



Pearson



**LEAP 2025 Biology
Technical Report: 2017–2018
Field Test**

Prepared by DRC, Pearson, and WestEd

**LEAP
2025**



FOREWORD

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

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2009) and in the new edition, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Table of Contents

FOREWORD	ii
1. Introduction	1
Summary of the 2017–2018 Activities Timeline	2
2. Assessment Framework	3
3. Overview of the Test Development Process	6
Proposal and Review of Topics and Sources	12
PE Bundling.....	12
Phenomena Selection and Outline Development	12
Matching Phenomena to Set Types	13
Outline and Stimuli Development.....	14
Item Writing and Review Process.....	15
4. Construction of Embedded Test Forms	21
5. Test Administration	24
Training of School Systems.....	24
Ancillary Materials.....	25
Time	27
Online Forms Administration	28
Accessibility and Accommodations.....	28
Testing Windows	29
Test Security Procedures.....	29

6. Scoring Activities	30
7. Data Analysis	41
Classical Item Statistics	41
Differential Item Functioning	45
Item Calibration	48
Measurement Models	49
Field Test Item Parameters.....	50
Item Fit.....	50
8. Data Review Process.....	56
References.....	58
Appendix A: Training Agendas.....	61

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 (LSS). 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) administration of the statewide summative science assessment for high school Biology. This report outlines the testing procedures, including forms construction, administration, calibration, and analyses.

Summary of the 2017–2018 Activities Timeline

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 2018 Biology embedded field test forms. Table 1.1 summarizes those key activities along with the months during which the activities were completed.

Table 1.1
Key Activities from February 2017 to August 2018

Date	Activity
February 2017	<ul style="list-style-type: none"> • Started item development planning for spring 2018 EFT • Item development plans approved by LDOE staff
March 2017	<ul style="list-style-type: none"> • Content development specifications and style guide prepared
March–April 2017	<ul style="list-style-type: none"> • WestEd began item writing and development • Topics and outlines approved by LDOE staff
May–August 2017	<ul style="list-style-type: none"> • LDOE staff reviewed proposed content
June 2017	<ul style="list-style-type: none"> • Spring 2018 Framework and Test Construction Document proposed
August 2017	<ul style="list-style-type: none"> • LDOE and WestEd Planning Meeting held
September 2017	<ul style="list-style-type: none"> • Spring 2018 Framework and Test Construction Document approved
October 2017	<ul style="list-style-type: none"> • Item Content/Bias Review Committee convened • Reconciliation meeting held between LDOE and WestEd staff • Test construction activities began
November 2017	<ul style="list-style-type: none"> • Technical Advisory Committee Meeting convened • LDOE staff reviewed proposed spring 2018 EFT selections
December 2017	<ul style="list-style-type: none"> • Online content delivered to administration vendor
January 2018	<ul style="list-style-type: none"> • Remaining spring 2018 materials delivered to administration vendor
April 2018	<ul style="list-style-type: none"> • Spring 2018 Embedded Field Test administered
May–June 2018	<ul style="list-style-type: none"> • Rangefinding meetings held
August 2018	<ul style="list-style-type: none"> • Data Review meeting held

2. Assessment Framework

The development of an assessment framework is one of the key deliverables for this scope of work. The Request for Proposal (RFP) specifies that the framework is to include the test design, test blueprint, range of standards covered, reporting categories, percentages of assessment items and score points by reporting category, projected testing times, and numbers of forms to be administered, as well as psychometric analysis activities.

Measuring student proficiency of the full depth and breadth of the Louisiana Student Standards for Science requires assessments built from a range of item types. As a general rule, specific item type usage depends on the most efficient and effective measurement of the target content. Multiple-choice (MC) and multiple-select (MS) item types, the most commonly used item types, provide students the opportunity to select the correct answer or answers from a set of answer choices. Multiple-select items allow students to demonstrate a greater depth of understanding than traditional MC items by requiring students to select more than one correct response, scored automatically through the vendor's scoring engine. Constructed-response (CR) and extended-response (ER) items allow students to demonstrate their ability 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). Because students write a response in their own words, teams of vendor-trained readers hand-score student responses. Technology-enhanced (TE) items allow students to demonstrate their ability to apply and communicate scientific knowledge and understanding as required by the SEPs and CCCs in a way not best achieved through MC or MS, but more cost-effective and less time-consuming than CR and ER because the vendor scores the TE items automatically within the scoring engine. TE items may ask students to develop models or 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 using different TE item types. The complexity of the TE items reduces the probability of guessing the correct answer. Two-part items allow students to apply their understanding of different but related knowledge to a concept or to support their assertions with evidence.

In two-part items, students may construct an explanation and then support that explanation with evidence, or engage in scientific argumentation by making a claim and evaluating the evidence to support that claim. Another useful application of two-part items is for students to develop a model in part A and evaluate the model in part B. Finally, a range of item types supports greater engagement on the part of the test takers, facilitating a more authentic assessment experience.

The test design includes item sets, a task, and standalone items. A stimulus that describes a scientific phenomenon provides the anchor for each item set or task. A focus that details some aspects of a phenomenon provides the anchor for standalone items. Item sets are made up of four items tied to a common stimulus. The item sets may include 1-point selected-response items (both single-select and multiple-select formats), 1- and 2-point technology-enhanced items, and 2-point two-part items (two-part independent and two-part dependent formats). Three of the item sets also include a 2-point constructed-response item. In addition to the item sets, the assessment contains one task set. The tasks are made up of five items tied to a common stimulus. Tasks may include 1-point selected-response items (both single-select and multiple-select formats), 1- and 2-point technology-enhanced items, 2-point two-part items (two-part independent and two-part dependent formats), and a 9-point extended-response (ER) item. Standalone items may be either 1-point selected-response items (both single-select and multiple-select formats), 1- and 2-point technology-enhanced items, or 2-point two-part items (two-part independent and two-part dependent formats). The standalone items provide greater flexibility to meet the test blueprint and greater coverage of the standards while still requiring students to make connections among the three dimensions of the Louisiana Student Standards for Science (LSSS). All points associated with the task set contribute to students' overall scores, but the 9-point ER does not contribute to the blueprint. This prevents the ER from impacting the proportional representation of content assessed by other parts of the test.

The assessment is administered primarily online, and technology-enhanced (TE) items are included in the test design. However, an accommodated paper version of the assessment is made available for students who are unable to test online. For that form, TE items are adapted to a paper format while still assessing the same content.

In addition to the test design and blueprint, the initial Assessment Framework contained a plan for field testing the newly developed items. The LDOE administered the newly

developed items for the LEAP 2025 Biology assessment embedded within the existing End-of-Course (EOC) Biology assessment. The embedded field test items included a task in session 2, and an item set and standalone items in sessions 1 and 3. Thus, the field test design included the range of item types (tasks, item sets, standalone items) that would appear on the LEAP 2025 Biology operational form beginning with the 2018–2019 school year. The construction of the embedded field test forms is addressed in Section 4 of this report.

The previous EOC Biology assessment was aligned to prior science academic content standards, and reported student performance according to four achievement levels. The Biology assessment developed for the LEAP 2025 aligns to the Louisiana Student Standards for Science adopted in 2017, and reports student performance according to five achievement levels.

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

3. Overview of the Test Development Process

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

Once the test design and the blueprint were approved, the item development plan was established. Following are the test blueprint components that guided initial item development projections.

Table 3.1b
Test Blueprint for LEAP 2025 Biology: DCI Domain Coverage

Biology: DCI Domain Coverage			
	# of PEs in LSS	Relative % in LSS	% 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 LSS	Relative % in LSS	% 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 LSS	Relative % in LSS	% 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 Subclaim Coverage

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

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

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		20	30
Standalone items	1	16	22	10	6		16	22
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	14	10	1	41	67

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

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

item development plan for the number of items developed by WestEd by reporting category.

Table 3.2

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

	Total Number of Sets	1 pt SRs	1pt TEs	2 pt TEs	TPD/TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	16	41	41	30	30	0	20	142
Tasks	6	12	12	12	12	6	0	48
Standalone Items	n/a	25	17	17	17	0	0	76

Note: assessment guide items and practice test items are not included in this table.

Proposal and Review of Topics and Sources

PE Bundling

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

LDOE approved 28 bundles for the Biology assessment. Of these bundles, 22 were targeted for development in the 2017–2018 cycle. Two were later put on hold for use in other contexts.

Phenomena Selection and Outline Development

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

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

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

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; and
- whether the phenomenon fit with the “PE bundles” developed earlier to provide meaningful, three-dimensional assessment of performance expectations.

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 as many PEs as possible in the field test. As part of the item development plan, all standards listed in the blueprint were aligned to at least one standalone item or an item in an item set or task.

Matching Phenomena to Set Types

Sets were purposefully designated as item sets or tasks, and the designation of the set (whether item or task) influenced the selection of phenomena. The tasks were based on stimuli that allowed students to delve deeply into a topic and were made up of four items that built upon each other to lead to a culminating ER item. The items in a task could require a specific order, and information in one item could be used in other items (although the items did not cue each other). Additionally, the items could be scaffolded to help discriminate student performance levels. The ER was three-dimensional; however, it could mix and match among the dimensions from the PE bundle to achieve this three-dimensionality. In total, six ERs were developed for six tasks. Like the tasks, the item sets were phenomena-based, but unlike the tasks, they comprised independent items that did not build upon each other. Although an item set does not need to contain a constructed-response (CR) item, for the 2017–2018 development cycle, WestEd developed CRs for all

item sets and for every reporting category. In total, 20 CRs were developed for 16 item sets.

For tasks and item sets, WestEd offered a document containing descriptions of 56 phenomena associated with bundles to the LDOE for its review prior to item development. Based on the list, the LDOE identified the 20 phenomena to be developed into stimuli for the task and item sets. 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 and item sets began.

In contrast to item sets and tasks, standalone items reflected independent content and are supported by a focus. A focus differs from a phenomenon in that it explores only certain key aspects of an event and is typically supported by less data. As stated previously, the standalone items were included within the blueprints to provide greater coverage of the standards assessed and to provide flexibility in meeting the blueprints and test characteristic curve targets across test administrations. The WestEd content lead developed the foci for standalone items, based on standards that lacked coverage across the item sets and tasks. Consequently, these items were developed last. For standalone items, WestEd submitted the items and corresponding foci simultaneously; there was no separate focus approval phase for these items.

Outline and Stimuli Development

WestEd used both experienced internal and external science assessment editors to develop the phenomena-based stimuli for item sets and tasks. 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 (2016–2017).

To support the outline development process, writers were given the Louisiana Student Standards for Science. They were also provided specific item set or task 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 and task 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 glossing was added for words that were above grade level and for words or phrases that were thought to be sources of potential confusion for students. The appropriateness of the stimuli for both content and readability was an explicit part of the content review process with Louisiana teachers.

Item Writing and Review Process

WestEd employed a cadre of item writers for the Biology assessment. All writers' resumes were reviewed and approved by the LDOE before engaging in any item development activities. As the first step in the item writing process, the WestEd content lead provided a webinar training to all writers in March 2017. For an outline of the information covered, see Appendix A for the LEAP 2025 Biology Item Training Agenda (March 2017). 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 tasks and 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 or task should be developed. The use of item set and task 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 set or task also specified the set type, the item types (e.g., SR, MS, TE, TPI, TPD, CR, ER), 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 and ER 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 bias/sensitivity training. 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 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 the LEAP 2025 Biology Training Agenda (2016–2017). 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 of 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. Initial style guidelines were based on documentation established with the LEAP 2025 Social Studies and U.S. History 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, tasks, and standalone items were completed, these batches were sent to the LDOE for review by the LDOE Science Assessment Coordinators; Director of Assessment Development for Math, Science, and Special Populations; Elementary Assessment Coordinator; Special Populations Assessment Coordinator; and Science Program Coordinator. Feedback from the LDOE review was implemented before the content and bias review meetings.

Content and Bias Review. After the completion of item development, WestEd coordinated face-to-face content and bias review meetings, convened in Baton Rouge. The meetings were led by facilitators from the LDOE and from WestEd. Participants included current classroom teachers, retired teachers, content specialists, and school administrators. For both content and bias review meetings, participants completed nondisclosure agreements as part of the activities. The recruitment process, conducted by LDOE staff, included participants from regions across the state. Participants represent the population of Louisiana students served—including special education, English learners, and students with disabilities—as well as the diverse geographic and demographic composition of the state. Because the content and bias review meeting took place over five days, committee members could not participate every day of the meeting. As a result, WestEd and the LDOE separated the meeting into two parts, and had participants attend the meeting in the first half of the week or the second half of the week. Consequently, most of the individual participants did not review or discuss every item, although every item was reviewed by a committee. Table 3.3 provides the demographic characteristics of the review committee.

Table 3.3

Representation of Educators Participating in 2017–2018 Content and Bias Reviews

Characteristic	Number of Participants
Classroom Teacher	12
Content/ Curriculum Specialist	1
School Administrator	
Other Staff	1
ELL Teacher	0
Special Education Teacher	0
Special Ed Teacher-Gifted	0
Visually or Hearing Impaired Teacher	1
Black or African American	2
Asian	0
Hispanic/ Latino	1
White	9
Male	3
Female	10
Total Participants	13

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

Before the committee members began the item review process, they received an orientation from the LDOE about the new 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 the items and voted in ABBI on whether to accept, accept with edits, or reject each item. (If participants skipped an item or chose not to record a decision for a given

item, the system registered the response as “No Vote” for that individual review. “No Vote” was recorded as the consensus rating when an initial group decision on an item was not reached, and the committee failed to return to that item and register a final vote to accept, revise, or reject the item.) Participants used personal laptops or laptops provided by WestEd to access ABBI. At the end of each day, WestEd made certain that the participants cleared their computer caches and deleted their download histories for the day. WestEd monitored participants to be sure that they did not use their cell phones at the table. WestEd also collected all materials at the end of each day, including notepads provided to the participants to write notes on as they reviewed the items.

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

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

Table 3.4
Vote Totals Based on Individual Votes Following Initial Review

Item Type	Number of Items	Votes to Accept	Votes to Accept with Edits	No Vote	Votes to Reject	Total Votes
CR	20	203	39	0	1	243
ER	6	70	3	2	0	75
MC	110	1091	240	12	8	1351
MS	21	205	44	4	1	254
TE	87	809	235	6	16	1066
TPD	31	292	80	5	1	378
TPI	13	117	38	2	1	158
All Biology	288	2787	679	31	28	3525

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, only four items were rejected. The others were either accepted as is or accepted with edits. None of the item sets or tasks were rejected by the committee.

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

4. Construction of Embedded Test Forms

As noted previously, the LEAP 2025 Biology items were embedded within an existing EOC Biology test form previously administered by DRC. The primary purpose of the field test was to obtain data to inform construction of the operational test forms, but the field test also provided exposure for students to the variety of item types and formats to be used in the LEAP 2025 Biology assessments.

Items were embedded within three sessions. Session two was designated as a separate session for the field test of the task set. Within sessions one and three, an item set and standalone items were field tested. To maximize the possibility of the item sets and task sets succeeding, two versions of each item set and task set were field tested. However, due to the number of field test versions and the limited number of task sets, each version of the task set appeared on two forms. Several of the standalone items were repeated across forms. The field test content was distributed across the forms such that the field test item sets and task sets did not reflect overlapping content. A total of 16 field test forms were created. Table 4.1 shows the sessions along with the types and numbers of operational and field test items that appeared in each session.

Table 4.1

Embedded Field Test Design for Biology

Test Session	Number of Items
Session 1: Operational standalone SR items	23 OP standalone SR items
One FT item set	1–3 FT item set SR item 0–3 FT item set TE item(s) 0–2 FT TPI/TPD item set TPI/TPD item 0–1 FT item set CR item
FT standalone SR items	2–3 FT standalone items
Session 2: FT task	1–4 FT task set SR items 0–3 FT task set TE item(s) 1 FT task set ER item
OP task	2 OP task set SR items 1 OP ER item
Session 3: Operational standalone SR items	23 OP standalone SR items
One FT item set	1–3 FT item set SR items 0–3 FT item set TE item(s) 0–2 FT TPI/TPD item set TPI/TPD item 0–1 FT item set CR item
FT standalone SR items	2–3 FT standalone items
Total Operational Items Tested across Forms for Biology	46 standalone SR items 2 OP task set SR items 1 OP ER item
Total Items Field Tested across Forms for Biology	6 task sets, 16 item sets, 16 CRs, 45 standalone SR items, 19 standalone TE items, 11 standalone TPD/TE items

The WestEd content lead made a concerted effort to avoid cueing and clanging between field test and operational items within sessions as the items were assigned to forms. Cueing occurs when content in one item provides clues to the answer of another item. Clanging refers to overlap or similarity of content. The content lead conducted a separate review of the forms to check for inadvertent cueing or clanging as part of the forms construction quality control process.

Following the final item placement by WestEd content leads, test maps containing each item's unique identification number (UIN) were created. The test maps captured details about each proposed form, including sessions, item sequences, UINs, 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. All constructed EFT forms were reviewed by the LDOE Science Assessment Coordinators and Research Analysts, and item changes and edits were implemented as requested. Item content was not delivered to DRC until approval by the LDOE was achieved.

5. Test Administration

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

Training of School Systems

To ensure that LEAP 2025 HS assessments are administered and scored in accordance with the Department’s mandates, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE offers monthly webinars, and weekly office hours to school system testing coordinators to communicate with and train school systems. The LDOE provides train-the-trainer opportunities for the school system test coordinators, who in turn convey test administration training to schools within their systems. The LDOE conducts quality-assurance visits during testing to ensure school system adherence to the standardized administration of the tests.

The school system test coordinators are responsible for the schools within their systems. They disseminate information to each school, offer assistance with test administration, and serve as liaisons between the LDOE and their school system. The LDOE also provides assistance with and interpretation of assessment data and test results.

Ancillary Materials

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

For the spring 2018 test administration, DRC produced an administration manual, the *High School Test Administration Manual* (TAM), which serves for the EOC and LEAP 2025 administrations.

DRC also produced a *Test Coordinator Manual* (TCM). LDOE assessment staff review, provide feedback, and give final approval for these manuals. The manuals are inclusive of all EOC and LEAP 2025 HS assessments in ELA, mathematics, social studies, and science. They provide detailed instructions for school systems and school test coordinators' responsibilities for distributing and collecting test materials for the following programs and for returning them to DRC when appropriate.

The test administration manual provides detailed instructions for administering the EOC and LEAP 2025 HS assessments. The manual includes instructions for test security, test administrator responsibilities, test preparation, administration of online tests, and post-test procedures.

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

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

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

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

Furthermore, the standardized procedures addressed in the TAM need to be followed, as the *Standards* state in Standard 6.1: “Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user” (114). To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the EOC was administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The test administration manuals included instructions for scheduling the test within the state testing window. The test administration manual also contained the schedule for timing each test session.

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

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

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

Test administration manuals outline the steps that teachers should take to prepare classroom environment testing for administering the LEAP 2025 online test. These include the following:

- Determine the layout of the classroom environment.
- Plan seating arrangements. Allow enough space between students to prevent the sharing of answers.

- Eliminate distractions such as bells or telephones.
- Use a Do Not Disturb sign on the door of the testing room.
- Make sure classroom maps, charts, and any other materials that relate to the content and processes of the test are covered or removed or are out of the students' view.

Standard 6.6. Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)

The test administration manuals present instructions for post-test activities to ensure that online tests are submitted and printed test materials are handled properly to maintain the integrity of student information and test scores. Detailed instructions guide test examiners in submitting all online test records. For students who were administered a braille version of the EOC/LEAP 2025 assessment, examiners are instructed to transcribe students' responses from the braille test book into the online testing system (INSIGHT) exactly as they responded in the braille test book.

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

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

Time

All sessions of the LEAP 2025 assessments were timed. Only students with an extended time accommodation were permitted to exceed the established time limits of any given session.

Online Forms Administration

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

Accessibility and Accommodations

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

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

Accommodations may be used with students who qualify under the Individuals with Disabilities Education Act (IDEA) and have an IEP or Section 504 of the Americans with Disabilities Act and have a Section 504 plan, or who are identified as an English learner (EL).

Accommodations must be specified in the qualifying student’s individual plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student information sheet at the time of test administration. AERA, APA, and NCME Standard 6.2 states:

When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (115)

In compliance with this standard, the LEAP Test Administration Manual contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP assessments. The following accommodations were offered for this administration:

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

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

Testing Windows

Field test items were administered during online testing, which was available from Monday, April 23, through Friday, May 18, 2018.

Test Security Procedures

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

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

6. Scoring Activities

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

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

Scoring of TE Items and Adjudication. TEs were processed through DRC's autoscoring engine and scored as tests were processed according to the assigned scoring rules as established during content creation. DRC's technology-enhanced scoring process included the following procedures:

- A scoring rubric was created for each technology-enhanced item. The rubric described the one and only correct answer for dichotomously scored items (i.e., items scored as either right or wrong). If partial credit was possible, the rubric described in detail the type of response that could receive credit for each score point.
- The information from the scoring rubric was entered into the scoring system within the item banking system so that the truth resided in one place along with the item image and other metadata. This scoring information designated specific information that varied by item type. For example, for a

drag-and-drop item, the information included which objects are to be placed in each drop region to receive credit.

- The information was then verified by another autoscoring expert.

After testing started, reports were generated that showed every response, how many students gave that response, and the score the scoring system provided for that response.

The scoring was then checked against the scoring rubric using two levels of verification. If any discrepancies were found, the scoring information was modified and verified again. The scoring process was then rerun. This checking and modification process continued until no other issues were found.

As a final check, a final report was generated that showed all student responses, their frequencies, and their received scores.

The adjudication process focuses on detecting possible errors in scoring for TE items. For adjudication, DRC provided a report listing the frequency distributions of TE item responses and an auto-frequency report detailing the multi-part multi-select items. Members of the LDOE and WestEd content staff examined the TE item response distributions and the auto-frequency reports to evaluate whether the items were scored appropriately.

No TE item scoring issues were identified. Had issues been identified, the recommended changes to the scoring algorithm would have been applied, and DRC would have rescored the item.

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

Selection of Scoring Evaluators. Standard 4.20 states the following:

The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy

and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring. (92)

The following sections explain how scorers were selected and trained for the LEAP 2025 Biology handscoring process and describe how the scorers were monitored throughout the handscoring process.

The Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2017–2018 LEAP 2025 high school Biology test responses had at least a four-year college degree in an appropriate field, such as a bachelor's degree in a STEM field.

DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, or recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to demonstrate their proficiency in writing by responding to a DRC writing topic and their proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and also demonstrate his or her ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

Each DRC scoring center is a secure facility. All employees are issued photo identification badges and are required to wear them in plain view at all times. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials may leave the scoring center and must sign legally binding confidentiality agreements before work begins. DRC retains these

agreements for the duration of the contract. To prevent the unauthorized duplication of secure materials, cell phone and camera use within the scoring rooms is strictly forbidden. Readers only have access to the student responses they are qualified to score. Each scorer is assigned a unique username and password to access the DRC imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

Handscoring Training Process. Standard 6.9 specifies:

Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (118)

Training Material Development. DRC scoring supervisors trained scorers using LDOE-approved training materials. These materials were developed by DRC and LDOE staff from a selection scored by Louisiana educators at rangefinding and include the following:

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

Training and Qualifying Procedures. Handscoring involves training and qualifying team leaders and evaluators, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities. LDOE visits the scoring centers to review training materials and oversee the training process. An explanation of the training and qualification procedures follows.

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

Table 6.1

Biology Training Set Composition

Set Type	Item Training	Annotated
Anchor Set	10 responses (2 responses per score points 2-4, 3 responses per score point 1, and one paper for score point 0)	Yes
Practice Set 1	10 responses representing the range of responses	Yes
Practice Set 2	10 responses representing the range of responses	Yes
Practice Set 3	10 responses representing the range of responses	
Practice Set 4	10 responses representing the range of responses	
Qualifying Set 1	10 responses comparable to the anchor set responses	No
Qualifying Set 2	10 responses comparable to the anchor set responses	No

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

Any scorer who did not qualify by the end of the qualifying process for an item was not allowed to score live student responses. The Qualifying Standard for the 0-4 point rubric was 80% on one of two sets.

Monitoring the Scoring Process. Standard 6.8 states:

Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (118)

The following section explains the monitoring procedures that DRC uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored. Detailed scoring rubrics, which specify the criteria for scoring, are available for all constructed-response items.

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

Validity Sets and Inter-Rater Reliability. In addition to the feedback that supervisors provided to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducted validity scoring using 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 pre-determined scores and were used to help detect potential room drift as well as individual scorer drift. This data was used to make decisions regarding the retraining and/or release of scorers, as well as the rescoring of responses.

Approximately 10% of all live student responses were scored by a second reader to establish inter-rater reliability statistics for all constructed-response items. This procedure is called a "double-blind read" because the second reader does not know the first reader's score. DRC monitored inter-rater reliability based on the responses that were scored by two readers. If a scorer fell below the expected rate of agreement, the team leader or scoring director retrained the scorer. If a scorer failed to improve after retraining and feedback, DRC removed the scorer from the project. In this situation, DRC removed all scores assigned by the scorer in question. The responses were then reassigned and rescored.

To monitor inter-rater reliability, DRC produced scoring summary reports on a daily basis. DRC’s scoring summary reports display exact, adjacent, and nonadjacent agreement rates for each reader. These rates are calculated based on responses that are scored by two readers.

- Percentage Exact (%EX)—total number of responses by reader where scores are the same, divided by the number of responses that were scored twice
- Percentage Adjacent (%AD)—total number of responses by reader where scores are one point apart, divided by the number of responses that were scored twice
- Percentage Nonadjacent (%NA)—total number of responses by reader where scores are more than one score point apart, divided by the number of responses that were scored twice

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

Table 6.2

Agreement Rate Requirements for Validity and Inter-Rater Reliability

Subject	Score Point Range	Perfect Agreement	Perfect Agreement + Adjacent
Biology	0-4	80%	100%

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 if a certain type of response was missing from initial training). These calibrations were designed to help refocus scorers on how to properly use the scoring guidelines. They were selected to help illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After a reader scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and the anchor responses to explain the reasoning behind each response’s score.

Reports and Reader Feedback. Reader performance and intervention information were recorded in reader feedback logs. These logs tracked information about actions taken with individual readers to ensure scoring consistency in regard to reliability, score point distribution, and validity performance. In addition to the Reader Feedback Logs, DRC provides LDOE with handscoring quality control reports for review throughout the scoring window.

Inter-Rater Reliability. A minimum of 10% of the constructed responses in Biology were scored independently by a second reader. The statistics for the inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the first and second score was examined.

Tables 6.3–6.6 provide the inter-rater reliability and score point distributions by grade level for the constructed-response and extended-response field test items administered in the spring 2018 forms.

Table 6.3

Constructed-Response Inter-Rater Reliability

Grade	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
HS	Item1	≥310	75	25	1
HS	Item2	≥300	85	13	1
HS	Item3	≥330	79	19	1
HS	Item4	≥320	84	15	1
HS	Item5	≥320	88	12	0
HS	Item6	≥330	81	19	1
HS	Item7	≥310	76	23	1
HS	Item8	≥350	84	16	0
HS	Item9	≥320	86	14	0
HS	Item10	≥350	93	7	0
HS	Item11	≥330	86	13	1
HS	Item12	≥330	87	13	0
HS	Item13	≥370	89	11	0
HS	Item14	≥360	96	4	0

*Total Exact+ Adjacent+ Non-adjacent does not always add up to 100% due to rounding

Table 6.4

Constructed-Response Score Point Distributions

Grade	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
HS	Item1	≥1,620	46	31	22	0
HS	Item2	≥1,570	60	19	20	0
HS	Item3	≥1,600	62	23	12	0
HS	Item4	≥1,620	65	20	14	0
HS	Item5	≥1,600	66	24	8	0
HS	Item6	≥1,560	58	29	10	0
HS	Item7	≥1,620	57	33	10	0
HS	Item8	≥1,590	73	20	3	0
HS	Item9	≥1,630	54	33	13	0
HS	Item10	≥1,590	64	28	4	0
HS	Item11	≥1,620	49	27	22	0
HS	Item12	≥1,600	34	24	40	0
HS	Item13	≥1,600	56	31	11	0
HS	Item14	≥1,600	82	14	3	0

Table 6.5

Extended-Response Inter-Rater Reliability

Item	Inter-Rater Reliability				
	2x	Part	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
Item1	≥4,880	N/A	61	25	14
Item2	≥4,850	N/A	69	28	3
Item3	≥4,890	Part A (0-5)	67	25	9
		Part B (0-4)	66	22	13
Item4	≥4,840	Part A (0-3)	69	29	2
		Part B (0-6)	61	26	14
Item5	≥5,140	Part A (0-6)	82	15	3
		Part B (0-3)	84	13	3

*Total Exact+ Adjacent+ Non-adjacent does not always add up to 100% due to rounding

Table 6.6

Extended-Response Score Point Distributions

Item	Score Point Distribution												
	Total	Part	% "0" Rating	% "1" Rating	% "2" Rating	% "3" Rating	% "4" Rating	% "5" Rating	% "6" Rating	% "7" Rating	% "8" Rating	% "9" Rating	% Blank
Item 1	≥4,880	N/A	31	26	16	9	7	6	3	1	1	0	0
Item 2	≥4,850	N/A	36	40	18	4	1	0	0	0	0	0	0
Item 3	≥4,890	Part A (0-5)	33	26	11	9	10	11					0
		Part B (0-4)	43	14	30	7	5						0
Item 4	≥4,840	Part A (0-3)	9	34	36	20							0
		Part B (0-6)	36	19	21	8	10	2	3				0
Item 5	≥5,140	Part A (0-6)	47	13	13	15	2	1	2				6
		Part B (0-3)	36	30	12	16							6

7. Data Analysis

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). The item-total correlation (point-biserial) is a measure of item discrimination. Items with higher item-total correlations provide better information about overall student ability (i.e., they discriminate between lower- and higher-ability students). Table 7.1 summarizes these item statistics by item type.

Table 7.1
Summary of Classical Statistics for Field Test Items for Biology

Item Type	N Items	p -value Mean	p -value SD	Item-Total Correlation Mean	Item-Total Correlation SD	Percent with B-level DIF	Percent with C-level DIF
MC	105	0.45	0.17	0.31	0.12	3%	0%
MS	20	0.17	0.11	0.22	0.12	1%	0%
CR	16	0.13	0.07	0.48	0.06	1%	0%
ER	6	0.09	0.07	0.56	0.17	0%	0%
TE	79	0.43	0.18	0.41	0.13	5%	1%
TPI	12	0.35	0.19	0.39	0.14	0%	0%
TPD	30	0.35	0.13	0.42	0.10	0%	0%

Table 7.2 summarizes the numbers of Biology field-tested items that were flagged according to defined criteria. The box plots that follow in Figures 7.1 and 7.2 illustrate the range of item p -values and good item–total discriminating power by item type.

Table 7.2
Number of Field Test Items Flagged for Item Statistics

Item Type	<i>N</i> Items	Flagged for p -value	Flagged for Mean	Flagged for Point-Biserial Correlation	Flagged for DIF
CR	16	0	16	2	2
ER	6	0	3	1	0
MC	105	12	0	19	8
MS	20	16	0	9	2
TE	79	9	2	4	15
TPD	30	0	6	0	1
TPI	12	0	6	2	0

Figure 7.1
Box Plot of Item P-Values

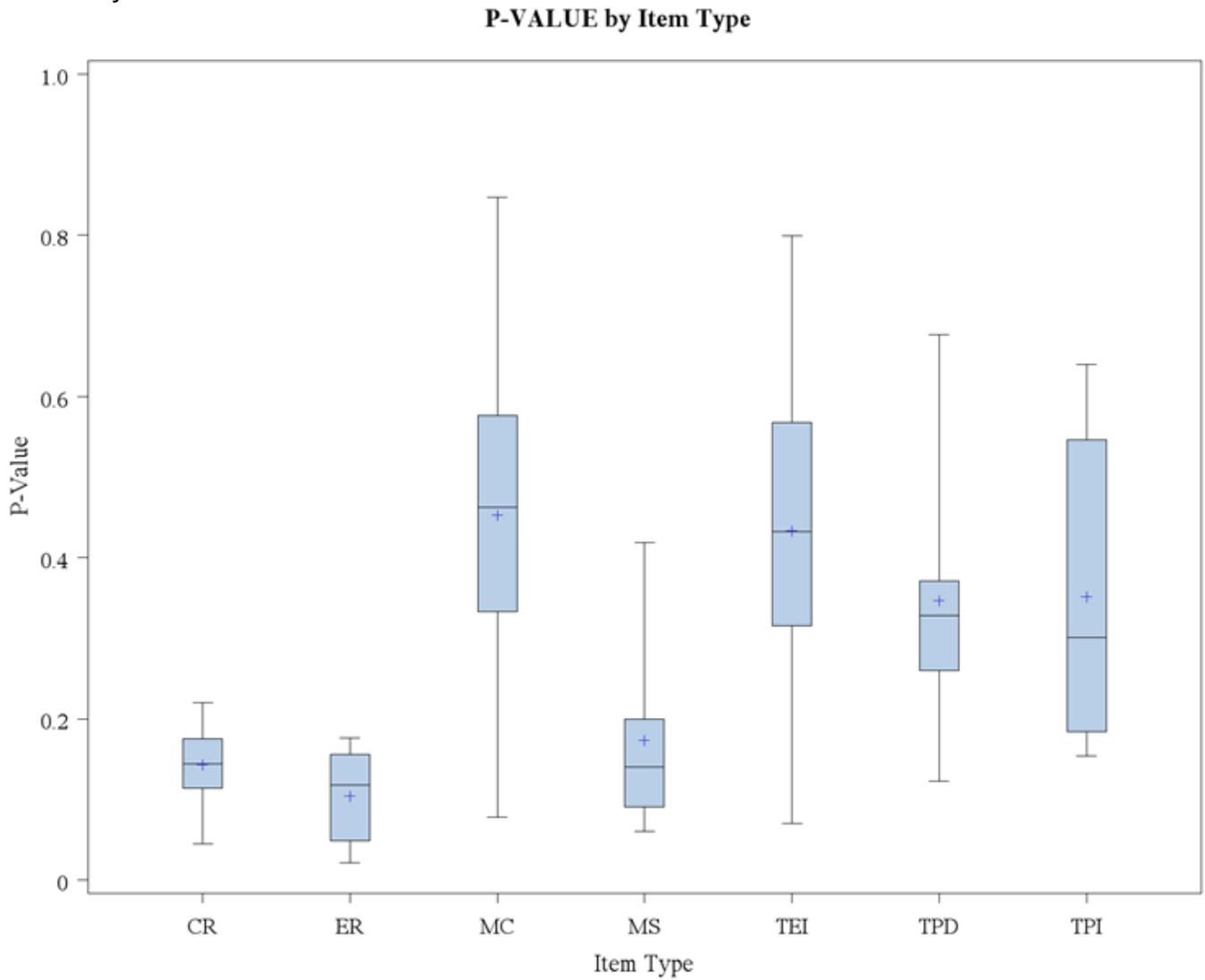
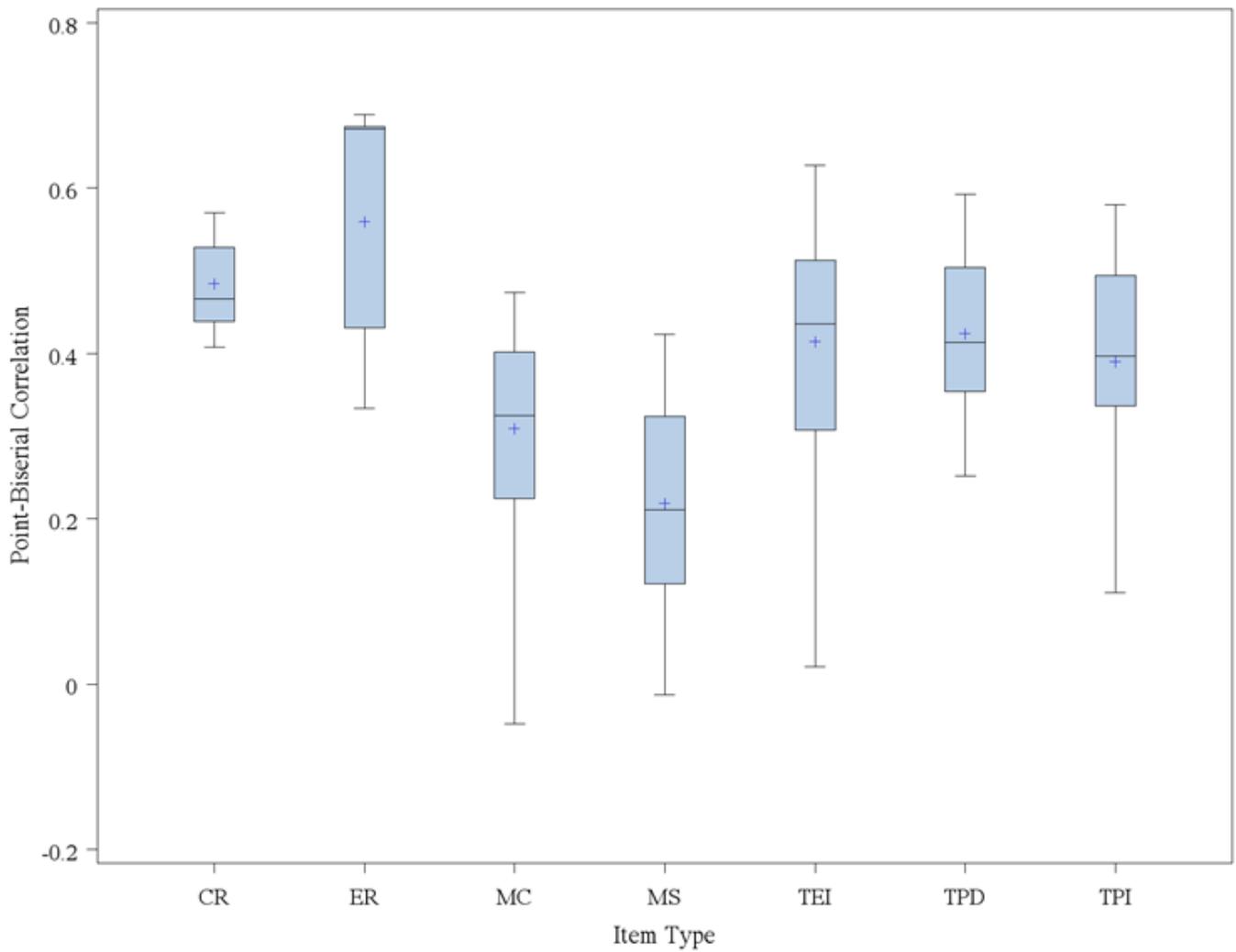


Figure 7.2

Box Plot of Item-Total Correlations/Point Biserial (PBIS)

Point-Biserial Correlation by Item Type



Differential Item Functioning

Differential item functioning (DIF) analyses are designed to detect statistical evidence of potential item bias. Because test scores can have many sources of variation, the test developers' task is to create assessments that measure the intended abilities and skills without introducing extraneous elements or construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores will reflect these unintended skills and knowledge, as well as what is purportedly assessed by the test. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). One of the factors that may render test scores as biased is differing cultural and socioeconomic experiences.

DIF is a statistical method to detect potential bias of an item. DIF is defined as a difference between groups (e.g., male and female) in the probability of getting an item correct. These analyses are conditioned on the ability that the assessment is intended to measure.

The DIF methodology for dichotomous items used the Mantel–Haenszel (*MH*) DIF statistic (Holland & Thayer, 1988; Mantel & Haenszel, 1959). The *MH* method is frequently used and is efficient in terms of statistical power (Clauser & Mazor, 1998). The Mantel–Haenszel chi-square statistic is computed as

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

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) was computed. The Educational Testing Service (ETS) first developed the ΔMH DIF statistic. To compute the ΔMH DIF, the *MH* alpha (the odds ratio) is first computed:

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

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 $2 \times 2 \times M$ (2 groups \times 2 item scores \times 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 $\Delta MH DIF$ is computed as

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

Positive values of $\Delta MH DIF$ indicate items that favor the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values of $\Delta MH DIF$ 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 $\Delta MH DIF$ are used to conduct statistical tests.

The *MH* chi-square statistic and the $\Delta MH DIF$ were used in combination to identify the field test items that exhibit strong, weak, or no DIF (Zieky, 1993). Table 7.3 defines the DIF categories for dichotomous items.

Table 7.3
DIF Categories for Dichotomous Items

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

For polytomous items, the standardized mean difference (*SMD*) (Dorans & Schmitt, 1991; Zwick, Thayer, & Mazzeo, 1997) and the Mantel χ^2 statistic (Mantel, 1963) are used to identify items with DIF. *SMD* estimates the average difference in performance between the reference group and the focal group while controlling for student ability. To calculate the *SMD*, let M represent the matching variable (total test score). For all $M = m$, identify the

students with raw score m and calculate the expected item score for the reference group (E_{rm}) and the focal group (E_{fm}). DIF is defined as $D_m = E_{fm} - E_{rm}$, and SMD is a weighted average of D_m using the weights $w_m = N_{fm}$ (the number of students in the focal group with raw score m), which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

The MH DIF statistic is based on a $2 \times (T+1) \times M$ (2 groups \times $T+1$ item scores \times 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:

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

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 7.4 defines the DIF categories for polytomous items.

Table 7.4
DIF Categories for Polytomous Items

DIF Category	Criteria
A (negligible)	Mantel χ^2 p -value > 0.05 or $ SMD/SD \leq 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 \geq 0.25$

Two DIF analyses were conducted for field test items: female/male and black/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The same methods were used to detect items on which black or white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Table 7.5 provide the percentages of items flagged for DIF. Items flagged with B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Very few field test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the Mantel–Haenszel statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 and *SMD* statistics. Table 7.5 summarizes the number of items showing strong DIF associated with any group comparison.

Table 7.5
Summary of DIF Flags for Field Test Items for Biology

Comparison Groups	A	B,[B-]	C,[C-]
Female – Male	256	4,[4]	0,[1]
African American – White	251	1,[11]	0,[2]

All items exhibiting DIF were reviewed by a committee of Louisiana teachers as well as LDOE and WestEd content staff. After review, no items were found to be exhibiting bias; therefore, no items were dropped during data review due to DIF analyses results and teacher committee reviews.

Item Calibration

LEAP 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. The LEAP 2025 Biology test was calibrated independent of the EOC Biology test.

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

Measurement Models

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

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

In that model, $p_i(\theta_j)$ is the probability that student j would earn a score of 1 on item i , b_i is the difficulty parameter for item i , a_i is the slope (or discrimination) parameter for item i , c_i is the pseudo-chance (or guessing) parameter for item i , and D is the constant 1.7.

This test also included five types of polytomous items: TE items scored 0–2, constructed response (CR) items scored 0–2, two-part independent (TPI) items scored 0–2, two-part dependent (TPD) items scored 0–2, and ER items scored 0–9. Data from polytomous items were used to estimate parameters for the generalized partial credit model (GPCM) (Muraki, 1992):

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$, $p_{im}(\theta_j)$ is the probability of an examinee with θ_j getting score m on item i , and M_i is the number of score categories of item i with possible item scores as consecutive integers from 0 to $M_i - 1$. 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.).

Field Test Item Parameters

The distributions of item parameters are summarized in Table 7.6. Figures 7.3–7.5 provide Box Plot displays of the distributions of IRT parameter estimates by item type. TPI, TPD, CR, and ER items have no c parameters because they are polytomous items and are therefore modeled using the GPCM.

It should be noted that a somewhat significant trend between classical item parameters (e.g., p -value) and IRT-based item parameters (e.g., b parameter) can be found. In addition, recommended ranges for IRT parameter estimates are functions of an assessment program and assessment results and will vary by large-scale assessment programs. As each of the LEAP 2025 assessments mature, however, desired targets/ranges (e.g., point-biserial higher than 0.3) can be defined in the annual Framework documents that the LDOE, Pearson, and WestEd use for annual test construction.

Item Fit

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q_1 statistic (Yen, 1981) is used as an index for how well theoretical item curves match observed item responses. Q_1 is computed by first conducting an IRT item parameter estimation, then estimating students’ achievement using the estimated item parameters, and, finally, using students’ achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and

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

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

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

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

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

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

The generalization of Q_1 for items with multiple response categories is

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

where

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

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 is displayed in Table 7.6.

Table 7.6
Summary of IRT Statistics for Field Test Items for Biology

Item Type	N Items	b Mean	b SD	a Mean	a SD	c Mean*	c SD*	% Fit (no model fit issues)
MC	105	2.23	12.03	0.67	0.41	0.16	0.13	65%
MS	20	4.33	4.88	0.44	0.44	0.03	0.04	50%
CR	16	1.39	0.61	0.69	0.21	-	-	79%
ER	6	1.51	0.57	0.35	0.11	-	-	71%
TE	79	0.91	2.71	0.52	0.30	0.07	0.09	80%
TPI	12	-0.49	8.83	0.32	0.21	-	-	73%
TPD	30	1.49	1.50	0.31	0.14	-	-	83%

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

*% Fit indicates % of items with no model fit issues.

*Note. TE items scored 0 and 1 have estimated c parameter.

Figure 7.3
Box Plot of IRT A Parameters

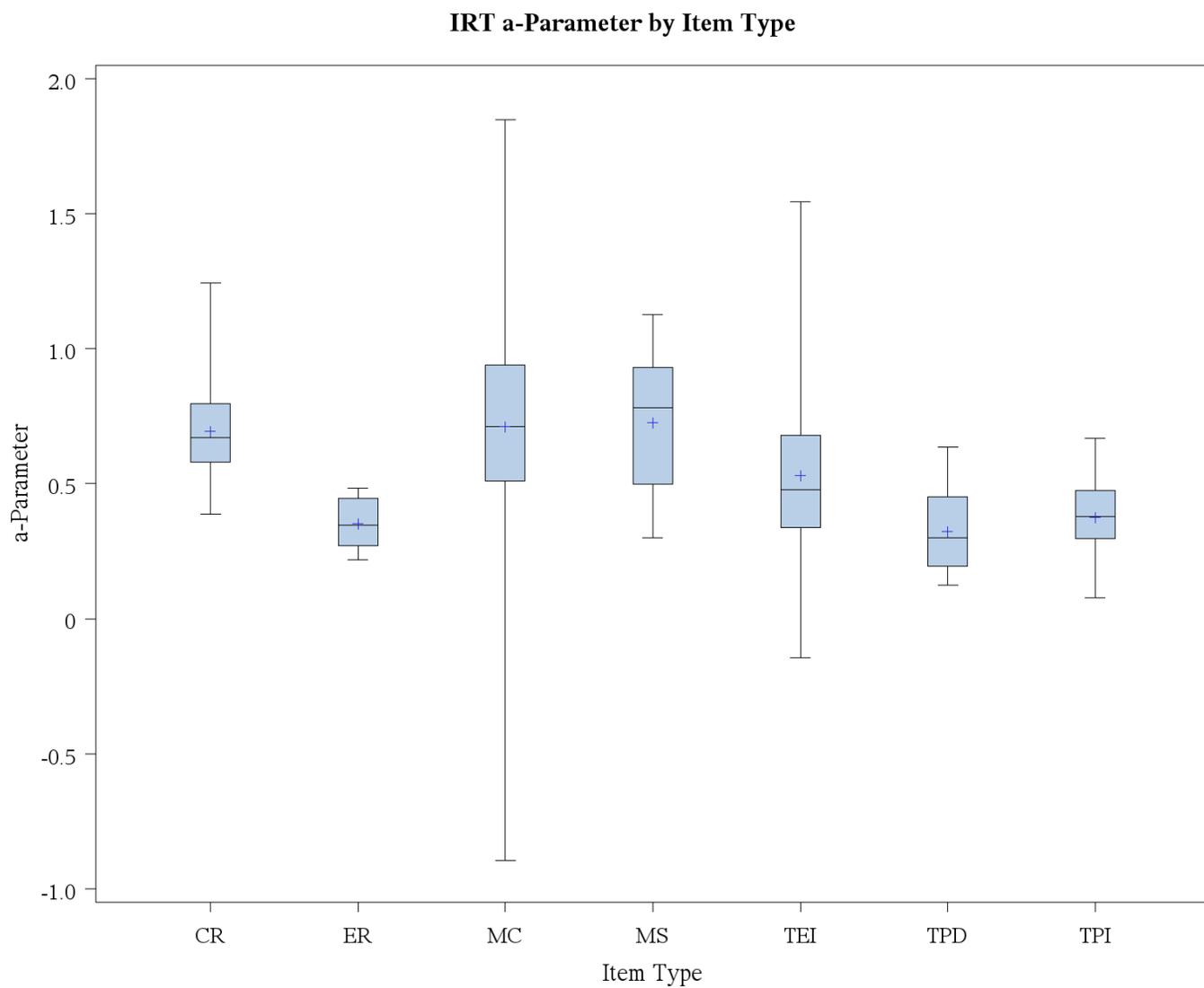


Figure 7.4
Box Plot of IRT B Parameters

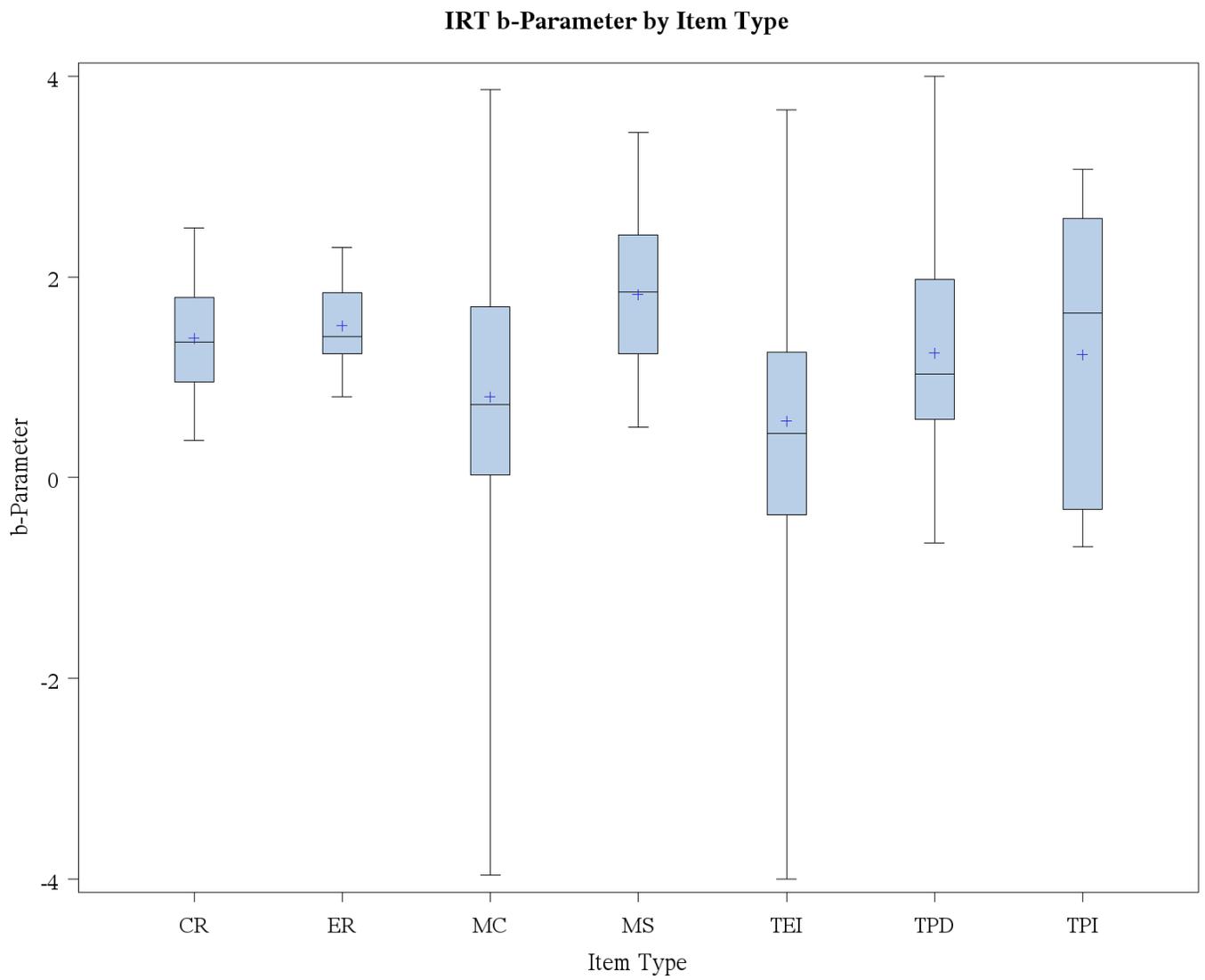
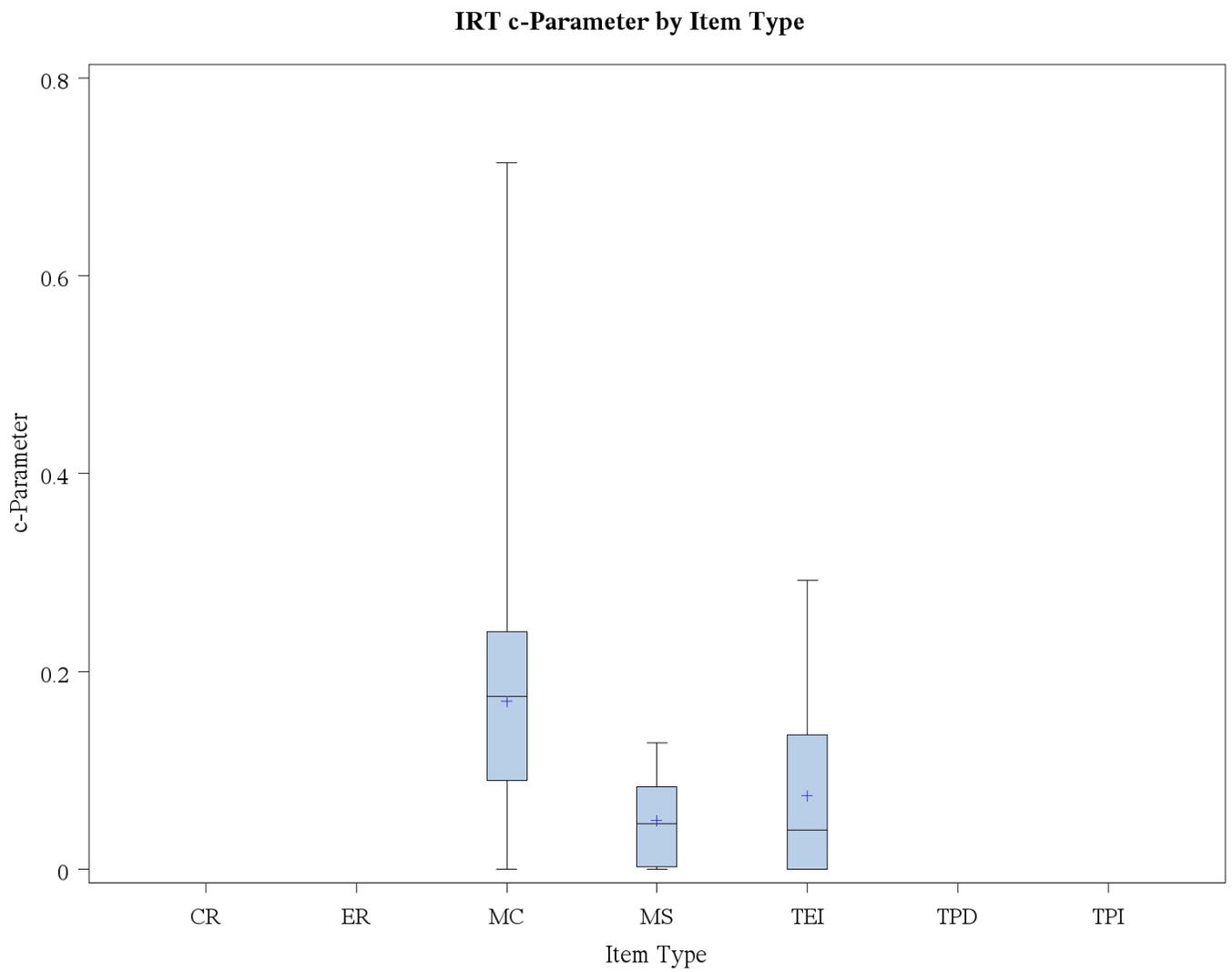


Figure 7.5
Box Plot of IRT C Parameters



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

8. Data Review Process

During data review, invited committee members review field-tested items with accompanying data, in order to make judgments about the appropriateness of items for use on operational test forms. As part of the data review process, participants are provided with item statistics that may indicate possible problems. Items are not automatically rejected on the sole basis of statistics; only items with concrete and identifiable flaws in their content are rejected.

The data review meeting for Biology began with a presentation and introduction to data review. The introductory training included a review of appropriate interpretations on item statistics (difficulty, discrimination, DIF, score distributions), what would be considered reasonable values, and how the values might differ across item types. To reinforce the training, participants were provided with a handout defining item statistics and a checklist including statistical and content considerations to keep in mind while reviewing items.

After signing a nondisclosure agreement, each participant was provided a computer to access Pearson's ABBI platform. Participants reviewed stimuli and statistics of standalone items and item sets on the Biology field test in ABBI. Content and psychometric representatives from the LDOE were present in the committee meeting.

Facilitators from Pearson and WestEd led the data review committee through the review of field-tested items by displaying on-screen stimuli and item statistics. Participants were instructed to evaluate the statistical information for each item and determine whether the item functioned as intended. Then, participants provided independent judgments regarding each item's suitability for future operational tests, in light of the field-test statistics. When an item exhibiting DIF was being reviewed, the facilitators specifically asked the committee members to review the DIF statistics and re-evaluate the items for any possible content problems that could lead to the item's possible differential performance. No items exhibiting DIF were identified to have flaws leading to the DIF flags. Judgments were followed by group discussion to reach consensus about each item, and consensus recommendations were then recorded. Specifically, the committee voted to accept, accept with edits (or "revise/re-field test"), or reject items. Table 8.1 summarizes the disposition of field-tested items from data review. If the committee's decision was to

edit or reject an item, additional information was captured to reflect the reason for the committee decision. Votes were compiled by the WestEd facilitator and recorded on one main judgment form.

Table 8.1
Summary of Biology Data Review Votes

Item Type	Number of Items		
	Accept	Accept w/Edits	Total
CR	14	2	16
ER	4	2	6
MC	99	6	105
MS	20	-	20
TE	78	1	79
TPD	30	-	30
TPI	12	-	12
Total	257	11	268

Following the data review meeting, LDOE content specialists reviewed items again, 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 and possibly administered operationally in the future. The reconciliation decisions were treated as the final decisions.

References

- AERA/APA/NCME. (2009/2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Warner (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, 63(4), 602–614.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31–44.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–47.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Educational Testing Service.
- Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

- Green, D. R. (1975, December). Procedures for assessing bias in achievement tests. Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.
- 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.
- 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.
- 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.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E.P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies: A revised core vocabulary*. Austin, TX: Steck-Vaughn.
- Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161–186). Hillsdale, NJ: Lawrence Erlbaum.

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

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

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

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

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

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

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

Appendix A: Training Agendas

LEAP 2025 Biology Item Outline Development Training Agenda Item Development Cycle for 2017–2018 Field Test

- I. **Item Development Process**
 - a. Overview
 - b. Steps in process
- II. **Outlines**
 - a. What outlines are
 - b. What outlines are not
 - c. Outline assignments
 - i. Tasks
 - ii. Item sets
 - iii. Standalone tasks
 - iv. Template
- III. **Considerations**
 - a. Tasks
 - b. Item sets
 - c. Phenomena list

LEAP 2025 Biology Item Writer Training Agenda Item Development Cycle for 2017–2018 Field Test

- I. **Project Overview: Outlines**
 - 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. **Item Development Materials on Box**

III. **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!

IV. **Anatomy of the Louisiana Student Standards for Science**

- 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

V. **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”

- VI. **Multidimensional Standards → Multidimensional Assessment**
 - a. Dimensions are never to be taught in isolation, and therefore are never tested in isolation.
 - b. 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
- VII. **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
- VIII. **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.
- IX. **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
- X. **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
 - d. Phenomena, contexts, and stimuli need to be the right grain size.
 - e. Goldilocks—provide only the information that is needed.
- XI. **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.

- XII. **Assessment Design: Item Groupings**
- a. The LSSS assessment will consist of three distinct “item groupings.”
 - i. Tasks (PE bundles; phenomena)
 - ii. Item sets (PE bundles; phenomena)
 - iii. Standalone items (single PE only; foci)
- XIII. **Item Grouping: 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
- XIV. **Item Grouping: 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.
- XV. **Item Grouping: 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.
- XVI. **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

XVII. 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. Used in tasks ONLY

XVIII. Item Types:

- a. Technology-enhanced (TE) item
 - i. TE items are worth 1 or 2 points
 - ii. Used in tasks, item sets, and standalone items
 - iii. TE item types (NO TE items 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
- XIX. **Development Process Overview**
- XX. **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.
- XXI. **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.
- XXII. **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.

XXIII. 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 forthcoming!

XXIV. Development Process Overview

XXV. 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 Item Development Training Agenda
Item Development Cycle for 2017–2018 Field Test**

I. Item Development Process

- a. Overview
- b. Steps in process

II. Approved Item Set Outline

- a. Example of sample item set
- b. Developed item set

III. Science and Engineering Practices (SEPs)

- a. K–12 Framework (pp. 42–79)
- b. SEP-DCI

IV. Crosscutting Concepts (CCCs)

- a. K–12 Framework (pp. 83–101)
- b. DCI-CCC

V. Dimensional Alignment

- a. SEP 2, Developing and Using Models
- b. CCC Systems and System Models

VI. Allowable Item Types

- a. Tasks
 - i. 1-point TE item
 - ii. 1-point SR
 - iii. 2-point TE item

- iv. 2-point EBSR
 - v. 9-point ER
 - b. Item sets
 - i. 1-point TE item
 - ii. 1-point SR
 - iii. 2-point TE item
 - iv. 2-point EBSR
 - v. 2-point CR
 - c. Standalone items
 - i. 1-point TE item
 - ii. 1-point SR
 - iii. 2-point TE item
 - iv. 2-point EBSR
- VII. **Reminders**
 - a. Stimulus and items developed per the outline
 - b. Every item has 2-D alignment minimum
 - c. CR and ER are text entry only
 - d. MC/MS (SR) are only ever 1 point
 - e. EBSR only TE/SR with Part A and Part B
 - f. Graphics reminders
 - g. Sources reminders

**LEAP 2025 Biology Editor Training Agenda
Item Development Cycle for 2017–2018 Field Test**

- I. **Item Set/Task/Standalone Item Overview**
 - a. Criteria for review
- II. **Item Development Process**
 - a. Four rounds of items slated for development in 2017
 - i. B1: Sample Assessment Guide Items
 - 1. “Operational” in development (four items per item set; five items/task; standalone items) but will never appear in a field test or form
 - 2. Developed for use in the Sample Assessment Guide and online training tool (OTT)
 - ii. B2–B4: 2018 Standalone Field Test
 - 3. Full-scale development
 - a. 10 items per itemset
 - b. 9 items in task (A and B versions)
 - c. Standalone items
 - 4. Items will appear on field test.
 - 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 in the fall
 - d. More editing and review rounds TBD
- 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**