.66	1.12
678	≥380	≥820	1.00	2.12
683	≥470	≥1290	1.23	3.35
688	≥650	≥1950	1.68	5.02
693	≥810	≥2760	2.11	7.13
696	≥910	≥3680	2.36	9.49
700	≥940	≥4630	2.44	11.93
703	≥1030	≥5660	2.67	14.60
706	≥1130	≥6800	2.93	17.53
709	≥1020	≥7830	2.64	20.17
712	≥1100	≥8930	2.85	23.02
714	≥1080	≥10020	2.80	25.82
716	≥1100	≥11120	2.84	28.66
719	≥1120	≥12250	2.91	31.57
721	≥1140	≥13390	2.94	34.51
723	≥1150	≥14550	2.97	37.48
726	≥1180	≥15730	3.04	40.53
728	≥1160	≥16890	2.99	43.52
730	≥1120	≥18020	2.91	46.42
732	≥1140	≥19160	2.94	49.37
734	≥1200	≥20370	3.10	52.47
736	≥1080	≥21450	2.80	55.26
738	≥1100	≥22560	2.85	58.11
740	≥1080	≥23640	2.79	60.91
742	≥1050	≥24700	2.72	63.62
744	≥1010	≥25710	2.60	66.22
745	≥990	≥26700	2.55	68.78
747	≥960	≥27660	2.49	71.26
749	≥950	≥28610	2.45	73.71
751	≥890	≥29500	2.29	76.01
753	≥790	≥30300	2.04	78.05
755	≥790	≥31100	2.06	80.11
756	≥730	≥31830	1.89	81.99
758	≥720	≥32550	1.86	83.85
760	≥690	≥33250	1.79	85.65
762	≥650	≥33900	1.69	87.33
763	≥600	≥34500	1.55	88.89
765	≥570	≥35080	1.48	90.36
767	≥530	≥35610	1.38	91.74
769	≥470	≥36090	1.23	92.97
771	≥440	≥36540	1.14	94.12
772	≥420	≥36960	1.09	95.20
774	≥340	≥37300	0.88	96.08
776	≥280	≥37590	0.74	96.83
778	≥240	≥37830	0.63	97.46
780	≥220	≥38050	0.57	98.03
783	≥170	≥38230	0.45	98.48
785	≥140	≥38370	0.37	98.85
787	≥120	≥38490	0.31	99.16
790	≥90	≥38590	0.24	99.41
793	≥70	≥38660	0.19	99.60
796	≥50	≥38720	0.14	99.74
799	≥40	≥38760	0.10	99.84
803	≥10	≥38780	0.04	99.89
807	≥20	≥38800	0.07	99.95
812	<10	≥38810	0.02	99.97
818	<10	≥38810	0.01	99.98
826	<10	≥38820	0.01	99.99
836	<10	≥38820	0.00	99.99
850	<10	≥38820	0.01	100.00



Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports Biology

Contents
Table F.1.Reliability and SEM for Overall and Subgroups: Spring 2022 Operational Biology
Table F.2.Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational Biology
Table F.3.Classification Accuracy and Decision Consistency: Spring 2022 Operational Biology

- Because the spring 2022 test was administered under the conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table F.1

Reliability and SEM for Overall and Subgroups: Spring 2022 Operational Biology

Subg	Reliability	SEM
All Students	0.900	3.72
Female	0.892	3.73
Male	0.870	4.38
African American	0.877	3.48
American Indian or Alaska Native	0.906	3.27
Asian	0.905	3.79
Hispanic/Latino	0.894	3.89
Multi-Racial	0.890	3.80
Native Hawaiian or Other Pacific Islander	0.886	3.94
White	0.870	4.03
Economically Disadvantaged: No	0.891	3.78
Economically Disadvantaged: Yes	0.886	3.65
English Learner: No	0.899	3.72
English Learner: Yes	0.824	3.31
Gifted or Talented	0.886	3.73
Regular Education	0.888	3.72
Special Education	0.861	3.37
Section 504: No	0.900	3.72
Section 504: Yes	0.893	3.60
Migrant: No	0.900	3.72
Migrant: Yes	0.877	3.60
Homeless: No	0.900	3.72
Homeless: Yes	0.885	3.61
Military Affiliation: No	0.900	3.71
Military Affiliation: Yes	0.895	3.76
Foster Care: No	0.900	3.72
Foster Care: Yes	0.873	3.53

Table F.2

Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational Biology

Administration	Cronbach's Alpha	Marginal Reliability
Spring 2022	0.900	0.91

Table F.3

Classification Accuracy and Decision Consistency: Spring 2022 Operational Biology

Accuracy Matrix: SPR 2022 Operational Biology

Form	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
D	1	0.14	0.02	0.00	0.00	0.00	0.17
	2	0.03	0.13	0.04	0.00	0.00	0.20
	3	0.00	0.05	0.28	0.05	0.00	0.37
	4	0.00	0.00	0.04	0.14	0.03	0.21
	5	0.00	0.00	0.00	0.01	0.03	0.04
	Total	0.18	0.20	0.36	0.20	0.06	1.00

Consistency Matrix: SPR 2022 Operational Biology

form	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
D	1	0.14	0.04	0.00	0.00	0.00	0.18
	2	0.04	0.10	0.06	0.00	0.00	0.20
	3	0.00	0.06	0.23	0.06	0.00	0.36
	4	0.00	0.00	0.06	0.12	0.03	0.20
	5	0.00	0.00	0.00	0.03	0.03	0.06
	Total	0.18	0.20	0.36	0.20	0.06	1.00

Table F.3.1

Estimates of Accuracy and Consistency of Achievement Level Classification

Form	Accuracy	Consistency	PChance	Kappa
D	0.723	0.62	0.245	0.497

Table F.3.2

Accuracy of Classification at Each Achievement Level

Form	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
D	0.852	0.626	0.741	0.681	0.725

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
D	0.941	0.908	0.91	0.961

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
D	0.917	0.871	0.874	0.945

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
D	0.715	0.725	0.674	0.506

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
D	0.025	0.044	0.049	0.026

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
D	0.034	0.048	0.041	0.012

Appendix G: Accommodated Print and Braille Creation

Guidelines for Accommodated Print and Braille

Louisiana believes that all students requiring test accommodations should be presented with the same rigor as students taking tests without accommodations. To ensure this, Louisiana accommodates the operational test form for each test administration, allowing all students to take the same items regardless of the need for an accommodated presentation. Careful consideration is given to all items that are used for Louisiana assessments for their ability to be faithfully represented in accommodated print (AP) and/or braille formats. Fairness for all populations, item integrity, and student-item interaction for technology-enhanced (TE) items are all factors when selecting the items that will appear on a Louisiana form. TE items are modified so that students who interact with an item on an AP or braille form will have a similar and equivalent experience to students who interact with that same item in the online environment. This maintains both the rigor and the content being assessed. Some examples of the modification process are provided below.

· Drag-and-drop items in the online environment require a student to place the answer options in an interactive table. For the AP and braille forms, the student is presented with a table with the same information as the interactive table (column or row headers, any completed cells, and blank spaces) and the answer options are listed below the table (similar to the online form in which the options are listed either below or to the right of the table). The directions are modified to ask the student to write the correct answer in its corresponding box. Students are also able to circle the text and draw arrows to indicate where it should be placed or add labels to the answer choices and write only the label in the box, as long as the intended response is clear to the test administrator who will transcribe the answers into the online system.

· Matching items in the online environment require a student to select a checkbox in one or more columns for each of multiple rows. In the AP and braille forms, the student is provided with a table and asked to mark an X in the correct places.

· Highlight-text items or item parts in the online environment require a student to click on the selected text, which highlights the selected word, phrase, or sentence. In the AP and braille forms, the text is presented in the same format and the student is asked to circle the answer. Where only certain words or phrases are selectable in the online system, those options are underlined in the AP and braille forms to indicate which words and/or phrases the student should select from.

· Drop-down menu items in the online environment have answer options in a drop-down menu format, oftentimes as part of a complete sentence. The AP and braille forms display the item with a blank line in place of the drop-down menu in the sentence, with all the answer options for the drop-down menu presented vertically below the sentence. The

directions are then modified to ask the student to circle the word/phrase that belongs in the blank.

· Short answer items in the online environment require a student to type the answer in a box. In the AP and braille forms, a box is provided for the student to write the response.

· Keypad input items in the online environment require a student to enter a numeric response including all rational and irrational numbers as well as expressions and equations. In the AP and braille forms, a box is provided for the student to write the response.

· Graphing items, including coordinate planes, number lines, line plots, and bar graphs, in the online environment require a student to complete a graph by plotting points, adding Xs to create a line plot, or raising/lowering bars to create a bar graph or histogram. In the AP and braille forms, the student is provided with the same coordinate plane, number line, line plot, or bar graph as in the online item, including titles, axis labels, and keys, and is asked to complete the graph.

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

AP forms are thoroughly reviewed by DRC and LDOE content experts, and braille forms are reviewed by an outside third-party braille expert. Students respond to their accommodated print and braille test using the same online test as used by the general population, either through use of a scribe or by themselves if able. This ensures a valid and reliable assessment for students who are unable to participate in the online assessment.

Appendix H: On-Going Quality Control

A system for monitoring, maintaining, and increasing the quality of its assessment system, including precise and technically sound criteria for the analyses of all of the assessments in its assessment system, is crucial and critical for keeping a high quality of assessments. The places where information about monitoring, maintaining, and improving quality is incorporated are included in the following table.

Related Information		Related Chapter/Source
Test Materials		
Item development quality procedures	Content alignment Cognitive complexity Bias, fairness, and sensitivity Technical design	Chapter 3
Form development quality procedures	Test specifications Review of statistical quality of items	Chapter 4
Test Administration		
Test administration training and procedures	Training and monitoring of test administrators Security Checklists Test Security Measurements	Chapter 5
Monitoring test administrations	LDOE site audits Data Forensics Analysis Response-Change Analysis Web Monitoring Plagiarism Detection	Chapter 5
Scoring		
Scorer recruitment, training and security procedures	Recruitment and interview process Security Training process, including material development and qualifying procedures.	Chapter 6
Monitoring scoring quality	Inter-rater reliability studies Validity Reader monitoring	Chapter 6
Psychometric Processes		
Psychometric quality procedures	Specifications document for operational analysis	Internal document between Pearson and the LDOE.
Monitoring psychometric quality	Key verification Calibration Scoring table generation Psychometric quality checks on the data	Chapter 7
Cuts based on Performance-Level Setting	Quality-controlled procedures for performance-level setting Derivation of the cut scores	Chapter 8