LEAP 2025 Science 3-8 Technical Report: 2017–2018 Field Test

Prepared by DRC, Pearson, and WestEd

FOREWORD

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

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 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 3

Summary of the 2017–2018 Activities Timeline 3

2. Assessment Framework 5

3. Overview of the Test Development Process 8

Proposal and Review of Topics and Sources 31

Performance Expectation Bundling 31

Phenomena Selection and Outline Development 32

Matching Phenomena to Set Types 33

Outline and Stimuli Development 35

Item Writing and Review Process 36

4. Construction of Field Test Forms 45

5. Test Administration 49

Training of School Systems 49

Ancillary Materials 49

Time 60

Online Forms Administration, Grades 3–8 60

Paper-Based Forms Administration, Grades 3 and 4 60

Accommodations 61

Testing Windows 62

Test Security Procedures....62

6. Scoring Activities 63

7. Data Analysis 74

Classical Item Statistics 74

Differential Item Functioning 81

Item Calibration....85

Measurement Models 85

Field Test Item Parameters....86

Item Fit....87

8. Data Review Process.... 95

References.... 99

Appendix A: Training Agendas.... 102

1. Introduction

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

The purpose of this Technical Report is to describe the process for the standalone field test administration of the statewide summative science assessment for grades 3–8. 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 science grade 3–8 field test forms.

For grades 3–8, field test items were administered using a standalone instrument. All tests were delivered in a computer-based format, with a paper-based option made available for grades 3 and 4. An accommodated paper-based format is made available for students in grades 5–8 who are not physically able to test on a computer.

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	• Started item development planning for spring 2018 field test • Item development plans approved by LDOE staff
March 2017	• Content development specifications and style guide prepared
March–April 2017	• WestEd began item writing and development • Topics and outlines approved by LDOE staff
May–August 2017	• LDOE staff reviewed proposed content
June 2017	• Spring 2018 Framework and Test Construction Document proposed
August 2017	• LDOE and WestEd planning meeting held
September 2017	• Spring 2018 Framework and Test Construction Document approved
October 2017	• Item Content/Bias Review Committee convened • Reconciliation meeting held between LDOE and WestEd staff • Test construction activities began
November 2017	• Technical Advisory Committee Meeting convened • LDOE staff reviewed proposed spring 2018 field test selections
December 2017	• Online content delivered to administration vendor
January 2018	• Remaining spring 2018 materials delivered to administration vendor
April 2018	• Spring 2018 Standalone Field Test administered
May–June 2018	• Rangefinding meetings held
August 2018	• 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, the test blueprint (included in Section 3), the 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) tied to a common stimulus. The item sets include 1-point selected-response items (both single-select and multiple-select formats), and 2-point two-part items (two-part independent and two-part dependent formats). For grades 5–8, item sets may have 1- or 2-point TE items in place of 1-point selected-response or 2-point two-part items. Three of the item sets include a two-point constructed-response item instead of a two-part item or 2-point TE item (grades 5–8 only). In addition to the item sets, the assessment contains one task. The tasks are made up of five items tied to a common stimulus. Tasks include 1-point selected-response items (both single-select and multiple-select formats), 2-point two-part items (two-part independent and two-part dependent formats), and a 6-point extended-response (ER) item for grades 3–4 or a 9-point ER item for grades 5–8. For grades 5–8, tasks may have 1- or 2-point TE items in place of 1-point selected-response or 2-point two-part items. Standalone items may be either 1-point selected-response items (both single-select and multiple-select formats), or 2-point two-part items (two-part independent and two-part dependent formats). For grades 5–8, standalone items may be 1- or 2-point TE items, as well. 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 (LSS) for Science. All points associated with the task contribute to students' overall scores, but the ER is not a component of the current blueprint. This prevents the ER from impacting the proportional representation of content assessed by other parts of the test.

Because the assessments at grades 3 and 4 were administered primarily on paper, the item types for these grades were limited to selected-response (i.e., multiple choice [MC] or multiple select [MS]), two-part (i.e., two-part independent [TPI] or two-part dependent [TPD]), constructed-response [CR], and task-based ER items. The assessments for grades 5–8 were administered primarily online, so technology-enhanced (TE) items were viable at these grades. However, paper-and-pencil versions of the assessments for grades 5–8 were made available as accommodated forms for students who were unable to test online. For those forms, TEs were adapted to reflect paper presentation while still addressing the same content. The extended-response items also differ for the assessments at grades 3–4 and 5–8: at grades 3 and 4, the ER is worth 6 points, while at grades 5–8 the ER is worth 9 points. The lower point value at grades 3 and 4 allows for fewer requirements per item to allow shorter written responses for these lower grades, where students' writing skills are still emerging.

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

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

Acronyms used in item and test development are presented in the following table.

Table 3.1a

Grades 3–8: Acronyms Used in Item and Test Development

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

The LSS for Science is a set of three-dimensional standards, written as performance expectations (PEs). Each PE comprises three dimensions: Science and Engineering Practices (SEPs), which are the practices or behaviors scientists and engineers use as they investigate real-world phenomena and design solutions to problems; Disciplinary Core Ideas (DCIs), which are key skills and content knowledge that students should master; and Crosscutting Concepts (CCCs), the connections across domains of science. There are eight science and engineering practices (seven of which are assessed), and seven cross-cutting concepts. The Disciplinary Core Ideas have been grouped into three domains for assessment purposes: Earth and Space Science (ESS), Life Science (LS), and Physical Science (PS).

The test design and blueprints developed for the LSS for Science incorporate the multidimensional nature of the LSS for Science. Once the test design and the blueprint were approved, the item development plan was established. The test blueprints that guided initial item development projections for grade 3 are presented in the following tables.

Table 3.1b

Test Blueprint for LEAP 2025 Grade 3: DCI Domain Coverage

Grade 3: DCI Domain Coverage			
	# of PEs in LSS	Relative % in LSS	% by points of all items
ESS	3	20%	15%–25%
LS	8	53%	48%–58%
PS	4	27%	22%–32%
Total	15	100%	

Table 3.1c

Test Blueprint for LEAP 2025 Grade 3: Minimal PE Coverage

Grade 3: Minimal PE Coverage Every PE will be included at least one time in a test			
	SEP	CCC	Min Items
03-ESS2-1	SEP 4 - DATA	CCC 1 - PAT	1
03-ESS2-2	SEP 8 - INFO	CCC 1 - PAT	1
03-ESS3-1	SEP 7 - ARG	CCC 2 - C/E	1
03-LS1-1	SEP 2 - MOD	CCC 1 - PAT	1
03-LS2-1	SEP 7 - ARG	CCC 4 - SYS	1
03-LS3-1	SEP 4 - DATA	CCC 1 - PAT	1
03-LS3-2	SEP 6 - E/S	CCC 2 - C/E	1
03-LS4-1	SEP 4 - DATA	CCC 3 - SPQ	1
03-LS4-2	SEP 6 - E/S	CCC 2 - C/E	1
03-LS4-3	SEP 7 - ARG	CCC 2 - C/E	1
03-LS4-4	SEP 7 - ARG	CCC 4 - SYS	1
03-PS2-1	SEP 3 - INV	CCC 2 - C/E	1
03-PS2-2	SEP 3 - INV	CCC 1 - PAT	1
03-PS2-3	SEP 1 - Q/P	CCC 2 - C/E	1
03-PS2-4	SEP 1 - Q/P	CCC 1 - PAT	1

Table 3.1d

Test Blueprint for LEAP 2025 Grade 3: CCC Coverage

Grade 3: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by Points of CCC Items
CCC 1 - PAT	6	40%	35%–45%
CCC 2 - C/E	6	40%	35%–45%
CCC 3 - SPQ	1	7%	5%–15%
CCC 4 - SYS	2	13%	8%–18%
CCC 5 - E/M	0	0%	0%
CCC 6 - S/F	0	0%	0%
CCC 7 - S/C	0	0%	0%
Total	15	100%	

Table 3.1e

Test Blueprint for LEAP 2025 Grade 3: SEP Coverage

Grade 3: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by Points of SEP Items
SEP 1 - Q/P	2	13%	8%–18%
SEP 2 - MOD	1	7%	5%–15%
SEP 3 - INV	2	13%	8%–20%
SEP 4 - DATA	3	20%	15%–25%
SEP 5 - MCT	0	0%	0%
SEP 6 - E/S	2	13%	8%–18%
SEP 7 - ARG	4	27%	22%–32%
SEP 8 - INFO	1	7%	5%–15%
Total	15	100%	

Table 3.1f

Test Blueprint for LEAP 2025 Grade 3: SEP Subclaim Coverage

Grade 3: SEP Subclaim Coverage				
Subclaim	# PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	4	29%	24%–34%	7
Subclaim 2 (SEPs 4, 5, 7)	7	50%	45%–55%	7
Subclaim 3 (SEPs 2 & 6)	3	21%	16%–26%	7
Total	14	100%		

Note: 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 Grade 3: SEP Compared to CCC Ratio

Grade 3: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

Table 3.1h

Test Blueprint for LEAP 2025 Grade 3: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR	CR, Two-Part	ER	Total Items	Total Points
4-Item Set	4	4	6	2	2		16	24
Standalone items	1	18	21	15	3		18	21
Task	1	5	12	2	2	1	5	12
Totals				19	7	1	39	57

The test blueprints that guided initial item development projections for grade 4 are presented in the following tables.

Table 3.1i

Test Blueprint for LEAP 2025 Grade 4: DCI Domain Coverage

Grade 4: DCI Domain Coverage			
Domain	# of PEs in LSS	Relative % in LSS	% by Points of All Items
ESS	6	43%	38%–48%
LS	2	14%	9%–19%
PS	6	43%	38%–48%
Total	14	100%	

Table 3.1j

Test Blueprint for LEAP 2025 Grade 4: Minimal PE Coverage

Grade 4: Minimal PE Coverage Every PE will be included at least one time in a test			
PE	SEP	CCC	Min Items
04-ESS1-1	SEP 6 - E/S	CCC 1 - PAT	1
04-ESS2-1	SEP 3 – INV	CCC 2 - C/E	1
04-ESS2-2	SEP 4 – DATA	CCC 1 - PAT	1
04-ESS2-3	SEP 1 - Q/P	CCC 2 - C/E	1
04-ESS3-1	SEP 8 – INFO	CCC 2 - C/E	1
04-ESS3-2	SEP 6 - E/S	CCC 2 - C/E	1
04-LS1-1	SEP 7 – ARG	CCC 4 - SYS	1
04-LS1-2	SEP 6 - E/S	CCC 2 - C/E	1
04-PS3-1	SEP 6 - E/S	E/M	1
04-PS3-2	SEP 3 – INV	E/M	1
04-PS3-3	SEP 1 - Q/P	E/M	1
04-PS3-4	SEP 6 - E/S	E/M	1
04-PS4-1	SEP 2 - MOD	CCC 1 - PAT	1
04-PS4-2	SEP 2 - MOD	CCC 2 - C/E	1

Table 3.1k

Test Blueprint for LEAP 2025 Grade 4: CCC Coverage

Grade 4: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by points of CCC items
CCC 1 - PAT	3	21%	16%–26%
CCC 2 - C/E	6	43%	38%–48%
CCC 3 - SPQ	0	0%	0%
CCC 4 - SYS	1	7%	5%–15%
CCC 5 - E/M	4	29%	24%–34%
CCC 6 - S/F	0	0%	0%
CCC 7 - S/C	0	0%	0%
Total	14	100%	

Table 3.1l

Test Blueprint for LEAP 2025 Grade 4: SEP Coverage

Grade 4: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by Points of SEP Items
SEP 1 - Q/P	2	14%	9%–19%
SEP 2 - MOD	2	14%	9%–19%
SEP 3 - INV	2	14%	9%–19%
SEP 4 - DATA	1	7%	5%–15%
SEP 5 - MCT	0	0%	0%
SEP 6 - E/S	5	36%	31%–41%
SEP 7 - ARG	1	7%	5%–15%
SEP 8 - INFO	1	7%	5%–15%
Total	14	100%	

Table 3.1m

Test Blueprint for LEAP 2025 Grade 4: SEP Subclaim Coverage

Grade 4: SEP Subclaim Coverage				
SEP Subclaim	# PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	4	31%	26%–36%	7
Subclaim 2 (SEPs 4, 5, 7)	2	15%	10%–20%	7
Subclaim 3 (SEPs 2 & 6)	7	54%	49%–59%	7
Total	13	100%		

Note: 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 3.1n

Test Blueprint for LEAP 2025 Grade 4: SEP Compared to CCC Ratio

Grade 4: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

Table 3.1o

Test Blueprint for LEAP 2025 Grade 4: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR	CR, Two-Part	ER	Total Items	Total Points
4-Item Set	5	4	6	2	2		20	30
Standalone items	1	17	20	14	3		17	20
Task	1	5	12	2	2	1	5	12
Totals				18	7	1	42	62

The test blueprints that guided initial item development projections for grade 5 are presented in the following tables.

Table 3.1p

Test Blueprint for LEAP 2025 Grade 5: DCI Domain Coverage

Grade 5: DCI Domain Coverage			
Domain	# of PEs in LSS	Relative % in LSS	% by Points of All Items
ESS	5	38%	33%–43%
LS	2	15%	9%–20%
PS	6	46%	41%–51%
Total	13	100%	

Table 3.1q

Test Blueprint for LEAP 2025 Grade 5: Minimal PE Coverage

Grade 5: Minimal PE Coverage Every PE will be included at least one time in a test			
PE	SEP	CCC	Min Items
05-ESS1-1	SEP 7 - ARG	CCC 3 - SPQ	1
05-ESS1-2	SEP 4 - DATA	CCC 1 - PAT	1
05-ESS2-1	SEP 2 - MOD	CCC 4 - SYS	1
05-ESS2-2	SEP 5 - MCT	CCC 3 - SPQ	1
05-ESS3-1	SEP 6 - E/S	CCC 4 - SYS	1
05-LS1-1	SEP 1 - Q/P	CCC 5 - E/M	1
05-LS2-1	SEP 2 - MOD	CCC 4 - SYS	1
05-PS1-1	SEP 2 - MOD	CCC 3 - SPQ	1
05-PS1-2	SEP 5 - MCT	CCC 5 - E/M	1
05-PS1-3	SEP 3 - INV	CCC 3 - SPQ	1
05-PS1-4	SEP 3 - INV	CCC 2 - C/E	1
05-PS2-1	SEP 7 - ARG	CCC 2 - C/E	1
05-PS3-1	SEP 2 - MOD	CCC 5 - E/M	1

Table 3.1r

Test Blueprint for LEAP 2025 Grade 5: CCC Coverage

Grade 5: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by Points of CCC Items
CCC 1 - PAT	1	8%	5%–15%
CCC 2 - C/E	2	15%	10%–22%
CCC 3 - SPQ	4	31%	22%–36%
CCC 4 - SYS	3	23%	18%–28%
CCC 5 - E/M	3	23%	18%–28%
CCC 6 - S/F	0	0%	0%
CCC 7 - S/C	0	0%	0%
Total	13	100%	

Table 3.1s

Test Blueprint for LEAP 2025 Grade 5: SEP Coverage

Grade 5: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by points of SEP items
SEP 1 - Q/P	1	8%	3%–13%
SEP 2 - MOD	4	31%	26%–36%
SEP 3 - INV	2	15%	10%–20%
SEP 4 - DATA	1	8%	3%–13%
SEP 5 - MCT	2	15%	10%–20%
SEP 6 - E/S	1	8%	3%–13%
SEP 7 - ARG	2	15%	10%–20%
SEP 8 - INFO	0	0%	–
Total	13	100%	

Table 3.1t

Test Blueprint for LEAP 2025 Grade 5: SEP Subclaim Coverage

Grade 5: SEP Subclaim Coverage				
	# of PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	3	23%	18%–28%	7
Subclaim 2 (SEPs 4, 5, 7)	5	38%	33%–43%	7
Subclaim 3 (SEPs 2 & 6)	5	38%	33%–43%	7
Total	13	100%		

Note: 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 3.1u

Test Blueprint for LEAP 2025 Grade 5: SEP Compared to CCC Ratio

Grade 5: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	25%
CCCs	50%	25%

Table 3.1v

Test Blueprint for LEAP 2025 Grade 5: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR, 1-pt TE	CR, 2-pt 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 test blueprints that guided initial item development projections for grade 6 are presented in the following tables.

Table 3.1w

Test Blueprint for LEAP 2025 Grade 6: DCI Domain Coverage

Grade 6: DCI Domain Coverage			
Domain	# of PEs in LSS	Relative % in LSS	% by Points of All Items
ESS	4	21%	15%–26%
LS	5	26%	21%–31%
PS	10	53%	48%–58%
Total	19	100%	

Table 3.1x

Test Blueprint for LEAP 2025 Grade 6: Minimal PE Coverage

Grade 6: Minimal PE Coverage Every PE will be included at least one time in a test			
PE	SEP	CCC	Min Items
06-MS-ESS1-1	SEP 2 - MOD	CCC 1 - PAT	1
06-MS-ESS1-2	SEP 2 - MOD	CCC 4 - SYS	1
06-MS-ESS1-3	SEP 4 - DATA	CCC 3 - SPQ	1
06-MS-ESS3-4	SEP 7 - ARG	CCC 2 - C/E	1
06-MS-LS1-1	SEP 3 - INV	CCC 3 - SPQ	1
06-MS-LS1-2	SEP 2 - MOD	CCC 6 - S/F	1
06-MS-LS2-1	SEP 4 - DATA	CCC 2 - C/E	1
06-MS-LS2-2	SEP 6 - E/S	CCC 1 - PAT	1
06-MS-LS2-3	SEP 2 - MOD	CCC 5 - E/M	1
06-MS-PS1-1	SEP 2 - MOD	CCC 3 - SPQ	1
06-MS-PS2-1	SEP 6 - E/S	CCC 4 - SYS	1
06-MS-PS2-2	SEP 3 - INV	CCC 7 - S/C	1
06-MS-PS2-3	SEP 1 - Q/P	CCC 2 - C/E	1
06-MS-PS2-4	SEP 7 - ARG	CCC 4 - SYS	1
06-MS-PS2-5	SEP 3 - INV	CCC 2 - C/E	1
06-MS-PS4-2	SEP 2 - MOD	CCC 6 - S/F	1
06-MS-PS3-1	SEP 4 - DATA	CCC 3 - SPQ	1
06-MS-PS3-2	SEP 2 - MOD	CCC 4 - SYS	1
06-MS-PS4-1	SEP 5 - MCT	CCC 1 - PAT	1

Table 3.1y

Test Blueprint for LEAP 2025 Grade 6: CCC Coverage

Grade 6: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by Points of CCC Items
CCC 1 - PAT	3	16%	11%–21%
CCC 2 - C/E	4	21%	16%–26%
CCC 3 - SPQ	4	21%	16%–26%
CCC 4 - SYS	4	21%	16%–26%
CCC 5 - E/M	1	5%	5–10%
CCC 6 - S/F	2	11%	6–16%
CCC 7 - S/C	1	5%	5–10%
Total	19	100%	

Table 3.1z

Test Blueprint for LEAP 2025 Grade 6: SEP Coverage

Grade 6: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by Points of SEP Items
SEP 1 - Q/P	1	5%	5%–10%
SEP 2 – MOD	7	37%	32%–42%
SEP 3 – INV	3	16%	11%–21%
SEP 4 – DATA	3	16%	11%–21%
SEP 5 – MCT	1	5%	5%–10%
SEP 6 - E/S	2	11%	5%–16%
SEP 7 – ARG	2	11%	5%–16%
SEP 8 – INFO	0	0%	0%
Total	19	100%	

Table 3.1aa

Test Blueprint for LEAP 2025 Grade 6: SEP Subclaim Coverage

Grade 6: SEP Subclaim Coverage				
SEP Subclaim	# PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	4	21%	16%–26%	7
Subclaim 2 (SEPs 4, 5, 7)	6	32%	27%–37%	7
Subclaim 3 (SEPs 2 & 6)	9	47%	42%–52%	7
Total	19	100%		

Note: 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 3.1bb

Test Blueprint for LEAP 2025 Grade 6: SEP Compared to CCC Ratio

Grade 6: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

Table 3.1cc

Test Blueprint for LEAP 2025 Grade 6: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR, 1-pt TE	CR, 2-pt 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 test blueprints that guided initial item development projections for grade 7 are presented in the following tables.

Table 3.1dd

Test Blueprint for LEAP 2025 Grade 7: DCI Domain Coverage

Grade 7: DCI Domain Coverage			
Domain	# of PEs in LSS	Relative % in LSS	% by Points of All Items
ESS	4	25%	20%–35%
LS	8	50%	45%–55%
PS	4	25%	20%–35%
Total	16	100%	

Table 3.1ee

Test Blueprint for LEAP 2025 Grade 7: Minimal PE Coverage

Grade 7: Minimal PE Coverage Every PE will be included at least one time in a test			
PE	SEP	CCC	Min Items
07-MS-ESS2-4	SEP 2 - MOD	CCC 5 - E/M	1
07-MS-ESS2-5	SEP 3 - INV	CCC 2 - C/E	1
07-MS-ESS2-6	SEP 2 - MOD	CCC 4 - SYS	1
07-MS-ESS3-5	SEP 1 - Q/P	CCC 7 - S/C	1
07-MS-LS1-3	SEP 7 - ARG	CCC 4 - SYS	1
07-MS-LS1-6	SEP 6 - E/S	CCC 5 - E/M	1
07-MS-LS1-7	SEP 2 - MOD	CCC 5 - E/M	1
07-MS-LS2-4	SEP 7 - ARG	CCC 7 - S/C	1
07-MS-LS2-5	SEP 6 - E/S	CCC 7 - S/C	1
07-MS-LS3-2	SEP 2 - MOD	CCC 2 - C/E	1
07-MS-LS4-4	SEP 6 - E/S	CCC 2 - C/E	1
07-MS-LS4-5	SEP 8 - INFO	CCC 2 - C/E	1
07-MS-PS1-2	SEP 4 - DATA	CCC 1 - PAT	1
07-MS-PS1-4	SEP 2 - MOD	CCC 2 - C/E	1
07-MS-PS1-5	SEP 2 - MOD	CCC 5 - E/M	1
07-MS-PS3-4	SEP 3 - INV	CCC 3 - SPQ	1

Table 3.1ff

Test Blueprint for LEAP 2025 Grade 7: CCC Coverage

Grade 7: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by Points of CCC Items
CCC 1 – PAT	1	6%	1%–11%
CCC 2 - C/E	5	31%	20%–36%
CCC 3 - SPQ	1	6%	1%–11%
CCC 4 - SYS	2	13%	8%–18%
CCC 5 - E/M	4	25%	20%–32%
CCC 6 - S/F	0	0%	0%
CCC 7 - S/C	3	19%	14%–24%
Total	16	100%	

Table 3.1gg

Test Blueprint for LEAP 2025 Grade 7: SEP Coverage

Grade 7: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by Points of SEP Items
SEP 1 - Q/P	1	6%	5%–15%
SEP 2 - MOD	6	38%	33%–43%
SEP 3 - INV	2	13%	8%–18%
SEP 4 - DATA	1	6%	5%–15%
SEP 5 - MCT	0	0%	0%
SEP 6 - E/S	3	19%	14%–24%
SEP 7 - ARG	2	13%	8%–18%
SEP 8 - INFO	1	6%	5%–15%
Total	16	100%	

Table 3.1hh

Test Blueprint for LEAP 2025 Grade 7: SEP Subclaim Coverage

Grade 7: SEP Subclaim Coverage				
SEP Subclaim	# PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	3	20%	15%–25%	7
Subclaim 2 (SEPs 4, 5, 7)	3	20%	15%–25%	7
Subclaim 3 (SEPs 2 & 6)	9	60%	55%–65%	7
Total	15	100%		

Note: 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 3.1ii

Test Blueprint for LEAP 2025 Grade 7: SEP Compared to CCC Ratio

Grade 7: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

Table 3.1jj

Test Blueprint for LEAP 2025 Grade 7: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR, 1-pt TE	CR, 2-pt 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 test blueprints that guided initial item development projections for grade 8 are presented in the following tables.

Table 3.1kk

Test Blueprint for LEAP 2025 Grade 8: DCI Domain Coverage

Grade 8: DCI Domain Coverage			
Domain	# of PEs in LSS	Relative % in LSS	% by Points of All Items
ESS	7	37%	32%–42%
LS	7	37%	32%–42%
PS	5	26%	21%–31%
Total	19	100%	

Table 3.1ll

Test Blueprint for LEAP 2025 Grade 8: Minimal PE Coverage

Grade 8: Minimal PE Coverage			
PE	SEP	CCC	Min Items
08-MS-ESS1-4	SEP 6 - E/S	CCC 3 - SPQ	1
08-MS-ESS2-1	SEP 2 - MOD	CCC 7 - S/C	1
08-MS-ESS2-2	SEP 6 - E/S	CCC 3 - SPQ	1
08-MS-ESS2-3	SEP 4 - DATA	CCC 1 - PAT	1
08-MS-ESS3-1	SEP 6 - E/S	CCC 2 - C/E	1
08-MS-ESS3-2	SEP 4 - DATA	CCC 1 - PAT	1
08-MS-ESS3-3	SEP 6 - E/S	CCC 2 - C/E	1
08-MS-LS1-4	SEP 7 - ARG	CCC 2 - C/E	1
08-MS-LS1-5	SEP 6 - E/S	CCC 2 - C/E	1
08-MS-LS3-1	SEP 2 - MOD	CCC 6 - S/F	1
08-MS-LS4-1	SEP 4 - DATA	CCC 1 - PAT	1
08-MS-LS4-2	SEP 6 - E/S	CCC 1 - PAT	1
08-MS-LS4-3	SEP 4 - DATA	CCC 1 - PAT	1
08-MS-LS4-6	SEP 5 - MCT	CCC 2 - C/E	1
08-MS-PS1-1	SEP 2 - MOD	CCC 3 - SPQ	1
08-MS-PS1-3	SEP 8 - INFO	CCC 6 - S/F	1
08-MS-PS1-6	SEP 6 - E/S	CCC 5 - E/M	1
08-MS-PS3-3	SEP 6 - E/S	CCC 5 - E/M	1
08-MS-PS3-5	SEP 7 - ARG	CCC 5 - E/M	1

Table 3.1mm

Test Blueprint for LEAP 2025 Grade 8: CCC Coverage

Grade 8: CCC Coverage			
CCC Overall	# in PEs in LSS	Relative % in LSS	% by Points of CCC Items
CCC 1 - PAT	5	26%	21%–31%
CCC 2 - C/E	5	26%	21%–31%
CCC 3 - SPQ	3	16%	11%–21%
CCC 4 - SYS	0	0%	0%
CCC 5 - E/M	3	16%	11%–21%
CCC 6 - S/F	2	11%	5%–16%
CCC 7 - S/C	1	5%	1%–11%
Total	19	100%	

Table 3.1nn

Test Blueprint for LEAP 2025 Grade 8: SEP Coverage

Grade 8: SEP Coverage			
SEP Overall	# in PEs in LSS	Relative % in LSS	% by Points of SEP Items
SEP 1 - Q/P	0	0%	0%
SEP 2 - MOD	3	16%	11%–21%
SEP 3 - INV	0	0%	0%
SEP 4 - DATA	4	21%	16%–26%
SEP 5 - MCT	1	5%	2%–15%
SEP 6 - E/S	8	42%	37%–42%
SEP 7 - ARG	2	11%	5%–16%
SEP 8 - INFO	1	5%	5%–15%
Total	19	100%	

Table 3.1oo

Test Blueprint for LEAP 2025 Grade 8: SEP Subclaim Coverage

Grade 8: SEP Subclaim Coverage				
SEP Subclaim	# PEs in LSS	Relative % in LSS	% by Points of SEP Items	Min Points
Subclaim 1 (SEPs 1 & 3)	0			7
Subclaim 2 (SEPs 4, 5, 7)	7	39%	34%–44%	7
Subclaim 3 (SEPs 2 & 6)	11	61%	56%–66%	7
Total	18	100%		

Note: 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 3.1pp

Test Blueprint for LEAP 2025 Grade 8: SEP Compared to CCC Ratio

Grade 8: SEP Compared to CCC Ratio		
	Relative Weight in LSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

Table 3.1qq

Test Blueprint for LEAP 2025 Grade 8: Operational Test Composition

Item Sets/Item Types	Total Sets	Total Items per Set	Total Points per Set	SR, 1-pt TE	CR, 2-pt 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 assessment item development plans were created in conjunction with LDOE content staff. The development plans 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 plans and the content distribution determined the focus of the item and tasks and standalone items to be developed. This section describes the processes used to develop the item sets, tasks, and standalone items. Tables 3.2a–f show the initial item development plans for the number of items developed by WestEd by reporting category for grades 3–8.

Table 3.2a
Number of Items Developed for Grade 3 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	46	0	0	33	0	12	79
Tasks	4	16	0	0	16	4	0	32
Standalone Items	n/a	23	0	0	11	0	0	34

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

Table 3.2b
Number of Items Developed for Grade 4 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	46	0	0	33	0	12	79
Tasks	4	16	0	0	16	4	0	32
Standalone Items	n/a	23	0	0	11	0	0	34

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

Table 3.2c

Number of Items Developed for Grade 5 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	23	23	17	16	0	12	79
Tasks	4	8	8	8	8	4	0	32
Standalone Items	n/a	15	5	6	5	0	0	31

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

Table 3.2d

Number of Items Developed for Grade 6 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	23	23	17	16	0	12	79
Tasks	4	8	8	8	8	4	0	32
Standalone Items	n/a	15	5	6	5	0	0	31

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

Table 3.2e

Number of Items Developed for Grade 7 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	23	23	17	16	0	12	79
Tasks	4	8	8	8	8	4	0	32
Standalone Items	n/a	15	5	6	5	0	0	31

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

Table 3.2f

Number of Items Developed for Grade 8 Assessment for Item Sets, Tasks, and Standalone Items

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	9	23	23	17	16	0	12	79
Tasks	4	8	8	8	8	4	0	32
Standalone Items	n/a	15	5	6	5	0	0	31

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

Proposal and Review of Topics and Sources

Performance Expectation Bundling

As a first step in the development process, WestEd used the 2017 Louisiana Student Standards for Science 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 is assessed in a meaningful way. Key to this bundling was the need to ensure that paired PEs and phenomena achieved a "natural fit." Therefore, not all PEs were bundled, some PEs appeared in more than one bundle, and some PEs were bundled across content domains. 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 for a "mix and match" approach in which the disciplinary core idea (DCI) and crosscutting concept (CCC) for one PE 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 one PE (DCI, SEP, and/or CCC) 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 worked together to assess the multidimensional nature of the performance expectations bundle.

Table 3.3a shows the bundles approved by the LDOE by grade, as well as the number of approved bundles that then were targeted for development in the 2017–2018 development cycle.

Table 3.3a
PE Bundling by Grade

Grade	Number of PE Bundles Approved	Number of Bundles Targeted for Development
3	16	13
4	15	13
5	17	13
6	15	13
7	17	13
8	17	13

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 grade 3–8 assessment are anchored on a scientific phenomenon described by 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 to an item in an item set or task.

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;
- 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 by 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 to 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 extended-response (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. For each grade, four ERs were developed for four tasks. Like the tasks, the item sets were phenomena-based, but unlike the tasks, they comprised independent items that did not necessarily build upon each other. For the 2017–2018 development cycle, WestEd developed nine item sets per grade. Although an item set does not need to contain a constructed-response (CR) item, WestEd developed CRs for all item sets and for every reporting category. In some cases, more than one CR was developed per item set. Table 3.3b shows the total number of CRs developed per grade.

Table 3.3b

Constructed-Response Item Development by Grade

Grade	Number of CRs Developed
3	14
4	11
5	12
6	10
7	12
8	10

For tasks and item sets, WestEd offered a document containing descriptions of phenomena associated with bundles to the LDOE to review prior to item development. Table 3.3c shows the number of phenomena submitted to the LDOE for grades 3–8.

Table 3.3c

Phenomena Submitted by Grade

Grade	Number of Phenomena Submitted
3	25
4	50
5	54
6	47
7	63
8	50

Based on the list, the LDOE identified 13 phenomena per grade to be developed into stimuli for the tasks 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 internet search for science articles on the LDOE's objectives, as well as training in universal design and bias and sensitivity issues. For an outline of the training, see Appendix A for the LEAP 2025 Grades 3–8 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, considers 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 Louisiana Student Standards for Science were used as an additional measure of grade-level appropriateness. WestEd and the LDOE also drew on the professional experience of educators, during Content and Bias Committee review, to verify that sources would be accessible to students, and made changes based on their feedback. Most of the stimuli developed for the assessments were found to be below or

at grade level; however, some of the science vocabulary was evaluated as above grade level. In those cases, additional support such as parenthetical definitions (glossing) was included for necessary science content 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 grades 3–8 assessment. All writers were 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 Grades 3–8 Item Training Agenda (March 2017). In the training, writers were provided context for the assessment, including LDOE expectations, the Louisiana Student Standards for Science, 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 for 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), and 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 distracters were incorrect but plausible, and to avoid cueing and outliers in the items. WestEd hosted an online training for writers in which

training in universal design and bias/sensitivity was provided. A variety of items were presented and reviewed using universal design and bias/sensitivity lenses. This training also included an overview of these topics (see Appendix A for the LEAP 2025 Grades 3–8 Item Training Agenda). WestEd provided training and feedback to the writers throughout the development cycle, as the LDOE and WestEd gained a clearer understanding of how the stimuli, items, and sets worked together.

WestEd provided additional training to a subset of editors outlining the specific responsibilities for those who served as editors for the grades 3–8 assessment. For an outline of the information covered, see the LEAP 2025 Grades 3–8 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 the content and bias review meeting, participants completed nondisclosure agreements as part of the activities. The recruitment process, conducted by LDOE staff, also included participants from regions across the state. Participants represent the population of Louisiana students served—including special education, English learners, 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 3.4 provides the demographic characteristics of the review committee.

Table 3.4

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

Grade Level	3	4	5	6-8 Group A	6-8 Group B
Classroom Teacher	7	8	4	10	10
Content/Curriculum Specialist	0	1	1	1	1
Instructional Lead	0	0	0	1	0
School Administrator	0	1	0	0	0
Other Staff	2	0	0	1	0
ELL Teacher	1	1	1	0	0
Language Immersion Teacher	0	0	0	0	1
Special Education Teacher	1	0	0	1	0
Special Ed Teacher_ Gifted	1	0	0	0	0
Visually or Hearing Impaired Teacher	0	1	0	1	1
Black or African American	1	2	3	2	3
Asian	0	1	1	0	0
Hispanic/ Latino	0	1	0	1	0
White	7	6	2	7	8
Male	2	2	1	2	1
Female	7	8	6	9	10
Total Participants	9	10	7	11	11

Note: As teachers may fulfill multiple roles, at some grades 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 science assessments, 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 SEP, DCI, and CCC alignment for each item and recorded the degree of alignment for each dimension and overall alignment on a worksheet on a scale of 0 (not aligned) to 3 (well aligned), referring to LSS for Science Appendix A (Learning Progressions). An item was considered to have a high degree of alignment if it aligned to the particular bullet listed in the PE. An item was considered to have a lower degree of alignment if it aligned to another bullet listed in the learning progression for that SEP or

CCC. Committee members also recorded whether or not the science for each item was accurate and whether each item was free of bias. Areas of concern considered included opportunity and access, portrayal of groups represented, and protecting privacy and avoiding offensive content.

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

Following the individual reviewers’ votes, the group came together to view and discuss each stimulus and item as it was projected on-screen with the goal of achieving consensus. The WestEd facilitators compiled detailed notes about committee decisions for implementation after the review. Because of the limited time available, there was not a review and discussion of 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. Tables 3.5a–g provide these results for grades 3–8, based on the participants’ individual votes on each item following their initial review.

Table 3.5a

Grade 3 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	14	109	14	0	0	123
ER	4	26	9	0	0	35
MC	65	482	90	3	1	576
MS	19	131	35	2	0	168
TPD	50	344	95	8	0	447
TPI	8	52	20	0	0	72
All Grade 3	**160**	**1144**	**263**	**13**	**1**	**1421**

Table 3.5b

Grade 4 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	11	82	11	1	0	94
ER	4	31	3	0	1	35
MC	64	519	38	5	2	564
MS	19	161	13	2	0	176
TPD	47	391	47	3	2	443
TPI	15	105	9	0	0	114
All Grade 4	**160**	**1289**	**121**	**11**	**5**	**1426**

Table 3.5c

Grade 5 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	12	68	5	0	0	73
ER	4	20	4	0	0	24
MC	34	197	5	0	0	202
MS	8	45	4	0	1	50
TE	68	359	36	0	1	396
TPD	18	96	8	0	0	104
TPI	13	67	7	0	0	74
All Grade 5	**157**	**852**	**69**	**0**	**2**	**923**

Table 3.5d

Grade 6 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	10	56	21	2	4	83
ER	4	15	12	1	2	30
MC	55	377	96	5	11	489
MS	13	66	25	5	2	98
TE	42	256	82	6	18	362
TPD	26	154	59	5	6	224
TPI	6	41	16	0	1	58
All Grade 6	**156**	**965**	**311**	**24**	**44**	**1344**

Table 3.5e

Grade 7 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	12	95	9	1	0	105
ER	4	33	6	0	0	39
MC	50	440	67	2	0	509
MS	15	136	16	3	0	155
TE	53	466	73	7	4	550
TPD	17	142	31	0	1	174
TPI	4	34	7	0	1	42
All Grade 7	**155**	**1346**	**209**	**13**	**6**	**1574**

Table 3.5f

Grade 8 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	10	89	7	2	0	98
ER	4	32	5	2	0	39
MC	58	550	50	7	2	609
MS	14	128	19	1	2	150
TE	52	471	76	5	3	555
TPD	13	124	17	1	0	142
TPI	6	55	9	0	0	64
All Grade 8	**157**	**1449**	**183**	**18**	**7**	**1657**

At the end of the meeting, consensus votes for each grade were compiled. The number of rejected items per grade is shown in the following table. All other items reviewed at each grade were either accepted as is, or accepted with edits. None of the item sets or tasks were rejected by the committee, as the majority of items within each item set or task were accepted.

Table 3.5g
Consensus Votes by Grade

Grade	Number of Rejected Items
3	0
4	0
5	1
6	8
7	1
8	0

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-check by content editors and two additional rounds of proofreading. Any changes resulting from these reviews were submitted to the LDOE for approval.

4. Construction of Field Test Forms

While the primary purpose of the field test was to obtain data to inform construction of the operational test forms, the field test also served as an opportunity to prepare the field for the multidimensional science assessments. To best achieve both purposes, it was desirable to have the individual field test forms reflect the blueprint as closely as possible, while taking into consideration that the field test forms were shorter in length than the operational test designs (see Tables 4.1–4.2 for the field test form designs for each grade). Each task appeared on two different forms, with different items assigned to each form. Similarly, each of the item sets was repeated across two forms, to allow all approved items to be field tested.

Based on these considerations, WestEd content leads selected a combination of item sets, tasks, and standalone items, across forms, that would ensure that the field test forms' relative distribution of score points by reporting category would approximate the blueprint for the operational assessment for the tested grade. Once item sets and tasks were determined, standalone items were chosen to achieve the final desired balance. Items were selected so that forms contained a similar number of points per session and similar quantities of item types per session.

Table 4.1

Field Test Design for Science Grades 3–4

Test Session	Numbers of Items
Session 1: One FT Item Set (no CR)	1–3 FT Item Set SR items 1–3 FT Item Set TPI/TPD items 1–3 FT Item Set SR items
One FT Item Set (plus CR)	1–3 FT Item Set TPI/TPD items 1 FT Item Set CR item
FT Standalone Items	2–4 FT Standalone SR items 0–2 FT Standalone TPI/TPD items
Session 2: FT Task	2 FT Task SR Items 2 FT Task TPI/TPD items 1 FT Task ER Item
Total Items Field Tested across Forms for Grade 3	4 tasks, 8 item sets, 8 CRs, 22 standalone SR items, 10 standalone TPI/TPD items
Total Items Field Tested across Forms for Grade 4	4 tasks, 8 item sets, 8 CRs, 23 standalone SR items, 9 standalone TPI/TPD items

Table 4.2
Field Test Design for Science Grades 5–8

Test Session	Numbers of Items
Session 1: One FT Item Set (no CR)	0–3 FT Item Set SR items 0–2 FT Item Set TPI/TPD items 0–3 FT Item Set TE items 1–4 FT Item Set SR items
One FT Item Set (plus CR)	0–2 FT Item Set TPI/TPD items 0–3 FT Item Set TE items 1 FT Item Set CR item
FT Standalone Items	1–4 FT Standalone SR items 0–2 FT Standalone TPI/TPD items 0–3 FT Standalone TE items
Session 2: FT Task	0–4 FT Task SR Items 0–2 FT Task TPI/TPD items 0–4 FT Task TE items 1 FT Task ER Item
Total Items Field Tested across Forms for Grades 5	4 tasks, 8 item sets, 8 CRs, 12 standalone SR items, 7 standalone TPI/TPD items, 11 standalone TE items
Total Items Field Tested across Forms for Grades 6	4 tasks, 8 item sets, 8 CRs, 14 standalone SR items, standalone 6 TPI/TPD items, 10 standalone TE items
Total Items Field Tested across Forms for Grades 7	4 tasks, 8 item sets, 8 CRs, 16 standalone SR items, 5 standalone TPI/TPD items, 9 standalone TE items
Total Items Field Tested across Forms for Grades 8	4 tasks, 8 item sets, 8 CRs, 17 standalone SR items, 5 standalone TPI/TPD items, 9 standalone TE items

The WestEd content lead made a concerted effort to avoid cueing and clanging between field test 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 field test forms were reviewed by 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*, "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 Education Assessment Program 2025 (LEAP 2025) 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 the LEAP 2025 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 provides train-the-trainer opportunities for the school system test coordinators, who in turn convey test-administration training to schools within their system. The LDOE conducts quality-assurance visits during testing to ensure school systems' adherence to the standardized administration of the tests.

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

Ancillary Materials

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

For the spring 2018 test administration, DRC produced two administration manuals:

1. *LEAP 2025 Grades 3–4 Paper-Based Test Administration Manual*
2. *LEAP 2025 Grades 3–8 Computer-Based Test Administration Manual*

DRC also produced Test Coordinators Manuals for paper-based test administrations and for computer-based test administration. LDOE assessment staff review, provide feedback, and give final approval for these manuals. The Test Coordinators Manuals are inclusive of grades 3–8 English Language Arts (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 for scoring.

Table of Contents for Paper-Based Testing Test Coordinators Manual

- Key Dates
- Spring 2018 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
- Test Security
    - Key Definitions
    - Violations of Test Security
    - Erasure Analysis
    - Voiding Student Tests
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions
    - Testing in Class-Sized Groups
    - Test Schedule
    - Extended Time for Testing

    - Extended Breaks
    - Makeup Testing
    - Test Administration Resources
    - Testing Times for Grades 3 and 4
- School System Test Coordinator
    - Conduct Training Session
    - Receive Test Materials
    - Large-Print, Braille, and CAS Test Materials
    - Accommodated Materials
    - Verify and Distribute Test Materials to School Test Coordinators
    - Request Additional Test Materials and Bar-Code Labels
    - Collect Materials from Schools After Testing
    - Used and Unused Answer Documents and Consumable Test Booklets (Defined)
    - Unscorable Documents and Unscorable Document Labels
- Directions for Returning Test Materials to DRC in May
    - Pickup 1: ELA, Math, Science, and Social Studies
    - Pickup 2: ELA, Math, Science, and Social Studies
    - Pickup 3: Nonscorable Materials for ELA, Math, Science, and Social Studies
    - Final Checklist for Returning Test Materials to DRC
- School Test Coordinator
    - Receive and Verify Test Materials
    - Conduct Test Administration and Security Training Session
    - Supervise Application of Bar-Code Labels and Coding of Answer Documents and Consumable Test Booklets
    - Soiled, Damaged, and Other Unscorable Answer Documents and Consumable Test Booklets
    - Verify and Distribute Materials to Test Administrators
    - Supervise Test Administration
    - Collect Test Materials

    - Used and Unused Answer Documents and Consumable Test Booklets (Defined)
    - Coding Responsibilities of Principals—Before Testing
    - Coding Responsibilities of Principals—Before or After Testing
    - Coding Responsibilities of Principals—After Testing
- Directions for Returning Test Materials to the DTC
    - Pickup 1: ELA, Math, Science, and Social Studies
    - Pickup 2: ELA, Math, Science, and Social Studies
    - Pickup 3: Nonscorables for ELA, Math, Science, and Social Studies
    - Final Checklist for Returning Test Materials to the DTC
- Void Notification
- Index

Table of Contents for Computer-Based Testing Test Coordinators Manual

- Key Dates Spring 2018
- Resources Available in eDIRECT Spring 2018
- Spring 2018 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - eDIRECT and INSIGHT
- Test Security
    - Key Definitions
    - Violations of Test Security
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions
    - Testing in Class-Sized Groups
    - Testing Schedule

    - Extended Time for Testing
    - Extended Breaks
    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grades 3 through 8
- Roles and Responsibilities
    - School System Test Coordinator
    - School Test Coordinator
    - Technology Coordinator
- Managing Test Tickets
    - Student Transfers
    - Locked Test Tickets
    - Technical Issues
    - Invalidating Test Tickets
- Resources for Online Testing
    - Test Administration Manuals
    - eDIRECT User Guides
    - LEAP 2025 Accommodations and Accessibility Features User Guide
    - INSIGHT Technology User Guide
    - Online Tools Training (OTT)
    - Student Tutorials

The test administration manuals provide detailed instructions for administering the LEAP 2025 assessments. The manuals include instructions for test security, test administrator responsibilities, test preparation, administration of tests (online or paper), and post-test procedures. Following is information included in the test administration manuals.

Table of Contents for LEAP 2025 Test Administration Manual (PBT)

- Spring 2018 Notes and Reminders

- Test Administrator Pre-Administration Oath of Security and Confidentiality Statement
- Test Administrator Post-Administration Oath of Security and Confidentiality Statement
- Overview
- Test Security
    - Secure Test Materials
    - Testing Irregularities and Security Breaches
    - Testing Environment
    - Violations of Test Security
    - Answer Change Analysis
    - Voiding Student Tests
- Test Administrator Responsibilities
- Test Administration Checklists
    - Before Testing
    - During Testing
    - After Testing (Daily)
    - After Testing (Last Day)
- Test Administrators' Frequently Asked Questions
- Test Materials
    - Receipt of Test Materials
- Testing Guidelines
    - Testing Eligibility
    - Test Schedule
    - Extended Time for Testing
- Testing Times for Grades 3 and 4
    - Makeup Testing
    - Testing Conditions
- Special Populations and Accommodations

    - IDEA Special Education Students
    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard-of-Hearing Students
    - English Learners (ELs)
- Hand-Coded Consumable Test Booklets and Answer Documents
- Students Absent from Testing
- Consumable Test Booklet and Answer Document Coding
    - Coding the Demographic Section
- Sample Grade 3 English Language Arts Consumable Test Booklet
- General Instructions
    - Student Marking/Erasing on Consumable Test Booklet
    - Reading Directions to Students
    - Special Instructions
- Directions for Administering LEAP 2025
- Post-Test Procedures
    - Test Administrator Oath of Security and Confidentiality Statement
    - Used and Unused Answer Documents and Consumable Test Booklets (Defined)
    - Transferring Student Responses
    - Returning Test Materials to the School Test Coordinator
- Index

Table of Contents for LEAP 2025 Test Administration Manual (CBT)

- Spring 2018 Notes and Reminders
- Test Administrator Pre-Administration Oath of Security and Confidentiality Statement

- Test Administrator Post-Administration Oath of Security and Confidentiality Statement
- Overview
- Test Security
    - Secure Test Materials
    - Testing Irregularities and Security Breaches
    - Testing Environment
    - Violations of Test Security
    - Voiding Student Tests
- Test Administrator Responsibilities
    - Software Tools and Features for Test Administrators
- Test Administration Checklists
    - Before Testing
    - During Testing
    - After Testing (Daily)
    - After Testing (Last Day)
- Test Administrators' Frequently Asked Questions
- Testing Guidelines
    - Testing Eligibility
    - Testing Schedule
    - Extended Time for Testing
    - Testing Times for Grades 3 through 8
    - Makeup Testing
    - Testing Conditions
- Online Tools Training
- Student Tutorials
- Directions for Administering the Grades 3–8 Computer-Based LEAP 2025 Tests
- Special Populations and Accommodations
    - IDEA Special Education Students

    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard-of-Hearing Students
    - English Learners (ELs)
- Students Absent from Testing
- Test Materials
    - Receipt of Test Materials
- General Instructions
    - Reading Directions to Students
- Post-Test Procedures
    - Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    - Returning Test Materials to the School Test Coordinator
- Index

The *Standards* contain multiple references relevant to test administration. Information in the LEAP 2025 test administration manuals addresses these in the following manner.

Directions for test administration found in the manual addresses Standard 4.15, which states: "The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented." (90)
The LEAP 2025 test administration manuals provide instructions for 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 paper and online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

Furthermore, the standardized procedures addressed in the test administration manual need to be followed, as the *Standards* state in Standard 6.1: "Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user" (114). To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the LEAP 2025 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 administer reports on testing concerns that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; allowing students to correct or complete answers after tests have been submitted; splitting sessions into two parts; ignoring the standardized directions in the online assessment; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the Individualized Education Program (IEP/IAP/EL plan); allowing accommodations for students who do not have an IEP/IAP/EL plan; or defining terms on the test.

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

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

- Determine the layout of the classroom environment.
- Plan seating arrangements. Allow enough space between students to prevent 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 large-print or braille version of the LEAP 2025, examiners are instructed to transcribe students' responses from the large-print test or braille test book into the online testing system (INSIGHT) exactly as they responded in the large-print or 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.

Return Material Forms and Guidelines. The Test Coordinators Manual instructs test coordinators regarding procedures for organizing and packing materials and returning them to DRC for secure inventory purposes. LDOE assessment staff have opportunities to review, provide feedback, and give final approval. The purpose of the instructions is to ensure that secure test materials are properly accounted for and organized appropriately for return shipment.

Security Checklists. As soon as printed test materials are received by a school system, the school system test coordinator ensures that the first and last security bar codes on

the tests match the packing list he or she received. The school system test coordinator then packages the tests to be sent to schools. Upon returning the test books to DRC, school and school system test coordinators are required to complete and submit an accountability form that details the number of test books or printed test forms returned. This form also requires that systems/schools document nonstandard situations, including lost, damaged, destroyed, extra, or missing test books.

Time

Each session of each content area test was timed to provide sufficient time for students to attempt all items. Only students with an extended time accommodation were permitted to exceed the established time limits of any given session. The test administration manuals provided examiners with timing guidelines for the assessments.

Online Forms Administration, Grades 3–8

The online forms were administered via DRC's INSIGHT online assessment system. School system and school personnel set up test sessions via DRC's online testing portal, eDIRECT, and printed test tickets. Students entered their ticket information to access the test in INSIGHT. In addition, students had access to Online Tools Training, which allowed them to practice using tools and features within INSIGHT. Students were required to experience the Online Tools Training (OTT) before the computer-based test administration. The OTT allows students to observe and practice features of the Online Assessment Software prior to an actual test administration. Students were also required to view the Student Tutorials, which present visual and verbal descriptions of the properties and features of the DRC INSIGHT Online Assessment Software.

Paper-Based Forms Administration, Grades 3 and 4

Schools with students in grades 3 and 4 had the option to administer the test via computer or paper-based testing. DRC printed and shipped paper materials to the sites that opted for paper-based testing. These materials were then returned to DRC after the field test, for processing and scoring with the online tests.

Accommodations

Accommodations that are allowed on the LEAP 2025 are listed in the LEAP 2025 Accommodations and Accessibility Features User Guide (https://www.louisianabelieves.com/docs/default-source/assessment/leap-accessibility-and-accommodations-manual.pdf?sfvrsn=6). Designated Supports 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)

The following are examples of accommodations offered for this administration:

- Text-to-speech for online testers
- Braille
- Large print (for grade 3 and 4 paper testers, as the online testing system has a magnification capacity for those testers)
- Kurzweil (a text-to-speech accommodation for grade 3 and 4 paper testers)
- Human reader (for grade 3 and 4 paper testers)

For more details about these accommodations, please refer to the *LEAP 2025 Accommodations and Accessibility Features User Guide*.

Testing Windows

The computer-based test window was available from Monday, April 2, through Friday, May 4, 2018. Paper-based testing occurred from April 30 through May 4, 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 LEAP 2025 assessments. Test security procedures are discussed throughout the Test Coordinators Manual and test administration manuals.

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

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

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 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 too few students selected the correct response (*p*-value less than 0.15), if the item did not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) selected a certain incorrect response. Lists of flagged items, with the reasons for flagging, were provided to WestEd content staff for key verification. Scoring of MS items was evaluated at data review.

Scoring of TEs and Adjudication. 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 TEs. For adjudication, DRC provided a report listing the frequency distributions of TE responses and an auto-frequency report detailing the multi-part multi-select items. LDOE and WestEd examined the TE response distributions and the auto-frequency reports to evaluate whether the items were scored appropriately. No TE 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 and Extended-Response Scoring. A full description of the methods used to score constructed and extended responses can be found in the *LEAP Processing Rules—Scoring (LEAP Spring 2018)* document. The document table of contents is listed below.

- Schedule, Locations, and Staffing
  - Training and Scoring Schedule
  - Scorer Degree Requirements
- Training
  - Social Studies Training Materials
- Qualifying
- Reader Monitoring Procedures
  - Team Leader Read-Behinds
  - Validity Responses
  - Recalibration Sets

    - Inter-Rater Reliability
    - Handscoring Quality Control Reports
        - Scoring Summary Report Sample—8-Point, Two-Trait Extended-Response Item
        - Scoring Summary Report Sample—2-Point Constructed-Response Item
    - Expected Agreement Rates (Inter-Rater Reliability and Validity)
        - Spring 2018 Operational Extended-Response Items—2017 Field Test Handscoring Data
        - Spring 2018 Operational Constructed-Response Items—2017 Field Test Handscoring Data
    - Reader Feedback Logs
- Handscoring Rules
    - AI Scoring of Grade 5–8 ER Items
    - Scoring of CR Items and Grade 3 & 4 ER Items
- Handling Unusual Responses
    - Nonscore Blanks
        - Grade 3–8 CR and ER Nonscore Code
    - Alerts
- Artificial Intelligence Scoring
    - Model Building
    - 2018 Model Enhancement Process
    - Evaluation Metric
    - Scoring Responses with the AI Engine
    - Quality Control of the AI Engine (MI)
    - Scoring (DRC)
    - Identifying Responses for Human Review
        - Alert Detection System
        - Identification of Non-Alert Responses Requiring Human Review
        - Identifying Copied Text and Plagiarism with the AI Engine
- Appendix

Tables 6.1–6.4 provide the inter-rater reliability and score point distributions by grade level for the constructed-response and extended-response items administered in the spring 2018 forms.

Table 6.1

Constructed-Response Inter-Rater Reliability

Grade	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
3	Item1	≥340	76	23	1
	Item2	≥320	96	3	1
	Item3	≥320	88	12	0
	Item4	≥350	87	13	0
	Item5	≥310	96	3	2
	Item6	≥320	87	13	0
	Item7	≥310	85	15	0
4	Item1	≥310	93	7	0
	Item2	≥330	85	15	0
	Item3	≥310	96	4	0
	Item4	≥330	76	24	0
	Item5	≥320	88	12	0
	Item6	≥320	90	10	0
	Item7	≥330	79	21	0
	Item8	≥310	82	18	0
5	Item1	≥330	87	13	0
	Item2	≥380	98	2	0
	Item3	≥320	96	4	1
	Item4	≥340	89	7	4
	Item5	≥350	93	6	1
	Item6	≥330	88	13	0
	Item7	≥330	91	7	2
6	Item1	≥310	88	11	1
	Item2	≥300	92	8	0
	Item3	≥320	86	14	0
	Item4	≥310	86	14	0
	Item5	≥320	91	9	0
	Item6	≥330	94	6	0
	Item7	≥310	96	4	0

Grade	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
7	Item1 (Two 0-1 parts)	≥320	96	4	0
			95	5	0
	Item2	≥320	97	3	0
	Item3	≥330	92	8	0
	Item4	≥330	92	8	0
	Item5	≥330	92	8	0
	Item6	≥340	91	9	0
	Item7	≥310	82	18	0
8	Item1	≥320	88	12	0
	Item2	≥330	74	24	1
	Item3	≥330	85	15	0
	Item4	≥320	86	13	1
	Item5	≥320	90	10	0
	Item6	≥300	80	19	1

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

Table 6.2

Constructed-Response Score Point Distributions

Grade	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
3	Item1	≥1,670	62	31	6	0
	Item2	≥1,660	87	5	5	2
	Item3	≥1,660	36	61	2	0
	Item4	≥1,670	40	40	19	1
	Item5	≥1,650	93	2	4	2
	Item6	≥1,660	58	33	7	1
	Item7	≥1,650	75	23	1	1
4	Item1	≥1,650	73	27	0	0
	Item2	≥1,660	52	46	0	1
	Item3	≥1,650	71	25	3	1
	Item4	≥1,660	46	37	16	1
	Item5	≥1,660	83	14	1	2
	Item6	≥1,660	75	20	5	0
	Item7	≥1,660	45	49	6	0
	Item8	≥1,650	63	36	0	1
5	Item1	≥1,660	59	34	6	0
	Item2	≥1,690	56	7	37	0
	Item3	≥1,650	69	12	19	0
	Item4	≥1,660	29	51	19	0
	Item5	≥1,670	66	19	14	0
	Item6	≥1,660	65	33	1	0
	Item7	≥1,660	54	30	16	0
6	Item1	≥1,650	68	29	3	0
	Item2	≥1,640	23	72	5	0
	Item3	≥1,650	63	32	4	0
	Item4	≥1,650	81	14	5	0
	Item5	≥1,640	74	20	6	0
	Item6	≥1,660	85	12	2	0
	Item7	≥1,650	82	11	7	0

Grade	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
7	Item1 (Two 0-1 parts)	≥1,630	96	3		0
			87	12		0
	Item2	≥1,640	84	12	3	0
	Item3	≥1,640	90	9	0	0
	Item4	≥1,650	39	42	19	0
	Item5	≥1,640	92	6	1	0
	Item6	≥1,660	89	9	0	0
	Item7	≥1,640	30	50	20	0
8	Item1	≥1,640	76	20	3	0
	Item2	≥1,650	55	34	11	0
	Item3	≥1,640	87	12	0	0
	Item4	≥1,630	42	51	6	0
	Item5	≥1,650	87	11	1	0
	Item6	≥1,630	9	63	28	0

Table 6.3

Extended-Response Inter-Rater Reliability

Grade	Inter-Rater Reliability					
	Item	2x	Part	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
3	Item 1	≥530	(0-6)	78	18	4
	Item 2	≥560	Part A (0-4)	89	9	3
			Part B (0-2)	96	4	1
	Item 3	≥540	Part A (0-4)	94	6	0
			Part B (0-2)	97	3	0
4	Item1	≥550	(0-6)	65	33	2
	Item2	≥540	(0-6)	91	8	1
	Item3	≥550	(0-6)	74	23	3
	Item4	≥540	(0-6)	75	17	9
5	Item 1	≥4,990	(0-9)	67	23	10
	Item 2	≥4,990	(0-9)	75	16	9
	Item 3	≥4,990	Part A (0-4)	81	17	2
			Part B (0-5)	78	19	3

Grade	Inter-Rater Reliability					
	Item	2x	Part	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
6	Item 1	≥4,990	Part A (0-2)	91	8	1
			Part B (0-7)	71	15	14
	Item 2	≥4,980	Part A (0-3)	84	13	2
			Part B (0-4)	87	11	2
			Part C (0-2)	91	9	0
	Item 3	≥4,980	Part A (0-3)	86	8	6
			Part B (0-3)	80	19	2
			Part C (0-3)	85	12	3
7	Item 1	≥4,950	Part A (0-3)	78	16	6
			Part B (0-4)	81	15	4
			Part C (0-2)	96	4	0
	Item 2	≥4,950	Part A (0-3)	68	30	2
			Part B (0-3)	83	15	1
			Part C (0-3)	92	8	1
	Item 3	≥4,990	(0-9)	58	26	15
8	Item 1	≥4,960	Part A (0-3)	83	14	2
			Part B (0-2)	86	12	2
			Part C (0-4)	79	12	9
	Item 2	≥4,970	Part A (0-2)	87	12	1
			Part B (0-6)	71	19	9
			Part C (0-1)	88	12	0
	Item 3	≥4,960	Part A (0-3)	93	6	0
			Part B (0-2)	87	12	0
			Part C (0-4)	74	22	4
	Item 4	≥4,950	Part A (0-3)	62	30	8
			Part B (0-6)	49	32	20

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

Table 6.4

Extended-Response Score Point Distributions

GR	Score Point Distribution													
	Item	Total	Part	% "0"	%"1"	% "2"	% "3"	% "4"	% "5"	% "6"	% "7"	% "8"	% "9"	% Blank
3	Item 1	≥2,760	(0-6)	63	13	15	4	4	0	0				1
	Item 2	≥2,770	Part A (0-4)	82	10	5	1	1						0
			Part B (0-2)	92	4	3								0
	Item 3	≥2,770	Part A (0-4)	87	9	2	1	0						1
			Part B (0-2)	94	5	1								1
4	Item 1	≥2,770	(0-6)	33	29	21	10	5	1	0				0
	Item 2	≥2,770	(0-6)	88	7	3	1	0	0	0				1
	Item 3	≥2,770	(0-6)	6	13	40	37	3	0	0				0
	Item 4	≥2,770	(0-6)	44	23	19	9	3	1	0				1
5	Item 1	≥4,990	(0-9)	42	12	11	9	9	6	5	3	2	1	0
	Item 2	≥4,990	(0-9)	47	18	12	9	7	4	2	2	0	0	0
	Item 3	≥4,990	Part A (0-4)	56	29	13	1	0						0
			Part B (0-5)	55	25	14	3	2	0					0

Table 6.4 continued

Extended-Response Score Point Distributions

GR	Score Point Distribution													
	Item	Total	Part	% "0"	%"1"	% "2"	% "3"	% "4"	% "5"	% "6"	% "7"	% "8"	% "9"	% Blank
6	Item 1	≥4,990	Part A (0-2)	75	16	9								0
			Part B (0-7)	35	5	18	7	14	7	9	4			0
	Item 2	≥4,980	Part A (0-3)	67	14	13	5							0
			Part B (0-4)	75	14	8	2	0						0
			Part C (0-2)	58	36	6								0
	Item 3	≥4,980	Part A (0-3)	68	19	0	13							0
			Part B (0-3)	58	29	11	2							0
			Part C (0-3)	62	17	19	2							0
7	Item 1	≥4,950	Part A (0-3)	71	17	10	2							0
			Part B (0-4)	71	19	8	1	0						0
			Part C (0-2)	88	10	1								0
	Item 2	≥4,950	Part A (0-3)	40	46	11	3							0
			Part B (0-3)	84	13	2	0							0
			Part C (0-3)	79	15	5	0							0
	Item 3	≥4,990	(0-9)	8	6	13	17	19	13	15	5	1	3	0

Table 6.4 continued

Extended-Response Score Point Distributions

GR	Score Point Distribution													
	Item	Total	Part	% "0"	%"1"	% "2"	% "3"	% "4"	% "5"	% "6"	% "7"	% "8"	% "9"	% Blank
8	Item 1	≥4,960	Part A (0-3)	72	19	8	0							0
			Part B (0-2)	77	16	7								0
			Part C (0-4)	67	13	17	1	2						0
	Item 2	≥4,970	Part A (0-2)	76	22	2								0
			Part B (0-6)	40	16	21	8	10	3	2				0
			Part C (0-1)	79	20									0
	Item 3	≥4,960	Part A (0-3)	89	9	2	0							0
			Part B (0-2)	67	26	7								0
			Part C (0-4)	44	24	17	10	4						0
	Item 4	≥4,950	Part A (0-3)	28	28	28	15							0
			Part B (0-6)	12	13	20	21	17	10	7				0

7. Data Analysis

Classical Item Statistics

A measure of item difficulty, *p* (or "the *p*-value"), indicates the average proportion of total points earned on an item. For example, if *p* = 0.50 on an MC item, then half of the examinees earned a score of 1. If *p* = 0.50 on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). The item-total correlation (point-biserial) is in general 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). Tables 7.1 through 7.6 provide summary item statistics by grade level–item type that were field tested.

Table 7.1
Summary of Classical Statistics for Field Test Items for Science Grade 3

Grade 3							
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	60	0.45	0.14	0.32	0.08	3%	0%
MS	17	0.23	0.10	0.33	0.14	6%	0%
CR	7	0.20	0.13	0.43	0.06	14%	0%
ER	6	0.07	0.04	0.55	0.08	0%	0%
TPI	5	0.43	0.14	0.45	0.10	0%	0%
TPD	46	0.35	0.15	0.46	0.12	2%	0%

*B-Level DIF is typically reviewed with lower priority than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.2

Summary of Classical Statistics for Field Test Items for Science Grade 4

Grade 4							
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	55	0.48	0.17	0.32	0.08	5%	2%
MS	15	0.35	0.19	0.32	0.11	7%	7%
CR	8	0.20	0.09	0.47	0.06	13%	0%
ER	8	0.20	0.13	0.56	0.12	0%	0%
TPI	13	0.41	0.13	0.44	0.10	8%	0%
TPD	45	0.37	0.14	0.45	0.09	7%	2%

*B-Level DIF is typically reviewed with lower priority than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.3

Summary of Classical Statistics for Field Test Items for Science Grade 05

Grade 5							
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	32	0.52	0.17	0.33	0.09	0%	0%
MS	7	0.24	0.18	0.28	0.13	0%	0%
CR	7	0.29	0.10	0.45	0.08	0%	0%
ER	6	0.17	0.04	0.71	0.06	0%	0%
TE	60	0.42	0.18	0.40	0.11	10%	2%
TPI	13	0.44	0.10	0.40	0.10	0%	0%
TPD	16	0.35	0.15	0.41	0.13	0%	0%

*B-Level DIF is typically reviewed with lower priority than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.4

Summary of Classical Statistics for Field Test Items for Science Grade 6

Grade 6							
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	47	0.44	0.14	0.30	0.11	4%	2%
MS	13	0.29	0.14	0.32	0.15	15%	0%
CR	7	0.18	0.11	0.43	0.03	29%	0%
ER	6	0.22	0.07	0.75	0.02	33%	0%
TE	37	0.38	0.20	0.35	0.14	14%	3%
TPI	6	0.37	0.15	0.39	0.16	17%	0%
TPD	25	0.33	0.15	0.39	0.11	4%	0%

*B-Level DIF is considered a more stringent criterion than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.5

Summary of Classical Statistics for Field Test Items for Science Grade 7

Grade 7							
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	46	0.39	0.14	0.27	0.12	9%	0%
MS	15	0.27	0.15	0.34	0.10	27%	0%
CR	7	0.16	0.18	0.36	0.13	0%	14%
ER	6	0.22	0.16	0.69	0.06	0%	0%
TE	46	0.46	0.22	0.38	0.14	15%	2%
TPI	4	0.41	0.10	0.38	0.05	0%	0%
TPD	17	0.33	0.19	0.36	0.13	0%	0%

*B-Level DIF is considered a more stringent criterion than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.6

Summary of Classical Statistics for Field Test Items for Science Grade 8

Grade 8							
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	53	0.42	0.15	0.29	0.11	11%	0%
MS	11	0.32	0.19	0.33	0.15	27%	0%
CR	6	0.23	0.20	0.41	0.10	17%	0%
ER	8	0.25	0.13	0.77	0.03	0%	0%
TE	46	0.46	0.19	0.39	0.11	13%	7%
TPI	6	0.36	0.07	0.40	0.11	0%	0%
TPD	12	0.32	0.11	0.39	0.09	0%	0%

*B-Level DIF is considered a more stringent criterion than C-Level DIF for large-scale educational assessments (Zwick, 2012).

Table 7.7 summarizes the numbers of field-tested items at each grade level that were flagged according to defined criteria (item mean < 0.50 for TE, TPD, TPI, or CR item types; item mean < 1.5 for ER item types at grades 3–4; item mean < 2.25 for ER item types for grades 5–8). The box plots that follow illustrate the range of item *p*-values by grade level and good item–total discriminating power exhibited overall by grade level.

Science items in all formats (e.g., TEI and CR) were newly developed and field-tested in 2017–2018. However, it seems like more MC and MS items are flagged compared to other item types. It may be members of a larger, theme-based set or task exhibiting different model-data fit than that specified by the IRT item fit model. Until further empirical research clearly answers this curiosity, we continue to pay attention to whether similar patterns are observed in coming years' test administrations.

Table 7.7

Number of Field Test Items Flagged for Item Statistics

Grade	Item Type	*N* Items	Flagged for *p*-value	Flagged for Mean	Flagged for Point-Biserial Correlation	Flagged for DIF	Flagged for IRT Fit	Flagged for Omit
3	CR	7	0	5	0	1	3	0
	ER	6	0	6	0	0	4	0
	MC	60	3	0	6	2	25	0
	MS	17	9	0	4	1	4	0
	TPD	46	0	13	2	1	8	0
	TPI	5	0	0	0	0	2	0
4	CR	8	0	6	0	1	2	0
	ER	8	0	6	0	0	2	0
	MC	55	2	0	6	4	25	0
	MS	15	6	0	3	2	4	0
	TPD	45	0	11	0	3	7	0
	TPI	13	0	3	0	1	5	0
5	CR	7	0	4	0	0	2	0
	ER	6	0	6	0	0	2	0
	MC	32	2	0	2	0	13	0
	MS	7	5	0	3	0	5	0
	TE	60	7	5	3	7	18	0
	TPD	16	0	4	1	0	3	0
	TPI	13	0	1	1	0	4	0
6	CR	7	0	6	0	2	2	0
	ER	6	0	4	0	2	0	0
	MC	47	3	0	10	3	16	0
	MS	13	6	0	3	2	3	0
	TE	37	6	5	5	6	14	0
	TPD	25	0	9	2	1	5	0
	TPI	6	0	1	1	1	1	0
7	CR	7	0	5	0	1	4	0
	ER	6	0	4	0	0	2	0
	MC	46	8	0	12	4	23	0
	MS	15	8	0	1	4	4	0
	TE	46	6	4	7	7	11	0
	TPD	17	0	6	1	0	4	0
	TPI	4	0	0	0	0	0	0
8	CR	6	0	3	0	1	3	0
	ER	8	0	6	0	0	0	0
	MC	53	7	0	12	6	24	0
	MS	11	4	0	3	3	5	0
	TE	46	1	4	3	8	13	0
	TPD	12	0	3	1	0	4	0
	TPI	6	0	0	0	0	2	0

Figure 7.1

Box Plot of Item p-Values by Grade

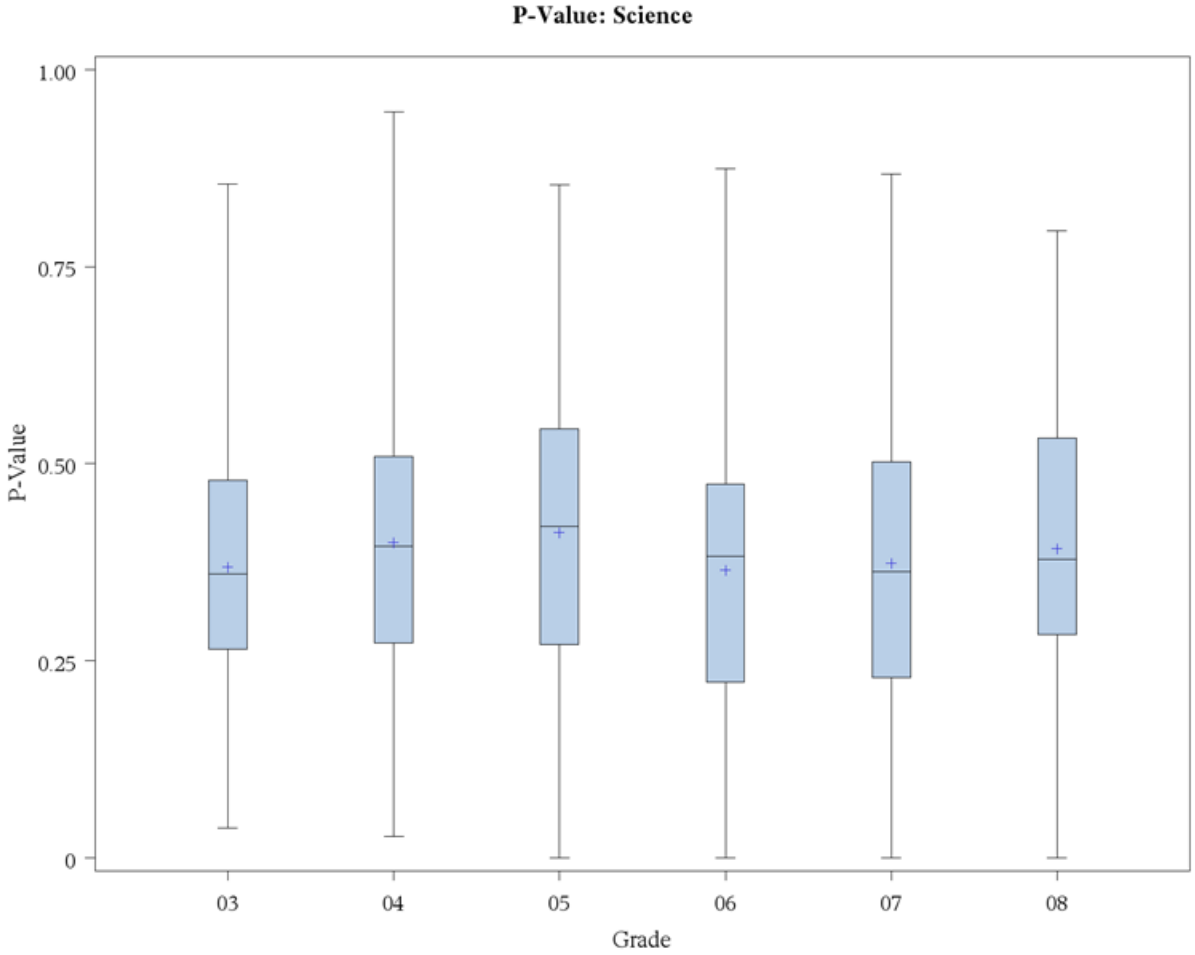
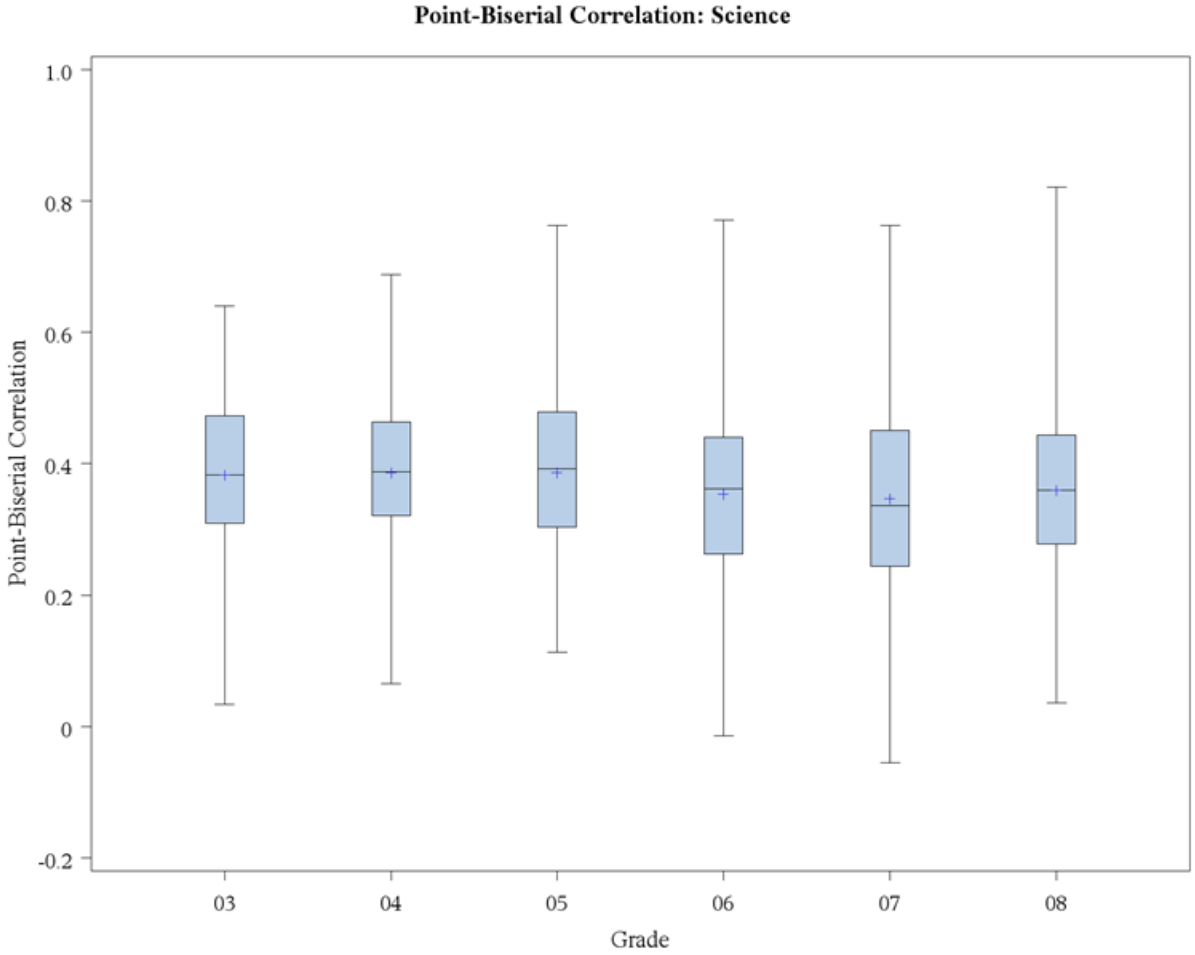


Figure 7.2

Box Plot of Item–Total Correlations/Point Biserial (PBIS) by Grade



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.

Analysis of Differential Item Functioning (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 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 2×2×*M* (2 groups × 2 item scores × *M* strata) frequency table, in which students in the reference (male or white) and focal (female or black) groups are matched on their total raw scores.

The *ΔMH DIF* is computed as

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

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

The MH chi-square statistic and the *ΔMH DIF* were used in combination to identify the field test items that exhibit strong, weak, or no DIF (Zieky, 1993). Table 7.8 defines the DIF categories for dichotomous items.

Table 7.8

DIF Categories for Dichotomous Items

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

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

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

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

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

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

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

The *p*-value associated with the Mantel χ^2 statistic and the *SMD* (on an effect-size metric) are used to determine DIF classifications. Table 7.9 defines the DIF categories for polytomous items.

Table 7.9

DIF Categories for Polytomous Items

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

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 Tables 7.10 and 7.11 provide the number 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 χP2 P and SMD statistics. Tables 7.10 and 7.11 summarize the number of items showing strong DIF associated with any group comparison.

Table 7.10

Summary of Female – Male DIF Flags for Field Test Items for Science by Grade

Female – Male			
Grade	A	B,[B-]	C,[C-]
03	140	0,[0]	0,[0]
04	138	0,[1]	0,[1]
05	135	1,[2]	0,[0]
06	132	2,[3]	0,[0]
07	134	0,[2]	1,[0]
08	134	1,[4]	0,[0]

Table 7.11

Summary of African American – White DIF Flags for Field Test Items for Science by Grade

African American – White			
Grade	A	B,[B-]	C,[C-]
03	135	0,[5]	0,[0]
04	131	0,[6]	0,[3]
05	134	0,[3]	0,[1]
06	127	1,[8]	0,[1]
07	124	3,[9]	0,[1]
08	127	1,[9]	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 science assessments are standards-based assessments that have been constructed to align to the Louisiana Student Standards for Science 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 science assessments. 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, 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 item response theory (IRT) parameters from LEAP 2025 assessment data. Multiple-choice (MC), multiple-select (MS), and some technology-enhanced (TE) items were scored dichotomously (0/1), so the 3-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.

The 2018 field test included five types of polytomous items: TEs scored 0–2, CR items scored 0–2, TPI items scored 0–2, TPD items scored 0–2, and ER items scored 0–6 for grades 3 and 4 or 0–9 for grades 5 through 8. Data from polytomous items were used to estimate parameters for the generalized partial credit model (GPCM) (Muraki, 1992):

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

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

Field Test Item Parameters

The distributions of item parameters are summarized by grade in Tables 7.12–7.17. 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 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 0.30) can be defined in the annual Framework documents that 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

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

where

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

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

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

and

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

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

A summary of IRT item parameter statistics and item fit by grade is displayed in Tables 7.12 through 7.17.

Table 7.12

Summary of IRT Statistics for Field Test Items for Science Grade 3

Grade 3								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	60	1.07	1.78	0.60	0.28	0.14	0.13	58
MS	17	3.47	5.04	0.63	0.34	0.03	0.04	76
CR	7	2.06	0.89	0.49	0.04			57
ER	6	2.36	0.64	0.51	0.06			33
TPI	5	0.95	1.32	0.40	0.20			60
TPD	46	1.33	3.47	0.33	0.15			83

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

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

Table 7.13

Summary of IRT Statistics for Field Test Items for Science Grade 4

Grade 4								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	55	2.27	10.6	0.63	0.34	0.15	0.13	55
MS	15	1.51	2.17	0.55	0.27	0.02	0.05	73
CR	8	2.36	1.04	0.64	0.14			75
ER	8	2.10	0.62	0.41	0.05			75
TPI	13	1.26	1.80	0.39	0.16			62
TPD	45	1.30	2.10	0.31	0.12			84

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

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

Table 7.14

Summary of IRT Statistics for Field Test Items for Science Grade 5

Grade 5								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	32	1.87	6.46	0.64	0.40	0.14	0.12	59
MS	7	2.41	1.72	0.54	0.11	0.05	0.04	29
CR	7	1.46	0.91	0.39	0.09			71
ER	6	2.08	0.40	0.29	0.04			67
TE	60	1.11	1.93	0.45	0.23	0.07	0.10	70
TPI	13	–12	45.1	0.31	0.14			69
TPD	16	2.00	2.40	0.28	0.16			81

*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.

Table 7.15

Summary of IRT Statistics for Field Test Items for Science Grade 6

Grade 6								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	47	1.76	2.56	0.65	0.39	0.16	0.13	66
MS	13	2.16	4.34	0.49	0.44	0.03	0.05	77
CR	7	2.03	0.58	0.54	0.14			71
ER	6	1.96	0.34	0.32	0.12			100
TE	37	1.42	7.43	0.35	0.25	0.07	0.09	62
TPI	6	9.61	27.6	0.30	0.25			83
TPD	25	7.06	30.4	0.23	0.13			80

*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.

Table 7.16

Summary of IRT Statistics for Field Test Items for Science Grade 7

Grade 7								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	46	3.13	11.3	0.61	0.43	0.18	0.11	50
MS	15	3.76	6.66	0.67	0.36	0.05	0.05	73
CR	7	2.79	1.70	0.55	0.21			43
ER	6	1.72	1.05	0.33	0.03			67
TE	46	0.06	6.10	0.44	0.33	0.07	0.06	76
TPI	4	0.99	1.27	0.26	0.06			100
TPD	17	4.05	5.53	0.25	0.16			76

*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.

Table 7.17

Summary of IRT Statistics for Field Test Items for Science Grade 8

Grade 8								
Item Type	*N* Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit (no model fit issues)
MC	53	1.93	2.82	0.63	0.36	0.15	0.13	55
MS	11	2.05	3.00	0.66	0.35	0.03	0.04	55
CR	6	2.07	1.79	0.56	0.12			50
ER	8	1.61	0.82	0.37	0.06			100
TE	46	0.98	2.51	0.49	0.29	0.18	0.20	72
TPI	6	1.35	0.92	0.35	0.19			67
TPD	12	2.20	2.55	0.21	0.09			67

*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 by Grade

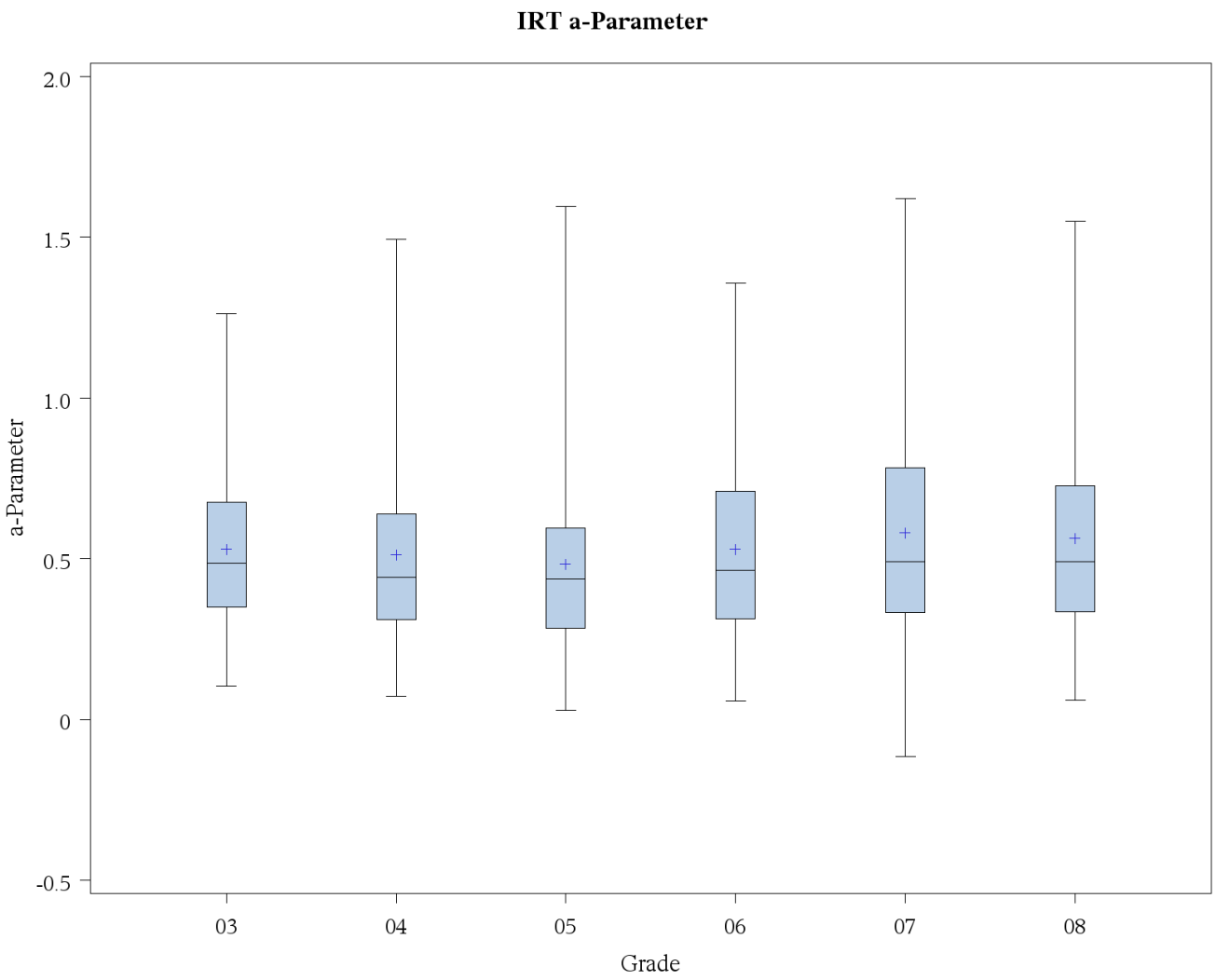


Figure 7.4

Box Plot of IRT B Parameters by Grade

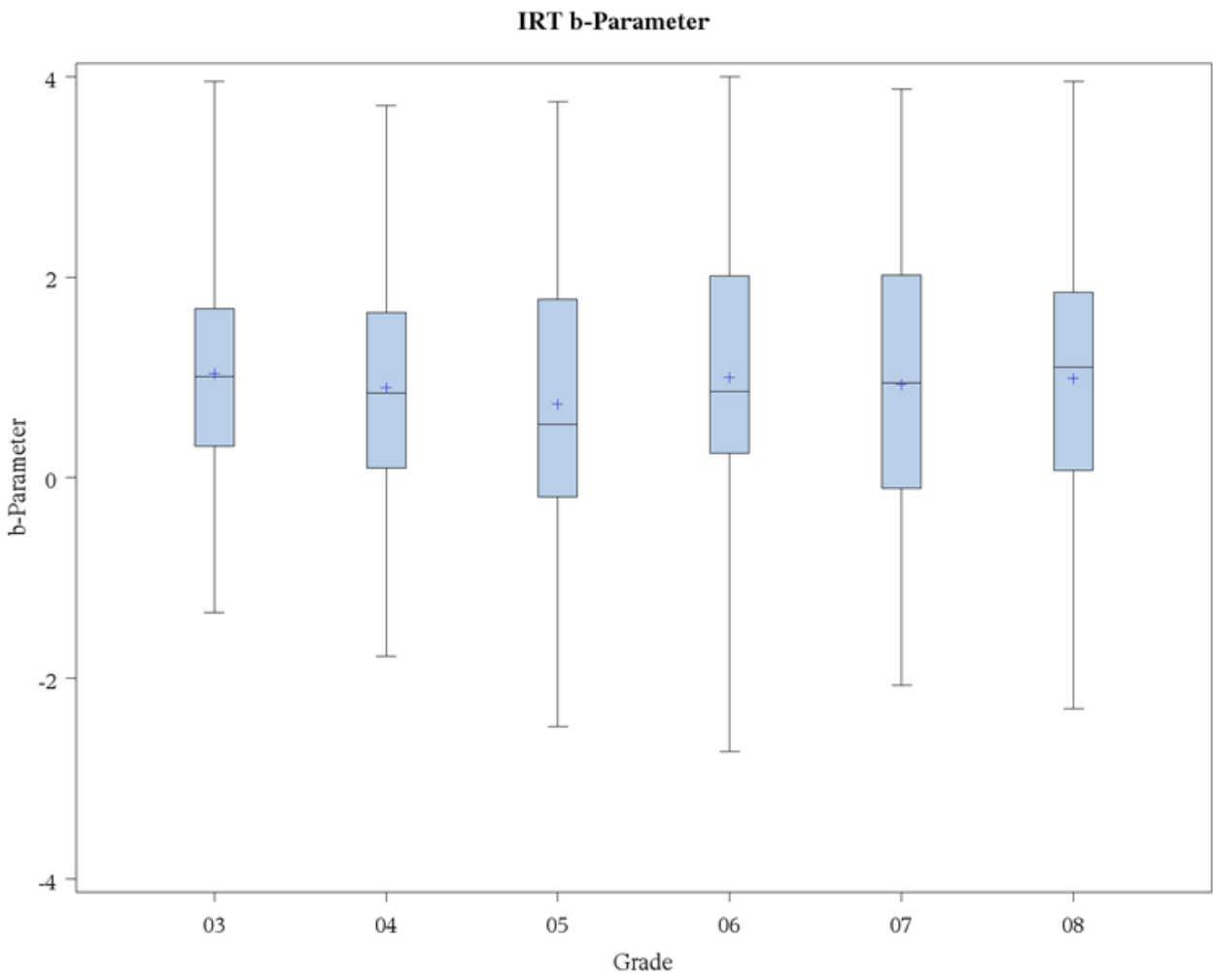
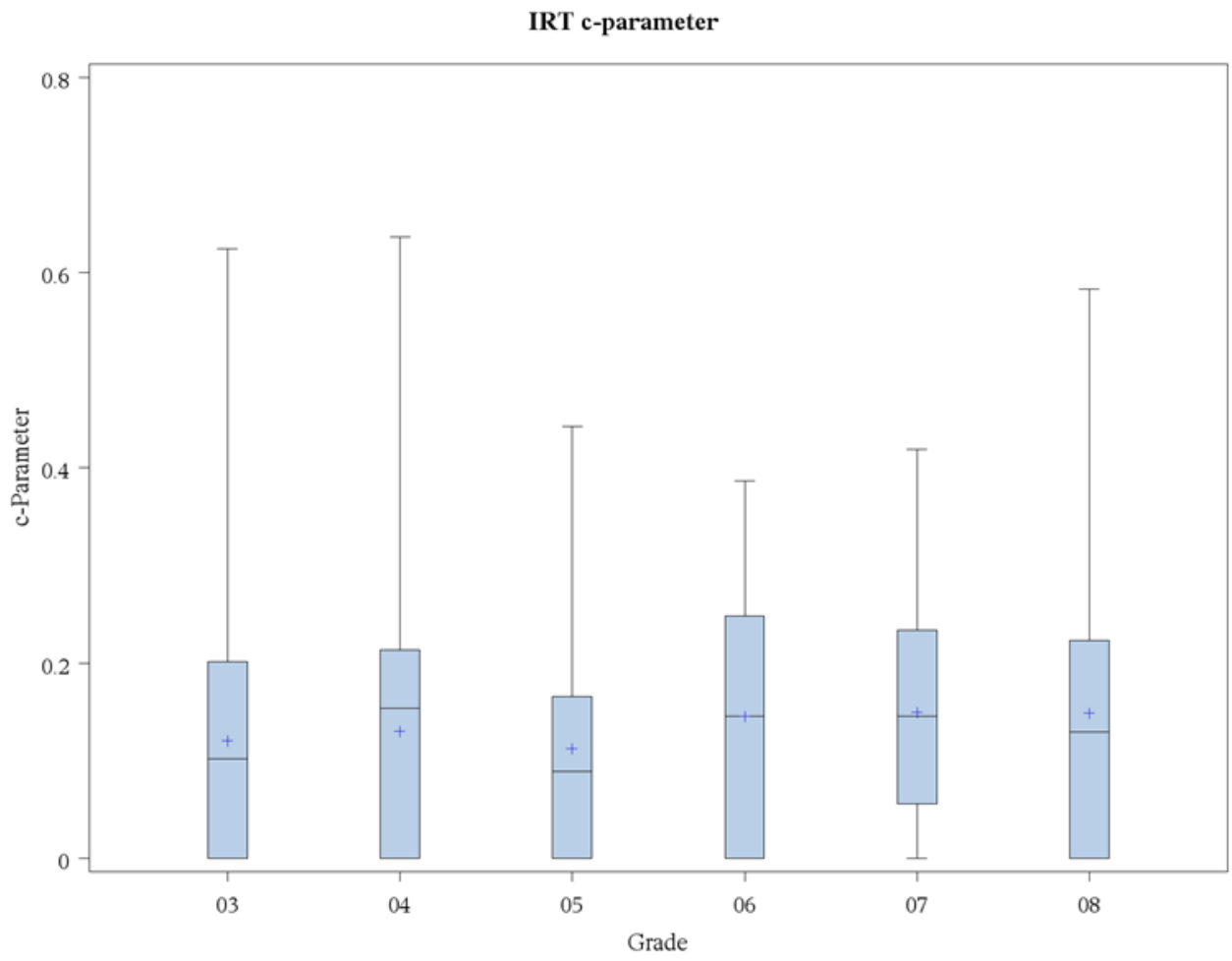


Figure 7.5

Box Plot of IRT C Parameters by Grade



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 meetings for science assessment items 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 with access to ABBI, Pearson's item-authoring tool. Participants reviewed items in a computer-based format through the ABBI platform. Content and psychometric representatives from the LDOE were present in the meetings.

Facilitators from Pearson and WestEd led the data review committees through the review of field-tested items. 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 statistical DIF were identified to have content-based flaws that could be attributed 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. 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 either the WestEd or Pearson facilitator and recorded on one main judgment form.

Tables 8.1 through 8.6 summarize 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 for each science grade level.

Table 8.1

Summary of Grade 3 Data Review Votes

Item Type	No. of Items		
	Accept	Accept w/Edits	Total
CR	5	2	7
ER	2	1	3
MC	57	2	59
MS	15	2	17
TPD	41	5	46
TPI	5	–	5
Total	125	12	137

Table 8.2

Summary of Grade 4 Data Review Votes

Item Type	No. of Items			
	Accept	Accept w/Edits	Reject	Total
CR	6	2	–	8
ER	4	–	–	4
MC	52	1	1	54
MS	14	–	–	14
TPD	41	2	–	43
TPI	12	1	–	13
Total	129	6	1	136

Table 8.3

Summary of Grade 5 Data Review Votes

Item Type	No. of Items			
	Accept	Accept w/Edits	Reject	Total
CR	7	–	–	7
ER	3	–	–	3
MC	28	3	–	31
MS	6	–	–	6
TE	50	10	–	60
TPD	15	–	1	16
TPI	12	1	–	13
Total	121	14	1	136

Table 8.4

Summary of Grade 6 Data Review Votes

Item Type	No. of Items		
	Accept	Accept w/Edits	Total
CR	7	–	7
ER	3	–	3
MC	45	1	46
MS	12	–	12
TE	36	1	37
TPD	24	–	24
TPI	6	–	6
Total	133	2	135

Table 8.5

Summary of Grade 7 Data Review Votes

Item Type	No. of Items		
	Accept	Accept w/Edits	Total
CR	7	–	7
ER	3	1	4
MC	41	3	44
MS	15	–	15
TE	45	1	46
TPD	16	–	16
TPI	4	–	4
Total	131	5	136

Table 8.6

Summary of Grade 8 Data Review Votes

Item Type	No. of Items		
	Accept	Accept w/Edits	Total
CR	6	–	6
ER	4	–	4
MC	51	2	53
MS	11	–	11
TE	44	1	45
TPD	12	–	12
TPI	6	–	6
Total	134	3	137

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 an additional opportunity to review item content and to consider possible revisions to re-field test items for possible future operational use. The reconciliation decisions were 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. (1963). Chi-square tests with one degree of freedom; extensions of the MantelHaenszel procedure. *Journal of the American Statistical Association*, 58, 690–700.

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

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

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 Grades 3–8 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. Standalones
      iv. Template
III. **Considerations**
   a. Tasks
   b. Item sets
   c. Phenomena list

LEAP 2025 Grades 3–8 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 as computer-based for all grades. Paper-based option available for grades 3 and 4.
II. **Item Development Materials on Box**
III. **Louisiana Student Standards for Science (LSS for Science)**
   a. New science standards were approved in early March 2017.

i. The LSS for Science represent the knowledge and skills needed for students to successfully transition to postsecondary educations 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 Mathematical 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, therefore 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.
        1. Develop a model; Analyze data; Construct an explanation

b. DCI

c. CCC
        1. Energy and Matter; Patterns; Scale, Proportion, and Quantity

VIII. **Phenomena: Keystone of 3D 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 typically 2 PEs, but 1-PE and 3-PE bundles are acceptable.
b. PE bundling is used in two of the three “item groupings” on LSS assessment.
c. See “Phenomena and Context Overview” and “Context and Stimuli” 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; 4 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 TEs, 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; 4 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 (likely an 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**

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, 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 item (TE)
    i. TEs are worth 1 or 2 points.
    ii. Use in tasks, item sets, and standalone
    iii. TE types (NO TEs 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
        1.
    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" for more information.
    v. Style guide forthcoming!

XXIV. **Information Security**

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

LEAP 2025 Grades 3–8 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 Systems Models

VI. **Allowable Item Types**
   a. Tasks
      i. 1-point TE
      ii. 1-point SR
      iii. 2-point TE
      iv. 2-point EBSR
      v. 9-point ER
   b. Item sets
      i. 1-point TE
      ii. 1-point SR
      iii. 2-point TE
      iv. 2-point EBSR
      v. 2-point CR
   c. Standalones
      i. 1-point TE
      ii. 1-point SR
      iii. 2-point TE
      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 Grades 3–8 Editor Training Agenda
Item Development Cycle for 2017–2018 Field Test

I. **Item Set/Task/Standalone 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 (4 items per item set; 5 items/task; standalones) 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/item set;
            b. 9 items in task (A and B versions)
            c. standalones
         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 TEs; 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 Submissions to Writers**
V. **Feedback to Writers**
VI. **Process Overview for Intake/E2**
VII. **Intake/E1 Rules for Returning Item Sets/Tasks/Standalone Submissions to E1 Writer**