LEAP 2025 Science 3–8
Technical Report: 2021–2022

Prepared by DRC, Pearson, and WestEd

EXECUTIVE SUMMARY

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

While this technical report and its associated materials have been produced in a way that can help educators understand the technical characteristics of the assessment used to measure student achievement, the information is primarily intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

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

Table of Contents

EXECUTIVE SUMMARY 2

1. Introduction 7

Summary of the 2021–2022 Activities 7

2. Assessment Frameworks 9

3. Overview of the Test Development Process 11

Item Development Plan 11

Proposal and Review of Topics and Sources 35

Performance Expectation Bundling 35

Phenomena Selection and Outline Development 36

Matching Phenomena to Item Sets and Tasks and Foci to Standalone Items 37

Outline and Stimuli Development 39

Item Writing and Review Process 41

Data Review Process and Results 51

4. Construction of Test Forms with Embedded Field Test 54

Test Design 54

Initial Construction 60

Operational Form 60

Field Test Versions 63

Revision and Review 64
Psychometric Approval of Operational Forms 64
LDOE Review 65
Test Forms and Accessible Versions 66
Online and Paper Forms 66
Accommodated Print Versions 67
Form Versions for Students with Visual Impairments 67
5. Test Administration 68
Training of School Systems 68
Ancillary Materials 69
Interpretive Guides 83
Time 83
Online Forms Administration, Grades 3–8 83
Paper-Based Forms Administration, Grade 3 83
Accessibility and Accommodations 84
Testing Windows 86
Test Security Procedures 86
Data Forensic Analyses 87
6. Scoring Activities 89
Constructed- and Extended-Response Item Scoring Process 91
7. Data Analysis 109
Classical Item Statistics 109

Differential Item Functioning 109
Measurement Models 115
Calibration and Linking 116
Operational Item Parameters 123
Item Fit 123
Dimensionality and Local Item Independence 125
Test Characteristic Curve (TCC) 128
Test Information Curve, Score Distribution, and IRT Difficulty Distribution 130
Field Test Data Review 138

8. Test Results and Score Reports 139

Demographic Characteristics of Students 139
Test Results 140
Effect Size 152
Score Reports 153
Achievement Level Policy Definitions 154

9. Reliability 156

Internal Consistency Reliability Estimation 156
Classical Standard Error of Measurement 157
Conditional Standard Error of Measurement and Cut Scores 158
Student Classification Accuracy and Consistency 161

10. Validity 163

Evidence for Construct-Related Validity 164

Internal Structure of Reporting Categories 164
Content-Related Evidence 164
Dimensionality and Principal Component Analysis 165
Evidence Based on Relations to Other Variables 166
Item Development and Field-Test Analysis 167

References 169

Appendix A: Training Agendas 173

Appendix B: Test Summary 189

Appendix C: Item Analysis Summary Report 205

Appendix D: Dimensionality 267

Appendix E: Scale Distribution and Statistical Report 276

Appendix F: Reliability and Classification Accuracy 289

Appendix G: Accommodated Print and Braille Creation 301

Appendix H: On-Going Quality Control 304

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 operational administration of the statewide summative science assessments for grades 3–8. This report outlines the testing procedures, forms construction, administration, statistical analyses, IRT (Item Response Theory) calibration, test results, reliability and validity, and reporting of scores.

Summary of the 2021–2022 Activities

WestEd and Pearson, in partnership with the LDOE and Data Recognition Corporation (DRC), the administration vendor, developed a timeline to capture the major activities necessary to produce the spring 2022 Science grades 3–8 operational forms with embedded field test items (EFT).

For grades 3–8, all tests were delivered in a computer-based format, with a paper-based option for grade 3. An accommodated paper-based format is available for students in grades 4–8 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 August 2019 to May 2022

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

2. Assessment Frameworks

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

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

For two-part items, students may construct an explanation and support the explanation with evidence or make a claim and evaluate evidence to support that claim. Another

application of two-part items is to develop a model in part A and to evaluate the model in part B. A range of item types and applications allows greater test-taker engagement and provides a more authentic assessment experience.

The test design includes item sets, a task, and standalone items. A stimulus that describes a scientific phenomenon anchors each item set or task. A focus that details some aspects of a phenomenon provides the common anchor for standalone items. Item sets are composed of four items associated with a common stimulus. The item sets may include 1-point selected-response items (single-select and/or MS formats), 1- and 2-point TE items, and 2-point two-part items (two-part independent [TPI] and/or two-part dependent [TPD] formats) tied to a common stimulus. For grades 5–8, item sets may include 1- or 2-point TE items. Three item sets include a two-point CR item. The assessment also includes one task. The task consists of five items tied to a common stimulus and includes 1-point selected-response items (both single-select and MS formats), 2-point two-part items (TPI and/or TPD formats), and a 9-point ER item for grades 5–8. The standalone items provide flexibility to meet the test blueprint and afford greater coverage of the standards while still requiring students to make connections among the three dimensions of the LSSS. All points associated with the task contribute to a student's overall score, but the ER item is not a component of the current blueprint and therefore not included in the proportional representation of content assessed by other parts of the test.

Because the assessment at grade 3 was administered primarily via paper, the item types were limited to selected-response (i.e., MC and MS), two-part (i.e., TPI and/or or TPD), and CR items. The assessments for grades 4–8 were administered primarily online, so TE items were viable at these grades. However, paper and pencil versions of the assessments for grades 4–8 were made available as accommodated forms for students who were unable to test online. For those forms, TE items were adapted for paper presentation to still address the same content.

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

3. Overview of the Test Development Process

Item Development Plan

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 test blueprints that guided 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 LSSS	Relative % in LSSS	% 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			
	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 LSSS	Relative % in LSSS	% 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 LSSS	Relative % in LSSS	% 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 Reporting Category Coverage

Grade 3: SEP reporting category Coverage				
Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (SEPs 1 & 3)	4	29%	24%–34%	7
Reporting Category 2 (SEPs 4, 5, 7)	7	50%	45%–55%	7
Reporting Category 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 reporting category (1–3), so SEP 8 is not being repeated across the reporting categories.

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 LSSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

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

Table 3.1h

Test Blueprint for LEAP 2025 Grade 4: DCI Domain Coverage

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

Table 3.1i

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	CCC 5 – E/M	1
04-PS3-2	SEP 3 – INV	CCC 5 – E/M	1
04-PS3-3	SEP 1 – Q/P	CCC 5 – E/M	1
04-PS3-4	SEP 6 – E/S	CCC 5 – 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 3.1j

Test Blueprint for LEAP 2025 Grade 4: CCC Coverage

Grade 4: CCC Coverage			
CCC Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1k

Test Blueprint for LEAP 2025 Grade 4: SEP Coverage

Grade 4: SEP Coverage			
SEP Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1l

Test Blueprint for LEAP 2025 Grade 4: SEP Reporting Category Coverage

Grade 4: SEP Reporting Category Coverage				
SEP Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (SEPs 1 & 3)	4	31%	26%–36%	7
Reporting Category 2 (SEPs 4, 5, 7)	2	15%	10%–20%	7
Reporting Category 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 reporting category (1–3), so SEP 8 is not being repeated across the reporting category.

Table 3.1m

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

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

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

Table 3.1n

Test Blueprint for LEAP 2025 Grade 5: DCI Domain Coverage

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

Table 3.1o

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 3.1p

Test Blueprint for LEAP 2025 Grade 5: CCC Coverage

Grade 5: CCC Coverage			
CCC Overall	# in PEs in LSSS	Relative % in LSSS	% by Points of CCC Items
CCC 1 – PAT	1	8%	5%–15%
CCC 2 – C/E	2	15%	9%–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 3.1q

Test Blueprint for LEAP 2025 Grade 5: SEP Coverage

Grade 5: SEP Coverage			
SEP Overall	# in PEs in LSSS	Relative % in LSSS	% 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%–15%
SEP 7 – ARG	2	15%	10%–20%
SEP 8 – INFO	0	0%	–
Total	13	100%	

Table 3.1r

Test Blueprint for LEAP 2025 Grade 5: SEP Reporting Category Coverage

Grade 5: SEP Reporting Category Coverage				
	# of PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (SEPs 1 & 3)	3	23%	18%–28%	7
Reporting Category 2 (SEPs 4, 5, 7)	5	38%	32%–43%	7
Reporting Category 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 reporting category (1–3), so SEP 8 is not being repeated across the reporting categories.

Table 3.1s

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

Grade 5: SEP Compared to CCC Ratio		
	Relative Weight in LSSS	Minimum %
SEPs	50%	30%
CCCs	50%	30%

The test blueprints that guided item development projections for grade 6 are presented in the following tables.

Table 3.1t

Test Blueprint for LEAP 2025 Grade 6: DCI Domain Coverage

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

Table 3.1u

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 3.1v

Test Blueprint for LEAP 2025 Grade 6: CCC Coverage

Grade 6: CCC Coverage			
CCC Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1w

Test Blueprint for LEAP 2025 Grade 6: SEP Coverage

Grade 6: SEP Coverage			
SEP Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1x

Test Blueprint for LEAP 2025 Grade 6: SEP Reporting Category Coverage

Grade 6: SEP Reporting Category Coverage				
SEP Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (SEPs 1 & 3)	4	21%	16%–26%	7
Reporting Category 2 (SEPs 4, 5, 7)	6	32%	27%–37%	7
Reporting Category 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 reporting category (1–3), so SEP 8 is not being repeated across the reporting categories.

Table 3.1y

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

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

The test blueprints that guided item development projections for grade 7 are presented in the following tables.

Table 3.1z

Test Blueprint for LEAP 2025 Grade 7: DCI Domain Coverage

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

Table 3.1aa

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 3.1bb

Test Blueprint for LEAP 2025 Grade 7: CCC Coverage

Grade 7: CCC Coverage			
CCC Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1cc

Test Blueprint for LEAP 2025 Grade 7: SEP Coverage

Grade 7: SEP Coverage			
SEP Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1dd

Test Blueprint for LEAP 2025 Grade 7: SEP Reporting Category Coverage

Grade 7: SEP Reporting Category Coverage				
SEP Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Reporting Category 1 (SEPs 1 & 3)	3	20%	15%–25%	7
Reporting Category 2 (SEPs 4, 5, 7)	3	20%	15%–25%	7
Reporting Category 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 reporting category (1–3), so SEP 8 is not being repeated across the reporting categories.

Table 3.1ee

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

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

The test blueprints that guided item development projections for grade 8 are presented in the following tables.

Table 3.1ff

Test Blueprint for LEAP 2025 Grade 8: DCI Domain Coverage

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

Table 3.1gg

Test Blueprint for LEAP 2025 Grade 8: Minimal PE Coverage

Grade 8: Minimal PE Coverage Everv PE will be included at least one time in a test			
PE	SEP	CCC	Min Items
08-MS-ESS1-4	SEP 6 – E/S	CCC 3 – SPO	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 3.1hh

Test Blueprint for LEAP 2025 Grade 8: CCC Coverage

Grade 8: CCC Coverage			
CCC Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1ii

Test Blueprint for LEAP 2025 Grade 8: SEP Coverage

Grade 8: SEP Coverage			
SEP Overall	# in PEs in LSSS	Relative % in LSSS	% 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 3.1jj

Test Blueprint for LEAP 2025 Grade 8: SEP Reporting Category Coverage

Grade 8: SEP Reporting Category Coverage				
SEP Reporting Category	# PEs in LSSS	Relative % in LSSS	% by Points of SEP Items	Min Points
Investigate (SEPs 4, 6, 8)	6	31.5%	27%–37%	7
Evaluate (SEPs 4, 5, 7)	6	31.5%	27%–37%	7
Reason Scientifically (SEPs 2 & 6)	7	37%	32%–42%	7
Total	19	100%		

Table 3.1kk

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

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

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 sets, tasks, and standalone items to be developed. Tables 3.2a–f show the item development plans for the number of items developed by WestEd by reporting category for grades 3–8.

Table 3.2a

Number of New 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	3	18	0	0	9	0	3	27
Tasks	0	0	0	0	0	0	0	0
Standalone Items	n/a	5	0	0	3	0	0	8

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	3	18	0	0	9	0	3	27
Tasks	0	0	0	0	0	0	0	0
Standalone Items	n/a	7	0	0	5	0	0	12

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	2	8	4	2	4	0	2	18
Tasks	1	2	3	1	4	2	0	10
Standalone Items	n/a	6	0	2	4	0	0	12

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	3	13	7	4	9	0	3	33
Tasks	0	0	0	0	0	0	0	0
Standalone Items	n/a	2	2	3	5	0	0	12

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	2	8	3	5	6	0	2	22
Tasks	1	3	1	4	2	2	0	10
Standalone Items	n/a	3	1	1	7	0	0	12

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	1	3	3	2	3	0	1	11
Tasks	2	6	4	6	4	4	0	20
Standalone Items	n/a	4	2	0	6	0	0	12

The development plans also included item sets and tasks that were revised and refield-tested. Item sets and tasks that were revised and refield-tested were chosen because some of the items deviated from psychometric criteria for item statistics based on student performance. Tables 3.3a–f show the item development plans for the revised and refield-tested item sets and tasks on the spring 2022 field test.

Table 3.3a

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 3 Assessment

	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*	2	8	0	0	7	0	2	15
Tasks	0	0	0	0	0	0	0	0

* Includes two tasks that were changed to Item Sets and for which CRs were written.

Table 3.3b

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 4 Assessment

	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*	5	20	0	0	16	0	7	36
Tasks	0	0	0	0	0	0	0	0

* Includes two tasks that were changed to Item Sets and for which CRs were written.

Table 3.3c

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 5 Assessment

	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	4	10	5	10	8	0	4	15
Tasks	2	5	3	5	3	2	0	16

Table 3.3d

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 6 Assessment

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	3	11	5	2	9	0	3	27
Tasks	2	6	2	0	8	2	0	16

Table 3.3e

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 7 Assessment

	Total Number of Sets	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
Item Sets	3	15	1	2	6	0	5	15
Tasks	2	5	1	6	4	2	0	16

Table 3.3f

Number of Revised and Refield-Tested Item Sets and Tasks for Grade 8 Assessment

	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	2	5	4	4	4	0	2	17
Tasks	2	7	1	3	5	2	0	16

Proposal and Review of Topics and Sources

Performance Expectation Bundling

In the previous item development cycle, WestEd used the 2017 LSSS to recommend how performance expectations could be bundled in a task or item set to ensure that the breadth of all dimensions of constituent PEs 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. In previous development, the LDOE and WestEd determined that some item sets and tasks would allow a "mix and match" approach in which the science and engineering practice (SEP) for one of the PEs in a bundle could be used to develop items aligned to the disciplinary core idea (DCI) and crosscutting concept (CCC) of the other PE in the bundle. This approach was discontinued beginning with the current cycle because it generated some items with a SEP alignment outside the reporting category for the PE the item aligned to and therefore did not fit the reporting category. Within each task or item set, each item was given a primary assignment to 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.

In the 2019–2022 item development cycle, additional PE bundles were proposed to the LDOE. Table 3.4a 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 2018–2019 development cycle.

Table 3.4a

PE Bundling by Grade

Grade	Total Number of PE Bundles Approved	Number of Bundles Targeted for Development
3	18	3
4	20	3
5	17	3
6	17	3
7	21	3
8	21	3

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 grades 3–8 assessments are anchored on scientific phenomena 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.

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

When identifying a phenomenon, the content lead considered:

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

Phenomena were chosen to represent the breadth of content described 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 PEs that had not been assessed in the previous field test.

Matching Phenomena to Item Sets and Tasks and Foci to Standalone Items

Both item sets and tasks were targeted for development for the 2019–2022 development cycle based on an analysis of the test bank for each grade. The development of item sets and tasks influenced the selection of phenomena. Like the tasks, the item sets are phenomena-based, but unlike the tasks, they are made up of independent items that do not necessarily build upon each other. Also, unlike the tasks, the items in the item sets do not scaffold to help discriminate student performance levels, do not require a specific order, and do not contain a three-dimensional extended-response (ER) item. Although an item set does not need to contain a constructed-response (CR) item, 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 3.4b shows the total number of CRs developed per grade.

Table 3.4b

Constructed-Response Item Development by Grade

Grade	Number of CRs Developed
3	3
4	3
5	2
6	3
7	2
8	1

WestEd developed two ERs for each task developed. Table 3.4c shows the total number of ERs developed per grade.

Table 3.4c

Extended Response Item Development by Grade.

Grade	Number of ERs Developed
3	0
4	0
5	2
6	0
7	2
8	4

For the item sets and tasks, WestEd offered a document containing descriptions of phenomena associated with bundles to the LDOE to review prior to item development. Table 3.4d shows the number of phenomena submitted to the LDOE for item sets and tasks at grades 3–8.

Table 3.4d

Phenomena Submitted by Grade

Grade	Number of Phenomena Submitted for Item Sets	Number of Phenomena Submitted for Tasks
3	11	0
4	7	0
5	9	2
6	9	0
7	5	2
8	2	4

For the item sets, the LDOE identified 3 phenomena at grades 3, 4, and 6; 2 phenomena at grades 5 and 7; and 1 phenomenon at grade 8 to be developed into stimuli. For the tasks, the LDOE identified 1 phenomenon at grades 5 and 7 and 2 phenomena at grade 8 to be developed into stimuli. Upon approval of the phenomena, WestEd submitted item outlines containing stimuli and item descriptions to the LDOE. Once the item outlines were approved, item development for the item sets began.

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

Outline and Stimuli Development

WestEd used both experienced internal and external science assessment editors to develop the phenomena-based stimuli for item sets. 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 (2019–2022).

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

Evaluating the Reading Level of Stimuli. WestEd performed Lexile and ATOS analyses on each stimulus to obtain quantitative measures of the readability of the texts. The Lexile Analyzer, developed by MetaMetrics, analyzes the semantic and syntactic features of a text and assigns it a Lexile measure. MetaMetrics also provides grade-level ranges corresponding to Lexile ranges. It should be noted that the grade-level ranges include overlap across grade levels. The ATOS text analysis tool, developed by Renaissance Learning, 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 LSSS were used as an additional measure of grade-level appropriateness. WestEd and the LDOE also drew on the professional experience of educators, during Content and Bias Committee review, to verify that sources would be accessible to students, and made changes based on their feedback. Most of the stimuli developed for the assessments were found to be below or at grade level; however, some of the science vocabulary was evaluated as above grade level. In those cases, additional support such as parenthetical definitions (glossing) was 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 assessments. All writers' resumes 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 February 2020. For an outline of the information covered, see Appendix A for the LEAP 2025 Grades 3–8 Item Outline Development Training Agenda. In the training, writers were provided context for the assessment, including LDOE expectations, the LSSS, and a review of best practices for item development. The item writers were provided the approved item topics and drafts of the stimuli, as well as item outlines that provided explanations of the phenomena underlying the item sets. Item writers were also provided with alignment to the Science and Engineering Practices, Crosscutting Concepts, and Disciplinary Core Ideas of the LSSS, and guidance on how each item set should be developed. The use of item set overviews allowed WestEd to provide direction 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 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 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. Writers were also provided training in universal design and bias/sensitivity. A variety of items were presented and reviewed using universal design and bias/sensitivity

lenses. This training also included an overview of these topics, (see Appendix A for the LEAP 2025 Grades 3–8 Item Writer Training Agenda). WestEd provided training and feedback to the writers throughout the development cycle, as the LDOE and WestEd gained a clearer understanding of how the stimuli, items, and 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 assessments. For an outline of the information covered, see Appendix A for the LEAP 2025 Grades 3–8 Editor Training Agenda. 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. Style guidelines continue to be based on documentation established with the LEAP 2025 Biology and Science assessments. This documentation was amended and updated as the development cycle progressed. When questions of style arose that were unanswered by existing documentation, WestEd consulted the LDOE, and approved changes were added to the project style guide.

LDOE Content Review. As writing and editing for batches of item sets and standalone items were completed, these batches were sent to the LDOE for review by the LDOE Science Assessment Coordinators; Assessment Content Supervisor for Math, Science, and Small Populations; Elementary Assessment Coordinator; Small Populations Assessment

Coordinator; and Science Program Coordinators. Feedback from the LDOE review was implemented before the content and bias review meetings.

Content and Bias Review. After the completion of item development, WestEd coordinated virtual content and bias review meetings, held using Zoom. The meetings were led by facilitators from the LDOE, WestEd, and Pearson. 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. Table 3.5 provides the demographic characteristics of the review committee.

Table 3.5

Representation of Educators Participating in 2021–2022 Content and Bias Reviews

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

Note: As teachers may fulfill multiple roles, at some grades representation of roles may exceed number of total participants.

Prior to joining the virtual review, committee members were required to watch a prerecorded training, including content training on how to evaluate the items as part of the committee review process. Participants were also provided meeting materials in advance of the review. At the start of the virtual committee, they received an orientation from the LDOE about the LEAP 2025 grades 3–8 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 PE, SEP, DCI, and CCC alignment for each item and recorded the degree of alignment for each dimension and overall alignment on a worksheet on a scale of 0 (not aligned) to 3 (well aligned), referring to LSSS Appendix A (Learning Progressions). An item was considered to have a high degree of alignment if it aligned to the particular bullet listed in the PE. An item was considered to have a lower degree of alignment if it aligned to another bullet listed in the learning progression for that SEP or CCC. Committee members also recorded whether the science for each item was accurate and whether each item was free of bias or sensitive content. Areas of concern considered included opportunity and access, portrayal of groups represented, and protecting privacy and avoiding offensive content.

After the review of each item, each member voted in ABBI on whether to accept, accept with edits, or reject each item, recording comments for any item where they noted issues with science accuracy or bias. (If participants skipped an item or chose not to record a decision for a given item, the system registered the response as “No Vote” for that individual review. “No Vote” was recorded as the consensus rating when an initial group decision on an item was not reached, and the committee failed to return to that item and register a final vote to accept, revise, or reject the item.) Participants used personal laptops to access ABBI and only had access to ABBI during meeting times. Participants were locked out of ABBI when the meeting was not in progress. At the end of each day, WestEd made certain that the participants cleared their computer caches and deleted their download histories for the day. WestEd required cameras to be on at all times during the meeting in order to monitor participants to be sure that participants were in a secure space and that they did not use their cell phones. Content security was stressed in the prerecorded training, during the meeting introduction, throughout the meetings, at the end of each day, and at the conclusion of each meeting.

Following the individual reviewers' votes, the group came together to view and discuss each stimulus and item as it was projected on-screen, with the goal of achieving consensus. The WestEd and Pearson facilitators compiled detailed notes about committee decisions for implementation after the review.

Results of Content Review. The results of the reviewers' individual judgments were captured in ABBI. Tables 3.6a–f provide these results, based on the participants' individual votes on each item following their initial review.

Table 3.6a

Grade 3 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	5	34	0	0	0	34
ER	0	0	0	0	0	0
MC	25	167	8	0	0	175
MS	4	28	0	0	0	28
TPD	13	85	4	0	1	90
TPI	6	38	4	0	0	42
Stimulus	5	1	0	0	0	1
All Grade 3	58	353	16	0	1	370

Table 3.6b

Grade 4 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	6	34	18	0	3	55
ER	0	0	0	0	0	0
MC	22	132	52	5	3	192
MS	5	32	13	0	0	45
TPD	7	47	14	2	0	63
TPI	11	62	33	2	1	98
Stimulus	5	16	8	1	1	26
All Grade 4	56	323	138	10	8	479

Table 3.6c

Grade 5 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	4	32	0	0	0	32
ER	2	14	2	0	0	16
MC	19	138	9	0	0	147
MS	3	10	0	0	0	10
TE	19	109	8	0	0	117
TPD	10	57	16	0	0	73
TPI	7	49	0	0	0	49
Stimulus	6	8	0	0	0	8
All Grade 5	70	417	35	0	0	452

Table 3.6d

Grade 6 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	5	32	17	1	0	50
ER	0	0	0	0	0	0
MC	16	121	33	2	2	158
MS	0	0	0	0	0	0
TE	18	148	27	2	0	177
TPD	9	69	19	0	0	88
TPI	11	86	22	2	0	110
Stimulus	4	21	9	1	0	31
All Grade 6	63	477	127	8	2	614

Table 3.6e

Grade 7 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	3	23	1	0	0	24
ER	2	10	6	0	0	16
MC	13	98	5	0	0	103
MS	3	23	1	0	0	24
TE	16	112	15	0	0	127
TPD	8	58	3	1	0	62
TPI	7	53	3	0	0	56
Stimulus	4	6	0	0	0	6
All Grade 7	56	383	34	1	0	418

Table 3.6f

Grade 8 Vote Totals Based on Individual Votes Following Initial Review

Item Type	*N* Items	Accept	Accept with Edits	No Vote	Reject	Total
CR	1	6	0	1	0	7
ER	4	26	4	1	0	31
MC	10	75	5	0	0	80
MS	3	23	0	0	1	24
TE	17	116	18	0	1	135
TPD	9	64	7	0	0	71
TPI	4	28	4	0	0	32
Stimulus	3	10	0	0	0	10
All Grade 8	51	348	38	2	2	390

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 were rejected by the committee. Table 3.6g shows the consensus votes for each grade.

Table 3.6g

Consensus Votes by Grade

Grade	Number of Rejected Items
3	0
4	0
5	0
6	0
7	0
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.

Data Review Process and Results

During data review of the spring 2022 FT items, content experts and psychometric support staff reviewed field-tested items with accompanying data to make judgments about the appropriateness of items for use on future operational test forms. Statistically flagged items were not rejected on the sole basis of statistics; only items with identifiable flaws based on content were rejected.

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

Facilitators from Pearson and WestEd led the data review. Statistical information was evaluated for each item to determine whether the item functioned as intended. Each item's suitability for future operational tests was then evaluated in the context of the field-test statistics. Judgments to accept, accept with edits (or "revise/refield-test"), or reject were then recorded for each item. If the decision was to edit or to reject an item, additional information was captured to document the reason for the decision. Table 3.7 summarizes the disposition of field-tested items from data review.

Table 9

FT Item Dispositions by Item Type, 2022 Data Review

Grade	Item Type	Number of Items				
		Accept	Accept with Edits	Reject	Total	% of Total
3	CR	3	1	1	5	10.20
	MC	19	0	2	21	42.86
	MS	4	0	0	4	8.16
	TE	0	0	0	0	0.00
	TPI	5	0	2	7	14.29
	TPD	9	0	3	12	24.49
	Total	40	1	8	49	100.00
4	CR	7	1	2	10	14.08
	MC	26	1	4	31	43.66
	MS	5	2	0	7	9.86
	TE	0	0	0	0	0.00
	TPI	9	1	0	10	14.08
	TPD	9	3	1	13	18.31
	Total	56	8	7	71	100.00
5	CR	4	1	0	5	5.49
	ER	2	1	0	3	3.30
	MC	15	5	5	25	27.47
	MS	1	1	1	3	3.30
	TE	25	9	1	35	38.46
	TPI	8	2	0	10	10.99
	TPD	5	4	1	10	10.99
	Total	60	23	8	91	100.00
6	CR	6	0	1	7	6.60
	ER	1	0	2	3	2.83
	MC	18	5	10	33	31.13
	MS	1	0	0	1	0.94
	TE	15	6	7	28	26.42
	TPI	11	0	4	15	14.15

Grade	Item Type	Number of Items				
		Accept	Accept with Edits	Reject	Total	% of Total
	TPD	11	1	7	19	17.92
	Total	63	12	31	106	100.00
7	CR	6	0	1	7	6.93
	ER	2	0	2	4	3.96
	MC	20	4	4	28	27.72
	MS	5	0	1	6	5.94
	TE	24	0	8	32	31.68
	TPI	7	0	4	11	10.89
	TPD	9	1	3	13	12.87
	Total	73	5	23	101	100.00
8	CR	3	1	0	4	4.26
	ER	1	0	4	5	5.32
	MC	14	5	3	22	23.40
	MS	5	2	0	7	7.45
	TE	19	4	10	33	35.11
	TPI	6	1	0	7	7.45
	TPD	10	2	4	16	17.02
	Total	58	15	21	94	100.00

Following the data review meeting, LDOE content specialists considered the item level data review outcomes to determine which sets and tasks could be used operationally or rejected unless revised/re-field tested. The reconciliation decisions were the final decisions. It should be noted that the training presentation agenda for data review is included in Appendix A: Training Agendas.

4. Construction of Test Forms with Embedded Field Test

Test Design

To assess the integrated nature of the content, practices, and crosscutting concepts of the LSSS, the LEAP 2025 3–8 science assessments involved set-based designs. The tests included item sets and, for grades 5–8, a task on each form, each anchored by a common stimulus or stimuli. Additionally, standalone items were included to support meeting the specific targets of the test blueprints. Table 4.1a shows the Test Design for Science Grade 3.

Table 4.1a

Test Design for Science Grade 3

Test Session	Numbers of Items
Session 1: One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
OP Standalone Items	4 OP Standalone SR Items 1 OP Standalone TPD/TPI Items
One FT Item Set	2 FT Item Set SR Items 1–2 FT Item Set TPD/TPI Item 0–1 FT Item Set CR Items
FT Standalone Items	0–2 FT Standalone SR Items 0–2 FT Standalone TPD/TPI Items
Session 2: One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
OP Standalone Items	6 OP Standalone SR Items 1 OP Standalone TPD/TPI Items
Total Items Field Tested Across Forms for Grade 3	5 FT Standalone SR Items 3 FT Standalone TPD/TPI Items 20 FT Item Set SR Items 16 FT Item Set TPD/TPI Items 5 Item Set CR Items

Note: Students do not complete more than one CR per item set. There were a total of 3 operational CR items per form.

Table 4.1b shows the Test Design for Science Grade 4.

Table 4.1b

Test Design for Science Grade 4

Test Session	Numbers of Items
Session 1: One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
OP Standalone Items	2 OP Standalone SR Items 1 OP Standalone TPD/TPI Items
FT Standalone Item	0–1 FT Standalone SR Items 0–1 FT Standalone TPD/TPI Items
Session 2: One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One FT Item Set	2 FT Task SR Items 2 FT Task TPD/TPI Items 0–1 FT Item Set CR Items
Total Items Field Tested Across Forms for Grade 4	7 FT Standalone SR items 5 FT Standalone TPD/TPI Items 21 FT Item Set SR Items 19 FT Item Set TPD/TPI Items 10 Item Set CR Items

Note: Students did not complete more than one CR per item set. There were a total of 3 operational CRs per form. Item sets field tested included one item set developed in 2018.

Table 4.1c shows the Test Design for Science Grades 5–8.

Table 4.1c
Test Design for Science Grades 5–8

Test Session	Numbers of Items
Session 1: One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
OP Standalone Items	2 OP Standalone SR Items 1 OP Standalone TPD/TPI Items
Session 2: One OP Task	2 OP Task SR Items 2 OP Task TPD/TPI Items 1 OP Task ER Item
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
One OP Item Set	2 OP Item Set SR Items 1–2 OP Item Set TPD/TPI Items 0–1 OP Item Set CR Items
OP Standalone Items	1 OP Standalone SR Item 2 OP Standalone TPD/TPI Items
Session 3: One FT Item Set or Task	2 FT Item Set SR Items 2 FT Item Set TPD/TPI Items 0–1 FT Item Set CR Items OR 2 FT Item Set SR Items 2 FT Item Set TPD/TPI Items 1 FT Item Set ER Item
FT Standalone Items	0–2 FT Standalone SR Items 0–2 FT Standalone TPD/TPI Items

Test Session	Numbers of Items
Total Items Field Tested Across Forms for Grade 5	6 FT Standalone SR Items 2 FT Standalone TE Items 4 FT Standalone TPD/TPI Items 17 FT Item Set SR Items 20 FT Item Set TE Items 10 FT Item Set TPD/TPI Items 7 FT Item Set CR Items 6 FT Task SR Items 12 FT Task TE Items 6 FT Task TPD/TPI Items 4 FT Task ER Items
Total Items Field Tested Across Forms for Grade 6	2 FT Standalone SR Items 5 FT Standalone TE Items 5 FT Standalone TPD/TPI Items 24 FT Item Set SR Items 18 FT Item Set TE Items 18 FT Item Set TPD/TPI Items 6 FT Item Set CR Items 6 FT Task SR Items 2 FT Task TE Items 8 FT Task TPD/TPI Items 2 FT Task ER Items
Total Items Field Tested Across Forms for Grade 7	2 FT Standalone SR Items 2 FT Standalone TE Items 5 FT Standalone TPD/TPI Items 23 FT Item Set SR Items 11 FT Item Set TE Items 12 FT Item Set TPD/TPI Items 7 FT Item Set CR Items 8 FT Task SR Items 12 FT Task TE Items 6 FT Task TPD/TPI Items 4 FT Task ER Items

Test Session	Numbers of Items
Total Items Field Tested Across Forms for Grade 8	4 FT Standalone SR Items 2 FT Standalone TE Items 6 FT Standalone TPD/TPI Items 8 FT Item Set SR Items 13 FT Item Set TE Items 7 FT Item Set TPD/TPI Items 3 FT Item Set CR Items 13 FT Task SR Items 14 FT Task TE Items 9 FT Task TPD/TPI Items 6 FT Task ER Items

Note: Students did not complete more than one CR per item set. There were a total of 3 operational CRs per form. For grades 5–8, item sets field tested included one item set developed in 2018.

Initial Construction

The purpose of the spring 2022 forms construction activities was to create operational forms using the spring 2018 and spring 2019 field test items that were approved for operational use and to embed field test items in the spring 2022 forms for potential use in future operational assessments. This section describes the process used to create operational and field test forms.

Operational Form

Data review-approved items, field tested in spring 2018 or 2019, were available for use on the spring 2022 operational assessments.

For each of grades 3 through 8, WestEd completed item selection for one operational (OP) form for the spring 2022 administration. WestEd worked with LDOE content staff to select items for the forms following the data review meeting in September and submitted these forms to Pearson psychometricians for consideration before formal submission to LDOE for approval. For grades 3 and 4, a combination of item sets and standalone items were chosen that would ensure that the relative distribution of score points by reporting category would meet the blueprints for the operational assessment while avoiding similar content and topics across the balance of items and item types. For grades 5-8, the WestEd content lead selected the task first and followed with a combination of item sets and standalone items that would ensure that the relative distribution of score points by reporting category would meet the blueprints for the operational assessment while avoiding similar content and topics across the balance of items and item types. Tables 4.2a–f provide the operational test composition for grades 3–8 for spring 2022.

Table 4.2a

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	Total Items	Total Points
4-Item Set	6	4	6	12	12	24	36
Standalone items	1	12	14	10	2	12	14
Totals	–	–	–	22	14	36	50

Table 4.2b

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	Total Items	Total Points
4-Item Set	7	4	6	14	16	28	42
Standalone items	1	8	10	16	2	8	10
Totals	–	–	–	20	18	36	52

Table 4.2c

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	10	10		20	30
Standalone items	1	12	16				12	16
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	12	12	1	37	61

Table 4.2d

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	10	10		20	30
Standalone items	1	12	16				12	16
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	12	12	1	37	61

Table 4.2e

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	10	10		20	30
Standalone items	1	12	16				12	16
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	12	12	1	37	61

Table 4.2f

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	10	10		20	30
Standalone items	1	12	16				12	16
Task	1	5	15	2	2	1	5	15
Totals	–	–	–	12	12	1	37	61

Field Test Versions

The number of field test versions administered in spring 2022 varied by grade. These data are shown in Table 4.3.

Table 4.3

Grade	Number of Field Test Versions
3	10
4	15
5	20
6	17
7	17
8	21

In some cases, the number of field test slots exceeded the number of items available for field testing. As a result, some items were repeated among field test versions. One or two versions of each item set or task were field tested as needed.

For grade 3, one field test item set and one field test standalone item were embedded within session 1 of the operational form. For grade 4, one field test standalone item was embedded in session 1 and a field test item set was embedded in session 2. For grade 5, a combination of either one standalone item and two item sets, one item set and five standalone items or one task and four standalone items were embedded in session 3. For grades 6–8, a combination of either one task and one item set, or two item sets and one standalone item to be field tested were in session 3 of the operational form.

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

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

Revision and Review

Psychometric Approval of Operational Forms

Prior to submitting the forms to LDOE staff for review, Pearson psychometricians and WestEd content specialists participated in an iterative process of reviewing and revising the forms. The psychometric review consisted of comparisons of the expected representation and the actual representation of reporting categories, science and engineering practices, disciplinary core ideas, crosscutting concepts, performance expectations, and item types—SR, CR, TPI, and TPD at grades 3 and 4; and SR, CR, TE, TPI, TPD, and ER at grades 5 through 8—on the operational forms.

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

Table 4.4

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

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

Note: Detailed information can be found from the 2021–2023 Framework and Test Construction Document. It should be noted that these values are psychometric recommendations. Actual item decision occurs by content staff based on these recommendation criteria.

LDOE Review

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

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

Test Forms and Accessible Versions

Online and Paper Forms

The LEAP 2025 science assessments for grades 3 through 8 are administered as computer-based tests (CBT) with a paper-based option for grade 3 (selected at the school system level) and an accommodated print form only for a student who requires a paper-based accommodation for grades 4–8.

Given that there were two modes for grade 3, the mode effect analysis and equating that are given below were intended to be used with the data from the 2022 Science operational test. The results, however, demonstrated that there was no modal effect between the two modes. As a result, equating was done without taking the mode effect into account.

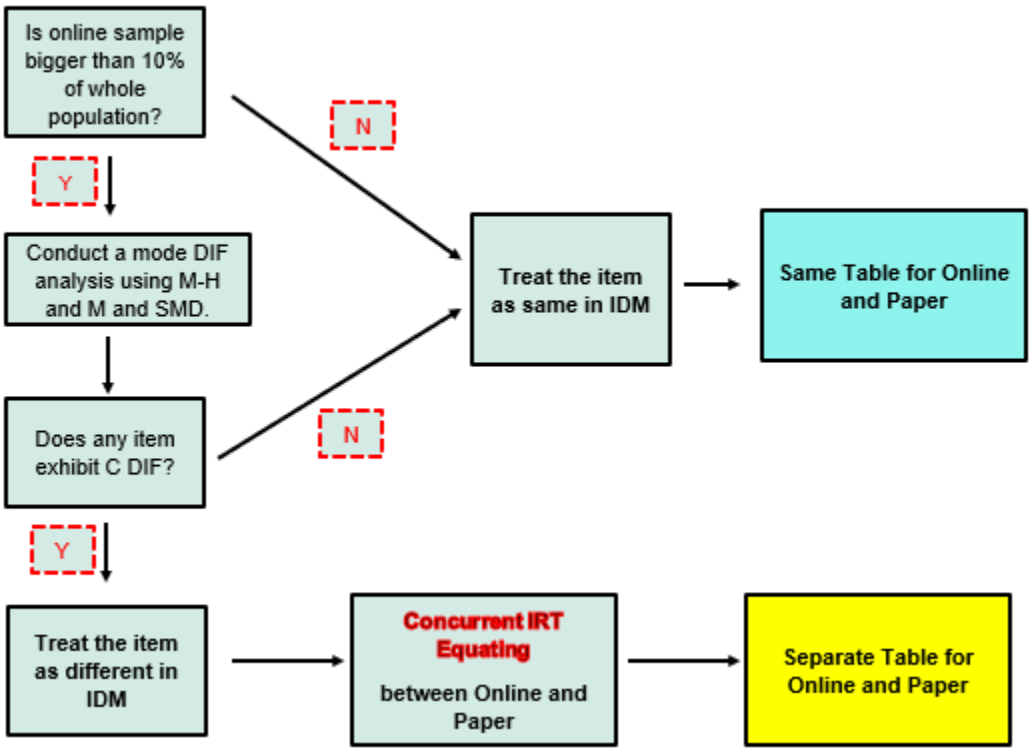


Figure 4.1. General overview of equating, including mode-effect analysis

Accommodated Print Versions

For grades 4–8, the accommodated print form was selected based on the field test version that contained the fewest and least complex technology-enhanced items. This version was identified as Version 1. The technology-enhanced items in this version were converted to a paper and pencil format that allowed students to record their responses, or have their responses transcribed into the test booklet. In addition, alternate text was written for all stimuli and items containing graphics. Detailed information can be found in Appendix G, Accommodated Print and Braille Creation.

Form Versions for Students with Visual Impairments

Braille and large-print test form versions were constructed for each grade to enable students with visual impairments to participate in the LEAP 2025 assessments. Version 1 of the grade 3 paper-based test form served as the basis for braille and large-print development. Braille forms for grades 4–8 were based on the accommodated print forms for operational items in Version 1. There are no large-print versions of the grades 4–8 accommodated print forms. Instead, students needing a large-print version in grades 4–8 use larger-sized monitors and/or the magnification features of the online testing system. All online test content has been developed to scale in relation to the available area on larger monitors while maintaining the correct aspect ratio. Specific recommendations on how to transcribe items into braille were provided by the braille publisher to produce the braille version of the LEAP 2025 assessments and the test administrator's notes that accompany the braille forms. The goal was to maximize the number of items that could be transcribed into braille.

For students who were administered a large-print or braille version, examiners were instructed to transcribe students' responses from the large-print or braille version into a consumable test booklet for grade 3, and the online testing system (INSIGHT) for grades 4 through 8, exactly as the students responded. Detailed information can be found in Appendix G, Accommodated Print and Braille Creation.

5. Test Administration

This chapter describes processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) *Standards for Educational and Psychological Testing* (hereafter the *Standards*), "The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions" (111). This chapter examines how test administration procedures implemented for the Louisiana 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 district test coordinators, who in turn convey test administration training to schools within their system. The LDOE conducts quality-assurance visits during testing to ensure adherence to the standardized administration of the tests.

The district 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 test administration, DRC produced two administration manuals:

1. *LEAP 2025 Grade 3 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 district and school test coordinators' responsibilities for distributing, collecting, and returning test materials to DRC for scoring.

Table of Contents for Paper-Based Testing Test Coordinators Manual

- Key Dates
- Spring 2022 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
    - Answer Change Analysis
    - Voiding Student Tests
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions

    - Test Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grade 3
- District Test Coordinator
    - Conduct Training Session
    - Receive Test Materials
    - Spanish Mathematics
    - Large-Print and Braille Test Materials and Communication Assistance Scripts (CAS)
    - 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 Consumable Test Booklets (Defined)
    - Unscorable Documents and Unscorable Document Labels
- Directions for Returning Test Materials to DRC in May
    - Pickup 1: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials
    - Pickup 3: Nonscorable Test Materials
    - 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 Consumable Test Booklets
    - Soiled, Damaged, and Other Unscorable Consumable Test Booklets
    - Verify and Distribute Materials to Test Administrators
    - Supervise Test Administration
    - Collect Test Materials
    - Used and Unused 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 District Test Coordinator
    - Pickup 1: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials
    - Pickup 3: Nonscorable Test Materials
    - Final Checklist for Returning Test Materials to the District Test Coordinator
- Void Notification—Spring 2022
- Index

Table of Contents for Computer-Based Testing Test Coordinators Manual

- Key Dates Spring 2022
- Resources Available in DRC INSIGHT Portal (eDIRECT) Spring 2022
- Spring 2022 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - DRC INSIGHT Portal (eDIRECT) and INSIGHT
- Test Security
    - Key Definitions
    - Violations of Test Security
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions
    - Testing Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Accommodations
    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grades 3 through 8
- Roles and Responsibilities
    - District 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
    - *DRC INSIGHT Portal (eDIRECT) User Guides*
    - *LEAP 2025 Accommodations and Accessibility Features User Guide*
    - *INSIGHT Technology User Guide*
    - Online Tools Training (OTT)
    - Student Tutorials
- Void Notification—Spring 2022

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 2022 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
    - 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
- Students Absent from Testing
- Consumable Test Booklet Coding
    - Coding the Demographic Section
- Sample Grade 3 English Language Arts Consumable Test Booklet
- General Instructions for LEAP 2025
    - Student Marking/Erasing on Consumable Test Booklet
    - Reading Directions to Students
    - Special Instructions
- Directions for Administering LEAP 2025 Tests
- Post-Test Procedures
    - Test Administrator Oath of Security and Confidentiality Statement
    - Used and Unused 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 2022 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
- Test Materials
    - Receipt of Test Materials
- 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
    - Student Tutorials
- 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)
- General Instructions
    - Reading Directions to Students
- LEAP 2025: Grades 3–8 English Language Arts (All Sessions)
- LEAP 2025: Grades 3–8 Mathematics (All Sessions)
- LEAP 2025: Grades 3–8 Science (Sessions 1–2)
- LEAP 2025: Grades 5–8 Science Session 3 Select Schools Only
- LEAP 2025: Grades 3–8 Social Studies (Grades 3–4 Sessions 1–2, Grades 5–8 Sessions 1–3)
- LEAP 2025: Grades 5–8 Social Studies Session 4 Select Schools Only
- 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 address Standard 4.15, which states:

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

The LEAP 2025 test administration manuals provide instructions for activities that happen before, during, and after testing with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the test administration manuals describe the following: general rules of 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 tests were administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The test coordinator’s manual included instructions for scheduling the test within the state testing window. The test coordinator’s manual and 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), Individual Accommodations Plan/504 Plan (IAP), or English Learner Plan (EL plan).
- allowing accommodations for students who do not have an IEP/IAP/EL plan.
- 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 tests. 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 tests, test administrators 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 administrators 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 district test coordinator ensures that the first and last security bar codes on the tests match the packing list they received. The district test coordinator then packages the tests to be sent to schools. Upon returning the test books to DRC, school and district 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.

Interpretive Guides

Essential to making valid interpretations of test scores is an understanding of what the test scores mean and how to interpret score reports. The Interpretive Guide is written for Louisiana teachers, administrators and parents who receive the LEAP 2025 score reports. https://www.louisianabelieves.com/resources/library/assessment-guidance

Time

Each session of each content area test was timed. Only students with an extended time accommodation were permitted to exceed the established time limits of any given session. The manuals provided test coordinators/administrators 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 INSIGHT portal (eDIRECT) and printed test tickets. Students entered their ticket information to access the test in INSIGHT. In addition, students have access to Online Tools Training before the testing window, which allows them to practice using tools and features within INSIGHT. Tutorials with online video clips that demonstrate features of the system are also available to students before testing.

Paper-Based Forms Administration, Grade 3

Schools with testers in grade 3 had the option to participate in either paper-based or computer-based testing for the spring 2022 test. DRC prints and ships paper materials to the sites that choose paper-based testing. These materials are returned to DRC after testing, for processing and scoring with the online tests.

Accessibility and Accommodations

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

- Access for All features are available to all students taking an assessment.
- Accessibility Features are available to students when deemed appropriate by a team of educators.
- Accommodations must appear in a student's IEP/IAP/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 an IAP, or who are identified as English Learners (ELs) and have an EL plan.

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

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

In compliance with this standard, the TAM contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP assessments. The following accommodations were provided by DRC for this administration:

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

The following additional access and accommodation features were also available:

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

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

Testing Windows

The computer-based test window was available from April 25 through May 25, 2022. Paper-based testing occurred from April 27 through May 3, 2022.

Test Security Procedures

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors, such as unauthorized exposure of test items that would affect the valid interpretation of test scores. Several test security measures are implemented for the LEAP 2025 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 tests and must account for all test materials and supervise the test administration at all times.

Data Forensic Analyses

Due to the importance of the LEAP 2025 assessments, it is prudent to ensure that the results from the assessments are based on effective instruction and true student achievement. While there are many ways to achieve meaningful understanding of student knowledge via test scores, there are also ways to obtain higher test scores that are not related to actual learning. To assist ensuring that assessment results are valid, data forensic analyses are conducted to help separate meaningful gains from spurious gains. It is important to note that although the results may be used to identify potential problems within a school, the identification of a problem is not an accusation of misconduct. Multiple methods were incorporated into the forensic analysis. The following methods were applied:

- Response Change Analysis
- Score Fluctuation Analysis
- Web Monitoring
- Plagiarism Detection

Response Change Analysis. Students make changes to answer choices when taking the LEAP 2025, and this is expected behavior. Unfortunately, changing student answers is also an opportunity for school personnel to improve classroom performance and, therefore, the response change analysis focuses on identifying school- and test-administrator-level response change patterns that are statistically improbable when compared to the expected pattern at the state level.

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

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

Plagiarism Detection. The LDOE monitors for two different plagiarism situations: copying from student to student and copying from an outside source, such as Wikipedia or another internet source. Instances of plagiarism are identified regardless of whether an item is scored by human scorers or artificial intelligence. Alerts are set to identify responses that may indicate the possibility of teacher interference, plagiarism, or disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). Alerted responses are given additional review so the appropriate response can be taken.

6. Scoring Activities

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

Selected-Response (SR) Item Keycheck. SR items for science include multiple-choice (MC) and multiple-select (MS) questions. Pearson calculates MC and MS item statistics and flags items if item statistics fall outside expected ranges. For example, items are flagged if few students select the correct response (*p*-value less than 0.15), if the item does not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) select a certain incorrect response. Lists of flagged MC and MS items, with the reasons for flagging, are provided to LDOE and WestEd content staff for key verification. The staff reviews the list of flagged MC and MS items to confirm that the answer keys are accurate. Scoring of MC and MS items is also evaluated at data review.

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

- A scoring rubric is created for each technology-enhanced item. The rubric describes the one and only correct answer for dichotomously scored items (i.e., items scored as either right or wrong). If partial credit is possible, the rubric

describes in detail the type of response that could receive credit for each score point.

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

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

Adjudication. TE items and other eligible items identified in the test map are automatically scored as tests are processed. TE items are scored according to scoring rules in the DOTS, which includes scoring information for all item types.

The adjudication process focuses on detecting possible errors in scoring for TE and MS items. For adjudication, DRC provides a report listing the frequency distributions of TE responses and multi-part multi-select items. Members of LDOE and WestEd content staff examine the TE and MS response distributions and the auto-frequency reports to evaluate whether the items are scored appropriately. In the event that scoring issues are identified, WestEd content staff and LDOE staff review and recommend changes to the scoring algorithm. Any changes to the scoring algorithm are based on the LDOE's decisions. DRC, in turn, applies the approved scoring changes to any affected items.

Constructed- and Extended-Response Item Scoring Process

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

Selection of Scoring Evaluators. Standard 4.20 states the following:

> The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring (92).

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

Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2021–2022 LEAP 2025 test responses have at least a four-year college degree.

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

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

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

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

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

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

occur, immediate action would be taken to secure materials, and the employee would be terminated. DRC has the same policy within our scoring sites.

Handscoring Training Process. Standard 6.9 specifies:

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

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

- Prompts and associated stimuli
- Rubrics
- Anchor sets
- Practice sets
- Qualifying sets

Training and Qualifying Procedures. Handscoring involves training and qualifying team leaders and evaluators, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities. The LDOE reviews training materials and oversees the training process.

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

Table 6.1
Science Training Set Composition

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

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

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

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

Table 6.2

Science Qualifying Standards

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

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

Monitoring the Scoring Process. Standard 6.8 states:

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

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

Reader Monitoring Procedures. Throughout the handscoring process, DRC project managers, scoring directors, and team leaders reviewed the statistics that were generated daily. DRC used one team leader for every 10 to 12 readers. If scoring concerns were apparent among individual scorers, team leaders dealt with those issues on an individual

basis. If a scorer appeared to need clarification of the scoring rules, DRC supervisors typically monitored one out of five of the scorer's readings, adjusting that ratio as needed. If a supervisor disagreed with a reader's scores during monitoring, the supervisor provided retraining in the form of direct feedback to the reader, using rubric language and applicable training responses.

Validity Sets and Inter-Rater Reliability. In addition to the feedback that supervisors provided to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducted validity scoring using validity responses. Validity responses were inserted among the live student responses.

The validity responses were added to DRC's image handscoring system prior to the beginning of scoring. Validity reports compared readers' scores to predetermined scores and were used to help detect potential room drift as well as individual scorer drift. This data was used to make decisions regarding the retraining and/or release of scorers, as well as the rescoring of responses.

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

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

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

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

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

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

Each reader was required to maintain a level of exact agreement on validity responses and on inter-rater reliability as shown under "Perfect Agreement" in the table above. Additionally, readers were required to maintain an acceptably low rate of nonadjacent agreement. To monitor this, DRC summed each reader's exact and adjacent agreement rates and required each reader to maintain the levels shown under "Perfect Agreement + Adjacent" in the table above.

Calibration Sets. DRC used these calibration sets to perform calibration across the entire scorer population for an item if trends were detected (e.g., low agreement between certain score points or if a certain type of response was missing from initial training). These calibrations were designed to help refocus scorers on how to properly use the

scoring guidelines. They were selected to help illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After readers scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and the anchor responses to explain the reasoning behind each response's score.

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

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

Tables 6.4–6.11 provide the inter-rater reliability and score point distributions by grade level for the constructed-response and extended-response items administered in the spring 2022 forms.

Table 6.4

Inter-Rater Reliability for Operational Constructed-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3	Item 1	≥14,150	98	2	0
	Item 2	≥15,540	91	8	0
	Item 3	≥16,410	95	5	0
4	Item 1	≥17,520	93	7	0
	Item 2	≥13,890	98	4	0
	Item 3	≥17,680	92	8	0
5	Item 1	≥12,320	85	13	2
	Item 2	≥11,930	93	6	1
	Item 3	≥13,800	91	8	1
6	Item 1	≥12,610	97	3	0
	Item 2	≥12,620	90	10	0
	Item 3	≥12,410	90	9	1
7	Item 1	≥13,760	86	13	0
	Item 2	≥14,140	93	4	3
	Item 3	≥15,810	95	5	0
8	Item 1	≥18,540	93	6	0
	Item 2	≥16,470	91	9	1
	Item 3	≥14,810	87	12	1

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

Table 6.5

Score Point Distributions for Operational Constructed-Response Items

Grade	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
3	Item 1	≥58,420	66	18	3	6	6
	Item 2	≥59,080	40	33	11	7	9
	Item 3	≥59,190	64	17	1	8	10
4	Item 1	≥57,340	57	22	8	0	12
	Item 2	≥55,250	66	19	5	0	10
	Item 3	≥57,430	43	43	2	0	13
5	Item 1	≥54,970	51	29	16	0	4
	Item 2	≥54,600	71	13	11	0	4
	Item 3	≥55,390	40	44	8	0	7
6	Item 1	≥55,020	79	9	7	0	5
	Item 2	≥55,160	64	27	3	0	5
	Item 3	≥54,980	74	17	3	0	4
7	Item 1	≥57,430	36	43	13	0	7
	Item 2	≥57,160	71	7	13	0	7
	Item 3	≥57,960	40	38	10	0	11
8	Item 1	≥58,950	58	23	3	0	15
	Item 2	≥58,070	51	35	3	0	11
	Item 3	≥57,270	53	27	11	0	8

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

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

Table 6.6

Inter-Rater Reliability for Operational-Extended Response Items

Grade	Inter-Rater Reliability*				
	2x	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
5	≥15,690	N/A	88	7	4
6	≥15,810	Part A	96	4	0
		Part B	87	8	5
7	≥14,790	N/A	90	10	0
8	≥17,690	Part A	86	13	2
		Part B	81	14	4

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

Table 6.7

Score Point Distributions for Operational Extended-Response Items

Grade	Score Point Distribution*													
	Total	Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	"6" (%)	"7" (%)	"8" (%)	"9" (%)	Blank (%)	Nonscore Codes (%)**
5	≥56,330	N/A	54	12	9	5	4	2	1	1	0	0	0	10
6	≥56,870	A	72	16	4								0	8
		B	38	6	14	6	12	5	7	3			0	8
7	≥57,720	N/A	10	8	14	17	18	11	11	1	0	1	0	8
8	≥58,620	A	30	27	20	8							0	15
		B	12	14	19	19	13	6	2				0	15

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

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

Table 6.8

Inter-Rater Reliability for Field Test Constructed-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3	Item 1	≥380	97	3	0
	Item 2	≥370	92	8	0
	Item 3	≥410	91	9	0
	Item 4	≥410	86	13	0
	Item 5	≥460	95	5	0
4	Item 1	≥400	85	14	1
	Item 2	≥410	88	11	0
	Item 3	≥430	94	6	0
	Item 4	≥350	96	4	0
	Item 5	≥370	88	11	2
	Item 6	≥400	95	4	0
	Item 7	≥460	94	6	0
	Item 8	≥350	94	6	1
	Item 9	≥410	94	6	0
	Item 10	≥370	83	10	7

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
5	Item 1	≥370	95	4	1
	Item 2	≥320	82	18	1
	Item 3	≥350	87	13	0
	Item 4	≥310	92	8	0
	Item 5	≥370	87	13	0
6	Item 1 Part A	≥340	88	12	0
	Item 1 Part B		83	17	0
	Item 2	≥410	90	8	2
	Item 3	≥400	87	13	0
	Item 4	≥420	96	4	0
	Item 5	≥390	88	12	0
	Item 6	≥390	81	13	6
	Item 7	≥420	98	2	0
7	Item 1	≥410	84	12	3
	Item 2	≥390	90	7	3
	Item 3	≥430	96	3	1
	Item 4	≥360	91	9	0
	Item 5	≥370	90	8	2
	Item 6	≥420	83	15	2
	Item 7	≥440	86	13	0
8	Item 1	≥430	84	15	1
	Item 2	≥340	90	9	1
	Item 3	≥470	85	14	1
	Item 4	≥390	88	11	1

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

Table 6.9

Score Point Distributions for Field Test Constructed-Response Items

Grade	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
3	Item 1	≥1,690	71	9	10	6	3
	Item 2	≥1,680	75	14	1	6	3
	Item 3	≥1,700	52	24	8	8	7
	Item 4	≥1,700	43	26	17	8	5
	Item 5	≥1,730	66	13	5	6	10
4	Item 1	≥1,700	65	24	7	0	5
	Item 2	≥1,700	73	18	3	0	5
	Item 3	≥1,710	78	12	3	0	7
	Item 4	≥1,670	40	44	15	0	1
	Item 5	≥1,680	57	26	14	0	3
	Item 6	≥1,700	78	15	1	0	4
	Item 7	≥1,730	53	29	9	0	10
	Item 8	≥1,670	77	13	8	0	1
	Item 9	≥1,700	70	20	4	0	6
	Item 10	≥1,680	55	15	27	0	3
5	Item 1	≥1,580	82	10	1	1	6
	Item 2	≥1,560	52	35	10	0	3
	Item 3	≥1,590	59	24	13	0	3
	Item 4	≥1,550	61	13	24	0	2
	Item 5	≥1,670	60	24	12	0	4
6	Item 1 Part A	≥1,670	30	68		0	2
	Item 1 Part B		64	34		0	2
	Item 2	≥1,690	79	9	5	0	6
	Item 3	≥1,700	52	37	5	0	6
	Item 4	≥1,700	65	20	7	0	6

Grade	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
	Item 5	≥1,680	16	40	39	0	5
	Item 6	≥1,690	65	21	9	0	4
	Item 7	≥1,700	85	4	3	0	7
7	Item 1	≥1,690	57	24	13	0	6
	Item 2	≥1,670	50	14	31	0	5
	Item 3	≥1,700	84	2	6	0	7
	Item 4	≥1,670	44	46	7	0	2
	Item 5	≥1,670	69	17	11	0	3
	Item 6	≥1,700	62	22	10	0	6
	Item 7	≥1,700	59	20	13	0	9
8	Item 1	≥1,690	29	34	30	0	7
	Item 2	≥1,580	70	15	11	0	4
	Item 3	≥1,640	43	33	13	0	11
	Item 4	≥1,610	62	25	5	0	7

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

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

Table 6.10

Inter-Rater Reliability for Field Test Extended-Response Items

Grade	Item	Inter-Rater Reliability*				
		2x	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
5	Item 1	≥600	Part A	86	12	3
			Part B	89	7	4
			Part C	86	9	5
	Item 2	≥350	Part A	89	11	0
			Part B	93	7	0
			Part C	94	6	0
	Item 3	≥350	Part A	85	15	0
			Part B	95	5	0
			Part C	91	9	0
6	Item 1	≥660	Part A	85	14	2
			Part B	84	13	3
7	Item 1	≥630	Part A	90	9	0
			Part B	89	9	2
			Part C	86	12	2
	Item 2	≥630	Part A	90	9	1
			Part B	76	12	12
8	Item 1	≥680	Part A	87	11	2
			Part B	92	8	1
			Part C	91	7	2

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

Table 6.11

Score Point Distributions for Field Test Extended-Response Items

Grade	Item	Score Point Distribution*													
		Total	Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	"6" (%)	"7" (%)	"8" (%)	"9" (%)	Blank (%)	Nonscore Codes (%)**
5	Item 1	≥2,790	A	33	42	9	12							0	5
			B	62	10	14	8							0	5
			C	75	10	8	2							0	5
	Item 2	≥1,580	A	67	23	5								0	5
			B	74	13	7	2							0	5
			C	84	7	4	1	0						0	5
	Item 3	≥1,590	A	50	20	16	10							0	3
			B	56	5	9	26							0	3
			C	49	18	17	13							0	3
6	Item 1	≥2,820	A	55	29	8	2	2						0	5
			B	53	15	12	7	4	3					0	5
7	Item 1	≥2,790	A	74	15	5	1							0	3
			B	64	20	11	1							0	3
			C	76	15	4	0							0	3
	Item 2	≥2,800	A	53	8	19	7	8						0	4
			B	46	16	9	7	9	8					0	4
8	Item 1	≥2,800	A	56	28	8	2							0	6
			B	69	20	4								0	6
			C	68	19	5	2	1						0	6

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

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

7. Data Analysis

Classical Item Statistics

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

Tables and figures that provide the additional information on classical item statistics for the spring 2022 test can be found in Appendix C: Item Analysis Summary Report. Tables C.1–C.5 show the summaries of classical item statistics. As a measure of item difficulty, p (or "the *p*-value") indicates the average proportion of total points earned on an item. For example, if p = 0.50 on an MC item, then half of the examinees earned a score of 1. If p = 0.50 on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). A measure of point-biserial correlation indicates a measure of item discrimination. Items with higher item-total correlations provide better information about how well items discriminate between lower- and higher-performing students. It should be noted that statistical analysis results for field test (FT) items are stored in Pearson's Assessment Banking and Building solutions for Interoperable assessment (ABBI) system.

Differential Item Functioning

Differential item functioning (DIF) analyses are intended to statistically signal potential item bias. DIF is defined as a difference between similar-ability groups' (e.g., males or females that attain the same total test score) probability of getting an item correct. Because test scores can reflect many sources of variation, the test developers' task is to create assessments that measure the intended knowledge and skills without introducing

construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores may reflect those extraneous elements in addition to what the test is purported to measure. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). Different cultural and socioeconomic experiences are among some factors that can confound test scores intended to reflect the measured construct.

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

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

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

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

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

where N_{r1k} is the number of correct responses in the reference group at ability level *k*, N_{f0k} is the number of incorrect responses in the focal group at ability level *k*, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level *k*, and N_{r0k} is the number of incorrect responses in the reference group at ability level *k*. The *MH DIF* statistic is based on a 2×2×*M* (2 groups × 2 item scores × *M*

strata) frequency table, in which students in the reference (male or white) and focal (female or black) groups are matched on their total raw scores.

The *ΔMH DIF* is then computed as

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

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

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

Table 7.1
DIF Categories for Dichotomous Items

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

For polytomous items, the standardized mean difference (*SMD*) (Dorans & Schmitt, 1991; Zwick, Thayer, & Mazzeo, 1997) and the Mantel χ^2 statistic (Mantel, 1963) are used to identify items with DIF. *SMD* estimates the average difference in performance between the reference group and the focal group while controlling for student ability. To calculate the *SMD*, let *M* represent the matching variable (total test score). For all $M = m$, identify the students with raw score *m* and calculate the expected item score for the reference group (E_{rm}) and the focal group (E_{fm}). DIF is defined as $D_m = E_{fm} - E_{rm}$, and *SMD* is a weighted average of D_m using the weights $w_m = N_{fm}$ (the number of students in the focal group with raw score *m*), which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

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

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

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

Table 7.2

DIF Categories for Polytomous Items

DIF Category	Criteria
A (negligible)	Mantel χ^2 p-value > 0.05 or $\lvert$ *SMD/SD* $\rvert$ ≤ 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* $\rvert$ ≥ 0.25

Three DIF analyses were conducted for the operational test items only: female/male, black/white, and Hispanic/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The same methods were used to detect items on which both black/white and Hispanic/white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Tables 7.3.1–7.3.3 provide the number of items flagged for DIF. Items flagged with A-DIF show negligible DIF, items flagged with B-DIF are said to exhibit slight to moderate DIF,and items with C-DIF are said to exhibit moderate to large DIF. Very few operational test items were flagged for C-DIF by either analysis.

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

Table 7.3.1

Summary of Female/Male DIF Flags by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	36	[0],[0]	[0],[0]
4	36	[0],[0]	[0],[0]
5	36	[0],[1]	[0],[0]
6	35	[0],[1]	[0],[0]
7	35	[0],[1]	[0],[0]
8	36	[0],[1]	[0],[0]

Table 7.3.2

Summary of African American/White DIF Flags by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	36	[0],[0]	[0],[0]
4	35	[0],[1]	[0],[0]
5	36	[0],[1]	[0],[0]
6	36	[0],[0]	[0],[0]
7	35	[0],[1]	[0],[0]
8	35	[0],[2]	[0],[0]

Table 7.3.3

Summary of Hispanic/White DIF Flags by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	36	[0],[0]	[0],[0]
4	35	[0],[1]	[0],[0]
5	36	[0],[1]	[0],[0]
6	36	[0],[0]	[0],[0]
7	35	[0],[1]	[0],[0]
8	35	[0],[2]	[0],[0]

Measurement Models

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

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

In that model, $p_i(\theta_j)$ is the probability that student *j* would earn a score of 1 on item *i*, b_i is the difficulty parameter for item *i*, a_i is the slope (or discrimination) parameter for item *i*, c_i is the pseudo-chance (or guessing) parameter for item *i*, and *D* is the constant 1.7. Since the Science tests also included polytomous items scored higher than 1 point, the generalized partial credit model (GPCM) (Muraki, 1992) was used to estimate the parameters of these items:

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

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

Calibration and Linking

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

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

The most common and straightforward way to score a test is to simply use the sum of points a student earned on the test, namely, the raw score. Although the raw score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations.

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

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

Tables 7.4.1–7.4.6 provide scale scores at selected percentiles that can be used to compare the distributional characteristics of the spring 2022 test form to previous administrations. Although these scale scores are rounded values, there were differences

in the scale score values for a given percentile across the forms. These variations could arise for several reasons: (1) differences in the proficiency (i.e., achievement) of the students in the samples or growth in student achievement across years; (2) unevenness in the respective distributions that combine with the number-correct-to-scale- score scoring method, leaving "gaps" in the scale; or (3) other sources of equating error. In general, however, the test characteristic function equating techniques will "level" the equated forms through the raw-to-scale- score adjustment.

Table 7.4.1

Comparisons of Scale Scores at Selected Percentiles: Grade 3 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	791	787	791
95	775	773	777
90	765	762	765
85	760	755	759
80	755	750	751
75	750	745	748
70	745	740	743
65	742	734	737
60	737	731	734
55	734	725	731
50	731	722	725
45	728	719	721
40	722	715	718
35	719	712	714
30	715	703	709
25	712	698	705
20	703	693	700
15	698	687	694
10	693	679	687
5	679	669	679
1	650	650	650

Table 7.4.2

Comparisons of Scale Scores at Selected Percentiles: Grade 4 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	798	798	803
95	782	779	782
90	774	770	771
85	766	762	764
80	764	756	759
75	758	751	754
70	753	748	749
65	751	742	747
60	748	739	741
55	743	734	739
50	740	731	733
45	737	725	730
40	734	721	727
35	728	718	723
30	725	712	720
25	722	707	716
20	716	703	711
15	708	695	701
10	704	690	695
5	690	678	687
1	668	651	664

Table 7.4.3

Comparisons of Scale Scores at Selected Percentiles: Grade 5 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	807	807	804
95	788	785	785
90	776	773	774
85	768	765	766
80	762	760	761
75	757	752	756
70	752	747	750
65	747	742	745
60	745	737	739
55	740	735	733
50	735	729	730
45	732	723	724
40	726	717	718
35	723	714	714
30	717	707	706
25	714	703	702
20	707	694	693
15	698	689	688
10	689	677	676
5	677	671	660
1	654	650	650

Table 7.4.4

Comparisons of Scale Scores at Selected Percentiles: Grade 6 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	797	794	800
95	779	776	778
90	769	767	766
85	763	758	758
80	758	753	753
75	753	749	747
70	749	744	741
65	744	739	736
60	742	734	730
55	736	731	727
50	734	725	721
45	728	722	717
40	725	719	714
35	722	716	706
30	719	709	702
25	712	704	697
20	709	700	692
15	704	695	687
10	695	683	680
5	683	676	665
1	657	650	650

Table 7.4.5

Comparisons of Scale Scores at Selected Percentiles: Grade 7 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	809	805	812
95	786	783	784
90	775	770	773
85	767	762	765
80	759	754	757
75	754	748	751
70	751	743	746
65	746	740	743
60	743	735	737
55	737	732	735
50	735	726	729
45	729	723	726
40	726	717	723
35	723	714	717
30	717	711	713
25	714	707	710
20	707	699	702
15	703	695	698
10	695	690	688
5	685	679	681
1	662	651	653

Table 7.4.6

Comparisons of Scale Scores at Selected Percentiles: Grade 8 Operational Forms

Percentile	2019 Spring Form A	2021 Spring Form A	2022 Spring Form B
99	803	799	802
95	784	778	781
90	773	768	773
85	766	761	765
80	761	756	758
75	756	750	754
70	752	745	749
65	747	743	744
60	743	738	740
55	741	733	735
50	736	729	730
45	731	726	728
40	729	721	723
35	723	718	717
30	721	712	711
25	715	708	708
20	708	701	701
15	705	697	697
10	697	687	687
5	682	675	682
1	658	650	658

Operational Item Parameters

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

Item Fit

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q_1 statistic (Yen, 1981) is used as an index for how well theoretical item curves match observed item responses. Q_1 is computed by first conducting an IRT item parameter estimation, then estimating students' achievement using the estimated item parameters, and, finally, using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at 10 selected equal intervals across the range of student achievement. Q_1 is computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-square (χ^2) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis (the expected item performance). Q_1 for each item type has varying degrees of freedom because the different item types have different numbers of IRT parameters. Therefore, Q_1 is not directly comparable across item types. An adjustment or linear transformation (translation to a *Z*-score, Z_{Q_1}) is made for different numbers of item parameters and sample size to create a more comparable statistic.

It should be noted that Yen's Q_1 statistic (Yen, 1981) was calculated to evaluate item fit for both operational and field test items by comparing observed and expected item

performance. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. For dichotomous items, Q_1 is computed as

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

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

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

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

The generalization of Q_1 for items with multiple response categories is

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

where

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

Both Q_1 and generalized Q_1 results are transformed to ZQ_1 and are compared to a criterion $ZQ_{1,\text{crit}}$ 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 for operational items is displayed in Appendix D: Dimensionality.

Dimensionality and Local Item Independence

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

Computation of the Q_3 statistic starts with expected student performance on each item, which is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference is the remainder in performance after accounting for underlying achievement. If performance on an item is driven by a predominant achievement construct, then the residual will be small (as tested by the Q_1 statistic), and the correlation between residuals of the item pairs will also be small. These correlations are analogous to partial correlations or the relationship between two variables (items) after accounting for the effects of a third variable (underlying achievement). The correlation among IRT residuals is the Q_3 statistic.

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

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

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

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

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

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

SCALING

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

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

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

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

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

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

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

which are rearranged as

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

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

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

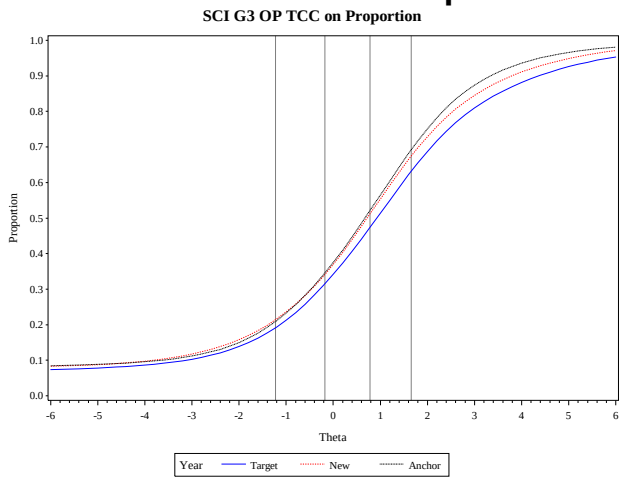
The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same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

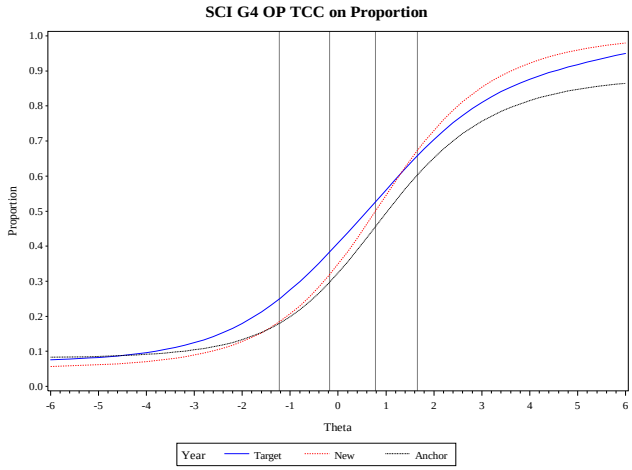
Test Characteristic Curve (TCC)

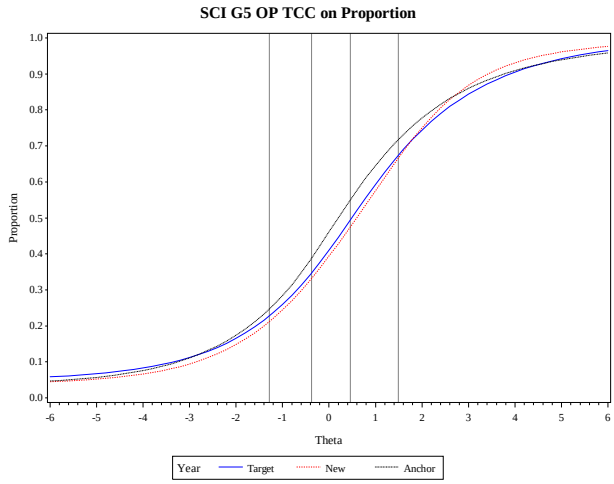
Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) across administrations of the LEAP 2025 Science assessments, as can be seen in the following figure. As seen from Plot 7.1, the TCCs between two years were similar across ability ranges. By the way, Plot 9.1 also indicates that the SEMs between two years are similar across ability ranges, especially in the middle ability ranges; each theta cut matches the scale score of each performance-level cut (e.g., 704, 725, 750, and 778 for Grade 4).

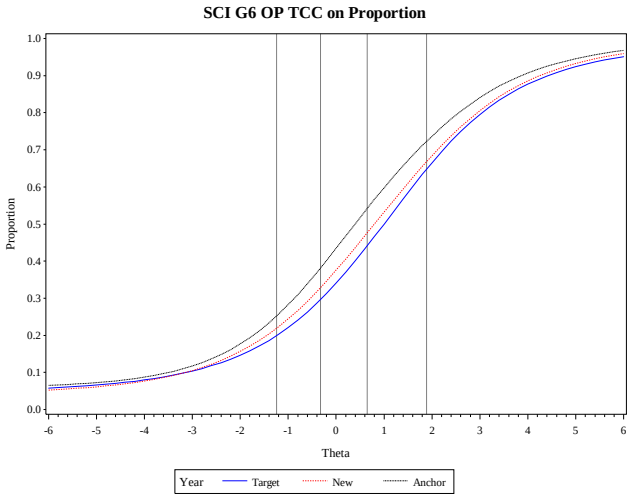
Plot 7.1

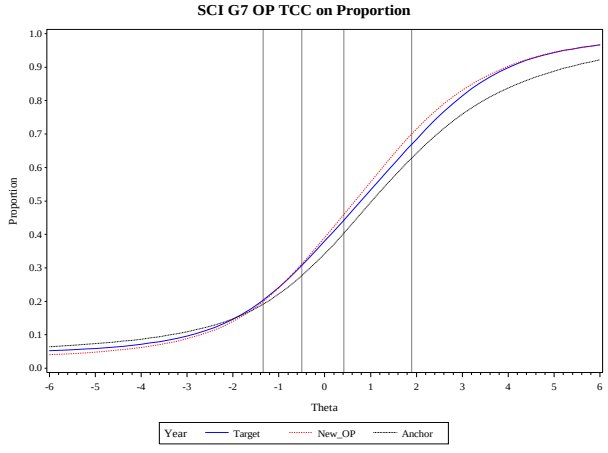
Test Characteristic Curve: Operational Science Gr3–8

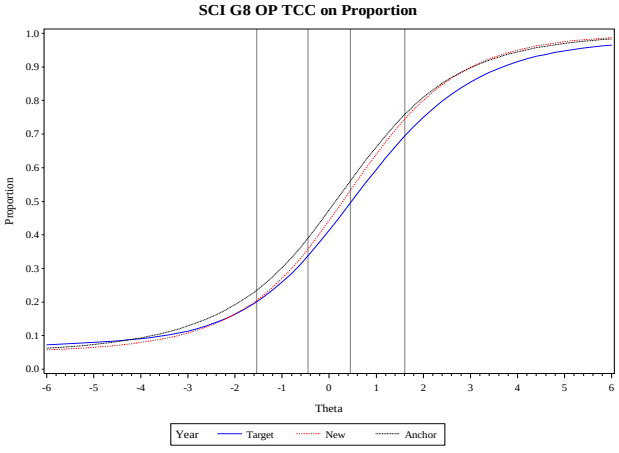












Test Information Curve, Score Distribution, and IRT Difficulty Distribution

In this section, student's Science test score distribution, IRT item difficulty (i.e., b-parameter) distribution, and item information curve are presented. Compared to the base year (i.e., 2019 Science test), the 2022 Science tests generally follow the shape of the base year's test information and provide more test information around the middle range of theta than other rages, as can be observed from Tables 7.5.1–7.5.6 and Plot 7.2.

Table 7.5.1
SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 3

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
1.23	theta < -3.5	0
0.00	-3.5 <= theta < -3.0	0
1.46	-3.0 <= theta < -2.5	0
2.12	-2.5 <= theta < -2.0	0
7.25	-2.0 <= theta < -1.5	0
10.15	-1.5 <= theta < -1.0	2
15.44	-1.0 <= theta < -0.5	0
17.17	-0.5 <= theta < 0.0	0
13.85	0.0 <= theta < 0.5	9
13.73	0.5 <= theta < 1.0	10
9.80	1.0 <= theta < 1.5	7
4.85	1.5 <= theta < 2.0	2
1.96	2.0 <= theta < 2.5	2
0.78	2.5 <= theta < 3.0	3
0.17	3.0 <= theta < 3.5	0
0.04	3.5 <= theta	1
-6.00	**Minimum**	-1.16
4.19	**Maximum**	3.56
-0.17	**Mean**	0.99
1.25	**SD**	0.98
≥49,320	**Total Number of Examinees**	36

Table 7.5.2

SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 4

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.87	theta < -3.5	0
0.00	-3.5 <= theta < -3.0	0
1.18	-3.0 <= theta < -2.5	0
2.05	-2.5 <= theta < -2.0	0
6.62	-2.0 <= theta < -1.5	0
9.16	-1.5 <= theta < -1.0	0
14.60	-1.0 <= theta < -0.5	0
19.98	-0.5 <= theta < 0.0	2
16.02	0.0 <= theta < 0.5	5
12.43	0.5 <= theta < 1.0	15
9.08	1.0 <= theta < 1.5	6
4.51	1.5 <= theta < 2.0	5
2.46	2.0 <= theta < 2.5	1
0.76	2.5 <= theta < 3.0	2
0.20	3.0 <= theta < 3.5	0
0.07	3.5 <= theta	0
-6.00	**Minimum**	-0.36
5.06	**Maximum**	2.99
-0.10	**Mean**	0.99
1.18	**SD**	0.71
≥48,910	**Total Number of Examinees**	36

Table 7.5.3

SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 5

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.74	theta < -3.5	0
0.91	-3.5 <= theta < -3.0	0
3.48	-3.0 <= theta < -2.5	0
5.42	-2.5 <= theta < -2.0	0
6.52	-2.0 <= theta < -1.5	0
10.08	-1.5 <= theta < -1.0	3
16.44	-1.0 <= theta < -0.5	3
11.77	-0.5 <= theta < 0.0	7
16.42	0.0 <= theta < 0.5	9
13.28	0.5 <= theta < 1.0	5
7.85	1.0 <= theta < 1.5	5
4.66	1.5 <= theta < 2.0	4
1.72	2.0 <= theta < 2.5	0
0.52	2.5 <= theta < 3.0	0
0.15	3.0 <= theta < 3.5	1
0.04	3.5 <= theta	0
-6.00	**Minimum**	-1.27
4.57	**Maximum**	3.25
-0.29	**Mean**	0.40
1.27	**SD**	0.99
≥48,900	**Total Number of Examinees**	37

Table 7.5.4

SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 6

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
1.42	theta < -3.5	0
1.63	-3.5 <= theta < -3.0	0
2.38	-3.0 <= theta < -2.5	0
7.27	-2.5 <= theta < -2.0	0
9.22	-2.0 <= theta < -1.5	0
13.80	-1.5 <= theta < -1.0	0
11.88	-1.0 <= theta < -0.5	2
16.31	-0.5 <= theta < 0.0	9
10.96	0.0 <= theta < 0.5	4
11.01	0.5 <= theta < 1.0	6
7.42	1.0 <= theta < 1.5	4
3.60	1.5 <= theta < 2.0	3
1.97	2.0 <= theta < 2.5	3
0.72	2.5 <= theta < 3.0	4
0.29	3.0 <= theta < 3.5	2
0.14	3.5 <= theta	0
-6.00	**Minimum**	-0.79
5.54	**Maximum**	3.28
-0.45	**Mean**	1.00
1.34	**SD**	1.21
≥49,300	**Total Number of Examinees**	37

Table 7.5.5

SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 7

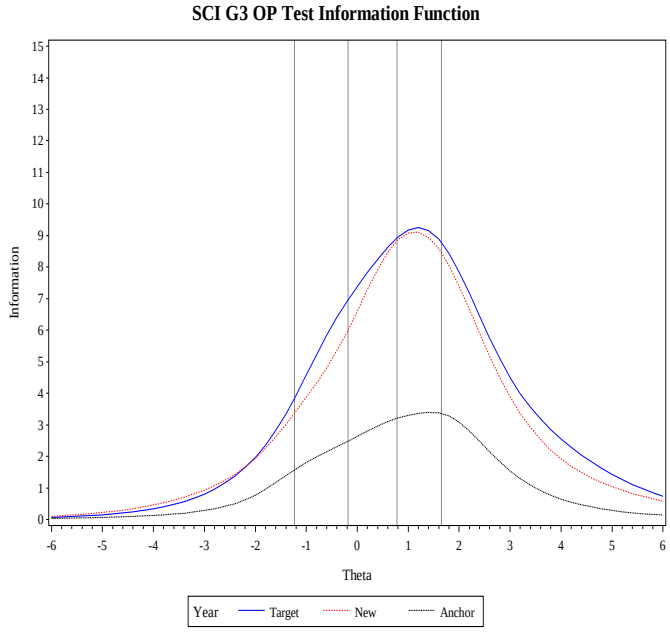
Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.79	theta < -3.5	0
0.86	-3.5 <= theta < -3.0	0
1.38	-3.0 <= theta < -2.5	0
4.30	-2.5 <= theta < -2.0	0
6.16	-2.0 <= theta < -1.5	0
14.79	-1.5 <= theta < -1.0	1
15.28	-1.0 <= theta < -0.5	2
17.89	-0.5 <= theta < 0.0	7
14.52	0.0 <= theta < 0.5	5
10.58	0.5 <= theta < 1.0	5
6.79	1.0 <= theta < 1.5	4
3.38	1.5 <= theta < 2.0	4
1.94	2.0 <= theta < 2.5	6
0.77	2.5 <= theta < 3.0	0
0.39	3.0 <= theta < 3.5	2
0.17	3.5 <= theta	0
-6.00	**Minimum**	-1.01
6.00	**Maximum**	3.39
-0.31	**Mean**	0.86
1.21	**SD**	1.09
≥50,990	**Total Number of Examinees**	36

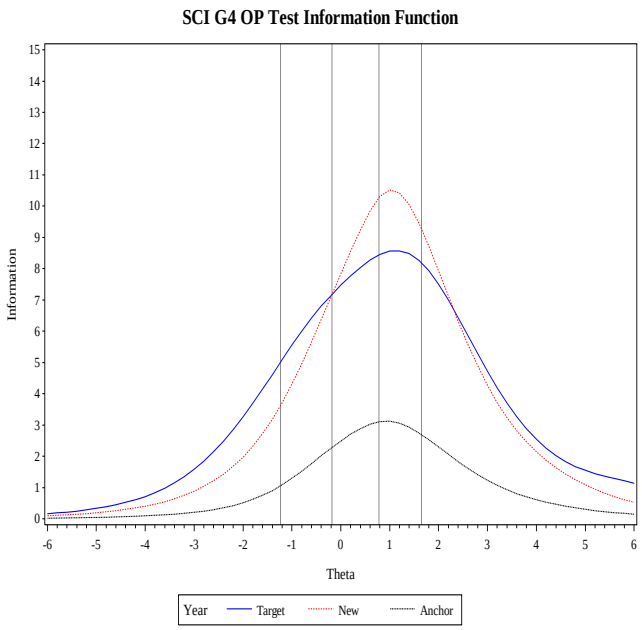
Table 7.5.6

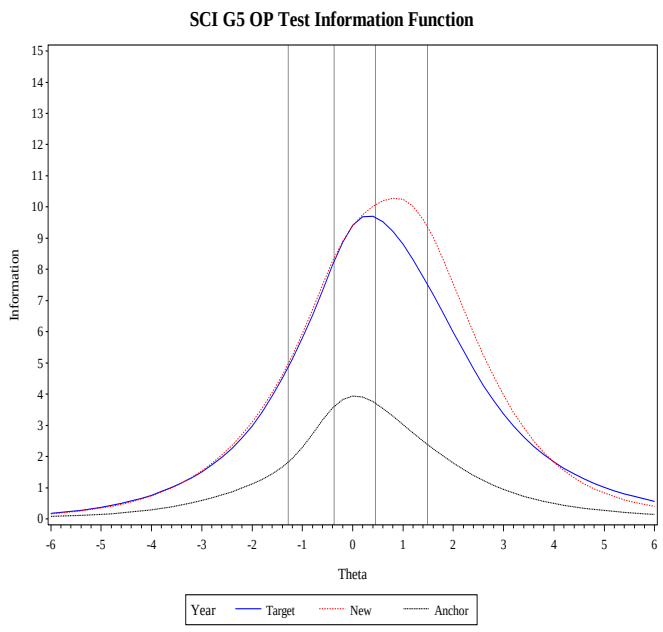
SPR 2022 Student's Score and IRT B-Parameter Distribution: Grade 8

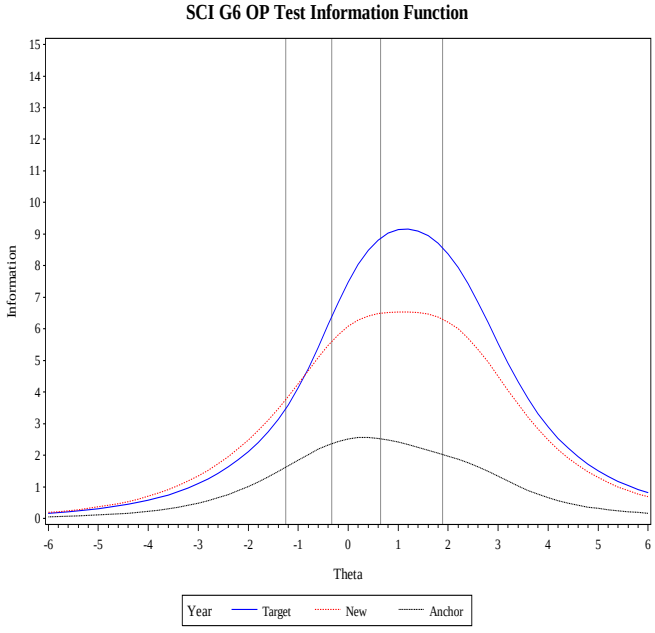
Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.36	theta < -3.5	0
0.45	-3.5 <= theta < -3.0	0
2.13	-3.0 <= theta < -2.5	0
1.93	-2.5 <= theta < -2.0	0
8.49	-2.0 <= theta < -1.5	0
13.66	-1.5 <= theta < -1.0	2
15.15	-1.0 <= theta < -0.5	2
13.73	-0.5 <= theta < 0.0	8
17.49	0.0 <= theta < 0.5	8
12.92	0.5 <= theta < 1.0	11
7.62	1.0 <= theta < 1.5	4
3.72	1.5 <= theta < 2.0	1
1.74	2.0 <= theta < 2.5	2
0.37	2.5 <= theta < 3.0	0
0.17	3.0 <= theta < 3.5	0
0.07	3.5 <= theta	0
-6.00	**Minimum**	-1.42
5.62	**Maximum**	2.30
-0.24	**Mean**	0.43
1.16	**SD**	0.84
≥50,720	**Total Number of Examinees**	38

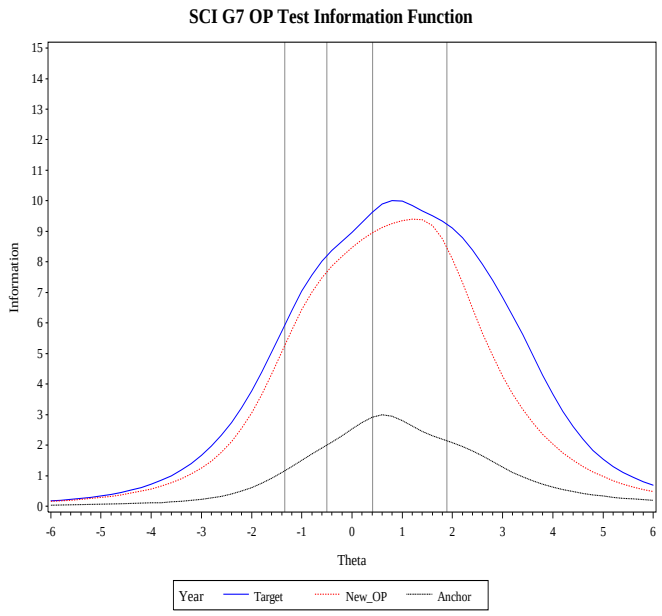
Plot 7.2
Test Information Curve; SPR 2022 Science G3-8

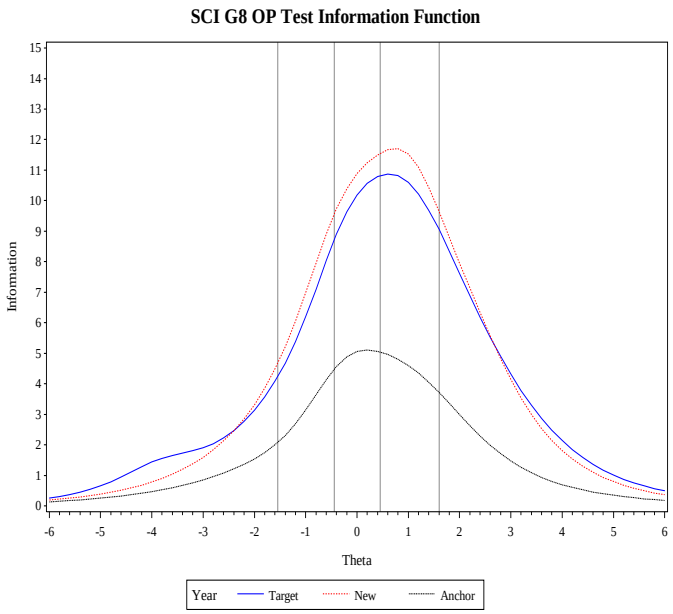









SCI G7 OP Test Information Function
Information
Theta
Year
Target
New_OP
Anchor


SCI G8 OP Test Information Function
Information
Theta
Year
Target
New
Anchor

Field Test Data Review

The process used to complete the field test item equating is an anchored item equating process. In this process the item parameters from the operational items from the 2022 administration were fixed as constant (i.e., to calculate Stocking-Lord equating constant) and the item parameters for the field test items were freely calibrated, placing the item parameters for the field test items on the same scale as the operational items.

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

8. Test Results and Score Reports

This chapter provides information on the results of the Spring LEAP 2025 Science tests. The scale score results and achievement level information are also presented here. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative description of student achievement. The levels are Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory. The results in the following tables are presented as evidence of the reliability and validity of the scores from the LEAP 2025 Science G3–8 tests.

Demographic Characteristics of Students

The operational Science tests were administered to all eligible students in the appropriate grade level during spring 2022. Spring 2022 operational score results were reviewed based on the following student characteristics:

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

Test Results

For the spring 2022 Science tests, the lowest obtainable scale score (LOSS) on the tests is 650 and the highest obtainable scale score (HOSS) is 850. Scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state and disaggregated into various demographic groups. In addition to the descriptive statistics presented in the following tables, scale score frequency distributions are presented in Appendix E: Scale Distribution and Statistical Report. Finally, because the spring 2022 tests were administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 8.1.1
LEAP 2025 State Test Results: Spring 2022 Grade 3

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥49,320	725.78	30.74	17	30	31	16	6
Gender								
Female	≥24,090	725.56	29.86	16	30	32	15	6
Male	≥25,230	725.99	31.56	17	30	29	17	7
Ethnicity								
African American	≥20,380	714.12	27.37	25	38	27	8	2
American Indian orAlaska Native	≥270	729.89	27.84	12	30	37	14	7
Asian	≥830	743.86	30.28	6	18	29	29	17
Hispanic/Latino	≥4,970	719.30	29.83	22	34	29	12	4
Multi-Racial	≥1,850	730.63	28.96	11	28	34	19	7
Native Hawaiian orOther Pacific Islander	≥30	728.38	28.05	19	19	35	24	3
White	≥20,930	737.48	29.45	9	22	34	24	11
Economically Disadvantaged*								
No	≥13,960	742.12	29.35	7	18	33	27	15
Yes	≥35,200	719.36	28.82	21	35	30	12	3

Table 8.1.1 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥46,480	726.89	30.66	16	29	31	17	7
Yes	≥2,840	707.59	25.89	32	41	21	5	NR
Education Classification								
Regular	≥43,010	727.71	30.42	15	29	32	17	7
Special	≥6,300	712.65	29.63	29	38	22	8	3
Section 504								
No	≥45,780	726.31	30.82	17	30	31	17	7
Yes	≥3,540	718.95	28.78	21	36	29	11	4
Migrant								
No	≥49,230	725.81	30.74	17	30	31	16	6
Yes	≥90	712.34	29.67	30	34	24	8	4
Homeless Status								
No	≥47,910	726.13	30.72	17	30	31	16	7
Yes	≥1,410	713.93	28.89	27	37	24	10	2
Military Affiliation								
No	≥48,370	725.49	30.69	17	30	31	16	6
Yes	≥950	740.31	29.47	8	19	34	26	14
Foster Care Status								
No	≥49,160	725.80	30.75	17	30	31	16	6
Yes	≥150	720.08	26.75	17	37	33	11	2

* Economic status was not available for all students.

Table 8.1.2

LEAP 2025 State Test Results: Spring 2022 Grade 4

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥48,910	733.86	29.73	15	24	32	23	7
Gender								
Female	≥23,900	732.45	28.79	16	25	32	22	5
Male	≥25,010	735.21	30.54	15	23	31	24	8
Ethnicity								
African American	≥20,500	721.94	26.20	23	32	32	12	2
American Indian orAlaska Native	≥260	739.32	27.02	9	22	33	29	6
Asian	≥800	755.06	30.03	5	11	25	36	23
Hispanic/Latino	≥5,080	728.96	29.14	19	27	30	19	5
Multi-Racial	≥1,680	739.01	28.26	10	21	35	26	8
Native Hawaiian orOther Pacific Islander	≥30	737.44	23.57	5	21	46	21	8
White	≥20,510	745.67	28.13	7	15	32	34	11
Economically Disadvantaged*								
No	≥13,980	749.79	28.00	6	13	30	37	15
Yes	≥34,630	727.55	27.94	19	28	32	17	3

Table 8.1.2 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥46,440	734.86	29.62	14	23	32	24	7
Yes	≥2,460	715.10	25.16	31	36	24	7	1
Education Classification								
Regular	≥42,940	736.10	29.22	13	22	33	25	7
Special	≥5,970	717.73	28.39	31	33	24	10	3
Section 504								
No	≥44,770	734.44	29.74	15	23	32	23	7
Yes	≥4,140	727.59	28.87	19	30	30	17	4
Migrant								
No	≥48,850	733.87	29.73	15	24	32	23	7
Yes	≥60	727.65	30.37	24	27	27	18	3
Homeless Status								
No	≥47,600	734.17	29.72	15	23	32	23	7
Yes	≥1,310	722.76	28.01	25	29	29	14	3
Military Affiliation								
No	≥47,960	733.55	29.68	15	24	32	23	7
Yes	≥950	749.57	27.95	5	14	29	37	15
Foster Care Status								
No	≥48,780	733.88	29.73	15	24	32	23	7
Yes	≥130	727.57	29.09	15	36	28	15	5

* Economic status was not available for all students.

Table 8.1.3

LEAP 2025 State Test Results: Spring 2022 Grade 5

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥48,900	727.90	36.92	20	26	23	24	7
Gender								
Female	≥23,820	728.75	35.54	18	27	24	24	7
Male	≥25,070	727.08	38.17	22	25	22	23	8
Ethnicity								
African American	≥20,660	713.20	33.51	31	33	21	13	2
American Indian orAlaska Native	≥250	729.42	33.39	14	33	22	27	4
Asian	≥740	755.48	35.65	6	13	20	37	24
Hispanic/Latino	≥4,790	722.64	36.20	24	28	23	20	5
Multi-Racial	≥1,640	734.62	35.48	15	22	25	29	8
Native Hawaiian orOther Pacific Islander	≥40	727.84	39.60	27	25	16	20	11
White	≥20,740	742.21	34.24	10	20	25	34	12
Economically Disadvantaged*								
No	≥14,580	747.31	34.12	8	17	23	36	16
Yes	≥34,000	719.73	34.92	26	30	22	18	3

Table 8.1.3 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥46,820	729.12	36.73	19	26	23	24	7
Yes	≥2,070	700.38	29.83	45	36	13	5	1
Education Classification								
Regular	≥42,890	731.48	35.73	17	26	24	26	8
Special	≥6,010	702.31	35.15	47	29	13	9	2
Section 504								
No	≥44,160	728.99	36.95	20	26	23	25	7
Yes	≥4,740	717.66	35.04	27	32	21	16	4
Migrant								
No	≥48,840	727.92	36.92	20	26	23	24	7
Yes	≥50	710.78	34.37	29	36	17	17	2
Homeless Status								
No	≥47,680	728.24	36.92	20	26	23	24	7
Yes	≥1,210	714.48	34.34	31	32	19	15	2
Military Affiliation								
No	≥48,000	727.54	36.88	21	26	23	23	7
Yes	≥890	747.07	34.30	8	17	22	39	14
Foster Care Status								
No	≥48,790	727.93	36.92	20	26	23	24	7
Yes	≥110	712.45	33.60	29	35	19	15	2

* Economic status was not available for all students.

Table 8.1.4

LEAP 2025 State Test Results: Spring 2022 Grade 6

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥49,300	722.14	33.65	27	28	23	19	4
Gender								
Female	≥23,950	722.35	32.27	25	29	24	18	3
Male	≥25,350	721.95	34.90	28	26	22	19	4
Ethnicity								
African American	≥20,700	709.84	29.36	38	33	19	9	1
American Indian orAlaska Native	≥280	725.45	30.08	19	28	31	18	3
Asian	≥790	747.47	37.86	11	17	23	32	18
Hispanic/Latino	≥5,070	715.52	33.11	34	28	21	15	2
Multi-Racial	≥1,620	726.63	31.64	20	27	27	21	4
Native Hawaiian orOther Pacific Islander	≥20	731.72	33.89	14	31	28	21	7
White	≥20,780	734.66	32.68	15	23	27	29	7
Economically Disadvantaged*								
No	≥14,460	740.18	32.25	11	20	27	33	9
Yes	≥34,560	714.74	31.29	33	31	21	13	2

Table 8.1.4 (continued)

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥47,230	723.36	33.41	25	28	23	20	4
Yes	≥2,070	694.43	26.37	60	28	9	3	NR
Education Classification								
Regular	≥43,740	724.97	33.13	23	28	24	21	4
Special	≥5,560	699.93	29.12	55	28	11	6	1
Section 504								
No	≥44,100	723.35	33.76	26	27	23	20	4
Yes	≥5,200	711.93	30.88	37	32	19	11	1
Migrant								
No	≥49,240	722.15	33.65	27	28	23	19	4
Yes	≥60	721.73	32.55	30	20	27	21	2
Homeless Status								
No	≥48,050	722.44	33.66	26	28	23	19	4
Yes	≥1,250	710.92	31.13	37	31	19	11	1
Military Affiliation								
No	≥48,460	721.86	33.59	27	28	23	19	4
Yes	≥840	738.42	33.09	13	21	26	31	9
Foster Care Status								
No	≥49,170	722.17	33.65	27	28	23	19	4
Yes	≥130	710.95	31.59	40	27	18	14	1

* Economic status was not available for all students.

Table 8.1.5

LEAP 2025 State Test Results: Spring 2022 Grade 7

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥50,990	730.41	32.50	17	27	30	23	4
Gender								
Female	≥25,080	731.49	30.60	14	27	32	23	3
Male	≥25,910	729.36	34.21	19	26	28	22	5
Ethnicity								
African American	≥21,880	718.73	28.20	25	35	28	12	1
American Indian orAlaska Native	≥290	733.66	28.78	13	22	37	27	2
Asian	≥740	757.69	35.02	5	13	22	42	18
Hispanic/Latino	≥4,840	724.24	33.23	24	27	28	19	3
Multi-Racial	≥1,690	735.13	31.89	13	24	31	27	5
Native Hawaiian orOther Pacific Islander	≥40	740.74	30.24	5	26	26	38	5
White	≥21,460	742.34	31.54	9	19	32	33	7
Economically Disadvantaged*								
No	≥15,140	746.59	31.76	7	16	31	37	9
Yes	≥35,550	723.67	30.29	21	31	30	17	2

Table 8.1.5 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥49,150	731.51	32.21	16	26	31	23	4
Yes	≥1,830	700.92	25.77	49	33	14	3	NR
Education Classification								
Regular	≥45,420	733.40	31.67	14	26	32	25	4
Special	≥5,560	706.04	28.75	43	34	16	6	1
Section 504								
No	≥45,600	731.65	32.51	16	26	30	24	4
Yes	≥5,380	719.89	30.44	25	34	26	13	2
Migrant								
No	≥50,930	730.42	32.50	17	27	30	23	4
Yes	≥60	719.20	28.75	21	36	28	13	2
Homeless Status								
No	≥49,820	730.65	32.49	17	27	30	23	4
Yes	≥1,160	720.06	31.43	25	31	28	14	2
Military Affiliation								
No	≥50,120	730.10	32.45	17	27	30	22	4
Yes	≥860	748.43	30.10	5	16	30	41	8
Foster Care Status								
No	≥50,860	730.44	32.51	17	27	30	23	4
Yes	≥120	718.31	28.15	24	37	24	14	1

* Economic status was not available for all students.

Table 8.1.6

LEAP 2025 State Test Results: Spring 2022 Grade 8

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥50,720	730.81	32.05	13	29	29	24	5
Gender								
Female	≥25,010	730.96	30.87	12	29	31	24	4
Male	≥25,700	730.66	33.17	15	28	27	24	6
Ethnicity								
African American	≥21,550	717.26	28.07	21	39	27	12	1
American Indian orAlaska Native	≥280	736.83	30.56	6	30	30	29	5
Asian	≥810	756.61	34.26	5	12	22	39	23
Hispanic/Latino	≥4,830	724.77	32.78	19	30	28	21	3
Multi-Racial	≥1,570	736.11	30.74	9	24	32	29	6
Native Hawaiian orOther Pacific Islander	≥40	742.77	35.33	9	19	23	40	9
White	≥21,610	744.20	29.31	5	19	32	36	8
Economically Disadvantaged*								
No	≥34,350	723.18	30.24	17	34	28	17	2
Yes	≥300	716.20	31.24	26	36	23	14	1

Table 8.1.6 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥48,900	731.95	31.70	12	28	30	25	5
Yes	≥1,810	700.02	25.40	42	42	13	4	NR
Education Classification								
Regular	≥45,500	733.60	31.34	11	27	30	26	5
Special	≥5,210	706.43	27.59	34	42	16	6	1
Section 504								
No	≥45,550	732.00	32.12	13	28	29	25	5
Yes	≥5,160	720.36	29.47	19	38	27	14	2
Migrant								
No	≥50,660	730.82	32.06	13	29	29	24	5
Yes	≥50	725.63	29.67	12	36	34	19	NR
Homeless Status								
No	≥49,600	731.04	32.03	13	29	29	24	5
Yes	≥1,110	720.75	31.45	20	37	24	15	3
Military Affiliation								
No	≥49,810	730.48	31.98	14	29	29	24	5
Yes	≥910	748.95	30.71	4	17	28	38	13
Foster Care Status								
No	≥50,580	730.85	32.05	13	29	29	24	5
Yes	≥130	715.29	29.37	27	37	25	10	2

* Economic status was not available for all students.

Effect Size

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

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

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

Cohen's *d*, then, expresses the difference in group means in terms of the standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the *d* statistic: *d* = 0.20 is a small ES, *d* = 0.50 is a medium ES, and *d* = 0.80 is a large ES. Based on Cohen's (1988) guidelines, certain trends are observable in Tables B.6.1–B.6.6. Although no big difference in Science tests was seen between females and males, mean raw scores and ESs show that Asian and White students tend to outperform other ethnicity groups. There were clear performance differences among regular education, gifted/talented education, and special education students in Education Classification and Non-English Learner and English Learner in EL status. Performance differences were also observed from Economically Disadvantaged status, Homeless status, Foster Care status, and Military Affiliation status.

Score Reports

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

School Roster Report. A School Roster Report, which provides summary information about student performance on the LEAP 2025 Grades 3–8 Science tests, is available to school systems and schools through eDIRECT. Total test scores and achievement level indicators are shown for the test of interest. Category and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by category and subcategory are summarized. More details can be found in the LEAP 2025 Grades 3-8 Interpretive Guide (iGUIDE) Spring 2022.

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

LEAP 2025 Grades 3-8 Interpretive Guide (iGUIDE) Spring 2022. The LEAP 2025 Grades 3-8 Interpretive Guide (iGUIDE) Spring 2022 was written to help Louisiana school system and school administrators, teachers, parents, and the general public understand the LEAP Science Grades 3–8 tests. The LEAP 2025 Grades 3-8 Interpretive Guide (iGUIDE) Spring

2022 was developed collaboratively by DRC and LDOE staff. LDOE staff had opportunities to review the guide, provide feedback, and give final approval. The elements of the table of contents are provided below:

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

Achievement Level Policy Definitions

Achievement level policy definitions for the LEAP 2025 Science tests are shown in Table 8.2. The titles and descriptions of the achievement levels were defined to be part of a cohesive assessment system, and the achievement levels indicate a student's ability to demonstrate proficiency on the LSSS defined for a specific course. The standard-setting section of the LEAP 2025 Biology 2018-2019 technical report contains comprehensive information

Table 8.2

Achievement Level Policy Definitions for LEAP 2025

Achievement Level	Achievement Level Policy Definition
Advanced	Students performing at this level have **exceeded** college and career readiness expectations and are well prepared for the next level of studies in this content area.
Mastery	Students performing at this level have **met** college and career readiness expectations and are prepared for the next level of studies in this content area.
Basic	Students performing at this level have **nearly met** college and career expectations and may need additional support to be fully prepared for the next level of studies in this content area.
Approaching Basic	Students performing at this level have **partially met** college and career readiness expectations and will need much support to be prepared for the next level of studies in this content area.
Unsatisfactory	Students performing at this level have **not yet met** the college and career readiness expectations and will need extensive support to be prepared for the next level of studies in this content area.

It should be noted that the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information address multiple best practices of the testing industry.

9. Reliability

Internal Consistency Reliability Estimation

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

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

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the i_{th} item or component, and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 Science tests is evidenced through a dimensionality analysis. Dimensionality analyses results are discussed in "Chapter 7. Data Analysis."
The reliability and classification accuracy reports in Appendix F: Reliability and Classification Accuracy provide coefficient alpha and IRT model-based or "marginal reliability" (Thissen, Chen, & Bock, 2003) for the total test.

While coefficient alpha values were between 0.84 and 0.89, the marginal alpha values were between 0.86 and 0.90 for the Science tests. Marginal reliability is described as "an average reliability over levels of θ or theta" (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 "posterior standard

deviations" (SEMs) in the IRTPRO output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31 θs. Marginal reliability is the average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals. The formula for marginal reliability is

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

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

Additional reliabilities were calculated on various demographics using the population of students. (Please refer to Table F.1.) Included with coefficient alpha in the tables are the number of students responding to the test, the mean score obtained by this group of students, and the standard deviation of the scores obtained for this group.

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

Classical Standard Error of Measurement

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

below his or her true score (the student's expected score). The SEM is calculated using both the standard deviation and the reliability of test scores, as follows:

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

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

It is important to note that the SEM provides an estimate of the average test score error for all students regardless of their individual proficiency levels. It is generally accepted that the SEM varies across the range of student proficiencies (Peterson, Kolen, & Hoover, 1989). For this reason, it is useful to report test-level SEM, and SEMs for 2022 Science between 3.26 and 3.89, as seen from Table B.4. In addition, SEMs by student group can be found in Appendix F.

Conditional Standard Error of Measurement and Cut Scores

It is important to note that the SEM index provides only an estimate of the average test score error for all students regardless of their individual levels of proficiency. By comparison, conditional standard error of measurement (CSEM) provides a reliability estimate at each score point on a test. Like the SEM, the CSEM reflects the amount of variance in a score resulting from random factors other than what the assessment is designed to measure, but it provides an estimate conditional on proficiency. The CSEM is usually smallest, and thus scores are most reliable, near the middle of the score distribution. Typically, achievement tests included relatively large numbers of moderately difficult items. Because these items are usually well matched to a majority of students' ability, they provide the most reliable estimates of ability. It is desirable, for an achievement test where students are classified into pass/fail categories, that the CSEM be lowest at the cut score for passing. The CSEMs at the four cut scores of each grade that define the performance levels are presented in Table 9.1. The standard-setting section of the LEAP 2025 Biology 2018-2019 technical report contains comprehensive information.

Table 9.1

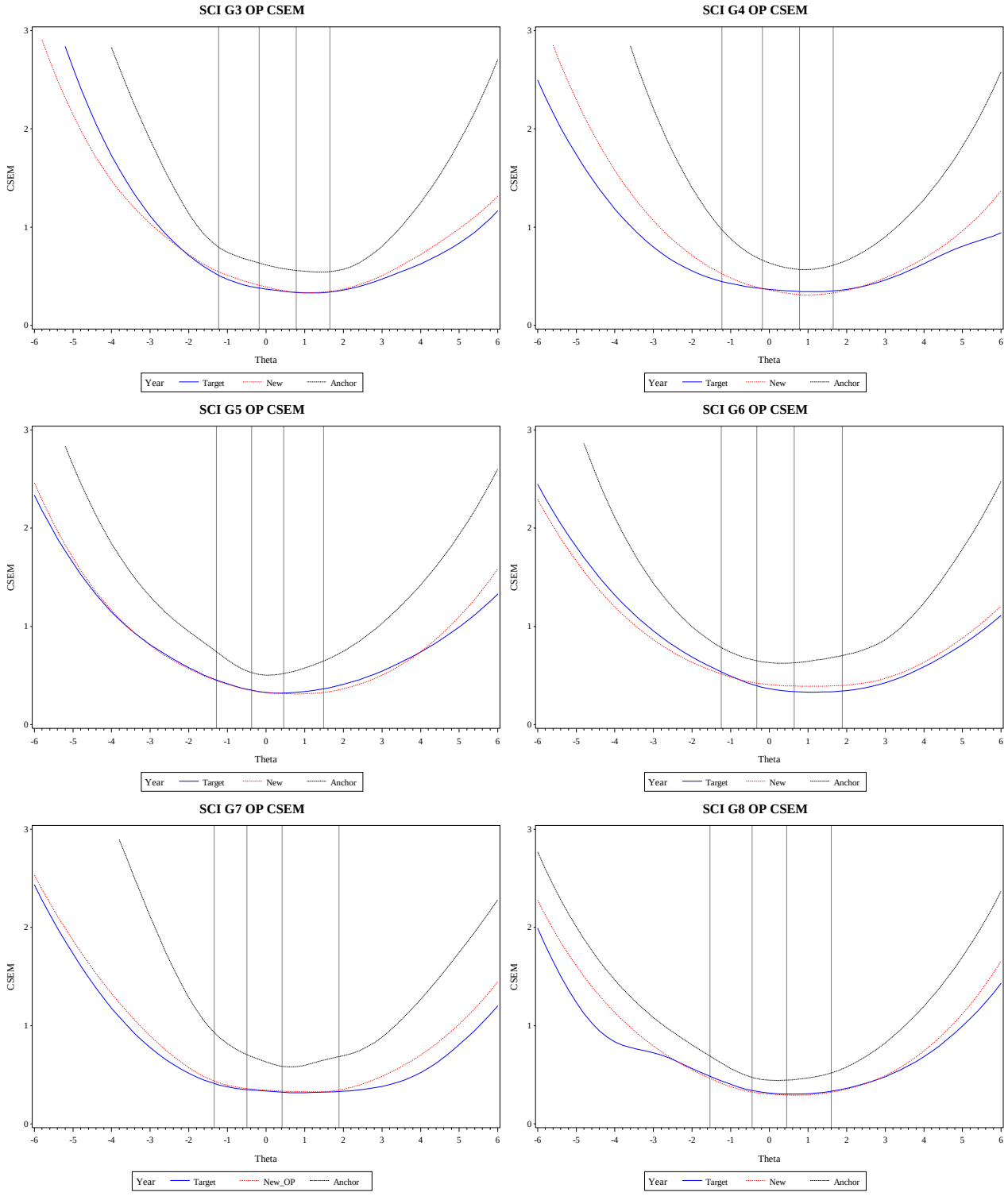
Conditional Standard Errors of Measurement at the Approaching Basic, Basic, Mastery, and Advanced Cut Scores: Operational 2022 LEAP Science

Grade	*Approaching Basic*		*Basic*		*Mastery*		*Advanced*	
	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
3	698	14	725	11	750	9	773	9
4	704	13	725	10	750	8	778	10
5	698	13	725	10	750	10	781	10
6	701	13	725	11	750	10	782	10
7	702	12	725	10	750	9	790	9
8	694	13	725	9	750	8	782	9

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

Plot 9.1

CSEM Curves: Science G3–8



Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their scale scores. It is important to know the reliability of student scores in any examination; assessing the reliability of the classification decisions based on these scores is of even greater importance. Classification decision reliability is estimated by the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive accuracy and consistency classification measures.

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

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

Accuracy and Consistency Indices. Three types of accuracy and consistency indices were generated: *overall*, *conditional-on-level*, and *cut point*, provided in Appendix F: Reliability and Classification Accuracy. The *overall accuracy* of performance-level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels. It is a proportion (or percentage) of correct classification across all the levels. While the overall accuracy indices were between 0.647 and 0.716, the overall consistency indices were 0.538 and 0.610 for the LEAP 2025 Science tests.

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

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

where P is the probability of consistent classification, and P_c is the probability of consistent classification by chance (Lee, Hanson, & Brennan, 2000). P is the sum of the diagonal elements, and P_c is the sum of the squared row totals. The PChance indices were between 0.223 and 0.242 for the 2022 Science tests.

Kappa is a measure of "how much agreement exists beyond chance alone" (Fleiss, 1973), which means that it provides the proportion of consistent classifications between two forms after removing the proportion of consistent classifications expected by chance alone. The Kappa indices were between 0.395 and 0.485 for the 2022 Science tests. *Consistency conditional-on-level* is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all the students classified into that level (marginal entry).

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

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points, the joint distribution of all the performance levels is collapsed into a dichotomized distribution around that specific cut point.

10. Validity

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

The 2021–2022 LEAP 2025 Science tests were designed and developed to provide fair and accurate scores that support appropriate, meaningful information for educational decisions. The knowledge, expertise, and professional judgment offered by Louisiana educators ultimately ensure that the content of the LEAP 2025 Science tests is an adequate and representative sample of appropriate content, and that the content is a legitimate basis upon which to derive valid conclusions about student achievement.

Chapters 2, 3, and 4 provide a general discussion of test book creation and the editing process, describing the selection of operational test items, the content distribution of embedded field test items, and the process to obtain approvals from the LDOE. The test design process and participation by Louisiana educators throughout the process—from item development, content review, and bias review to test selection—reinforce confidence in the content and design of LEAP 2025 to derive valid inferences about Louisiana student performance. The data review process and results are also discussed. Chapter 5 of the technical report describes the process, procedures, and policies that guide the administration of the LEAP 2025 assessments, including accommodations, test security, and detailed written procedures provided to test administrators and school personnel. Chapter 6 describes scoring processes and activities for the LEAP 2025 Science tests.

Chapter 7 describes classical data analysis and item response theoretic calibration, scaling, and equating methods, as well as processes and procedures to clean data to ensure replicable, iterative calibrations and scaling of the 2022 Science tests to derive

scale scores from students' raw scores. Some references to introductory and advanced discussions of IRT are provided. Chapter 7 also describes an analysis of DIF. Complete tables of gender and ethnicity DIF results for all 2022 Science operational items are presented in Appendix C. Chapter 8 of the technical report summarizes the test results, score distributions, score reports, and achievement level information. Chapter 9 addresses Cronbach's alpha and marginal alpha as measures of internal consistency and describes analysis procedures for classification consistency and classification accuracy. In addition, test validity is addressed in this chapter.

Evidence for Construct-Related Validity

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

Internal Structure of Reporting Categories

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

Content-Related Evidence

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

contained both content and statistical information about both operational and field-tested items.

Dimensionality and Principal Component Analysis

Appendix D: Dimensionality provides information about principal component analysis of the Science tests. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, in the case of scholastic achievement, a one-dimensional scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 2004; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983).

To check the unidimensionality of the spring 2022 assessment, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the Statistical Analysis System (SAS) program. The first and second principal component eigenvalues were compared *without rotation*. Table D.2 and Plot D.1 summarize the results of the first and second principal component eigenvalues of the assessments. A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the tables and figures, the first component is substantially larger than the second eigenvalue for the 2022 Science tests. Because the spring test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Evidence Based on Relations to Other Variables

Evidence based on *relations to other variables* is a typical utility of criterion-related validity evidence to measure concurrent or predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). Thus, external variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores on other tests) to evaluate the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades).

Most students who completed LEAP Science also took Mathematics and ELA tests. Thus, per grade, correlations between Science and Mathematics and between Science and EAL were calculated, overall, and by demographic groups. In general, high correlation coefficients were observed between LEAP Science and ELA tests and between LEAP Science and Mathematics tests regardless of grades. However, there were some variations in the sizes of association across the various demographic categories. For instance, regardless of grades, the group of EL students displayed lower correlation coefficients than other groups. The specific details can be obtained in a separate report called External Validity Study: SPR 2022 that was provided to the LDOE.

Item Development and Field-Test Analysis

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

Various field test forms were used to administer the test items. Once these items were scored, the LDOE and WestEd/Pearson conducted additional item analysis and content review. Any field test items that exhibited statistical results that suggested potential problems were carefully reviewed by both LDOE and WestEd/Pearson content specialists. A determination was then made as to whether an item should be accepted, rejected, and revised/refield-tested. Information on statistical analyses for field test items can be obtained in Chapter 6, Data Analysis.

In summary, additional, corroborating evidence consistent with the validity, reliability, and consistency of the LEAP 2025 Science assessment has been documented in the LEAP Grades 3–8 Science framework, test development plans, and the 2019 Science standard-setting technical report. Table 10.1 summarizes the sources of validity evidence and indicates where the evidence can be found in the technical report.

Table 10.1

Evidence of Validity and the Corresponding Technical Report Chapter

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A: Training Agendas

LEAP 2025 Grades 3–8 Item Outline Development Training Agenda
Item Development Cycle for 2019–2022 LEAP 2025 Assessment in Science

I. Item Development Process
   a. Overview
   b. Steps in process
II. Louisiana Student Standards for Science (LSSS)
   a. New science standards were approved in early March 2017.
      i. The LSSS represent the knowledge and skills needed for students to successfully transition to postsecondary education and the workplace. The standards call for students to:
         1. Apply content knowledge to real-world phenomena and to design solutions;
         2. Demonstrate the practices of scientists and engineers;
         3. Connect scientific learning to all disciplines of science; and
         4. Express ideas grounded in scientific evidence.
   b. The Louisiana Student Standards are not the NGSS!
III. Anatomy of the LSSS
   a. Descriptor
   b. Grade level
   c. Standard
   d. Domain
   e. Topic number
   f. Performance Expectation
      i. Science and Engineering Practices
      ii. Disciplinary Core Ideas
      iii. Crosscutting Concepts
IV. Outlines
   a. What outlines are
      i. Definition and purpose
      ii. Components
   b. What outlines are not
      i. Characteristics
      ii. Non-examples
   c. Outline assignments

i. Tasks
    Components
        a. Stimulus
            i. Purpose of graphics, data tables, and graphs
            ii. Reading level
        b. Item types (G3, 4 vs. 5–EOC/Bio)
        c. Bundling of PEs
ii. Item sets
    Components
        a. Stimulus
        b. Item types (G3, 4 vs. 5–EOC/Bio)
        c. Bundling of PEs
iii. Standalones
        a. Purpose
        b. Use of graphics, data tables, and graphs
        c. Item types
        d. Single PEs
iv. Template

V. Considerations

a. Tasks
    i. Needed number of items and ERs
    ii. Dimensionality
    iii. Number of items seen by students vs. number of items developed
    iv. Use of PEs
    v. Use of scaffolding within the task
b. Item sets
    i. Needed number of items and ERs
    ii. Dimensionality
    iii. Interchangeability
    iv. Use of PEs (mix and match)
    v. Number of items seen by students vs. number of items developed
c. Phenomena list (topics to avoid)
d. Bias and sensitivity
    i. Definitions
        1. Bias
        2. Sensitivity
        3. Stereotyping
        4. Fairness
    ii. Rationale for removing bias and sensitivity

1. Portrayal of groups within Louisiana's diverse population
2. Protection of privacy and avoidance of offensive content

iii. Potential sources of bias
1. Ethnicity
2. Culture
3. Religion
4. Disability
5. Gender/age stereotypes
6. Geography
7. Socioeconomic status
8. Controversial issues or contexts
9. English language proficiency

iv. Strategies to avoid bias
1. Include non-DCI-related information needed to understand stimulus/make stimulus accessible to students regardless of background.
2. Use familiar language and contexts to avoid accessibility bias.
3. Avoid issues and themes that demean, offend, or inaccurately portray any religion, ethnicity, culture, gender, social group, or disability.
4. Avoid topics that will offend the privacy of values and beliefs of students, parents, or the public.

LEAP 2025 Grades 3–8 Item Writer Training Agenda

Item Development Cycle for 2019–2022 LEAP 2025 Assessment in Science

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

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

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

IV. More Acronyms
   a. SEP key
      i. 1. Q/P = Asking Questions and Defining Problems

ii. 2. MOD = Developing and Using Models
iii. 3. INV = Planning and Carrying Out Investigations
iv. 4. DATA = Analyzing and Interpreting Data
v. 5. MCT = Using Mathematics and Computational Thinking
vi. 6. E/S = Constructing Explanations and Designing Solutions
vii. 7. ARG = Engaging in Argument from Evidence
viii. 8. INFO = Obtaining, Evaluating, and Communicating Information

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

c. "Acronyms Cheat Sheet"

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

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

VII. Phenomena: Keystone of 3-D Assessments
a. Phenomena: Observable events that students can use the three dimensions to explain or make sense of
i. Links to phenomena websites are available in the "LEAP Phenomena and Context" document.

VIII. 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

IX. Contexts and Stimuli
a. Stimuli contain contexts in which phenomena are presented.
b. Contexts and stimuli should be unique and novel.
    i. Non-textbook
    ii. Think outside the box
c. Stimuli must be student friendly and grade appropriate.
    i. Engaging to students
    ii. Free of bias and sensitivity issues
        1. Definitions
            a. Bias
            b. Sensitivity
            c. Stereotyping
            d. Fairness
        2. Rationale for Removing Bias and Sensitivity
            a. Portrayal of groups within Louisiana's diverse population
            b. Protection of privacy and avoidance of offensive content
        3. Potential Sources of Bias
            a. Ethnicity
            b. Culture
            c. Religion
            d. Disability
            e. Gender/age stereotypes
            f. Geography
            g. Socioeconomic status
            h. Controversial issues or contexts
            i. English language proficiency
        4. Strategies to Avoid Bias
            a. Include non-DCI related information needed to understand stimulus/make stimulus accessible to students regardless of background.
            b. Use familiar language and contexts to avoid accessibility bias.
            c. Avoid issues and themes that demean, offend, or inaccurately portray any religion, ethnicity, culture, gender, social group, or disability.
            d. Avoid topics that will offend the privacy of values and beliefs of students, parents, or the public.
d. Phenomena, contexts, and stimuli need to be the right grain size.
e. Goldilocks—provide only the information that is needed

X. Phenomena and PE Bundles

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

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

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

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

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

XV. 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

XVI. 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

XVII. Item Types:

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

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

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

XVIII. Development Process Overview

XIX. 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.

XX. 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.

XXI. Cognitive Complexity*

a. Need for a range of items of varied cognitive complexity
b. Existing models of cognitive complexity (e.g., DOK)
c. Development of a model to address three-dimensional items of LEAP assessment*
d. (*As the TAGS-M model was in development during the early portion of the 2018–2019 development cycle, item writers used their understanding of cognitive complexity to develop two- and three-dimensional items aligned to the PEs of the LSSS, targeting a broad range of cognitive complexities. These items were then coded by WestEd staff after the TAGS-M model was complete.)

XXII. Sourcing
    a. Sources are required for specific information, such as species, planets, stars, elements, or designs of existing solutions.
        i. Sources are not needed for commonly known facts.
            1. Formula for photosynthesis
            2. The definition of speed
        ii. If in doubt, source!
        iii. Use reputable sources
        iv. See “Sources” for more information.
XXIII. Graphics
    a. Graphics are used to convey ideas, data, and/or concepts in a simplified visual form.
        i. Graphics are essential components of science and include:
            1. Tables, diagrams, models, graphs, images
        ii. All graphics must be introduced appropriately with an introductory statement. Some graphics require only a brief introduction; some require a bit more, e.g.:
            1. The students’ results are shown in the table below.
            2. Students made a scale drawing of their prototype. The scale drawing is shown below.
        iii. Be aware that some graphics may be changed during production to control for colorblindness.
        iv. See “General Guidelines for Graphics” document for more information.
        v. Style guide
XXIV. Development Process Overview
XXV. Information Security
    a. Do NOT email!
    b. We will send/receive items and assignments using a secure system.
    c. General questions about processes OK

LEAP 2025 Grades 3–8 Editor Training Agenda

Item Development Cycle for the 2018–2019 LEAP 2025 Science Assessment

I. Item Set/Task/Standalone Item Overview
   a. Criteria for review
II. Item Development Process
   a. One round of items slated for development in 2018–2019
   b. All batches will go through four rounds of LDOE review at different stages of development before committee:
      i. Outline review (item descriptions; graphic roughs)
      ii. Item development
         1. R1 (fully fleshed-out items; functional TE items; graphics; sources)
         2. R2 (implementation of LDOE feedback; rewrites possible; revisions expected)
         3. R3 (final look before committee review—no editing, all comments are for committee review)
   c. Committee review
III. Process Overview for Intake/E1
IV. Intake/E1 Rules for Returning Item Sets/Tasks/Standalone Item Submissions to Writers
V. Feedback to Writers
VI. Process Overview for Intake/E2
VII. Intake/E1 Rules for Returning Item Sets/Tasks/Standalone Item Submissions to E1 Writer
VIII. Use of the Style Guides
   a. Social Studies/Science Content Style Guide
   b. TEI Guide
   c. Graphics Style Guide

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

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

a. Alignment is the key element of content review.
    i. Is the item providing an appropriate measure of the PE and its related dimensions?
    ii. Item content alignment is the degree to which an item measures the intended PE and its related dimensions.
    iii. Put another way: An item is determined to be aligned if the item allows the student to provide evidence of his or her understanding of the specified PE and its related dimensions.
b. Additional considerations include:
    i. Scoring/key accuracy
    ii. Scientific accuracy

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

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

ii.) Avoid regional language or words with different meanings in different groups.
iii.) Avoid idioms and figurative language.

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

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

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

XII. Content Review Decisions
a. Yes ("Accept")
i. Item is acceptable as is
ii. Aligned
iii. Scientifically accurate
iv. Scoring information correct
v. Free of bias concerns
b. No ("Accept with Edits" or "Reject")
i. Due to content concerns
ii. Metadata alignment with explanation
iii. Science accuracy concern with explanation
iv. Due to bias concerns
v. With explanation

c. Reject when:
    i. Complete alignment mismatch
    ii. Unfixable context flaws
d. Revise when:
    i. Fixes can be made
    ii. Item Alignment Information

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

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

LEAP 2025 Grades 3–8 Data Review Training Agenda

I. What is a Data Review?
   a. Statistical Definition: Classical Test Theory
      1. P-value
      2. Point-Biserial
      3. Option/Distribution Analysis
      4. Differential Item Function (DIF)
      5. Flagging Value

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

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

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

II. Judgement Task in ABBI
   a. Accept
   b. Accept with Edits
   c. Reject

Appendix B: Test Summary

Science G3–8

Contents
Table B.1 Percentage of Points by Reporting Category (includes Task Items): Spring 2022 Operational SC G3–8
Table B.2 Standard Coverage: Spring 2022 Operational SC G3–8
Table B.3 Item Type Summary: Spring 2022 Operational SC G3–8
Table B.4 Raw Score Summary: Spring 2022 Operational SC G3–8
Table B.5 Raw Score Summary by Reporting Category: Spring 2022 Operational SC G3–8
Table B.6 Scale Score and Raw Score Summary: Spring 2022 Operational SC G3–8

- Because the spring 2022 test was administered during the 2022 COVID-19 pandemic, greatcaution should be applied when any statistical inference is drawn.

Table B.1

Percentage of Points by Reporting Category (includes Task Items): Spring 2022 Operational SC G3–8

Reporting Category	G3	G4	G5	G6	G7	G8
N/A	6.0%	9.6%	-	1.7%	8.3%	-
1 Investigate	30.0%	28.8%	19.7%	15.3%	15.0%	24.6%
2 Evaluate	48.0%	15.4%	31.1%	27.1%	31.7%	31.1%
3 Reason Scientifically	16.0%	46.2%	49.2%	55.9%	45.0%	44.3%

* N/A indicates no reporting category.

Table B.2
Standard Coverage: Spring 2022 Operational SC G3–8

Grade 3

Reporting Categories		No. of Items					% of Test
		TPI	TPD	MS	MC	CR	
		N	N	N	N	N	
N/A	3-ESS2-2	1			1		5.56
	Sub-Total	**1**			**1**		**5.56**
1 Investigate	3-PS2-1		2		2		11.11
	3-PS2-2		1		1		5.56
	3-PS2-3				2		5.56
	3-PS2-4	1			2		8.33
	Sub-Total	**1**	**3**		**7**		**30.56**
2 Evaluate	3-ESS2-1	1		1	1		8.33
	3-ESS3-1		1				2.78
	3-LS2-1		1		2		8.33
	3-LS3-1				1		2.78
	3-LS4-1	1			2		8.33
	3-LS4-3				1	1	5.56
	3-LS4-4		1		2	1	11.11
	Sub-Total	**2**	**3**	**1**	**9**	**2**	**47.22**
3 Reason Scientifically	3-LS1-1	1			1		5.56
	3-LS3-2				1	1	5.56
	3-LS4-2				2		5.56
	Sub-Total	**1**			**4**	**1**	**16.67**
Total		**5**	**6**	**1**	**21**	**3**	**100.00**

***** N/A indicates no reporting category.

Grade 4

Reporting Categories		No. of Items					% of Test
		TPI	TPD	MS	MC	CR	
		N	N	N	N	N	
N/A	4-ESS2-1				1	1	5.56
	4-ESS3-1	1					2.78
	Sub-Total	**1**			**1**	**1**	**8.33**
1 Investigate	4-ESS2-1	1	1		2		11.11
	4-ESS2-3				1	1	5.56
	4-PS3-2			1			2.78
	4-PS3-3			1	2	1	11.11
	Sub-Total	**1**	**1**	**2**	**5**	**2**	**30.56**
2 Evaluate	4-ESS2-2	1			2		8.33
	4-LS1-1		1		2		8.33
	Sub-Total	**1**	**1**		**4**		**16.67**
3 Reason Scientifically	4-ESS1-1		1				2.78
	4-ESS3-2	1	1	1			8.33
	4-LS1-2				1		2.78
	4-PS3-1		3	1	1		13.89
	4-PS3-4	1		2			8.33
	4-PS4-1		1		1		5.56
	4-PS4-2				1		2.78
	Sub-Total	**2**	**6**	**4**	**4**		**44.44**
Total		**5**	**8**	**6**	**14**	**3**	**100.00**

* N/A indicates no reporting category.

Grade 5

Reporting Categories		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
1 Investigate	5-LS1-1			1	1			1	8.11
	5-PS1-3					1			2.70
	5-PS1-4		1	1		2			10.81
	Sub-Total		**1**	**2**	**1**	**3**		**1**	**21.62**
2 Evaluate	5-ESS1-1			2	1	1			10.81
	5-ESS1-2		1	1					5.41
	5-ESS2-2			2	1				8.11
	5-PS1-2	1		1					5.41
	5-PS2-1		1			1			5.41
	Sub-Total	**1**	**2**	**6**	**2**	**2**			**35.14**
3 Reason Scientifically	5-ESS2-1			1		2			8.11
	5-ESS3-1	1		2			1		10.81
	5-LS2-1		1	1				1	8.11
	5-PS1-1	1		1					5.41
	5-PS3-1	1				2		1	10.81
	Sub-Total	**3**	**1**	**5**		**4**	**1**	**2**	**43.24**
Total		**4**	**4**	**13**	**3**	**9**	**1**	**3**	**100.00**

Grade 6

Reporting Categories		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
N/A	6-MS-ESS1-2			1					2.78
	Sub-Total			**1**					**2.78**
1 Investigate	6-MS-LS1-1				1			1	5.56
	6-MS-PS2-2			1		1			5.56
	6-MS-PS2-3	1							2.78
	6-MS-PS2-5					1			2.78
	Sub-Total	**1**		**1**	**1**	**2**		**1**	**16.67**
2 Evaluate	6-MS-ESS1-3			1					2.78
	6-MS-ESS3-4				1				2.78
	6-MS-LS2-1			1					2.78
	6-MS-PS2-4		1	1		1			8.33
	6-MS-PS3-1		1			2			8.33
	6-MS-PS4-1			2					5.56
	Sub-Total		**2**	**5**	**1**	**3**			**30.56**
3 Reason Scientifically	6-MS-ESS1-1		1	1				1	8.33
	6-MS-ESS1-2					1			2.78
	6-MS-LS1-2		1			1			5.56
	6-MS-LS2-2		1		2		1		11.11
	6-MS-LS2-3					1			2.78
	6-MS-PS1-1			1					2.78
	6-MS-PS2-1		1			1			5.56
	6-MS-PS3-2			2				1	8.33
	6-MS-PS4-2			1					2.78
	Sub-Total		**4**	**5**	**2**	**4**	**1**	**2**	**50.00**
Total		**1**	**6**	**12**	**4**	**9**	**1**	**3**	**100.00**

* N/A indicates no reporting category.

Grade 7

Reporting Categories		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
N/A	7-MS-LS1-3				1				2.78
	7-MS-LS4-5	2							5.56
	Sub-Total	**2**			**1**				**8.33**
1 Investigate	7-MS-ESS2-5			1					2.78
	7-MS-ESS3-5	1			2				8.33
	7-MS-PS3-4			2					5.56
	Sub-Total	**1**		**3**	**2**				**16.67**
2 Evaluate	7-MS-LS1-3					1	1		5.56
	7-MS-LS2-4			1		2		1	11.11
	7-MS-PS1-2				1	2			8.33
	Sub-Total			**1**	**1**	**5**	**1**	**1**	**25.00**
3 Reason Scientifically	7-MS-ESS2-4			1	1			1	8.33
	7-MS-ESS2-6			2		1			8.33
	7-MS-LS1-6				1				2.78
	7-MS-LS1-7			2					5.56
	7-MS-LS2-5			1					2.78
	7-MS-LS3-2		1						2.78
	7-MS-LS4-4			2		1		1	11.11
	7-MS-PS1-4			1		1			5.56
	7-MS-PS1-5			1					2.78
	Sub-Total		**1**	**10**	**2**	**3**		**2**	**50.00**
Total		**3**	**1**	**14**	**6**	**8**	**1**	**3**	**100.00**

* N/A indicates no reporting category.

Grade 8

Reporting Categories		No. of Items							% of Test
		TPI	TPD	TEI	MS	MC	ER	CR	
		N	N	N	N	N	N	N	
1 Investigate	8-MS-ESS3-2					1			2.70
	8-MS-ESS3-3			2		1			8.11
	8-MS-LS1-5			1					2.70
	8-MS-PS1-3			2		2			10.81
	8-MS-PS1-6				1				2.70
	8-MS-PS3-3					1			2.70
	Sub-Total			**5**	**1**	**5**			**29.73**
2 Evaluate	8-MS-ESS2-3	1							2.70
	8-MS-LS1-4				1				2.70
	8-MS-LS4-1			1		1			5.41
	8-MS-LS4-3		1						2.70
	8-MS-LS4-6	1		2		1			10.81
	8-MS-PS3-5		1			1		1	8.11
	Sub-Total	**2**	**2**	**3**	**1**	**3**		**1**	**32.43**
3 Reason Scientifically	8-MS-ESS1-4	1							2.70
	8-MS-ESS2-1			1					2.70
	8-MS-ESS2-2			1		1		1	8.11
	8-MS-ESS3-1			1		1	1		8.11
	8-MS-LS3-1					1			2.70
	8-MS-LS4-2			1				1	5.41
	8-MS-PS1-1					3			8.11
	Sub-Total	**1**		**4**		**6**	**1**	**2**	**37.84**
Total		**3**	**2**	**12**	**2**	**14**	**1**	**3**	**100.00**

Table B.3

Item Type Summary: Spring 2022 Operational SC G3–8

Grade	MC	MS	TEI	CR	ER	TPD	TPI
3	21	1	-	3	-	6	5
4	14	6	-	3	-	8	5
5	9	3	13	3	1	4	4
6	9	4	12	3	1	6	1
7	8	6	14	3	1	1	3
8	14	2	12	3	1	2	3

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

Table B.4

Raw Score Summary: Spring 2022 Operational SC G3–8

Grade	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
3	≥49,320	18.79	8.43	0	46	0.39	0.40	0.85	3.23
4	≥48,910	18.88	9.10	0	50	0.37	0.42	0.87	3.31
5	≥48,900	22.85	10.93	0	58	0.43	0.46	0.89	3.66
6	≥49,300	19.98	9.69	1	56	0.35	0.39	0.84	3.86
7	≥50,990	21.52	10.05	0	58	0.35	0.41	0.85	3.88
8	≥50,720	25.52	11.38	1	60	0.43	0.45	0.89	3.76

* Reliability is Cronbach's alpha.

Table B.5

Raw Score Summary by Reporting Category: Spring 2022 Operational SC G3–8

Admin	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
3	Investigate	5.97	3.21	0	15	0.39	0.40	0.65	1.90
	Evaluate	8.74	4.24	0	24	0.38	0.40	0.74	2.16
	Reason Scientifically	2.83	1.56	0	8	0.39	0.35	0.41	1.20
4	Investigate	4.71	3.00	0	15	0.33	0.42	0.68	1.70
	Evaluate	3.04	1.80	0	8	0.38	0.38	0.47	1.31
	Reason Scientifically	9.23	4.62	0	24	0.38	0.42	0.74	2.36
5	Investigate	4.93	2.53	0	12	0.44	0.43	0.61	1.58
	Evaluate	7.81	4.31	0	19	0.42	0.48	0.76	2.11
	Reason Scientifically	10.10	5.21	0	29	0.44	0.45	0.76	2.55
6	Investigate	3.08	1.96	0	9	0.36	0.45	0.58	1.27
	Evaluate	6.14	3.24	0	16	0.38	0.41	0.64	1.94
	Reason Scientifically	10.39	5.44	0	32	0.33	0.37	0.68	3.08
7	Investigate	2.65	1.79	0	9	0.27	0.38	0.44	1.34
	Evaluate	6.78	3.56	0	19	0.36	0.41	0.54	2.41
	Reason Scientifically	10.49	5.51	0	27	0.39	0.44	0.78	2.58
8	Investigate	6.82	3.33	0	15	0.39	0.40	0.65	1.90
	Evaluate	8.60	3.69	0	19	0.38	0.40	0.74	2.16
	Reason Scientifically	10.10	5.54	0	27	0.39	0.35	0.41	1.20

Table B.6.1

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 3

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥49,320	100.00	725.78	30.74	18.79	8.43	-
Female	≥24,090	48.85	725.56	29.86	18.66	8.19	**0.03**
Male	≥25,230	51.15	725.99	31.56	18.91	8.66	-
African American	≥20,380	41.33	714.12	27.37	15.50	6.79	**0.85**
American Indian or Alaska Native	≥270	0.55	729.89	27.84	19.66	8.14	**0.28**
Asian	≥830	1.69	743.86	30.28	24.01	9.08	**-0.22**
Hispanic/Latino	≥4,970	10.09	719.30	29.83	17.01	7.79	**0.60**
Multi-Racial	≥1,850	3.77	730.63	28.96	20.01	8.25	**0.24**
Native Hawaiian or Other PacificIslander	≥30	0.08	728.38	28.05	19.41	8.02	**0.31**
White	≥20,930	42.45	737.48	29.45	22.08	8.63	-
Economically Diaantaged: No	≥13,960	28.31	742.12	29.35	23.49	8.74	**-0.83**
Economically Disadvantaged: Yes	≥35,200	71.37	719.36	28.82	16.94	7.54	-
EL: No	≥46,480	94.23	726.89	30.66	19.09	8.47	**-0.63**
EL: Yes	≥2,840	5.77	707.59	25.89	13.86	5.94	-
Regular Education	≥43,010	87.21	727.71	30.42	19.30	8.45	**-0.48**
Special Education	≥6,300	12.79	712.65	29.63	15.30	7.45	-
Section 504: No	≥45,780	92.81	726.31	30.82	18.94	8.47	**-0.26**
Section 504: Yes	≥3,540	7.19	718.95	28.78	16.79	7.59	-
Migrant: No	≥49,230	99.81	725.81	30.74	18.79	8.43	**-0.42**
Migrant: Yes	≥90	0.19	712.34	29.67	15.25	7.51	-
Homeless: No	≥47,910	97.13	726.13	30.72	18.88	8.45	**-0.39**
Homeless: Yes	≥1,410	2.87	713.93	28.89	15.58	7.23	-
Military Affiliation: No	≥48,370	98.07	725.49	30.69	18.71	8.41	**0.51**
Military Affiliation: Yes	≥950	1.93	740.31	29.47	22.97	8.69	-
Foster Care: No	≥49,160	99.68	725.80	30.75	18.79	8.44	**-0.21**
Foster Care: Yes	≥150	0.32	720.08	26.75	17.00	6.97	-

Table B.6.2

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 4

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥48,910	100.00	733.86	29.73	18.88	9.10	-
Female	≥23,900	48.87	732.45	28.79	18.38	8.76	**0.11**
Male	≥25,010	51.13	735.21	30.54	19.36	9.40	-
African American	≥20,500	41.91	721.94	26.20	15.14	7.17	**0.90**
American Indian or Alaska Native	≥260	0.54	739.32	27.02	20.45	8.80	**0.23**
Asian	≥800	1.64	755.06	30.03	25.84	10.03	**-0.35**
Hispanic/Latino	≥5,080	10.40	728.96	29.14	17.38	8.68	**0.57**
Multi-Racial	≥1,680	3.43	739.01	28.26	20.35	8.98	**0.24**
Native Hawaiian or Other PacificIslander	≥30	0.08	737.44	23.57	19.46	7.97	**0.34**
White	≥20,510	41.95	745.67	28.13	22.58	9.24	-
Economically Disadvantaged: No	≥13,980	28.59	749.79	28.00	23.97	9.34	**-0.83**
Economically Disadvantaged: Yes	≥34,630	70.80	727.55	27.94	16.86	8.16	-
EL: No	≥46,440	94.95	734.86	29.62	19.18	9.13	**-0.66**
EL: Yes	≥2,460	5.05	715.10	25.16	13.27	6.45	-
Regular Education	≥42,940	87.79	736.10	29.22	19.53	9.09	**-0.60**
Special Education	≥5,970	12.21	717.73	28.39	14.21	7.70	-
Section 504: No	≥44,770	91.53	734.44	29.74	19.06	9.14	**-0.24**
Section 504: Yes	≥4,140	8.47	727.59	28.87	16.92	8.50	-
Migrant: No	≥48,850	99.87	733.87	29.73	18.88	9.10	**-0.21**
Migrant: Yes	≥60	0.13	727.65	30.37	16.98	9.17	-
Homeless: No	≥47,600	97.32	734.17	29.72	18.97	9.12	**-0.38**
Homeless: Yes	≥1,310	2.68	722.76	28.01	15.53	7.86	-
Military Affiliation: No	≥47,960	98.05	733.55	29.68	18.78	9.07	**0.57**
Military Affiliation: Yes	≥950	1.95	749.57	27.95	23.92	9.31	-
Foster Care: No	≥48,780	99.73	733.88	29.73	18.89	9.11	**-0.22**
Foster Care: Yes	≥130	0.27	727.57	29.09	16.89	8.36	-

Table B.6.3

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 5

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥48,900	100.00	727.90	36.92	22.85	10.93	-
Female	≥23,820	48.72	728.75	35.54	22.99	10.62	**-0.02**
Male	≥25,070	51.28	727.08	38.17	22.72	11.22	-
African American	≥20,660	42.25	713.20	33.51	18.44	9.20	**0.87**
American Indian or Alaska Native	≥250	0.52	729.42	33.39	23.05	10.08	**0.38**
Asian	≥740	1.53	755.48	35.65	31.53	11.46	**-0.41**
Hispanic/Latino	≥4,790	9.81	722.64	36.20	21.27	10.40	**0.55**
Multi-Racial	≥1,640	3.36	734.62	35.48	24.82	10.79	**0.21**
Native Hawaiian or Other PacificIslander	≥40	0.09	727.84	39.60	22.89	12.04	**0.39**
White	≥20,740	42.41	742.21	34.24	27.13	10.76	-
Economically Disadvantaged: No	≥14,580	29.83	747.31	34.12	28.78	10.86	**-0.82**
Economically Disadvantaged: Yes	≥34,000	69.53	719.73	34.92	20.35	9.95	-
EL: No	≥46,820	95.75	729.12	36.73	23.20	10.93	**-0.77**
EL: Yes	≥2,070	4.25	700.38	29.83	14.93	7.41	-
Regular Education	≥42,890	87.70	731.48	35.73	23.83	10.79	**-0.75**
Special Education	≥6,010	12.30	702.31	35.15	15.82	9.27	-
Section 504: No	≥44,160	90.30	728.99	36.95	23.18	10.98	**-0.31**
Section 504: Yes	≥4,740	9.70	717.66	35.04	19.76	9.97	-
Migrant: No	≥48,840	99.88	727.92	36.92	22.85	10.93	**-0.45**
Migrant: Yes	≥50	0.12	710.78	34.37	17.92	9.12	-
Homeless: No	≥47,680	97.51	728.24	36.92	22.95	10.95	**-0.38**
Homeless: Yes	≥1,210	2.49	714.48	34.34	18.84	9.54	-
Military Affiliation: No	≥48,000	98.17	727.54	36.88	22.74	10.91	**0.55**
Military Affiliation: Yes	≥890	1.83	747.07	34.30	28.76	10.85	-
Foster Care: No	≥48,790	99.77	727.93	36.92	22.86	10.94	**-0.42**
Foster Care: Yes	≥110	0.23	712.45	33.60	18.28	9.00	-

Table B.6.4

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 6

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥49,300	100.00	722.14	33.65	19.98	9.69	-
Female	≥23,950	48.59	722.35	32.27	19.94	9.32	**0.01**
Male	≥25,350	51.41	721.95	34.90	20.01	10.03	-
African American	≥20,700	41.99	709.84	29.36	16.36	7.82	**0.81**
American Indian or Alaska Native	≥280	0.57	725.45	30.08	20.68	8.84	**0.30**
Asian	≥790	1.61	747.47	37.86	27.72	11.60	**-0.41**
Hispanic/Latino	≥5,070	10.28	715.52	33.11	18.16	9.12	**0.56**
Multi-Racial	≥1,620	3.29	726.63	31.64	21.15	9.31	**0.25**
Native Hawaiian or Other PacificIslander	≥20	0.06	731.72	33.89	22.69	10.22	**0.09**
White	≥20,780	42.15	734.66	32.68	23.62	9.94	-
Economically Disadvantaged: No	≥14,460	29.34	740.18	32.25	25.30	10.01	**-0.83**
Economically Disadvantaged: Yes	≥34,560	70.09	714.74	31.29	17.78	8.64	-
EL: No	≥47,230	95.80	723.36	33.41	20.30	9.69	**-0.81**
EL: Yes	≥2,070	4.20	694.43	26.37	12.58	6.06	-
Regular Education	≥43,740	88.71	724.97	33.13	20.74	9.68	**-0.72**
Special Education	≥5,560	11.29	699.93	29.12	13.95	7.38	-
Section 504: No	≥44,100	89.44	723.35	33.76	20.33	9.77	**-0.35**
Section 504: Yes	≥5,200	10.56	711.93	30.88	16.99	8.44	-
Migrant: No	≥49,240	99.87	722.15	33.65	19.98	9.69	**-0.01**
Migrant: Yes	≥60	0.13	721.73	32.55	19.88	9.43	-
Homeless: No	≥48,050	97.46	722.44	33.66	20.06	9.71	**-0.34**
Homeless: Yes	≥1,250	2.54	710.92	31.13	16.77	8.39	-
Military Affiliation: No	≥48,460	98.28	721.86	33.59	19.89	9.66	**0.50**
Military Affiliation: Yes	≥840	1.72	738.42	33.09	24.76	10.20	-
Foster Care: No	≥49,170	99.73	722.17	33.65	19.98	9.69	**-0.32**
Foster Care: Yes	≥130	0.27	710.95	31.59	16.87	8.44	-

Table B.6.5

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 7

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥50,990	100.00	730.41	32.50	21.52	10.05	-
Female	≥25,080	49.18	731.49	30.60	21.75	9.58	**-0.04**
Male	≥25,910	50.82	729.36	34.21	21.30	10.49	-
African American	≥21,880	42.92	718.73	28.20	17.81	8.23	**0.81**
American Indian or Alaska Native	≥290	0.58	733.66	28.78	22.42	9.01	**0.28**
Asian	≥740	1.47	757.69	35.02	30.35	11.34	**-0.50**
Hispanic/Latino	≥4,840	9.50	724.24	33.23	19.77	9.86	**0.54**
Multi-Racial	≥1,690	3.32	735.13	31.89	22.94	10.03	**0.23**
Native Hawaiian or Other PacificIslander	≥40	0.08	740.74	30.24	24.79	9.79	**0.05**
White	≥21,460	42.09	742.34	31.54	25.27	10.18	-
Economically Disadvantaged: No	≥15,140	29.69	746.59	31.76	26.66	10.33	**-0.77**
Economically Disadvantaged: Yes	≥35,550	69.73	723.67	30.29	19.37	9.10	-
EL: No	≥49,150	96.41	731.51	32.21	21.84	10.03	**-0.89**
EL: Yes	≥1,830	3.59	700.92	25.77	12.99	6.33	-
Regular Education	≥45,420	89.09	733.40	31.67	22.39	9.95	**-0.82**
Special Education	≥5,560	10.91	706.04	28.75	14.40	7.76	-
Section 504: No	≥45,600	89.43	731.65	32.51	21.91	10.10	**-0.37**
Section 504: Yes	≥5,380	10.57	719.89	30.44	18.24	8.99	-
Migrant: No	≥50,930	99.88	730.42	32.50	21.52	10.05	**-0.35**
Migrant: Yes	≥60	0.12	719.20	28.75	18.03	8.21	-
Homeless: No	≥49,820	97.71	730.65	32.49	21.59	10.06	**-0.32**
Homeless: Yes	≥1,160	2.29	720.06	31.43	18.39	9.19	-
Military Affiliation: No	≥50,120	98.31	730.10	32.45	21.42	10.03	**0.58**
Military Affiliation: Yes	≥860	1.69	748.43	30.10	27.28	9.92	-
Foster Care: No	≥50,860	99.76	730.44	32.51	21.53	10.05	**-0.38**
Foster Care: Yes	≥120	0.24	718.31	28.15	17.72	8.21	-

Table B.6.6

Scale Score and Raw Score Summary: Spring 2022 Operational Science: Grade 8

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥50,720	100.00	730.81	32.05	25.52	11.38	-
Female	≥25,010	49.32	730.96	30.87	25.50	11.03	**0.00**
Male	≥25,700	50.68	730.66	33.17	25.55	11.71	-
African American	≥21,550	42.49	717.26	28.07	20.61	9.42	**0.96**
American Indian or Alaska Native	≥280	0.55	736.83	30.56	27.45	11.19	**0.27**
Asian	≥810	1.60	756.61	34.26	34.95	12.30	**-0.42**
Hispanic/Latino	≥4,830	9.52	724.77	32.78	23.56	11.23	**0.62**
Multi-Racial	≥1,570	3.11	736.11	30.74	27.38	11.10	**0.27**
Native Hawaiian or Other PacificIslander	≥40	0.08	742.77	35.33	30.07	12.65	**0.02**
White	≥21,610	42.62	744.20	29.31	30.34	10.90	-
Economically Disadvantaged: No	≥16,050	31.65	747.42	29.42	31.53	10.97	**-0.83**
Economically Disadvantaged: Yes	≥34,350	67.74	723.18	30.24	22.76	10.45	-
EL: No	≥48,900	96.42	731.95	31.70	25.90	11.33	**-0.95**
EL: Yes	≥1,810	3.58	700.02	25.40	15.29	7.30	-
Regular Education	≥45,500	89.72	733.60	31.34	26.47	11.27	**-0.84**
Special Education	≥5,210	10.28	706.43	27.59	17.21	8.59	-
Section 504: No	≥45,550	89.81	732.00	32.12	25.96	11.43	**-0.38**
Section 504: Yes	≥5,160	10.19	720.36	29.47	21.71	10.14	-
Migrant: No	≥50,660	99.88	730.82	32.06	25.53	11.38	**-0.18**
Migrant: Yes	≥50	0.12	725.63	29.67	23.53	10.63	-
Homeless: No	≥49,600	97.80	731.04	32.03	25.60	11.38	**-0.32**
Homeless: Yes	≥1,110	2.20	720.75	31.45	21.98	10.82	-
Military Affiliation: No	≥49,810	98.20	730.48	31.98	25.40	11.35	**0.59**
Military Affiliation: Yes	≥910	1.80	748.95	30.71	32.12	11.31	-
Foster Care: No	≥50,580	99.74	730.85	32.05	25.54	11.38	**-0.48**
Foster Care: Yes	≥130	0.26	715.29	29.37	20.04	9.87	-

Appendix C: Item Analysis Summary Report

Summary Statistics Reports

Contents
Table C.1 P-Value Summary by Grade: Spring 2022 Operational SC G3–8 Plot C.1 P-Value Summary by Grade: Spring 2022 Operational SC G3–8
Table C.2 Item-Total Correlation Summary by Grade: Spring 2022 Operational SC G3–8 Plot C.2 Item-Total Correlation Summary by Grade: Spring 2022 Operational SC G3–8
Table C.3 Corrected Point-Biserial Correlation Summary by Grade: Spring 2022 Operational SC G3–8 Plot C.3 Corrected Point-Biserial Correlation Summary by Grade: Spring 2022 Operational SC G3–8
Table C.4 Item-Total Correlation Summary by Reporting Category and Grade: Spring 2022 Operational SC G3–8
Table C.5.1 IRT-A Parameter Summary by Reporting Category: SC G3-8 Table C.5.2 IRT-B Parameter Summary by Reporting Category: SC G3-8 Table C.5.3 IRT Parameter Summary: Spring 2022 Operational SC G3–8 Plot C.5.1 IRT Parameter Summary: Spring 2022 Operational SC G3–8: A-Parameter Plot C.5.2 IRT Parameter Summary: Spring 2022 Operational SC G3–8: B-Parameter Plot C.5.3 IRT Parameter Summary: Spring 2022 Operational SC G3–8: C-Parameter
Table C.6 Statistically Flagged Items by Item Type: Spring 2022 Operational SC G3–8

- Because the spring 2022 test was administered during the 2022 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table C.1.1

P-Value Summary by Grade: Spring 2022 Operational SC G3–8

Grade	No. of Items	0 le p lt 0.2	0.2 le p lt 0.4	0.4 le p lt 0.6	0.6 le p lt 0.8	0.8 le p le 1.0
3	36	2	17	15	2	0
4	36	3	20	12	1	0
5	37	2	15	12	7	1
6*	37	6	17	13	1	0
7	36	8	13	14	1	0
8*	38	2	12	19	5	0

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

Plot C.1.1

P-Value Summary by Grade: Spring 2022 Operational SC G3–8

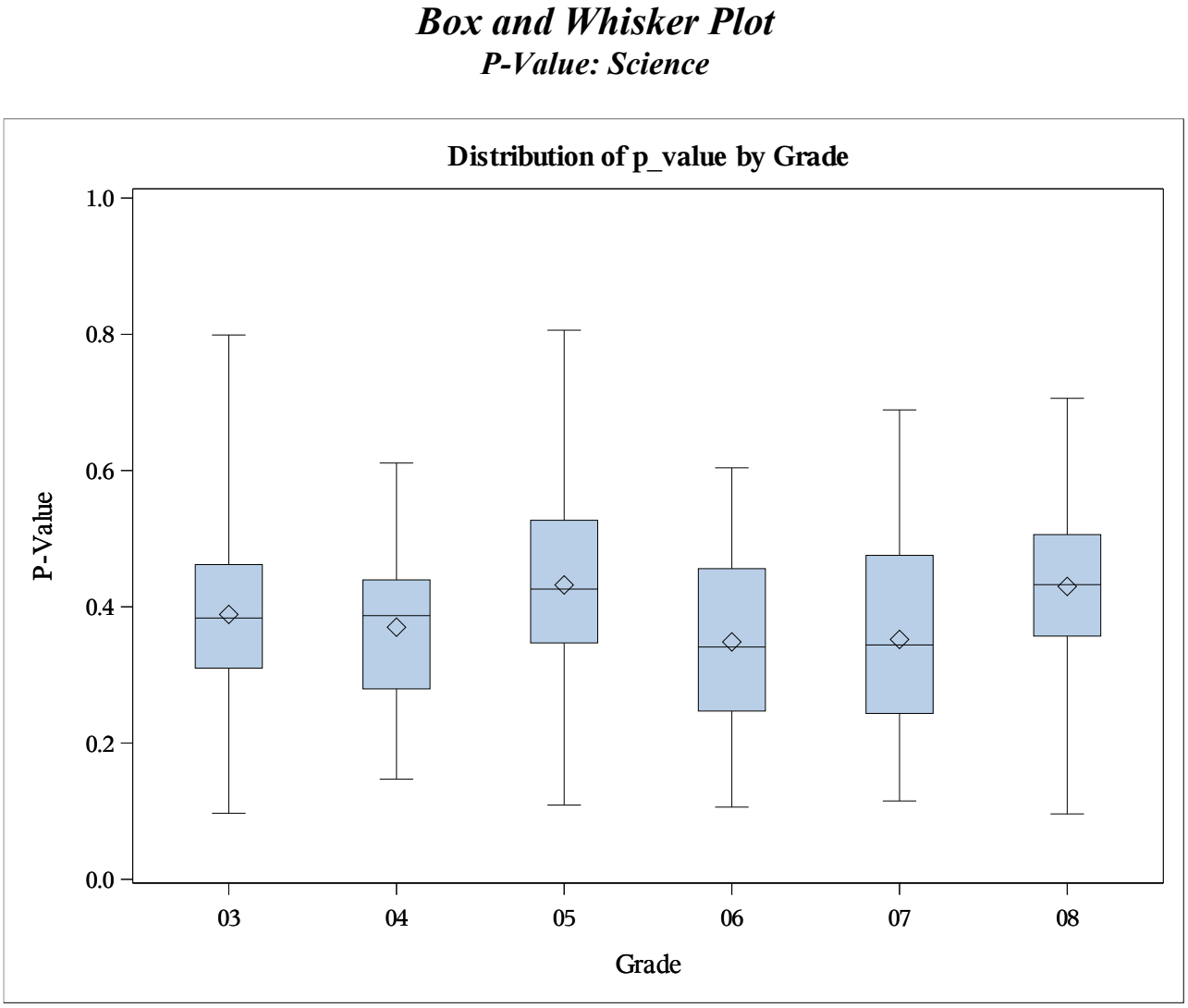


Table C.1.2

P-Value Summary by Item Type: Spring 2022 Operational SC G3–8

Grade 3

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.097	0.097	0.133	0.298	0.298
MC	21	0.254	0.333	0.429	0.476	0.799
MS	1	0.264	0.264	0.264	0.264	0.264
TPD	6	0.206	0.293	0.378	0.386	0.479
TPI	5	0.342	0.357	0.357	0.421	0.672

Grade 4

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.147	0.147	0.197	0.249	0.249
MC	14	0.239	0.379	0.434	0.486	0.611
MS	6	0.191	0.268	0.291	0.388	0.460
TPD	8	0.297	0.338	0.372	0.418	0.555
TPI	5	0.271	0.394	0.395	0.398	0.411

Grade 5

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.180	0.180	0.308	0.311	0.311
ER	1	0.109	0.109	0.109	0.109	0.109
MC	9	0.353	0.607	0.619	0.637	0.806
MS	3	0.266	0.266	0.363	0.452	0.452
TEI	13	0.205	0.355	0.426	0.507	0.692
TPD	4	0.262	0.280	0.324	0.394	0.438
TPI	4	0.377	0.417	0.463	0.481	0.492

Grade 6

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.111	0.111	0.119	0.168	0.168
ER	2	0.115	0.115	0.193	0.271	0.271
MC	9	0.247	0.298	0.398	0.490	0.604
MS	4	0.219	0.277	0.397	0.482	0.505
TEI	12	0.106	0.284	0.355	0.435	0.574
TPD	6	0.156	0.288	0.389	0.427	0.558
TPI	1	0.486	0.486	0.486	0.486	0.486

Grade 7

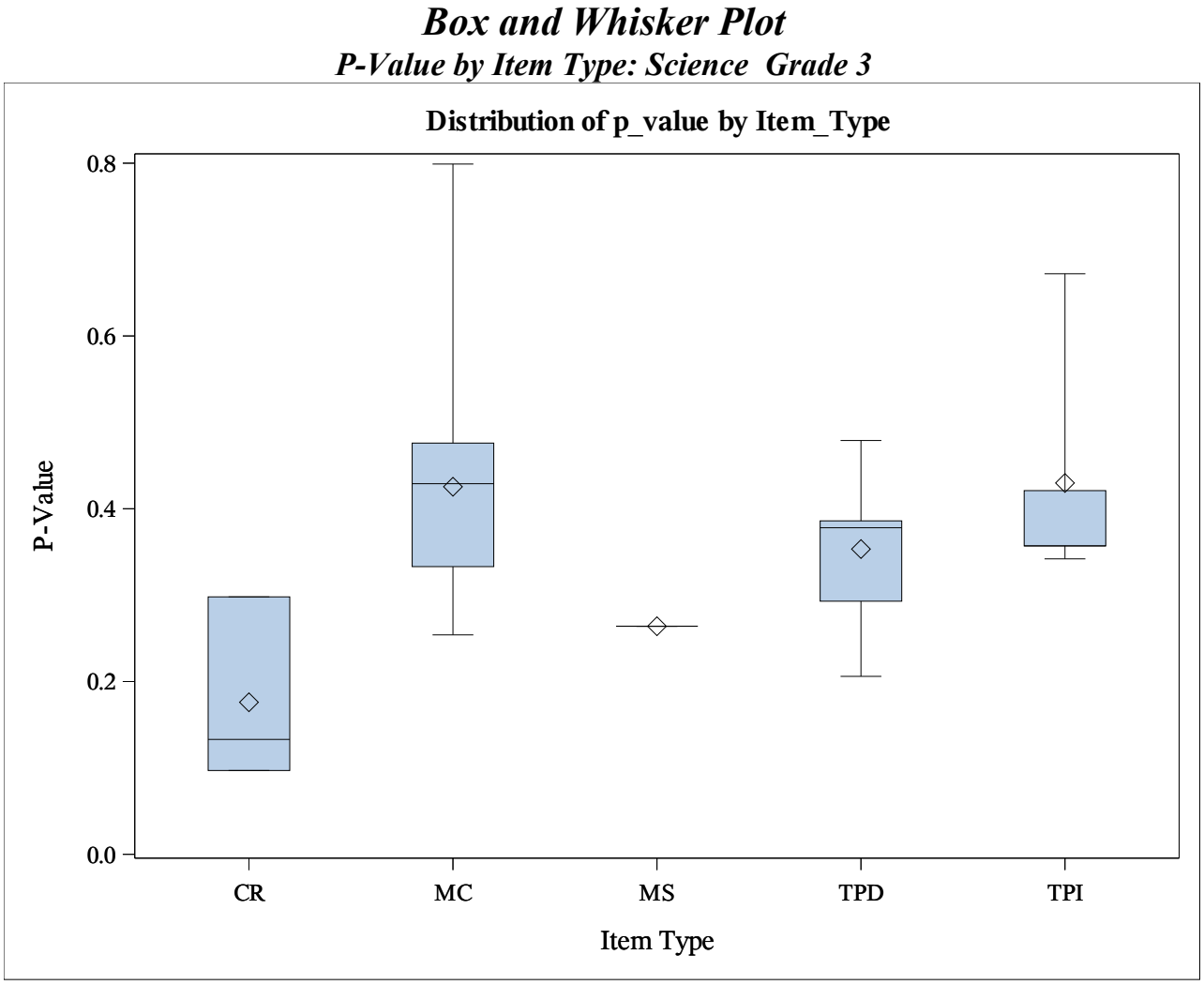
Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.174	0.174	0.300	0.350	0.350
ER	1	0.338	0.338	0.338	0.338	0.338
MC	8	0.237	0.283	0.366	0.484	0.556
MS	6	0.149	0.159	0.220	0.314	0.461
TEI	14	0.115	0.192	0.400	0.533	0.689
TPD	1	0.484	0.484	0.484	0.484	0.484
TPI	3	0.260	0.260	0.434	0.465	0.465

Grade 8

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.154	0.154	0.212	0.254	0.254
ER	2	0.318	0.318	0.334	0.350	0.350
MC	14	0.340	0.373	0.431	0.511	0.703
MS	2	0.361	0.361	0.405	0.448	0.448
TEI	12	0.096	0.428	0.485	0.569	0.706
TPD	2	0.406	0.406	0.436	0.466	0.466
TPI	3	0.270	0.270	0.451	0.536	0.536

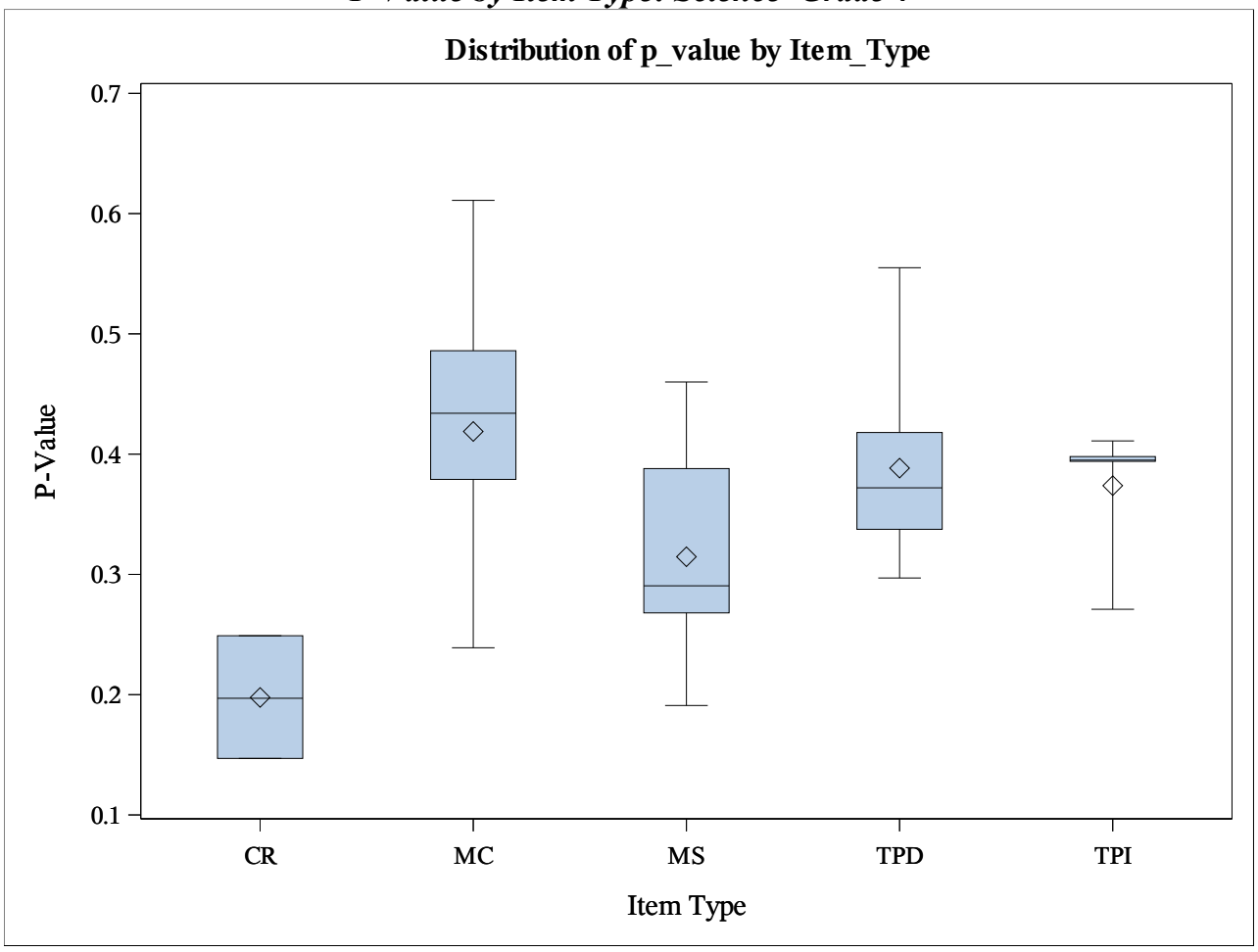
Plot C.1.2

P-Value Summary by Item Type: Spring 2022 Operational SC G3–8



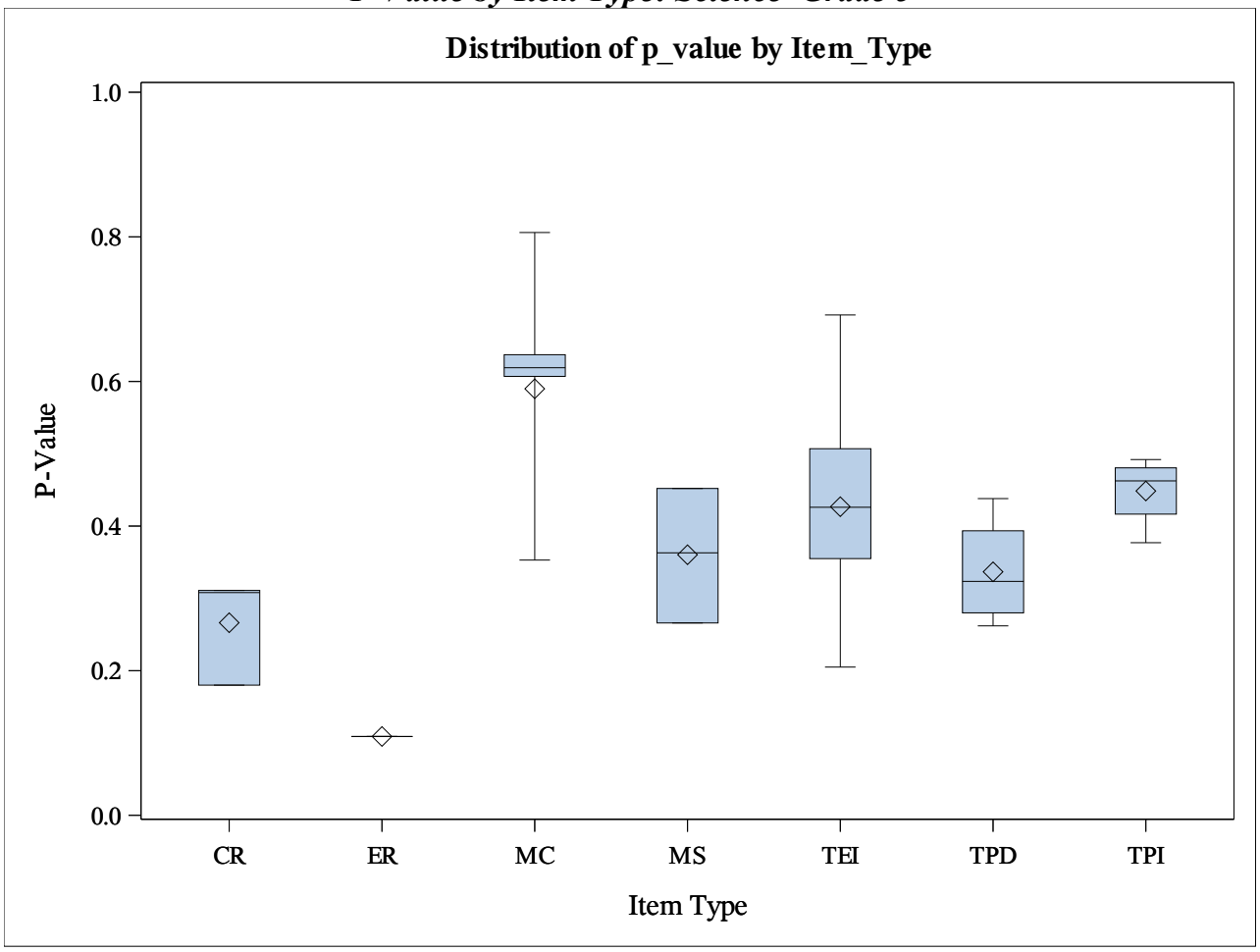
Box and Whisker Plot

P-Value by Item Type: Science Grade 4



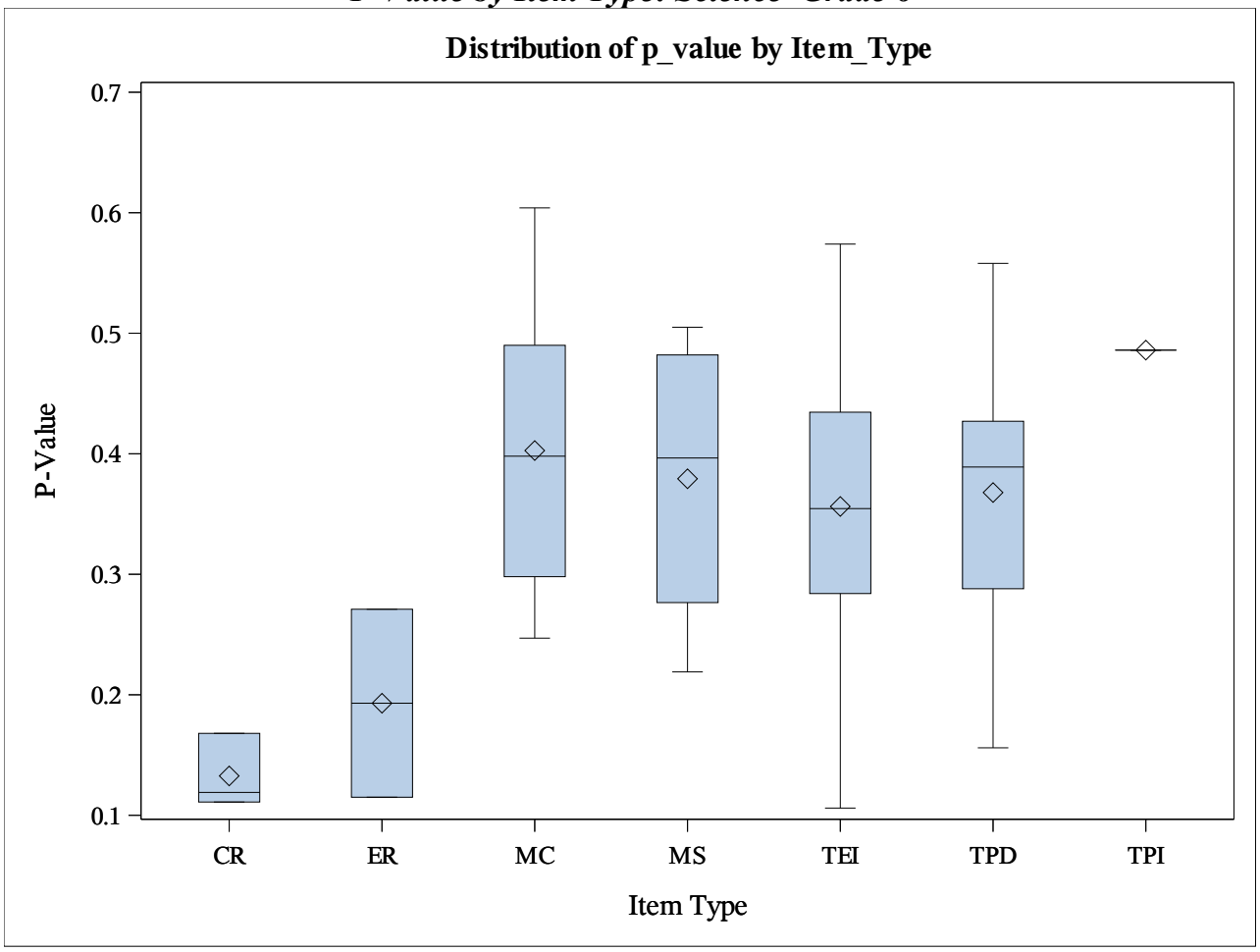
Box and Whisker Plot

P-Value by Item Type: Science Grade 5



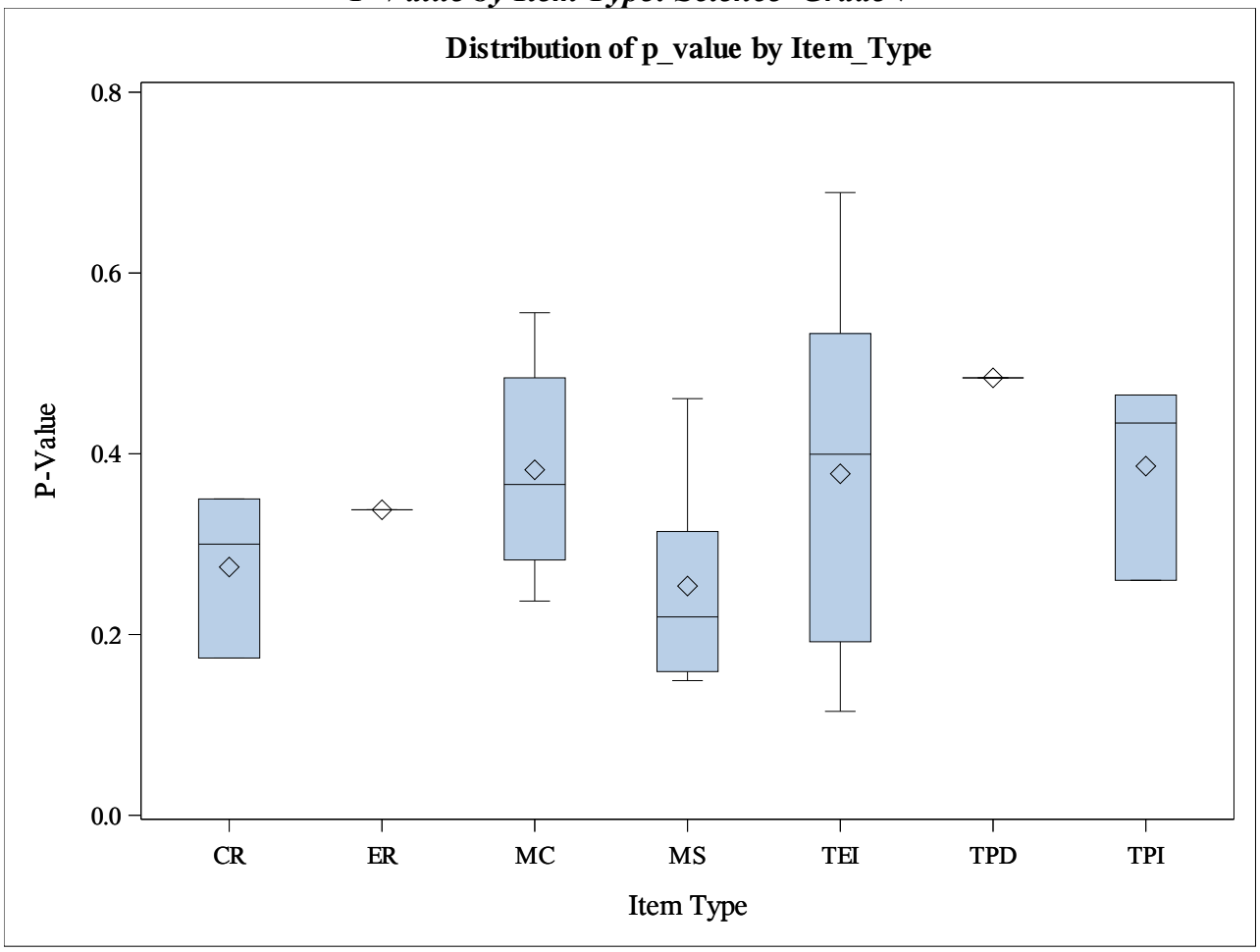
Box and Whisker Plot

P-Value by Item Type: Science Grade 6



Box and Whisker Plot

P-Value by Item Type: Science Grade 7



Box and Whisker Plot

P-Value by Item Type: Science Grade 8

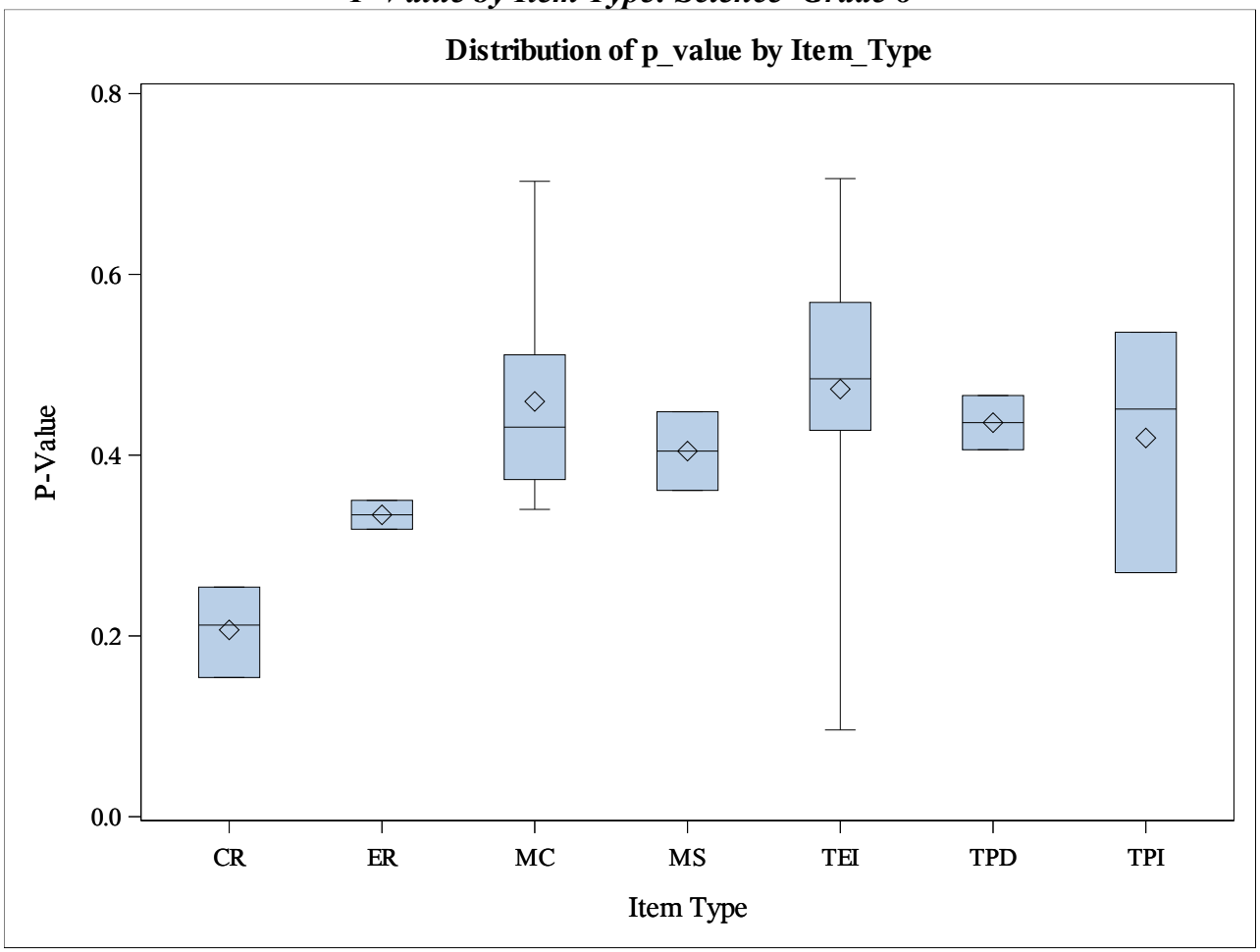


Table C.2.1

Item-Total Correlation by Grade: Spring 2022 Operational SC G3–8

Grade	No. of Items	r lt 0	0.0 le r lt 0.2	0.2 le r lt 0.3	0.3 le r lt 0.4	0.4 le r lt 0.5
3	36	0	1	4	14	12
4	36	0	1	5	8	14
5	37	0	0	2	7	18
6*	37	0	3	3	11	13
7	36	0	2	5	11	7
8*	38	0	1	3	9	12

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

Plot C.2.1

Item-Total Correlation by Grade: Spring 2022 Operational SC G3–8

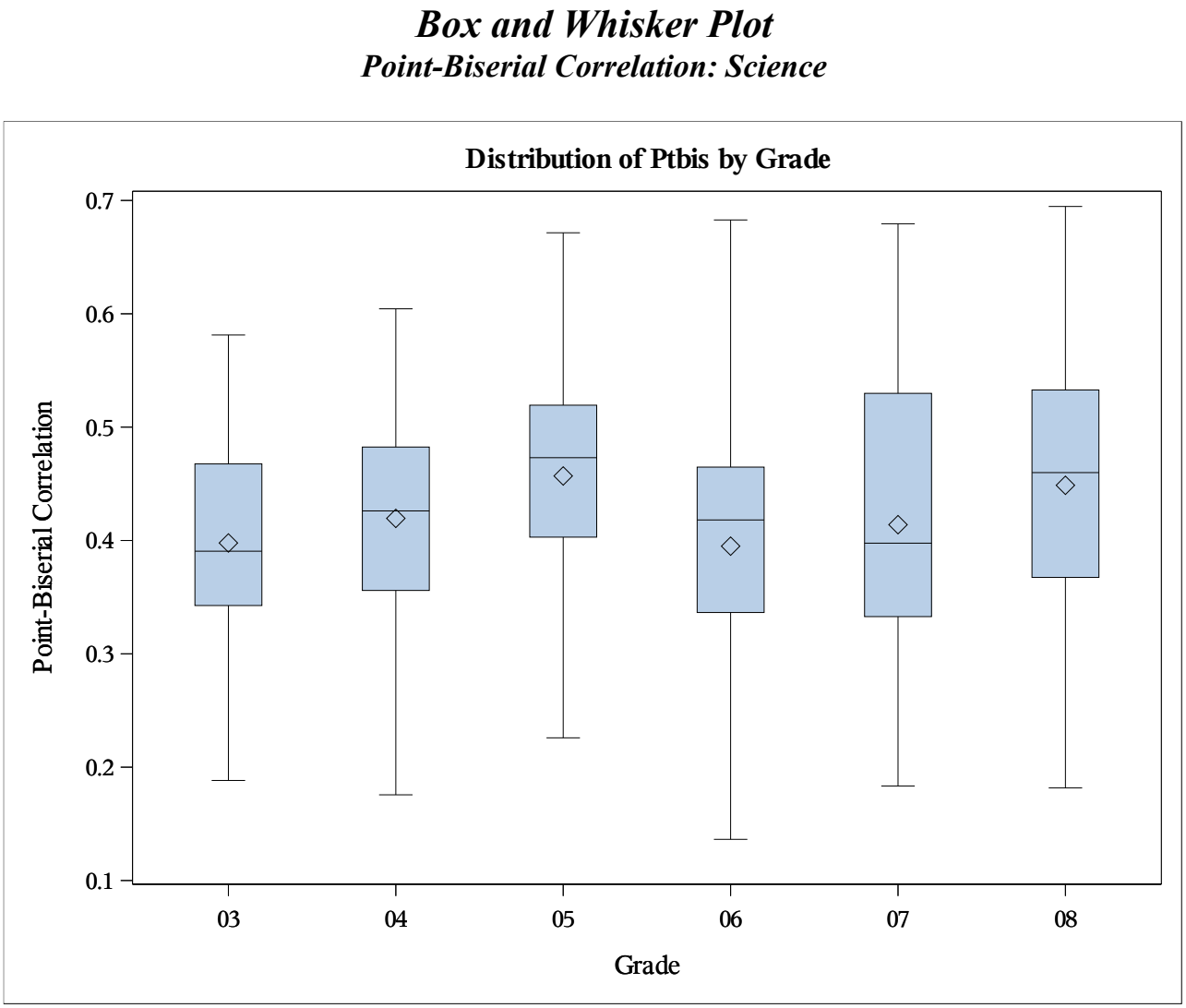


Table C.2.2

Item-Total Correlation Summary by Item Type: Spring 2022 Operational SC G3–8

Grade 3

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.339	0.339	0.347	0.513	0.513
MC	21	0.188	0.302	0.370	0.427	0.466
MS	1	0.331	0.331	0.331	0.331	0.331
TPD	6	0.353	0.424	0.494	0.567	0.579
TPI	5	0.371	0.469	0.495	0.544	0.581

Grade 4

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.461	0.461	0.479	0.583	0.583
MC	14	0.176	0.295	0.356	0.416	0.463
MS	6	0.236	0.334	0.400	0.470	0.515
TPD	8	0.408	0.434	0.473	0.531	0.597
TPI	5	0.378	0.438	0.501	0.532	0.604

Grade 5

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.460	0.460	0.479	0.550	0.550
ER	1	0.671	0.671	0.671	0.671	0.671
MC	9	0.237	0.355	0.393	0.412	0.560
MS	3	0.317	0.317	0.488	0.547	0.547
TEI	13	0.380	0.415	0.487	0.520	0.644
TPD	4	0.226	0.298	0.427	0.519	0.556
TPI	4	0.473	0.473	0.480	0.489	0.492

Grade 6

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.380	0.380	0.421	0.452	0.452
ER	2	0.375	0.375	0.529	0.683	0.683
MC	9	0.144	0.327	0.367	0.424	0.456
MS	4	0.250	0.335	0.454	0.503	0.519
TEI	12	0.136	0.263	0.427	0.488	0.559
TPD	6	0.312	0.336	0.356	0.466	0.515
TPI	1	0.576	0.576	0.576	0.576	0.576

Grade 7

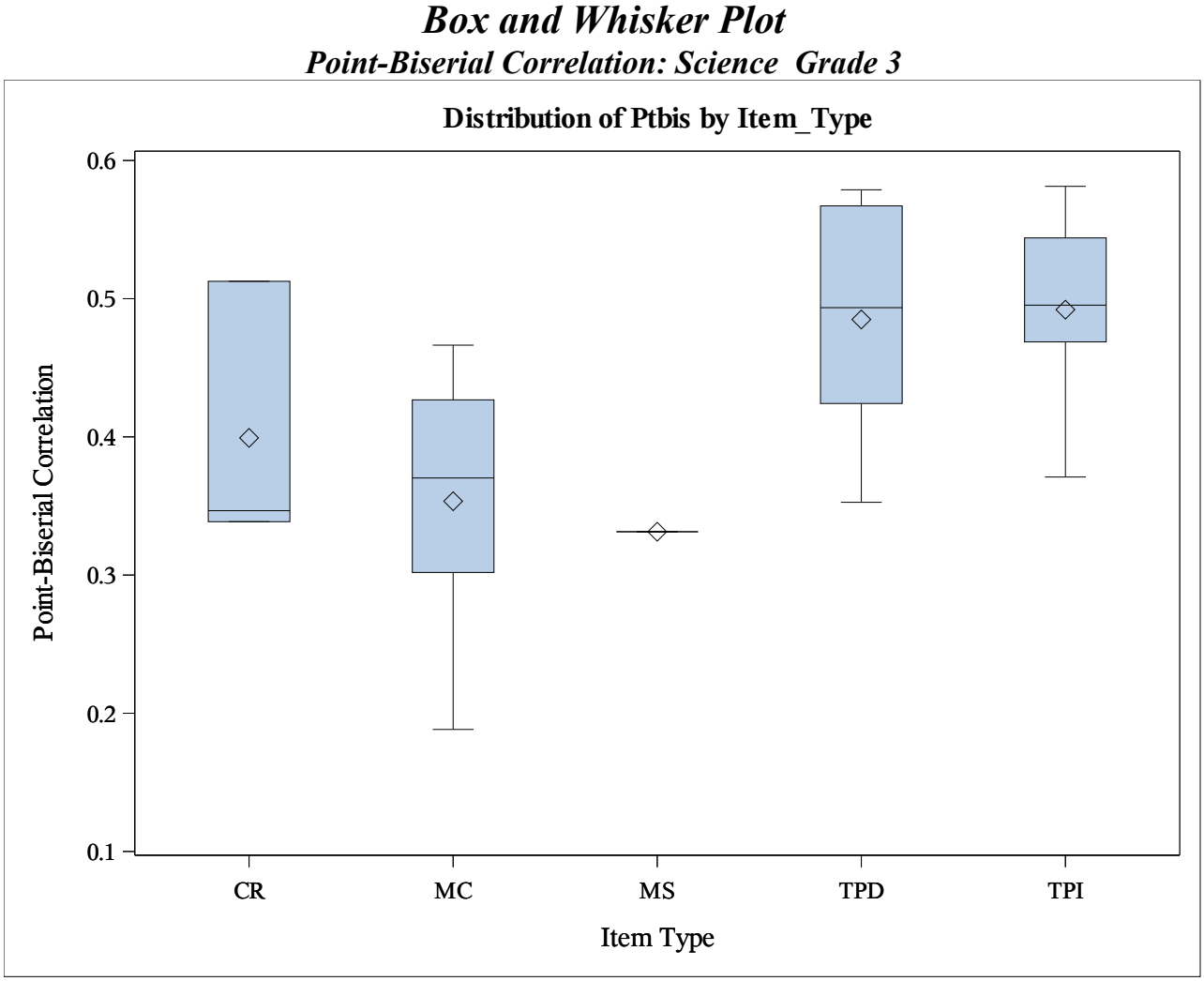
Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.512	0.512	0.527	0.546	0.546
ER	1	0.679	0.679	0.679	0.679	0.679
MC	8	0.183	0.202	0.333	0.373	0.533
MS	6	0.287	0.368	0.390	0.551	0.618
TEI	14	0.208	0.338	0.425	0.490	0.617
TPD	1	0.550	0.550	0.550	0.550	0.550
TPI	3	0.251	0.251	0.362	0.445	0.445

Grade 8

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.434	0.434	0.544	0.611	0.611
ER	2	0.658	0.658	0.676	0.695	0.695
MC	14	0.257	0.310	0.357	0.396	0.535
MS	2	0.487	0.487	0.510	0.533	0.533
TEI	12	0.182	0.450	0.475	0.551	0.633
TPD	2	0.408	0.408	0.455	0.501	0.501
TPI	3	0.411	0.411	0.471	0.505	0.505

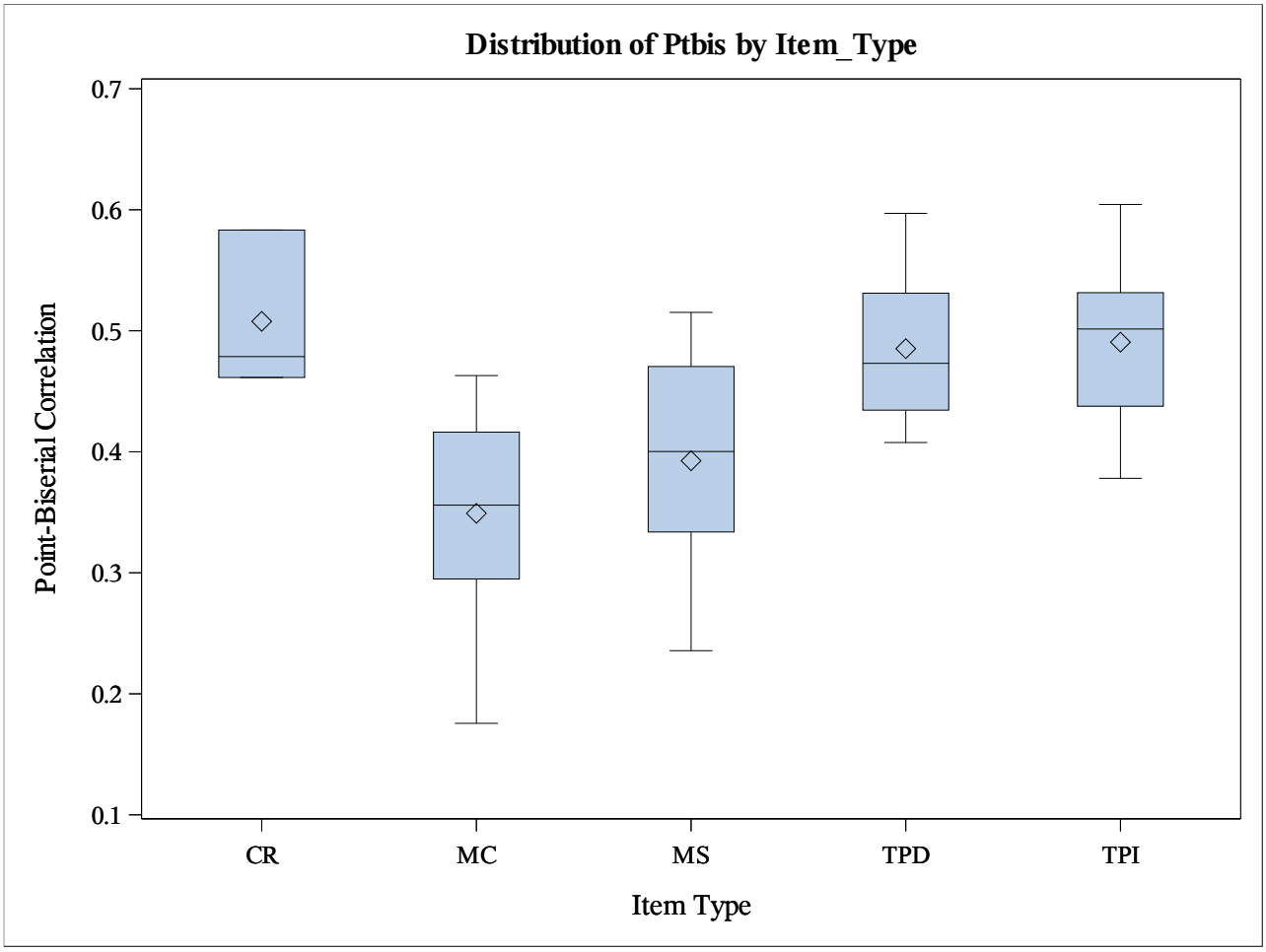
Plot C.2.2

Item-Total Correlation Summary by Item Type: Spring 2022 Operational SC G3–8



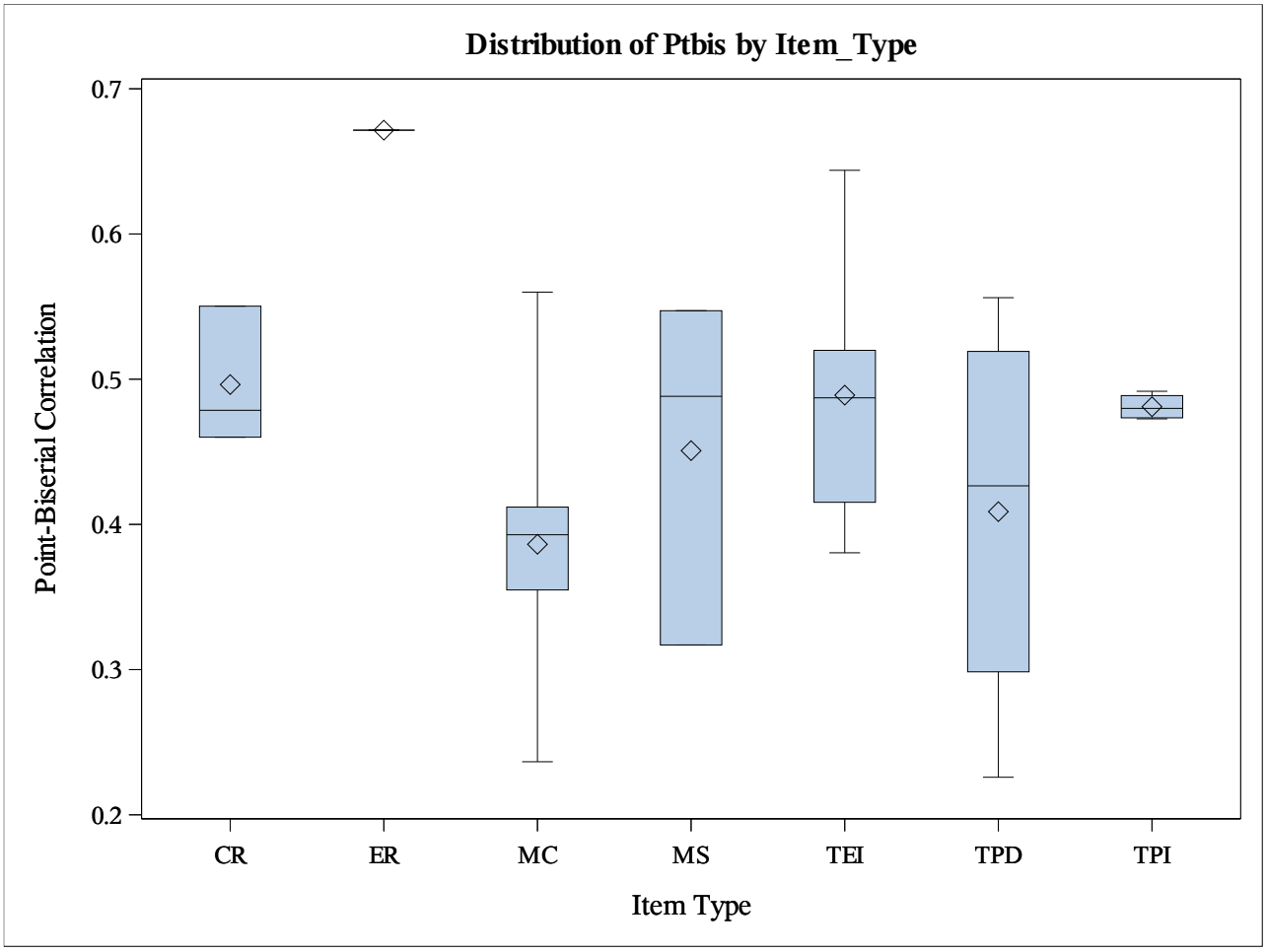
Box and Whisker Plot

Point-Biserial Correlation: Science Grade 4



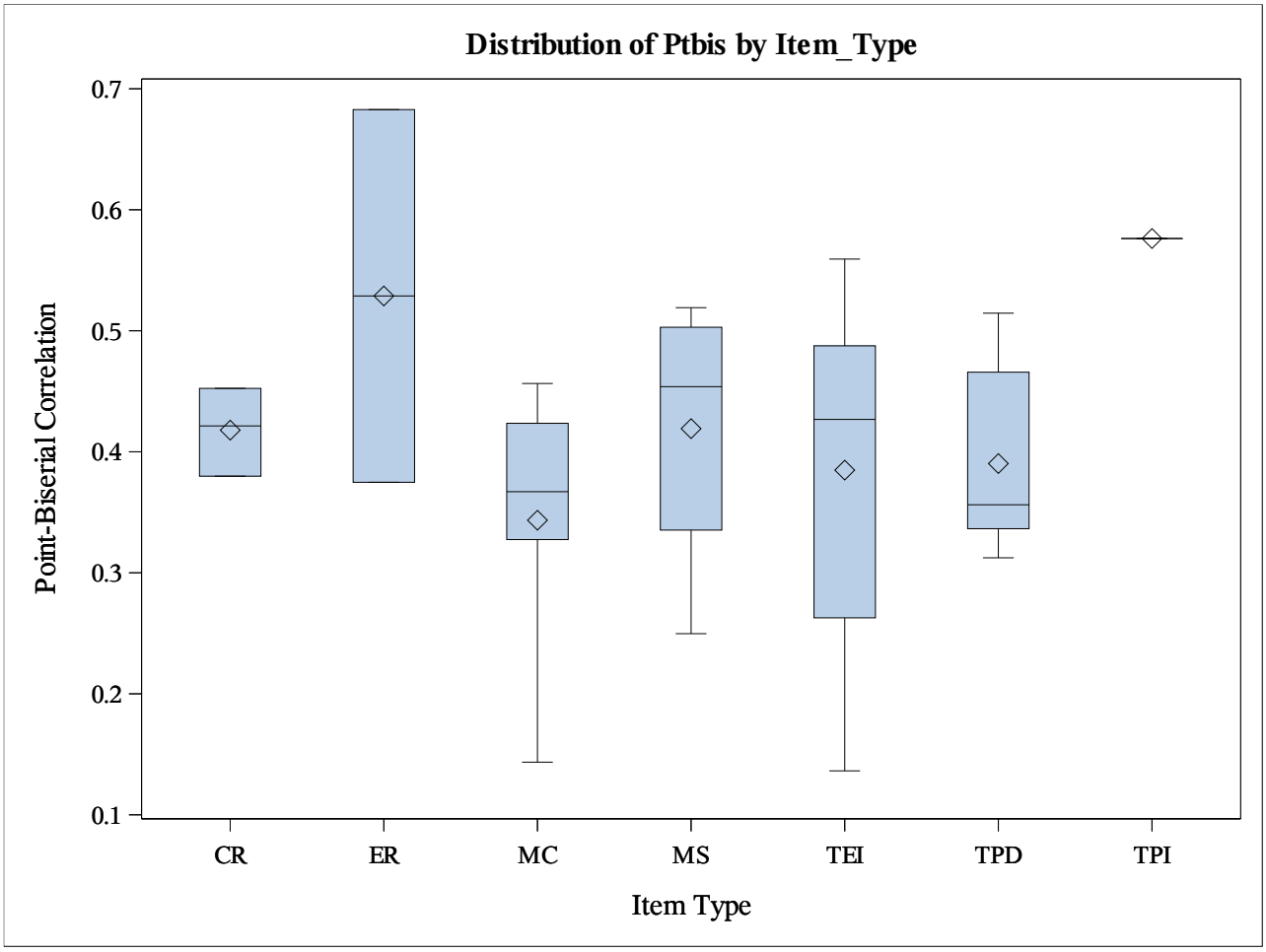
Box and Whisker Plot

Point-Biserial Correlation: Science Grade 5



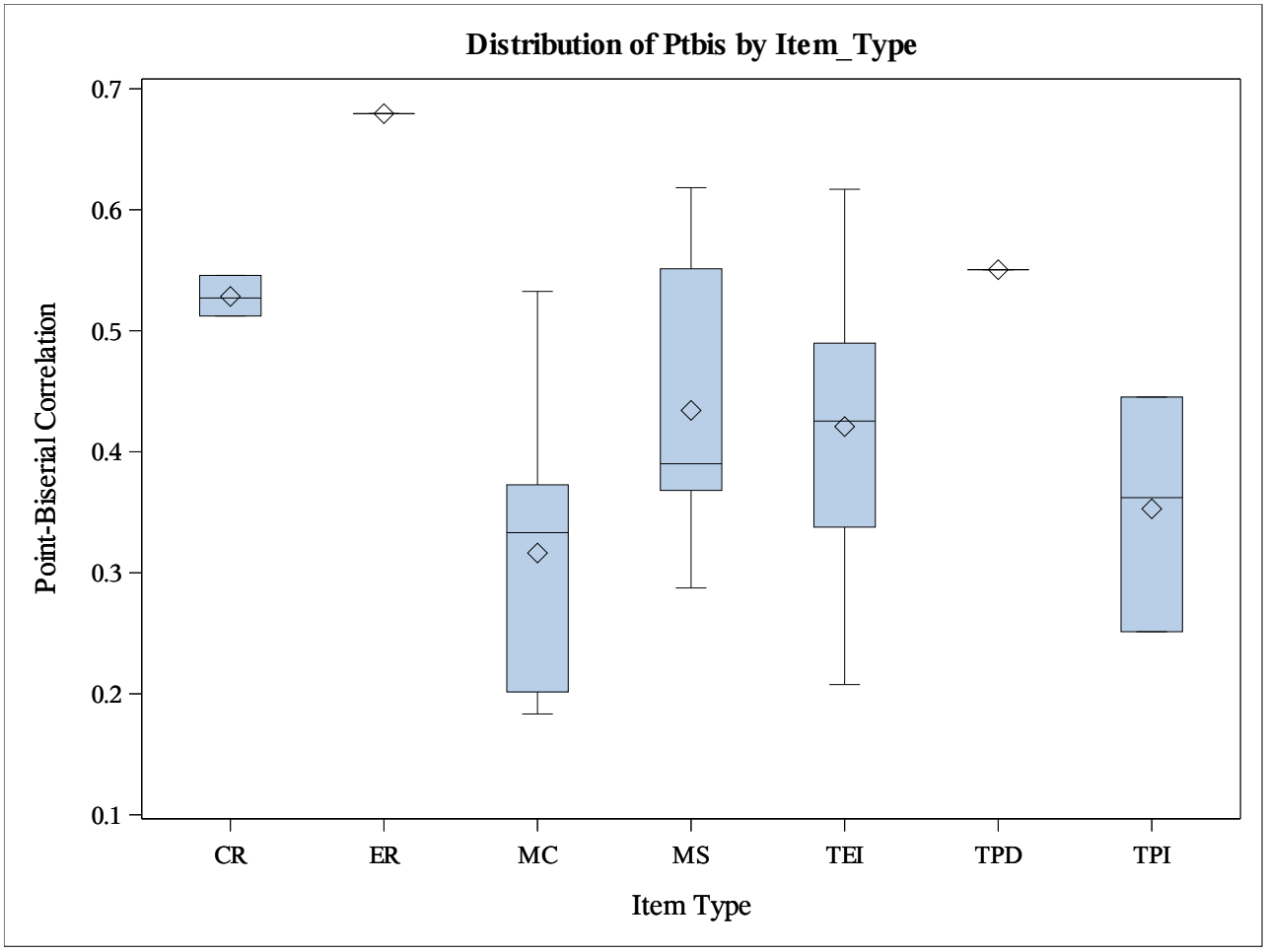
Box and Whisker Plot

Point-Biserial Correlation: Science Grade 6



Box and Whisker Plot

Point-Biserial Correlation: Science Grade 7



Box and Whisker Plot

Point-Biserial Correlation: Science Grade 8

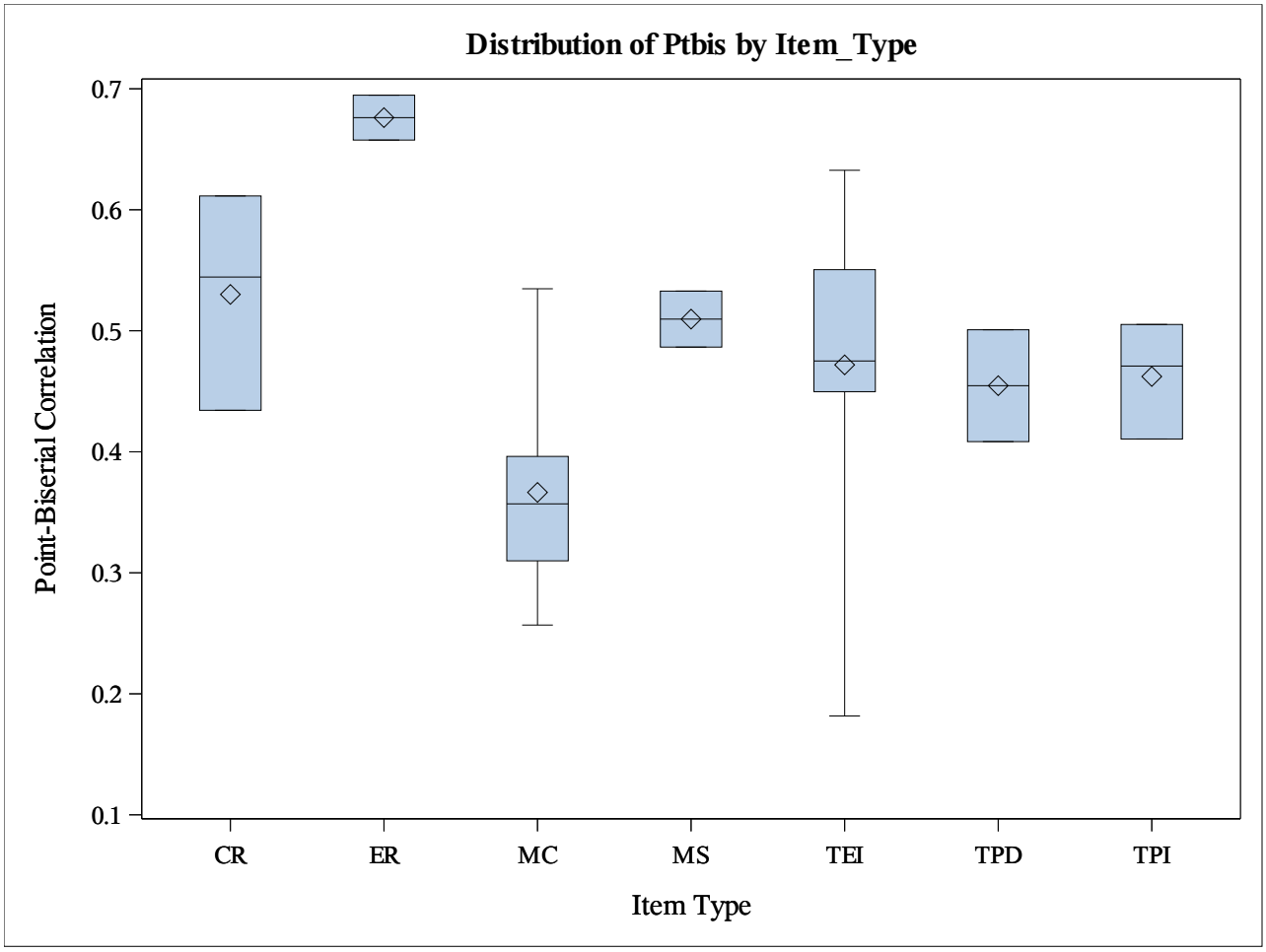


Table C.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2022 Operational SC G3–8*

Grade	No. of Items	r lt 0	0.0 le r lt 0.2	0.2 le r lt 0.3	0.3 le r lt 0.4	0.4 le r lt 0.5
3	36	0	3	9	15	7
4	36	0	2	7	13	11
5	37	0	2	2	11	14
6**	37	0	5	7	14	9
7	36	0	5	8	10	8
8**	38	0	1	5	12	11

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

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

Plot C.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2022 Operational SC G3–8*

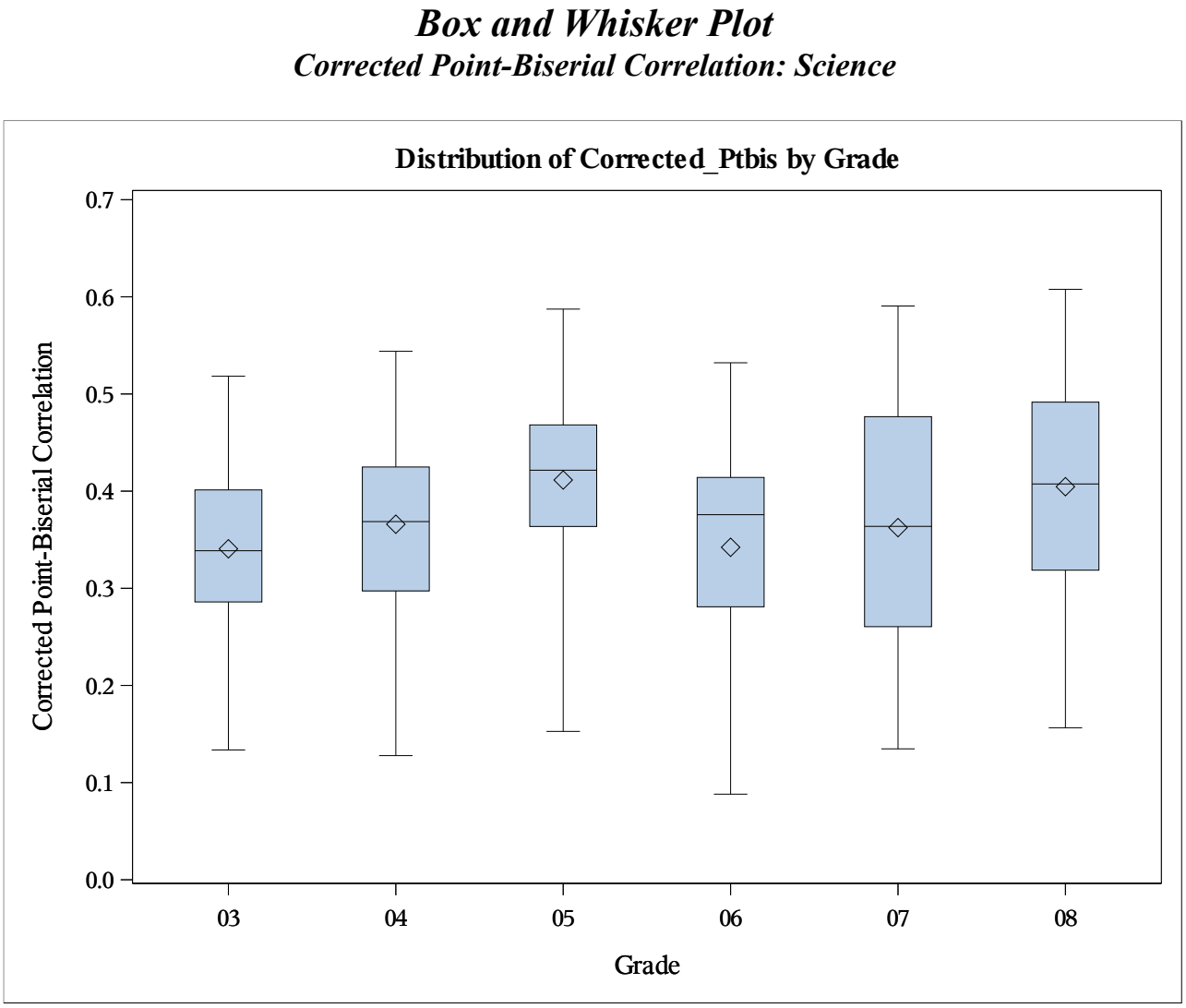


Table C.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational SC G3–8*

Grade 3

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.283	0.283	0.303	0.448	0.448
MC	21	0.134	0.248	0.317	0.377	0.418
MS	1	0.284	0.284	0.284	0.284	0.284
TPD	6	0.288	0.355	0.408	0.497	0.510
TPI	5	0.296	0.399	0.434	0.480	0.518

Grade 4

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.413	0.413	0.423	0.541	0.541
MC	14	0.128	0.245	0.307	0.370	0.419
MS	6	0.194	0.287	0.358	0.427	0.473
TPD	8	0.344	0.359	0.404	0.464	0.539
TPI	5	0.307	0.375	0.437	0.476	0.544

Grade 5

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.412	0.412	0.423	0.505	0.505
ER	1	0.573	0.573	0.573	0.573	0.573
MC	9	0.194	0.315	0.354	0.375	0.528
MS	3	0.280	0.280	0.452	0.515	0.515
TEI	13	0.348	0.379	0.452	0.477	0.587
TPD	4	0.153	0.234	0.368	0.465	0.508
TPI	4	0.416	0.419	0.427	0.435	0.439

Grade 6

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.331	0.331	0.378	0.405	0.405
ER	2	0.329	0.329	0.431	0.532	0.532
MC	9	0.099	0.281	0.324	0.384	0.415
MS	4	0.209	0.293	0.412	0.464	0.482
TEI	12	0.088	0.222	0.380	0.431	0.495
TPD	6	0.250	0.258	0.289	0.388	0.455
TPI	1	0.527	0.527	0.527	0.527	0.527

Grade 7

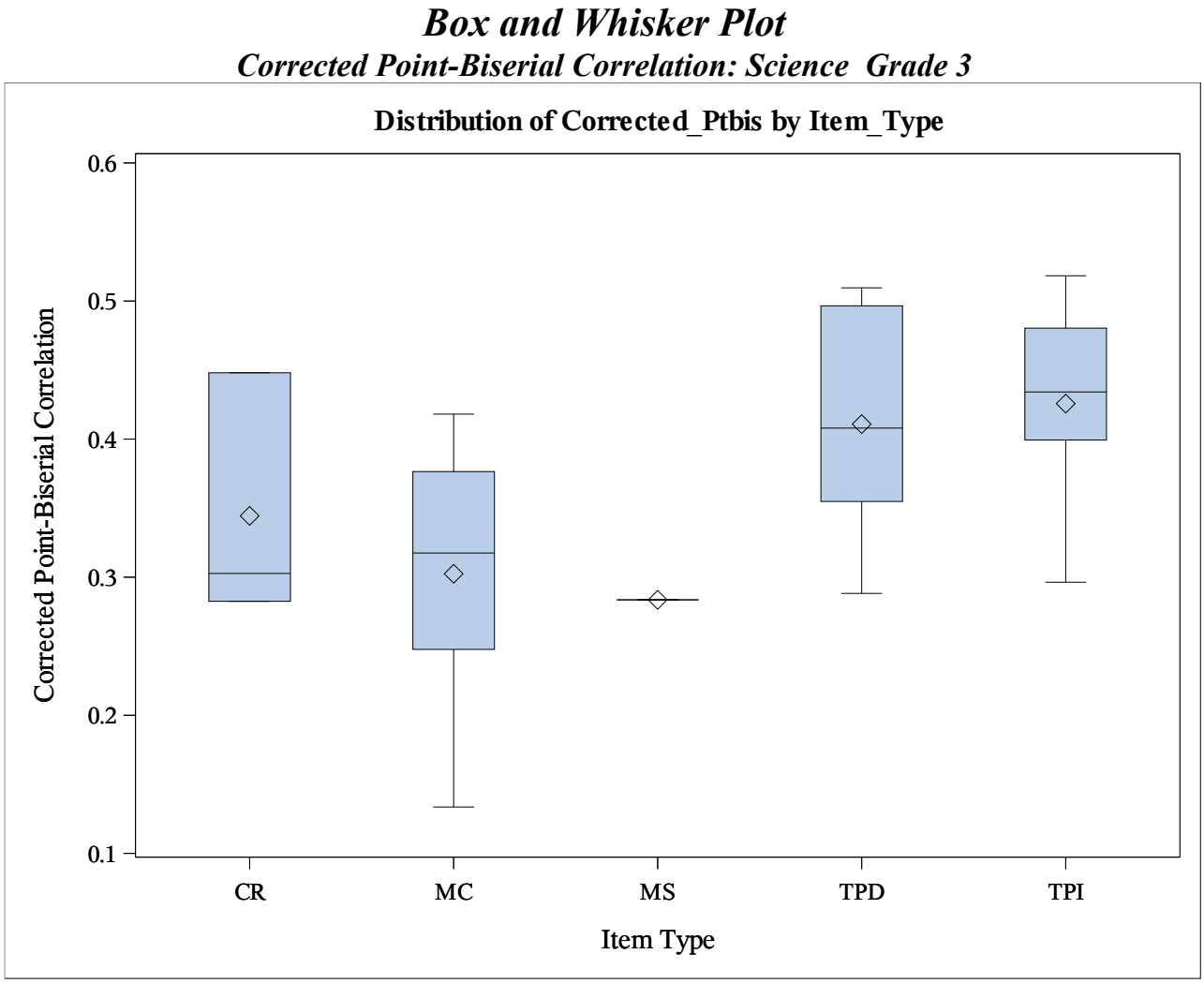
Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.461	0.461	0.473	0.495	0.495
ER	1	0.544	0.544	0.544	0.544	0.544
MC	8	0.135	0.158	0.288	0.334	0.496
MS	6	0.251	0.327	0.359	0.515	0.591
TEI	14	0.170	0.259	0.378	0.451	0.562
TPD	1	0.480	0.480	0.480	0.480	0.480
TPI	3	0.192	0.192	0.306	0.385	0.385

Grade 8

Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.395	0.395	0.509	0.571	0.571
ER	2	0.604	0.604	0.606	0.608	0.608
MC	14	0.215	0.271	0.318	0.360	0.502
MS	2	0.453	0.453	0.477	0.500	0.500
TEI	12	0.156	0.398	0.424	0.501	0.586
TPD	2	0.338	0.338	0.389	0.439	0.439
TPI	3	0.359	0.359	0.422	0.452	0.452

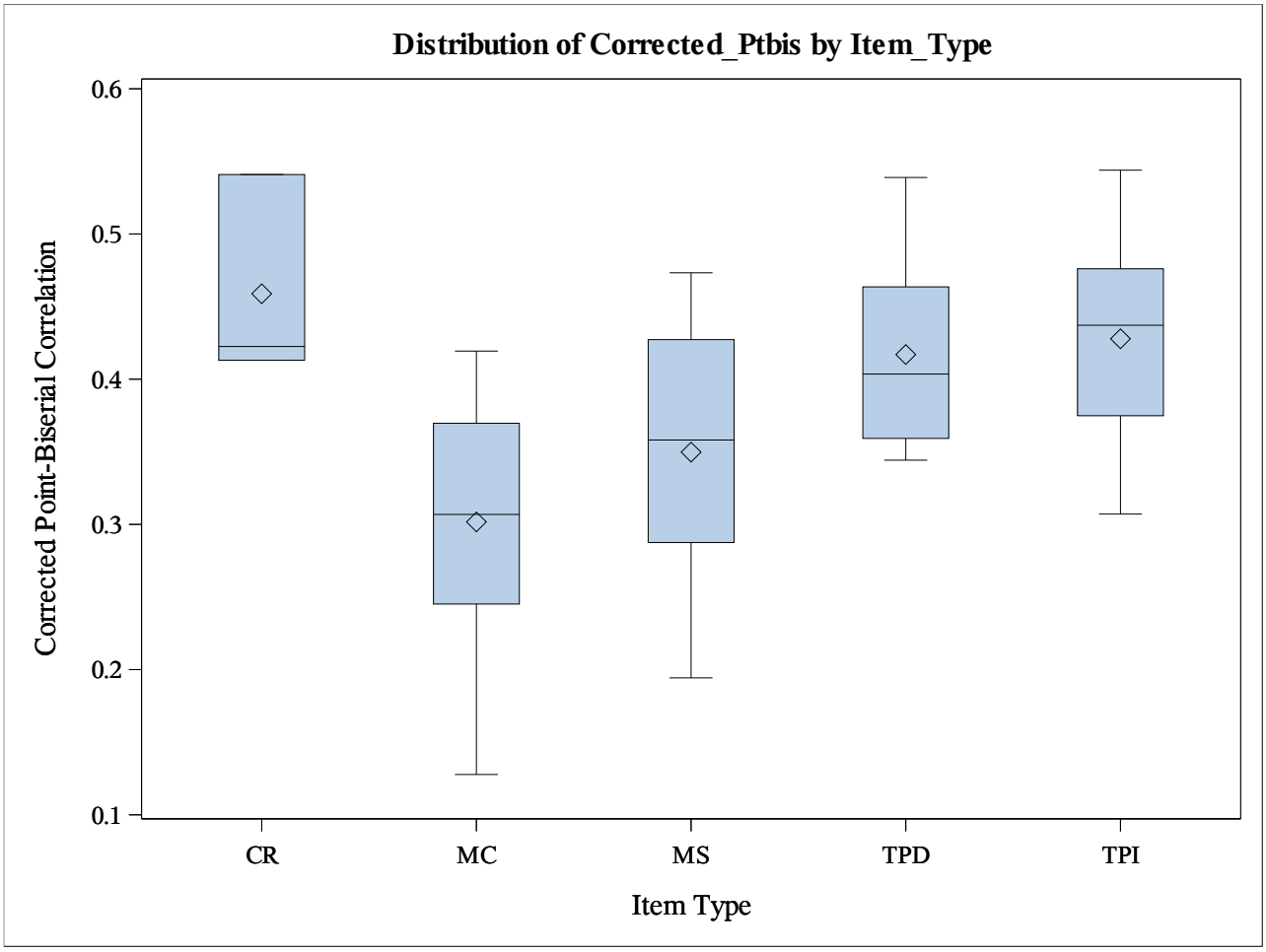
Plot C.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational SC G3–8



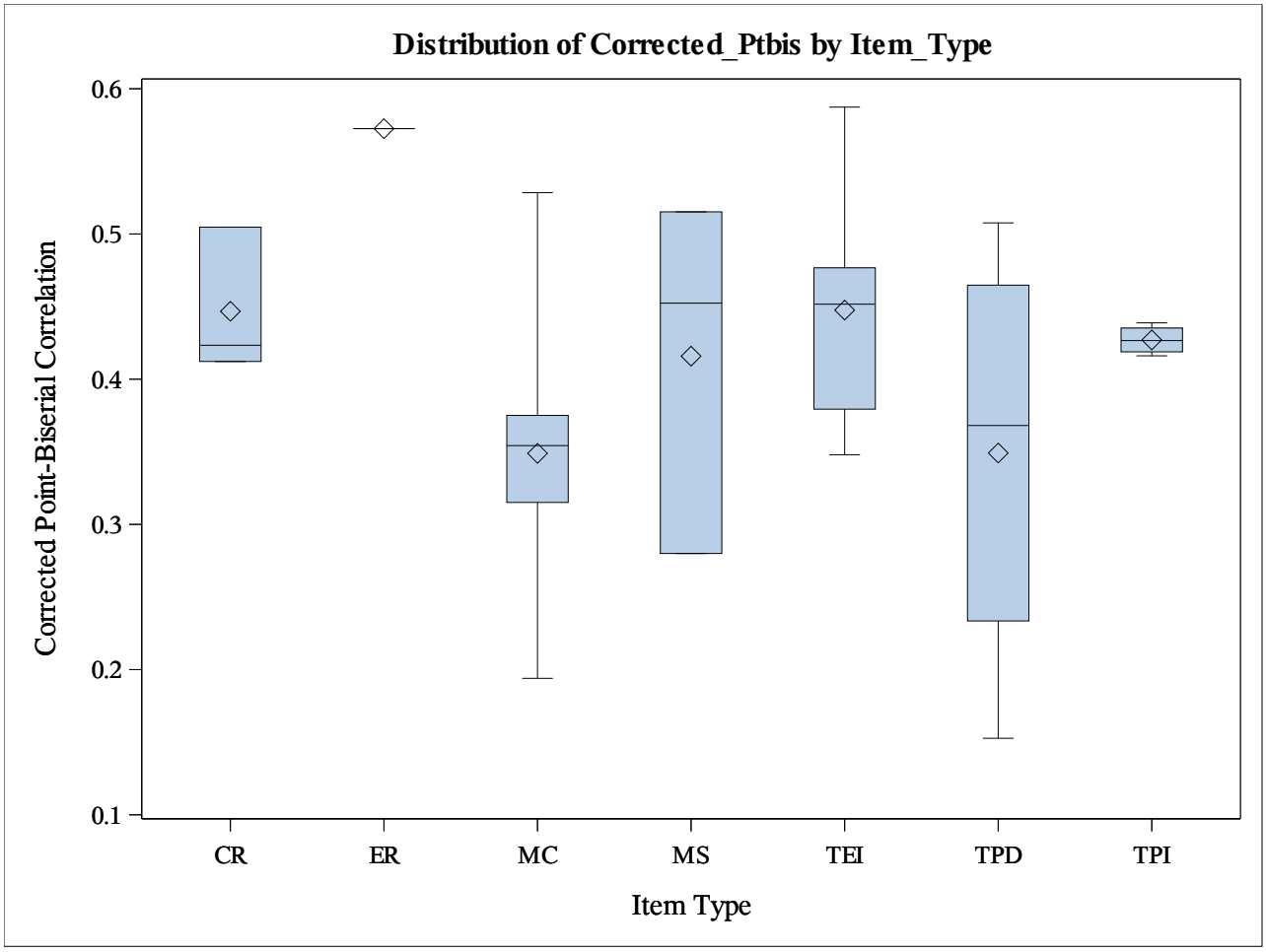
Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 4



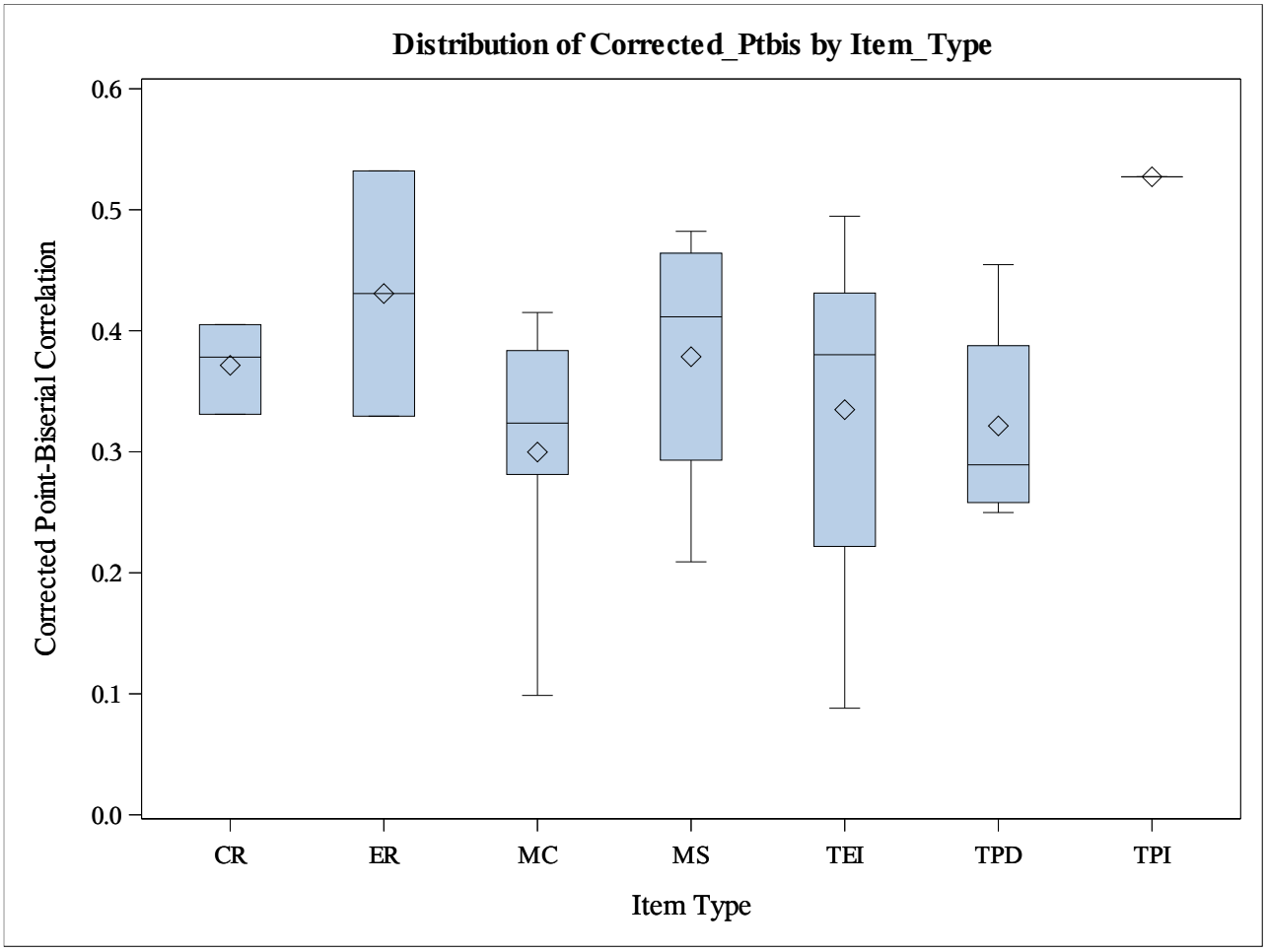
Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 5



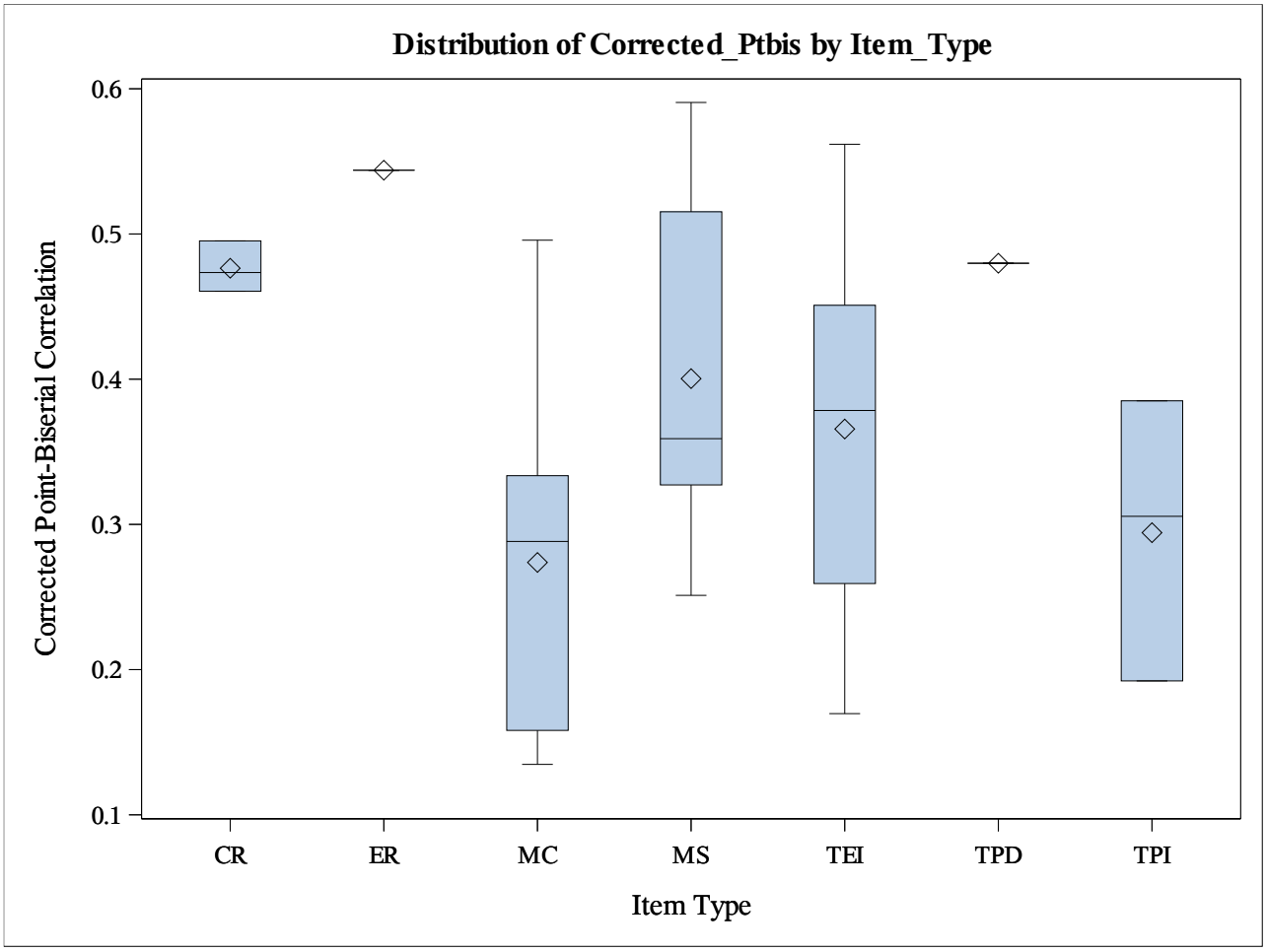
Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 6



Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 7



Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 8

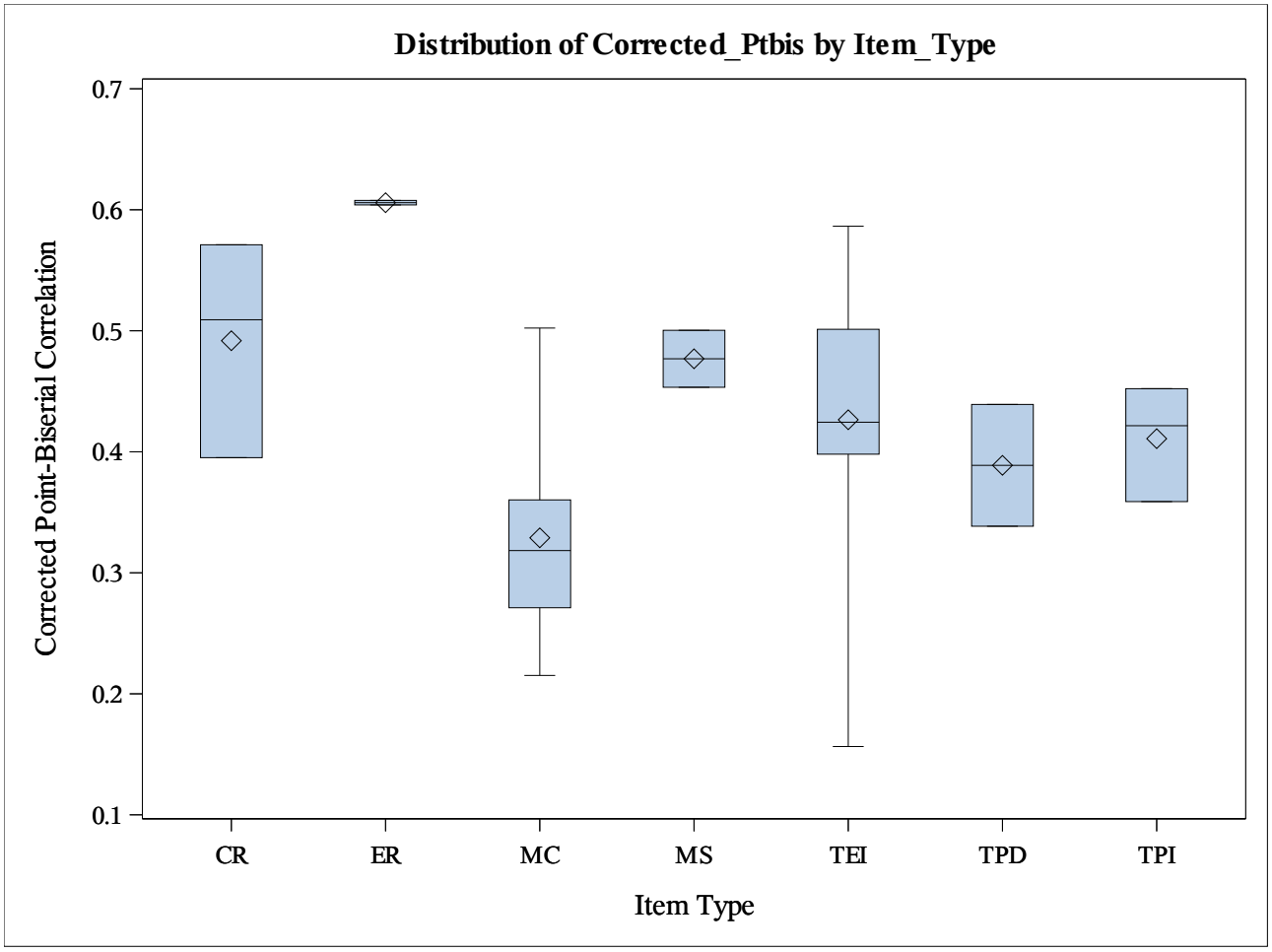


Table C.4.1

Item-Total Correlation Summary by Reporting Category: Spring 2022 Operational SC G3–8

Gr	ReportingCategory	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	Investigate	11	0.225	0.302	0.387	0.500	0.579
	Evaluate	17	0.242	0.353	0.394	0.466	0.544
	Reason Scientifically	6	0.188	0.339	0.364	0.407	0.427
4	Investigate	11	0.298	0.376	0.406	0.479	0.583
	Evaluate	6	0.176	0.276	0.372	0.458	0.604
	Reason Scientifically	16	0.236	0.356	0.426	0.501	0.597
5	Investigate	8	0.317	0.379	0.402	0.499	0.565
	Evaluate	13	0.226	0.460	0.482	0.547	0.644
	Reason Scientifically	16	0.237	0.408	0.467	0.494	0.671
6	Investigate	6	0.327	0.367	0.454	0.510	0.576
	Evaluate	11	0.225	0.354	0.424	0.465	0.559
	Reason Scientifically*	19	0.136	0.250	0.375	0.456	0.683
7	Investigate	6	0.287	0.296	0.329	0.445	0.618
	Evaluate	9	0.183	0.328	0.391	0.546	0.679
	Reason Scientifically	18	0.208	0.350	0.456	0.533	0.605
8	Investigate	11	0.257	0.384	0.458	0.533	0.568
	Evaluate	12	0.182	0.359	0.440	0.494	0.544
	Reason Scientifically*	15	0.275	0.357	0.472	0.611	0.695

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

Table C.4.2

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2022 Operational SC G3–8

Grade 3

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2 Evaluate	2	0.347	0.347	0.430	0.513	0.513
	3 Reason Scientifically	1	0.339	0.339	0.339	0.339	0.339
MC	1 Investigate	7	0.225	0.257	0.304	0.387	0.440
	2 Evaluate	9	0.242	0.355	0.378	0.427	0.466
	3 Reason Scientifically	4	0.188	0.272	0.382	0.417	0.427
MS	2 Evaluate	1	0.331	0.331	0.331	0.331	0.331
TPC	1 Investigate	3	0.500	0.500	0.567	0.579	0.579
	2 Evaluate	3	0.353	0.353	0.424	0.487	0.487
TPI	1 Investigate	1	0.469	0.469	0.469	0.469	0.469
	2 Evaluate	2	0.495	0.495	0.520	0.544	0.544
	3 Reason Scientifically	1	0.371	0.371	0.371	0.371	0.371

Grade 4

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	1 Investigate	2	0.479	0.479	0.531	0.583	0.583
MC	1 Investigate	5	0.298	0.319	0.376	0.391	0.399
	2 Evaluate	4	0.176	0.226	0.306	0.397	0.458
	3 Reason Scientifically	4	0.262	0.279	0.356	0.420	0.423
MS	1 Investigate	2	0.406	0.406	0.438	0.470	0.470
	3 Reason Scientifically	4	0.236	0.285	0.364	0.455	0.515
TPC	1 Investigate	1	0.440	0.440	0.440	0.440	0.440
	2 Evaluate	1	0.408	0.408	0.408	0.408	0.408
	3 Reason Scientifically	6	0.429	0.460	0.502	0.543	0.597
TPI	1 Investigate	1	0.501	0.501	0.501	0.501	0.501
	2 Evaluate	1	0.604	0.604	0.604	0.604	0.604
	3 Reason Scientifically	2	0.378	0.378	0.408	0.438	0.438

Grade 5

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	1 Investigate	1	0.479	0.479	0.479	0.479	0.479
	3 Reason Scientifically	2	0.460	0.460	0.505	0.550	0.550
ER	3 Reason Scientifically	1	0.671	0.671	0.671	0.671	0.671
MC	1 Investigate	3	0.387	0.387	0.393	0.412	0.412
	2 Evaluate	2	0.412	0.412	0.486	0.560	0.560
	3 Reason Scientifically	4	0.237	0.270	0.329	0.387	0.420
MS	1 Investigate	1	0.317	0.317	0.317	0.317	0.317
	2 Evaluate	2	0.488	0.488	0.518	0.547	0.547
TEI	1 Investigate	2	0.520	0.520	0.542	0.565	0.565
	2 Evaluate	6	0.380	0.460	0.496	0.580	0.644
	3 Reason Scientifically	5	0.403	0.414	0.415	0.487	0.497
TPD	1 Investigate	1	0.371	0.371	0.371	0.371	0.371
	2 Evaluate	2	0.226	0.226	0.354	0.482	0.482
	3 Reason Scientifically	1	0.556	0.556	0.556	0.556	0.556
TPI	2 Evaluate	1	0.473	0.473	0.473	0.473	0.473
	3 Reason Scientifically	3	0.474	0.474	0.486	0.492	0.492

Grade 6

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	1 Investigate	1	0.421	0.421	0.421	0.421	0.421
	3 Reason Scientifically	2	0.380	0.380	0.416	0.452	0.452
ER	3 Reason Scientifically	2	0.375	0.375	0.529	0.683	0.683
MC	1 Investigate	2	0.327	0.327	0.347	0.367	0.367
	2 Evaluate	3	0.354	0.354	0.405	0.424	0.424
	3 Reason Scientifically	4	0.144	0.160	0.307	0.447	0.456
MS	1 Investigate	1	0.487	0.487	0.487	0.487	0.487
	2 Evaluate	1	0.519	0.519	0.519	0.519	0.519
	3 Reason Scientifically	2	0.250	0.250	0.335	0.421	0.421
TEI	1 Investigate	1	0.510	0.510	0.510	0.510	0.510
	2 Evaluate	5	0.225	0.435	0.454	0.465	0.559
	3 Reason Scientifically	5	0.136	0.214	0.301	0.359	0.541
TPD	2 Evaluate	2	0.336	0.336	0.349	0.361	0.361
	3 Reason Scientifically	4	0.312	0.332	0.409	0.490	0.515
TPI	1 Investigate	1	0.576	0.576	0.576	0.576	0.576

Grade 7

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2 Evaluate	1	0.546	0.546	0.546	0.546	0.546
	3 Reason Scientifically	2	0.512	0.512	0.520	0.527	0.527
ER	2 Evaluate	1	0.679	0.679	0.679	0.679	0.679
MC	2 Evaluate	5	0.183	0.189	0.328	0.338	0.404
	3 Reason Scientifically	3	0.214	0.214	0.341	0.533	0.533
MS	1 Investigate	2	0.287	0.287	0.453	0.618	0.618
	2 Evaluate	1	0.391	0.391	0.391	0.391	0.391
	3 Reason Scientifically	2	0.368	0.368	0.460	0.551	0.551
TEI	1 Investigate	3	0.296	0.296	0.309	0.348	0.348
	2 Evaluate	1	0.617	0.617	0.617	0.617	0.617
	3 Reason Scientifically	10	0.208	0.350	0.443	0.490	0.605
TPD	3 Reason Scientifically	1	0.550	0.550	0.550	0.550	0.550
TPI	1 Investigate	1	0.445	0.445	0.445	0.445	0.445

Grade 8

Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2 Evaluate	1	0.544	0.544	0.544	0.544	0.544
	3 Reason Scientifically	2	0.434	0.434	0.523	0.611	0.611
ER	3 Reason Scientifically	2	0.658	0.658	0.676	0.695	0.695
MC	1 Investigate	5	0.257	0.357	0.384	0.396	0.535
	2 Evaluate	3	0.335	0.335	0.346	0.400	0.400
	3 Reason Scientifically	6	0.275	0.287	0.334	0.367	0.523
MS	1 Investigate	1	0.533	0.533	0.533	0.533	0.533
	2 Evaluate	1	0.487	0.487	0.487	0.487	0.487
TEI	1 Investigate	5	0.441	0.458	0.462	0.533	0.568
	2 Evaluate	3	0.182	0.182	0.371	0.478	0.478
	3 Reason Scientifically	4	0.472	0.483	0.532	0.602	0.633
TPD	2 Evaluate	2	0.408	0.408	0.455	0.501	0.501
TPI	2 Evaluate	2	0.471	0.471	0.488	0.505	0.505
	3 Reason Scientifically	1	0.411	0.411	0.411	0.411	0.411

Table C.5.1.1

IRT-A Parameter Summary by Reporting Category: Grade 3

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	0	0	0
0.2 <= a < 0.4	1	3	2	6
0.4 <= a < 0.6	5	7	1	14
0.6 <= a < 0.8	2	4	1	7
0.8 <= a < 1.0	1	1	0	2
1.0 <= a < 1.2	1	2	2	6
1.2 <= a < 1.4	1	0	0	1
1.4 <= a < 1.6	0	0	0	0
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.34	0.30	0.27	0.27
Maximum	1.23	1.17	1.11	1.23
Mean	0.66	0.62	0.65	0.65
SD	0.29	0.25	0.36	0.27
Number of Items	11	17	6	36

Table C.5.2.1

IRT-B Parameter Summary by Reporting Category: Grade 3

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

Table C.5.1.2

IRT-A Parameter Summary by Reporting Category: Grade 4

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	0	0	0
0.2 <= a < 0.4	2	1	6	9
0.4 <= a < 0.6	2	1	6	11
0.6 <= a < 0.8	5	2	2	9
0.8 <= a < 1.0	1	2	1	5
1.0 <= a < 1.2	0	0	1	1
1.2 <= a < 1.4	0	0	0	0
1.4 <= a < 1.6	1	0	0	1
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.33	0.37	0.28	0.28
Maximum	1.41	0.82	1.06	1.41
Mean	0.69	0.63	0.53	0.61
SD	0.30	0.19	0.22	0.25
Number of Items	11	6	16	36

Table C.5.2.2

IRT-B Parameter Summary by Reporting Category: Grade 4

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0
-1.5 <= b < -1.0	0	0	0	0
-1.0 <= b < -0.5	0	0	0	0
-0.5 <= b < 0.0	0	0	1	2
0.0 <= b < 0.5	0	2	3	5
0.5 <= b < 1.0	6	1	7	15
1.0 <= b < 1.5	2	1	3	6
1.5 <= b < 2.0	3	1	1	5
2.0 <= b < 2.5	0	0	0	1
2.5 <= b < 3.0	0	1	1	2
3.0 <= b < 3.5	0	0	0	0
3.5 <= b	0	0	0	0
Minimum	0.54	0.10	-0.36	-0.36
Maximum	1.74	2.99	2.67	2.99
Mean	1.07	1.24	0.87	0.99
SD	0.44	1.06	0.68	0.71
Number of Items	11	6	16	36

Table C.5.1.3

IRT-A Parameter Summary by Reporting Category: Grade 5

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	1	0	1
0.2 <= a < 0.4	2	3	3	8
0.4 <= a < 0.6	5	3	8	16
0.6 <= a < 0.8	1	2	3	6
0.8 <= a < 1.0	0	2	1	3
1.0 <= a < 1.2	0	1	0	1
1.2 <= a < 1.4	0	1	1	2
1.4 <= a < 1.6	0	0	0	0
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.28	0.10	0.38	0.10
Maximum	0.69	1.36	1.23	1.36
Mean	0.50	0.64	0.59	0.58
SD	0.13	0.35	0.23	0.26
Number of Items	8	13	16	37

Table C.5.2.3

IRT-B Parameter Summary by Reporting Category: Grade 5

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0
-1.5 <= b < -1.0	1	0	2	3
-1.0 <= b < -0.5	1	1	1	3
-0.5 <= b < 0.0	1	3	3	7
0.0 <= b < 0.5	1	5	3	9
0.5 <= b < 1.0	2	2	1	5
1.0 <= b < 1.5	0	1	4	5
1.5 <= b < 2.0	2	0	2	4
2.0 <= b < 2.5	0	0	0	0
2.5 <= b < 3.0	0	0	0	0
3.0 <= b < 3.5	0	1	0	1
3.5 <= b	0	0	0	0
Minimum	-1.09	-0.83	-1.27	-1.27
Maximum	1.61	3.25	1.91	3.25
Mean	0.36	0.47	0.37	0.40
SD	1.03	1.02	1.01	0.99
Number of Items	8	13	16	37

Table C.5.1.4

IRT-A Parameter Summary by Reporting Category: Grade 6

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	2	1	3
0.2 <= a < 0.4	1	1	6	8
0.4 <= a < 0.6	2	5	6	14
0.6 <= a < 0.8	3	3	1	7
0.8 <= a < 1.0	0	0	4	4
1.0 <= a < 1.2	0	0	1	1
1.2 <= a < 1.4	0	0	0	0
1.4 <= a < 1.6	0	0	0	0
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.38	0.18	0.19	0.18
Maximum	0.78	0.78	1.01	1.01
Mean	0.57	0.46	0.53	0.52
SD	0.14	0.19	0.26	0.22
Number of Items	6	11	19	37

Table C.5.2.4

IRT-B Parameter Summary by Reporting Category: Grade 6

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0
-1.5 <= b < -1.0	0	0	0	0
-1.0 <= b < -0.5	0	0	2	2
-0.5 <= b < 0.0	2	3	4	9
0.0 <= b < 0.5	0	3	1	4
0.5 <= b < 1.0	1	3	1	6
1.0 <= b < 1.5	2	1	1	4
1.5 <= b < 2.0	0	0	3	3
2.0 <= b < 2.5	1	1	1	3
2.5 <= b < 3.0	0	0	4	4
3.0 <= b < 3.5	0	0	2	2
3.5 <= b	0	0	0	0
Minimum	-0.31	-0.29	-0.79	-0.79
Maximum	2.40	2.47	3.28	3.28
Mean	0.78	0.58	1.33	1.00
SD	0.99	0.81	1.43	1.21
Number of Items	6	11	19	37

Table C.5.1.5

IRT-A Parameter Summary by Reporting Category: Grade 7

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	1	0	2	4
0.2 <= a < 0.4	2	3	1	7
0.4 <= a < 0.6	0	2	7	9
0.6 <= a < 0.8	1	2	2	6
0.8 <= a < 1.0	0	1	4	5
1.0 <= a < 1.2	0	1	2	3
1.2 <= a < 1.4	1	0	0	1
1.4 <= a < 1.6	1	0	0	1
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.16	0.24	0.17	0.16
Maximum	1.51	1.12	1.07	1.51
Mean	0.71	0.61	0.62	0.61
SD	0.57	0.29	0.28	0.34
Number of Items	6	9	18	36

Table C.5.2.5

IRT-B Parameter Summary by Reporting Category: Grade 7

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0
-1.5 <= b < -1.0	0	0	1	1
-1.0 <= b < -0.5	0	0	2	2
-0.5 <= b < 0.0	1	2	4	7
0.0 <= b < 0.5	1	1	2	5
0.5 <= b < 1.0	1	1	3	5
1.0 <= b < 1.5	0	2	2	4
1.5 <= b < 2.0	1	1	1	4
2.0 <= b < 2.5	1	2	3	6
2.5 <= b < 3.0	0	0	0	0
3.0 <= b < 3.5	1	0	0	2
3.5 <= b	0	0	0	0
Minimum	-0.04	-0.29	-1.01	-1.01
Maximum	3.18	2.01	2.08	3.39
Mean	1.27	0.93	0.55	0.86
SD	1.24	0.86	1.01	1.09
Number of Items	6	9	18	36

Table C.5.1.6

IRT-A Parameter Summary by Reporting Category: Grade 8

IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
a < 0.0	0	0	0	0
0.0 <= a < 0.2	0	0	0	0
0.2 <= a < 0.4	1	3	4	8
0.4 <= a < 0.6	4	4	3	11
0.6 <= a < 0.8	3	4	3	10
0.8 <= a < 1.0	2	0	2	4
1.0 <= a < 1.2	0	1	1	2
1.2 <= a < 1.4	1	0	2	3
1.4 <= a < 1.6	0	0	0	0
1.6 <= a < 1.8	0	0	0	0
1.8 <= a < 2.0	0	0	0	0
2.0 <= a	0	0	0	0
Minimum	0.38	0.24	0.33	0.24
Maximum	1.30	1.17	1.28	1.30
Mean	0.72	0.57	0.71	0.67
SD	0.28	0.27	0.32	0.29
Number of Items	11	12	15	38

Table C.5.2.6

IRT-B Parameter Summary by Reporting Category: Grade 8

IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
b < -3.5	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0
-1.5 <= b < -1.0	0	2	0	2
-1.0 <= b < -0.5	1	0	1	2
-0.5 <= b < 0.0	2	3	3	8
0.0 <= b < 0.5	4	3	1	8
0.5 <= b < 1.0	3	2	6	11
1.0 <= b < 1.5	1	0	3	4
1.5 <= b < 2.0	0	1	0	1
2.0 <= b < 2.5	0	1	1	2
2.5 <= b < 3.0	0	0	0	0
3.0 <= b < 3.5	0	0	0	0
3.5 <= b	0	0	0	0
Minimum	-0.94	-1.42	-0.58	-1.42
Maximum	1.26	2.30	2.06	2.30
Mean	0.34	0.25	0.64	0.43
SD	0.66	1.07	0.76	0.84
Number of Items	11	12	15	38

Table C.5.3

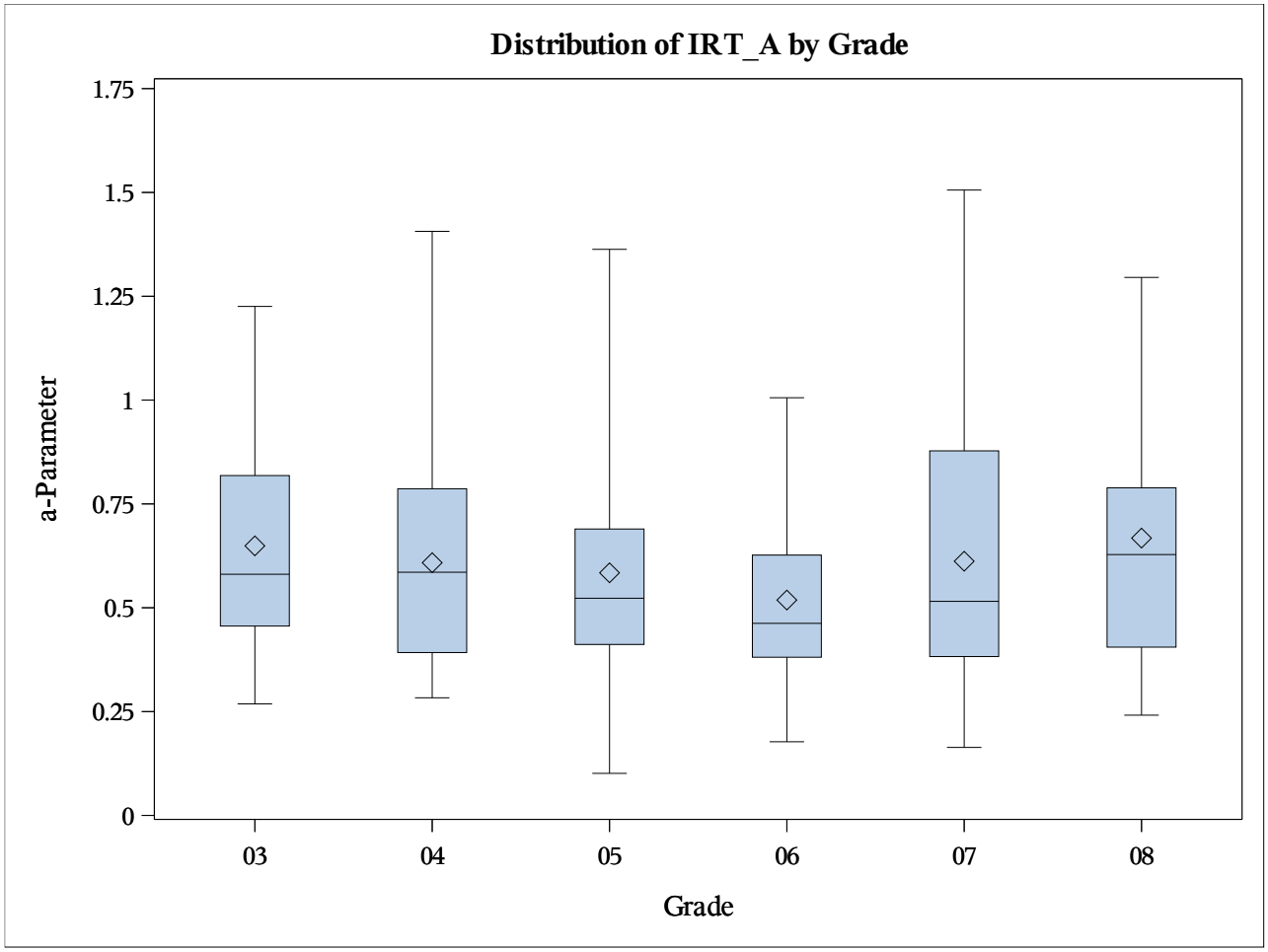
IRT Parameter Summary: Spring 2022 Operational SC G3–8

Grade	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	a	36	0.269	0.456	0.581	0.818	1.226
	b	36	-1.162	0.463	0.677	1.459	3.557
	c	22	0	0.146	0.193	0.231	0.27
4	a	36	0.283	0.392	0.586	0.787	1.406
	b	36	-0.36	0.57	0.909	1.341	2.995
	c	20	0.007	0.056	0.138	0.196	0.301
5	a	37	0.101	0.412	0.523	0.69	1.363
	b	37	-1.273	-0.339	0.201	1.129	3.248
	c	20	0.002	0.03	0.07	0.207	0.346
6	a*	37	0.177	0.381	0.462	0.627	1.006
	b*	37	-0.792	-0.07	0.838	1.989	3.281
	c	20	0.001	0.043	0.136	0.186	0.304
7	a	36	0.164	0.383	0.515	0.878	1.506
	b	36	-1.012	-0.024	0.856	1.671	3.386
	c	19	0.001	0.015	0.062	0.145	0.346
8	a*	38	0.241	0.405	0.628	0.789	1.295
	b*	38	-1.42	-0.054	0.444	0.943	2.299
	C	20	0.009	0.077	0.139	0.239	0.311

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

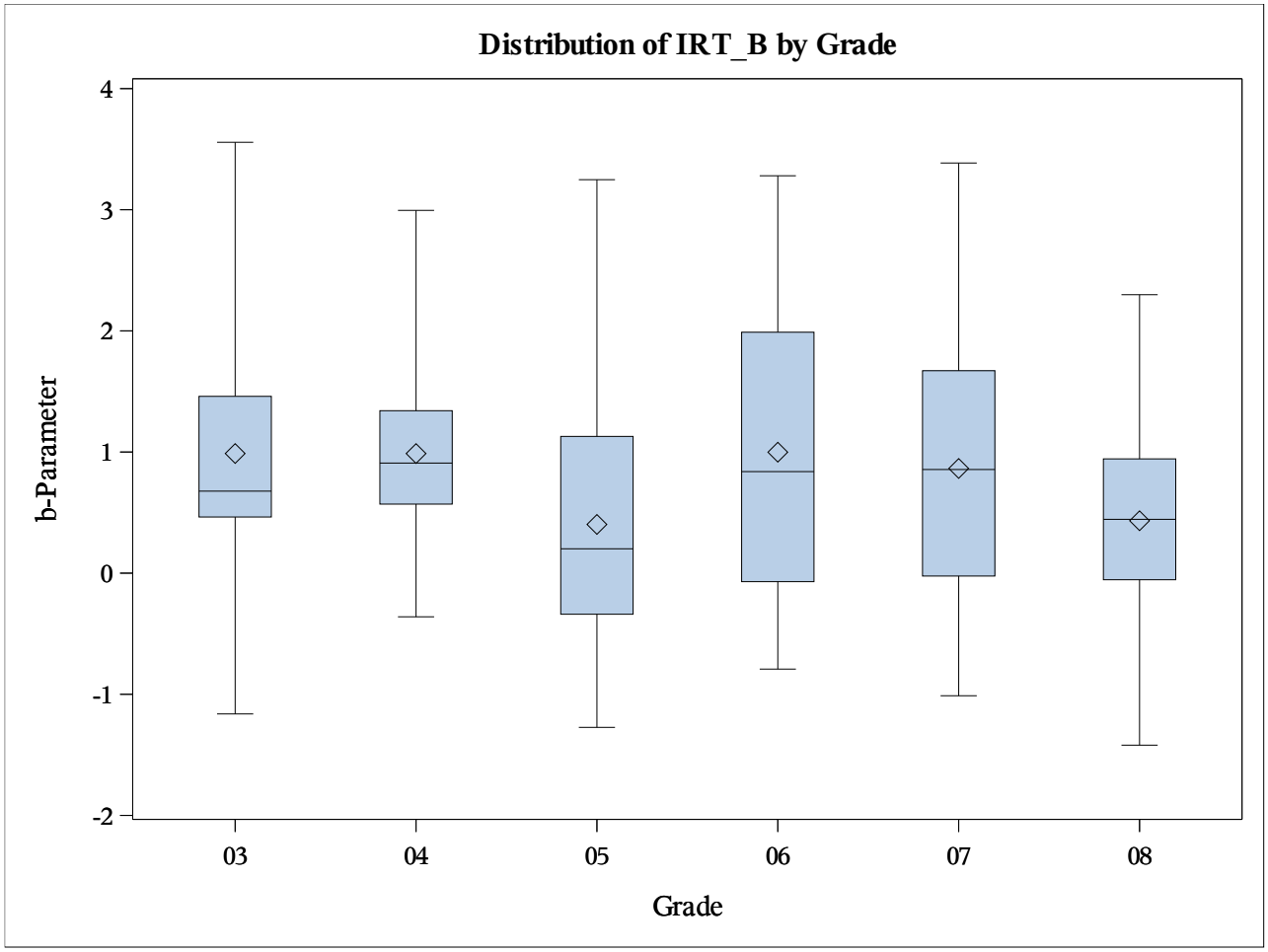
Plot C.5.1

IRT Item Parameter Summary for Spring 2022 Operational SC G3–8: A-Parameter



Plot C.5.2

IRT Item Parameter Summary for Spring 2022 Operational SC G3–8: B-Parameter



Plot C.5.3

IRT Item Parameter Summary for Spring 2022 Operational SC G3–8: C-Parameter

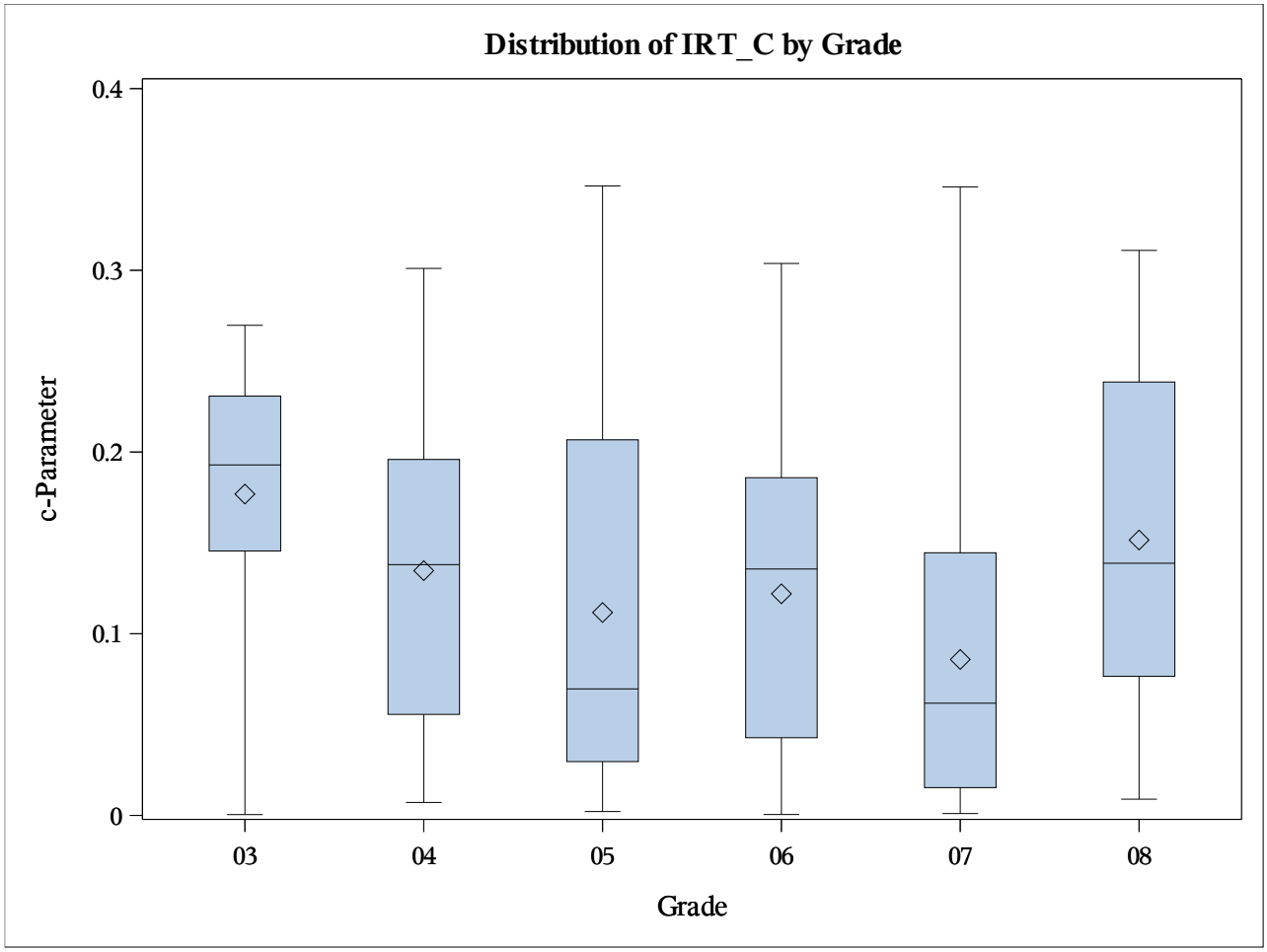


Table C.5.4

IRT Parameter Summary by Item Type: Spring 2022 Operational SC G3–8

Grade 3

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.345	0.345	0.477	0.507	0.507
	b	3	1.057	1.057	2.937	3.557	3.557
MC	a	21	0.415	0.627	0.728	1.008	1.226
	b	21	-1.103	0.467	0.696	1.49	2.667
	c	21	0.023	0.149	0.199	0.231	0.27
MS	a	1	0.491	0.491	0.491	0.491	0.491
	b	1	0.659	0.659	0.659	0.659	0.659
	c*	1	0	0	0	0	0
TPD	a	6	0.304	0.334	0.351	0.501	0.517
	b	6	0.006	0.451	0.527	1.429	2.547
TPI	a	5	0.269	0.413	0.529	0.576	0.577
	b	5	-1.162	0.444	0.62	0.706	1.368

*The value of c parameter is 0.00046.

Grade 4

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.511	0.511	0.588	0.904	0.904
	b	3	1.56	1.56	1.62	2.014	2.014
MC	a	14	0.35	0.458	0.768	0.879	1.406
	b	14	-0.166	0.75	0.909	1.005	2.995
	c	14	0.033	0.117	0.186	0.202	0.301
MS	a	6	0.455	0.542	0.621	0.79	0.795
	b	6	0.113	0.617	1.244	1.408	2.669
	c	6	0.007	0.02	0.056	0.076	0.091
TPD	a	8	0.297	0.35	0.385	0.497	0.656
	b	8	-0.36	0.328	0.713	1.078	1.275
TPI	a	5	0.283	0.383	0.436	0.583	0.609
	b	5	0.332	0.558	0.582	0.847	1.507

Grade 5

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.374	0.374	0.436	0.594	0.594
	b	3	0.981	0.981	1.187	1.28	1.28
ER	a	1	0.377	0.377	0.377	0.377	0.377
	b	1	1.913	1.913	1.913	1.913	1.913
MC	a	9	0.46	0.505	0.585	0.682	1.363
	b	9	-1.216	-0.834	-0.411	0.151	1.836
	c	9	0.011	0.067	0.21	0.229	0.346
MS	a	3	0.69	0.69	0.705	0.869	0.869
	b	3	0.16	0.16	0.419	1.61	1.61
	c	3	0.027	0.027	0.067	0.126	0.126
TEI	a	13	0.393	0.476	0.553	0.786	1.231
	b	13	-1.273	-0.339	0.201	0.791	1.476
	c	8	0.002	0.016	0.042	0.08	0.116
TPD	a	4	0.101	0.19	0.309	0.435	0.531
	b	4	0.121	0.4	1.105	2.39	3.248
TPI	a	4	0.349	0.366	0.39	0.407	0.416
	b	4	-0.19	-0.11	0.016	0.345	0.626

Grade 6

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.356	0.356	0.457	0.462	0.462
	b	3	1.945	1.945	2.402	2.656	2.656
ER	a	2	0.188	0.188	0.282	0.376	0.376
	b	2	1.365	1.365	2.026	2.687	2.687
MC	a	9	0.384	0.556	0.593	0.734	1.006
	b	9	-0.177	-0.05	0.915	1.126	3.281
	c	9	0.055	0.135	0.176	0.197	0.294
MS	a	4	0.506	0.571	0.706	0.859	0.942
	b	4	-0.367	-0.267	-0.119	0.959	1.989
	c	4	0.001	0.007	0.018	0.089	0.157
TEI	a	12	0.33	0.41	0.482	0.705	0.923
	b	12	-0.792	-0.02	0.55	2.118	2.648
	c	7	0.002	0.032	0.067	0.146	0.304
TPD	a	6	0.177	0.187	0.239	0.303	0.459
	b	6	-0.76	0.068	0.86	1.386	2.78
TPI	a	1	0.627	0.627	0.627	0.627	0.627
	b	1	-0.306	-0.306	-0.306	-0.306	-0.306

Grade 7

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.499	0.499	0.521	0.583	0.583
	b	3	0.58	0.58	0.932	1.252	1.252
ER	a	1	0.24	0.24	0.24	0.24	0.24
	b	1	1.03	1.03	1.03	1.03	1.03
MC	a	8	0.391	0.402	0.591	0.972	1.124
	b	8	-0.37	0.115	1.128	2.008	2.053
	c	8	0.013	0.032	0.09	0.219	0.346
MS	a	6	0.509	0.603	0.722	1.01	1.336
	b	6	-0.009	0.629	1.31	1.701	2.036
	c	6	0.004	0.015	0.034	0.065	0.075
TEI	a	14	0.164	0.257	0.537	0.932	1.506
	b	14	-1.012	-0.421	0.305	1.641	3.177
	c	5	0.001	0.008	0.062	0.145	0.148
TPD	a	1	0.404	0.404	0.404	0.404	0.404
	b	1	-0.214	-0.214	-0.214	-0.214	-0.214
TPI	a	3	0.176	0.176	0.324	0.374	0.374
	b	3	0.087	0.087	0.199	3.386	3.386

Grade 8

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	a	3	0.516	0.516	0.737	0.74	0.74
	b	3	0.899	0.899	1.571	2.062	2.062
ER	a	2	0.372	0.372	0.479	0.586	0.586
	b	2	0.687	0.687	0.743	0.799	0.799
MC	a	14	0.39	0.578	0.846	0.975	1.295
	b	14	-1.42	-0.037	0.936	0.996	1.399
	c	14	0.02	0.111	0.208	0.247	0.311
MS	a	2	0.783	0.783	0.786	0.789	0.789
	b	2	0.014	0.014	0.272	0.529	0.529
	c	2	0.009	0.009	0.031	0.052	0.052
TEI	a	12	0.301	0.392	0.522	0.724	1.276
	b	12	-1.356	-0.375	0.015	0.33	2.299
	c	4	0.01	0.037	0.094	0.138	0.15
TPD	a	2	0.241	0.241	0.299	0.357	0.357
	b	2	-0.043	-0.043	0.176	0.396	0.396
TPI	a	3	0.347	0.347	0.427	0.436	0.436
	b	3	-0.436	-0.436	0.123	1.446	1.446

Table C.6

Statistically Flagged Operational Items: Spring 2022 Operational SC G3–8

Grade 3

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	2	0	0	1
MC	21	0	1	0	0
MS	1	0	0	0	0
TEI	6	1	0	0	0
TPD	5	0	0	0	0
TPI	3	2	0	0	1

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

Grade 4

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	3	0	0	0
MC	14	1	1	0	0
MS	6	1	0	0	0
TEI	8	0	0	0	0
TPD	5	0	0	0	0
TPI	3	3	0	0	0

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

Grade 5

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	1	0	0	0
ER	1	1	0	0	0
MC	9	0	0	0	0
MS	3	0	0	0	0
TEI	13	2	0	1	0
TPD	4	0	0	0	0
TPI	4	0	0	0	0

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

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

Grade 6

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	3	0	0	0
ER**	1	1	0	0	0
MC	9	1	2	0	0
MS	4	1	0	0	0
TEI	12	3	1	1	0
TPD	6	1	0	0	0
TPI	1	0	0	0	0

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

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

Grade 7

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	1	0	0	0
ER**	1	0	0	0	0
MC	8	1	2	0	0
MS	6	4	0	1	0
TEI	14	4	0	0	0
TPD	1	0	0	0	0
TPI	3	0	0	0	0

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

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

Grade 8

Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
CR	3	2	0	0	0
ER**	1	0	0	0	0
MC	14	0	0	0	0
MS	2	0	0	0	0
TEI	12	1	1	1	0
TPD	2	0	0	0	0
TPI	3	0	0	0	0

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

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

Appendix D: Dimensionality

Dimensionality Reports: Science

Contents
Table D.1 Zq1 Statistics and Summary Data: Spring 2022 Operational SC G3–8
Table D.2 Q3 Statistics and Summary Data: Spring 2022 Operational SC G3–8
Table D.3 Reporting Category Intercorrelation Coefficients: Spring 2022 Operational SC G3–8
Table D.4 First and Second Eigenvalues: Spring 2022 Operational SC G3–8 Plot D.1 Principal Component Analysis: Spring 2022 Operational SC G3–8

- Because the spring 2022 test was administered during the 2022 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table D.1

Zq1 Statistics and Summary Data: Spring 2022 Operational SC G3–8

Grade 3

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

Grade 4

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

Grade 5

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

Grade 6

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

Grade 7

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

Grade 8

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

Table D.2

Q3 Statistics and Summary Data: Spring 2022 Operational SC G3–8

Grade	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.135	-0.080	-0.050	-0.022	0.047	0.120
4	0.154	-0.067	-0.044	-0.021	0.059	0.187
5	0.189	-0.082	-0.045	-0.018	0.078	0.244
6	0.137	-0.143	-0.043	-0.016	0.050	0.153
7	0.153	-0.108	-0.049	-0.016	0.072	0.192
8	0.182	-0.069	-0.044	-0.016	0.087	0.399

Table D.3

Reporting Category Intercorrelation Coefficients: Spring 2022 Operational SC G3–8

Grade	Reporting_Category	1 Investigate	2 Evaluate	3 Reason Scientifically
3	1 Investigate	1.00		
	2 Evaluate	0.68	1.00	
	3 Reason Scientifically	0.49	0.53	1.00
4	1 Investigate	1.00		
	2 Evaluate	0.56	1.00	
	3 Reason Scientifically	0.70	0.60	1.00
5	1 Investigate	1.00		
	2 Evaluate	0.68	1.00	
	3 Reason Scientifically	0.70	0.76	1.00
6	1 Investigate	1.00		
	2 Evaluate	0.63	1.00	
	3 Reason Scientifically	0.65	0.69	1.00
7	1 Investigate	1.00		
	2 Evaluate	0.50	1.00	
	3 Reason Scientifically	0.56	0.71	1.00
8	1 Investigate	1.00		
	2 Evaluate	0.69	1.00	
	3 Reason Scientifically	0.74	0.74	1.00

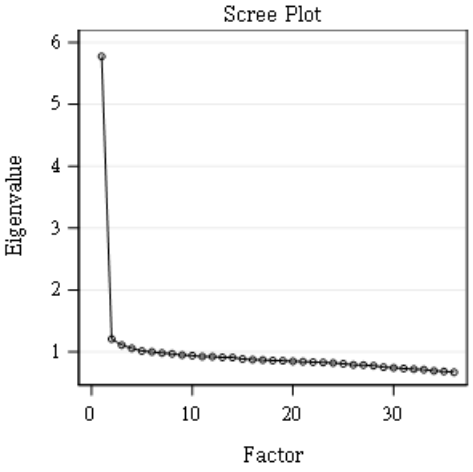
Table D.4

First and Second Eigenvalue: Spring 2022 Operational SC G3–8

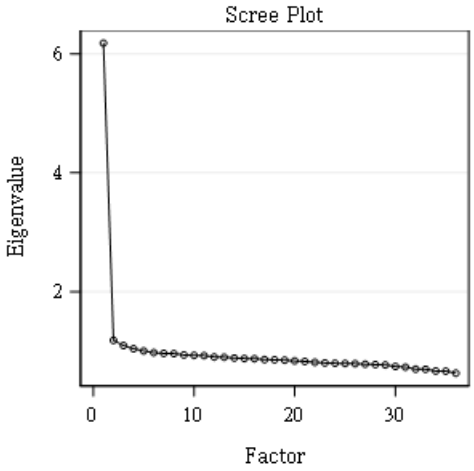
Grade	Mode	First Eigenvalue	Second Eigenvalue	Ratio
3	Online	5.775	1.206	4.789
	Paper	6.181	1.182	5.229
4	Online	6.740	1.130	5.965
5	Online	8.162	1.188	6.870
6	Online	6.433	1.172	5.489
7	Online	6.927	1.414	4.899
8	Online	8.201	1.177	6.968

Plot D.1

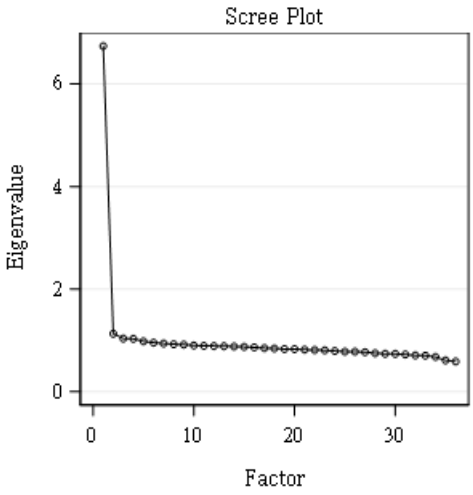
Principal Component Analysis Plot: Spring 2022 Operational SC G3-8



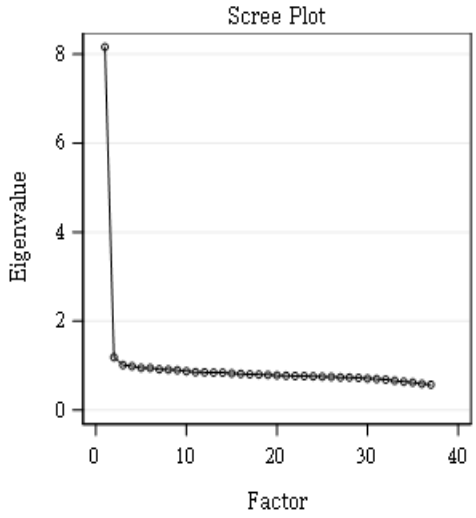
Grade 3: Online



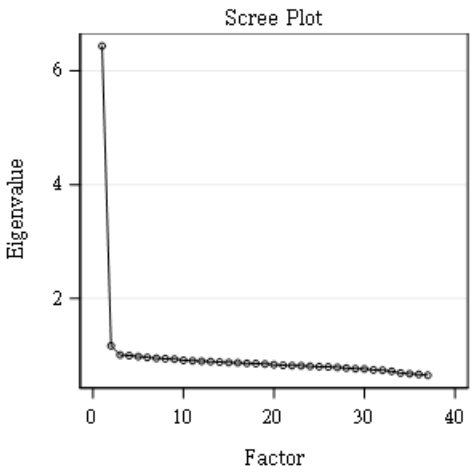
Grade 3: Paper



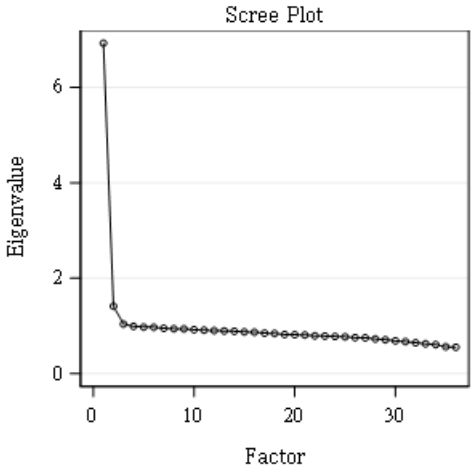
Grade 4



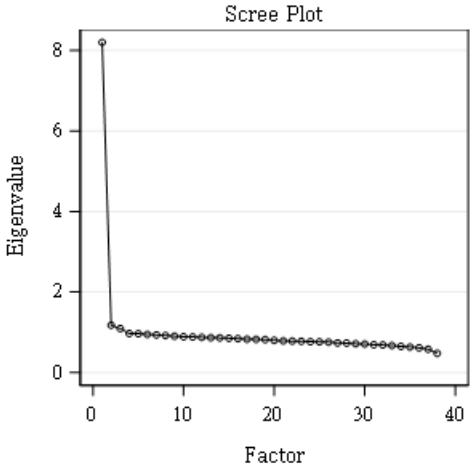
Grade 5



Grade 6



Grade 7



Grade 8

Appendix E: Scale Distribution and Statistical Report

Science

Contents
Table E.1.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 3 Table E.1.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 3
Table E.2.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 4 Table E.2.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 4
Table E.3.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 5 Table E.3.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 5
Table E.4.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 6 Table E.4.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 6
Table E.5.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 7 Table E.5.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 7
Table E.6.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science Grade 8 Table E.6.2 Frequency Distribution of Scale Scores: Spring 2022 Operational Science Grade 8

- Because the spring 2022 test was administered during the 2022 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table E.1.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 3

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

N	≥49320		
Mean	725.78	Median	725.00
Std deviation	30.74	Variance	944.84
Skewness	-0.0678	Kurtosis	-0.1498
Mode	700.00	Std Error Mean	0.1384
Range	189.00	Interquartile Range	43.00

Quantile	Estimate
100% Max	839
99%	791
95%	777
90%	765
75% Q3	748
50% Median	725
25% Q1	705
10%	687
5%	679
1%	650
0% Min	650

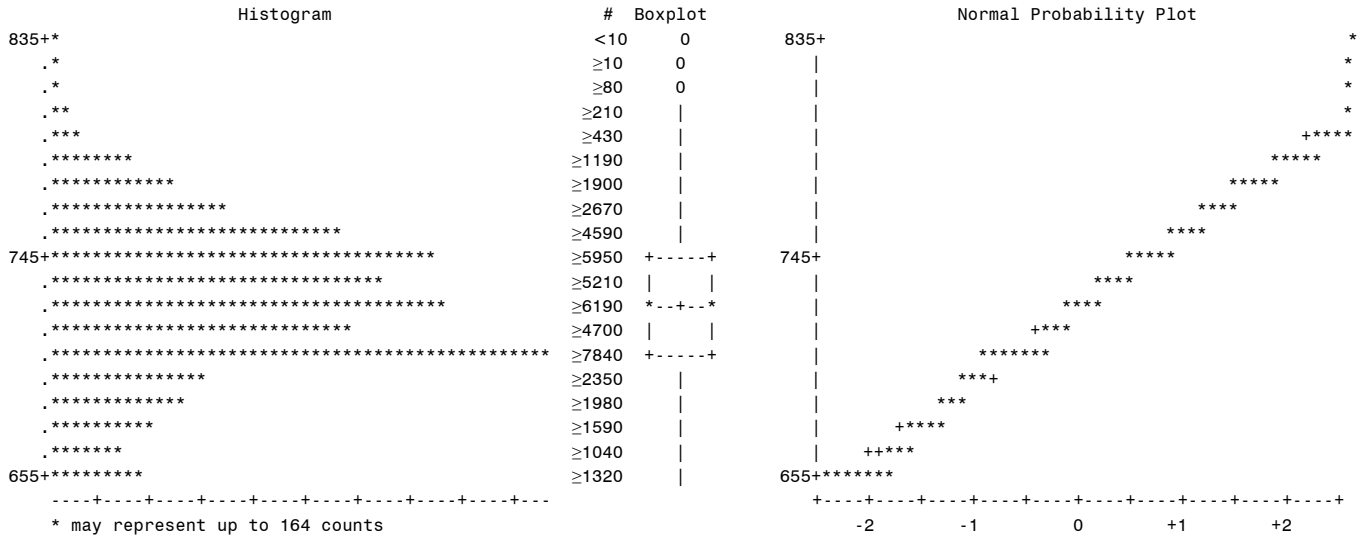


Table E.1.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 3

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥600	≥600	1.23	1.23
654	**********	≥720	≥1320	1.46	2.69
668	**********	≥1040	≥2370	2.12	4.80
679	**********	≥1590	≥3960	3.22	8.03
687	**********	≥1980	≥5940	4.03	12.06
694	**********	≥2350	≥8290	4.76	16.82
700	**********	≥2650	≥10950	5.38	22.20
705	**********	≥2610	≥13570	5.31	27.51
709	**********	≥2570	≥16140	5.21	32.72
714	**********	≥2420	≥18570	4.92	37.65
718	**********	≥2270	≥20840	4.61	42.26
721	**********	≥2280	≥23120	4.63	46.89
725	**********	≥2020	≥25150	4.11	51.00
728	**********	≥1880	≥27030	3.82	54.82
731	**********	≥1910	≥28950	3.89	58.70
734	**********	≥1690	≥30640	3.43	62.13
737	**********	≥1610	≥32250	3.27	65.40
740	**********	≥1610	≥33860	3.27	68.66
743	**********	≥1530	≥35400	3.11	71.77
746	**********	≥1420	≥36830	2.90	74.67
748	**********	≥1380	≥38210	2.80	77.47
751	**********	≥1260	≥39470	2.56	80.03
754	**********	≥1160	≥40640	2.36	82.39
757	**********	≥1100	≥41740	2.24	84.63
759	**********	≥1050	≥42800	2.14	86.77
762	**********	≥990	≥43800	2.02	88.80
765	**********	≥860	≥44660	1.76	90.55
768	**********	≥810	≥45470	1.64	92.20
771	**********	≥670	≥46150	1.37	93.57
774	**********	≥650	≥46810	1.34	94.90
777	**********	≥560	≥47370	1.15	96.05
780	**********	≥490	≥47870	1.00	97.05
784	**********	≥400	≥48270	0.82	97.87
787	**********	≥300	≥48570	0.61	98.47
791	**********	≥260	≥48830	0.54	99.01
796	**********	≥170	≥49010	0.35	99.36
801	**********	≥130	≥49140	0.27	99.63
806	**********	≥70	≥49220	0.16	99.79
812	**********	≥50	≥49270	0.11	99.89
819	**********	≥30	≥49300	0.06	99.96
828	**********	≥10	≥49320	0.03	99.99
839	**********	<10	≥49320	0.01	100.00

400 800 1200 1600 2000 2400

Frequency

Table E.2.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 4

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

N	≥48910		
Mean	733.86	Median	733.00
Std deviation	29.73	Variance	883.97
Skewness	-0.0546	Kurtosis	-0.0067
Mode	711.00	Std Error Mean	0.1344
Range	200.00	Interquartile Range	38.00

Quantile	Estimate
100% Max	850
99%	803
95%	782
90%	771
75% Q3	754
50% Median	733
25% Q1	716
10%	695
5%	687
1%	664
0% Min	650

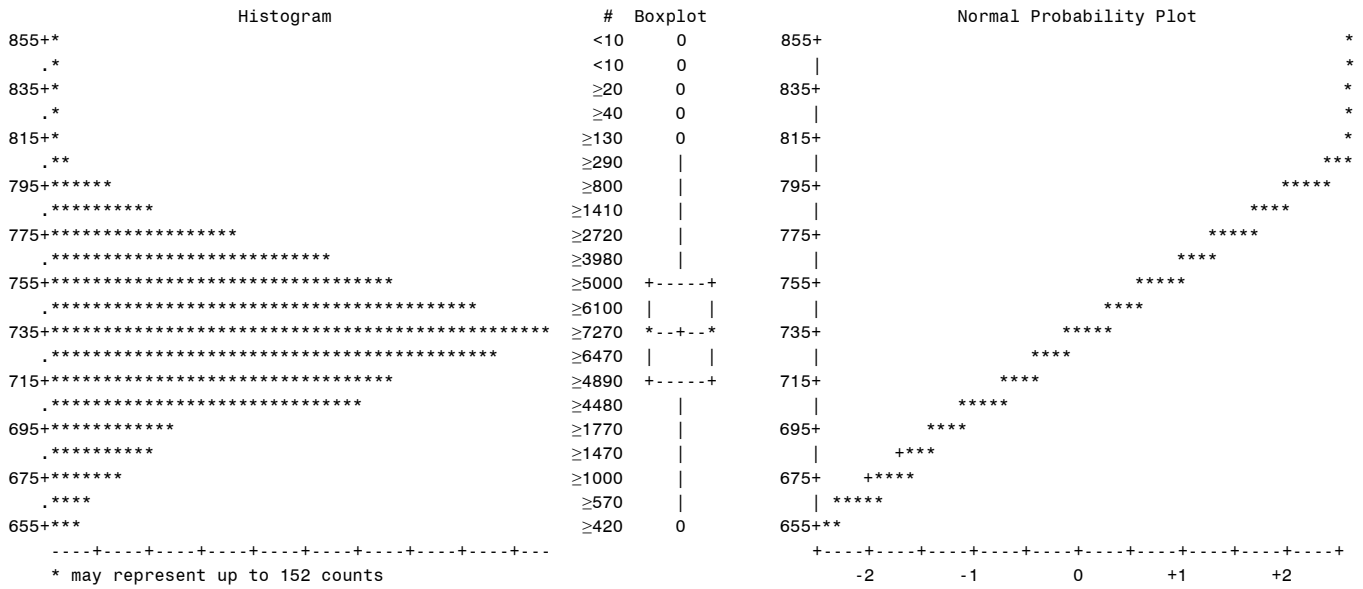


Table E.2.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 4

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	*********	≥420	≥420	0.87	0.87
664	**********	≥570	≥1000	1.18	2.05
678	********************	≥1000	≥2000	2.05	4.11
687	*****************************	≥1470	≥3470	3.01	7.11
695	**********************************	≥1770	≥5240	3.62	10.73
701	***	≥2170	≥7410	4.44	15.17
706	**	≥2310	≥9730	4.72	19.89
711	**	≥2480	≥12210	5.07	24.96
716	**	≥2410	≥14620	4.94	29.90
720	***	≥2240	≥16870	4.60	34.50
723	**	≥2110	≥18980	4.32	38.82
727	**	≥2110	≥21100	4.32	43.14
730	**********************************	≥1920	≥23020	3.93	47.07
733	**********************************	≥1890	≥24910	3.87	50.93
736	**********************************	≥1730	≥26640	3.54	54.47
739	**********************************	≥1730	≥28370	3.54	58.01
741	********************************	≥1650	≥30030	3.38	61.39
744	********************************	≥1580	≥31620	3.25	64.64
747	******************************	≥1500	≥33120	3.07	67.71
749	**************************	≥1360	≥34480	2.78	70.49
752	**************************	≥1370	≥35860	2.82	73.31
754	************************	≥1230	≥37090	2.53	75.84
757	************************	≥1190	≥38290	2.44	78.28
759	************************	≥1190	≥39480	2.43	80.71
761	**********************	≥1080	≥40560	2.21	82.92
764	**********************	≥1060	≥41620	2.17	85.09
766	********************	≥960	≥42590	1.98	87.08
769	*****************	≥870	≥43460	1.79	88.86
771	*****************	≥790	≥44250	1.62	90.48
774	*****************	≥740	≥45000	1.52	92.00
777	*************	≥630	≥45630	1.29	93.29
779	**********	≥550	≥46190	1.14	94.43
782	**********	≥530	≥46720	1.09	95.52
785	**********	≥480	≥47200	0.99	96.51
788	********	≥400	≥47600	0.82	97.33
792	*******	≥350	≥47960	0.72	98.05
795	*****	≥230	≥48190	0.48	98.53
799	****	≥210	≥48400	0.44	98.96
803	****	≥180	≥48580	0.37	99.33
807	**	≥110	≥48700	0.23	99.56
812	**	≥80	≥48780	0.16	99.73
818	*	≥50	≥48830	0.11	99.84
824	*	≥40	≥48880	0.09	99.93
832		≥20	≥48900	0.04	99.97
841		<10	≥48910	0.02	99.99
850		<10	≥48910	0.01	100.00

400 800 1200 1600 2000 2400

Frequency

Table E.3.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 5

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 05

N	≥48900		
Mean	727.90	Median	730.00
Std deviation	36.92	Variance	1363.25
Skewness	-0.0745	Kurtosis	-0.5081
Mode	698.00	Std Error Mean	0.1670
Range	200.00	Interquartile Range	54.00

Quantile	Estimate
100% Max	850
99%	804
95%	785
90%	774
75% Q3	756
50% Median	730
25% Q1	702
10%	676
5%	660
1%	650
0% Min	650

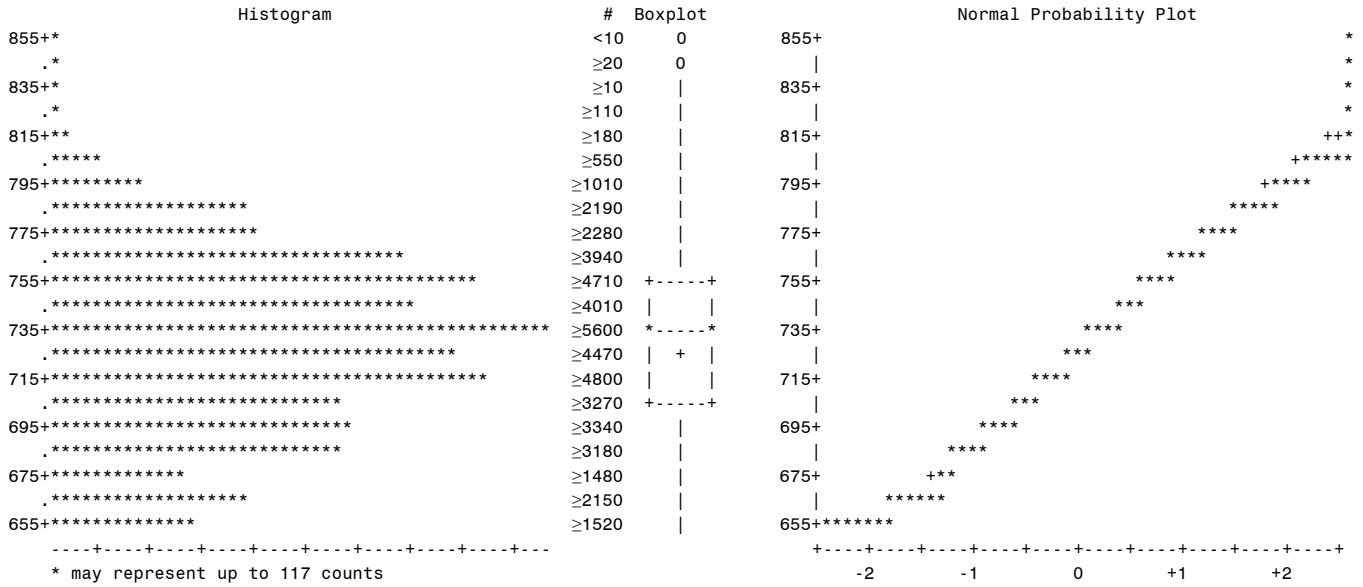


Table E.3.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 5

FREQUENCY DISTRIBUTION - SCALE SCORES
Science

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥1520	≥1520	3.12	3.12
660	*******	≥980	≥2510	2.01	5.13
668	**********	≥1170	≥3680	2.40	7.53
676	**********	≥1480	≥5160	3.03	10.56
682	**********	≥1550	≥6710	3.18	13.73
688	**********	≥1630	≥8350	3.35	17.08
693	**********	≥1610	≥9960	3.29	20.37
698	**********	≥1730	≥11690	3.55	23.92
702	**********	≥1580	≥13280	3.24	27.16
706	**********	≥1680	≥14960	3.45	30.60
710	**********	≥1670	≥16640	3.43	34.04
714	**********	≥1580	≥18230	3.24	37.28
718	**********	≥1530	≥19760	3.14	40.42
721	**********	≥1550	≥21310	3.17	43.59
724	**********	≥1460	≥22780	2.99	46.58
727	**********	≥1460	≥24240	2.99	49.57
730	**********	≥1430	≥25670	2.92	52.50
733	**********	≥1390	≥27070	2.86	55.36
736	**********	≥1390	≥28460	2.85	58.21
739	**********	≥1370	≥29840	2.82	61.03
742	**********	≥1360	≥31210	2.79	63.82
745	**********	≥1330	≥32540	2.73	66.55
747	**********	≥1310	≥33860	2.69	69.24
750	**********	≥1240	≥35100	2.54	71.78
753	**********	≥1230	≥36340	2.53	74.31
756	**********	≥1160	≥37500	2.37	76.68
758	**********	≥1070	≥38570	2.19	78.87
761	**********	≥1060	≥39640	2.18	81.06
764	**********	≥1000	≥40640	2.06	83.12
766	**********	≥950	≥41590	1.94	85.06
769	**********	≥920	≥42510	1.88	86.94
771	**********	≥860	≥43380	1.76	88.71
774	**********	≥760	≥44140	1.55	90.26
777	**********	≥650	≥44790	1.35	91.61
780	**********	≥640	≥45430	1.31	92.91
782	**********	≥560	≥46000	1.15	94.06
785	**********	≥530	≥46530	1.09	95.15
788	**********	≥460	≥46990	0.94	96.09
791	**********	≥380	≥47380	0.80	96.89
794	**********	≥330	≥47710	0.69	97.58
797	**********	≥280	≥48000	0.59	98.17
801	**********	≥230	≥48240	0.48	98.65
804	**********	≥170	≥48410	0.36	99.01
808	**********	≥140	≥48550	0.29	99.29
812	**********	≥100	≥48660	0.21	99.51
817	**********	≥70	≥48740	0.16	99.67
822	**********	≥70	≥48810	0.15	99.82
827	**********	≥40	≥48850	0.08	99.90
833	**********	≥10	≥48870	0.03	99.93
840	**********	≥10	≥48880	0.03	99.96
849	**********	≥10	≥48890	0.03	99.99
850	**********	<10	≥48900	0.01	100.00

100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700

Frequency

Table E.4.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 6

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

N	≥49300		
Mean	722.14	Median	721.00
Std deviation	33.65	Variance	1132.42
Skewness	0.1339	Kurtosis	-0.3152
Mode	697.00	Std Error Mean	0.1515
Range	200.00	Interquartile Range	50.00

Quantile	Estimate
100% Max	850
99%	800
95%	778
90%	766
75% Q3	747
50% Median	721
25% Q1	697
10%	680
5%	665
1%	650
0% Min	650

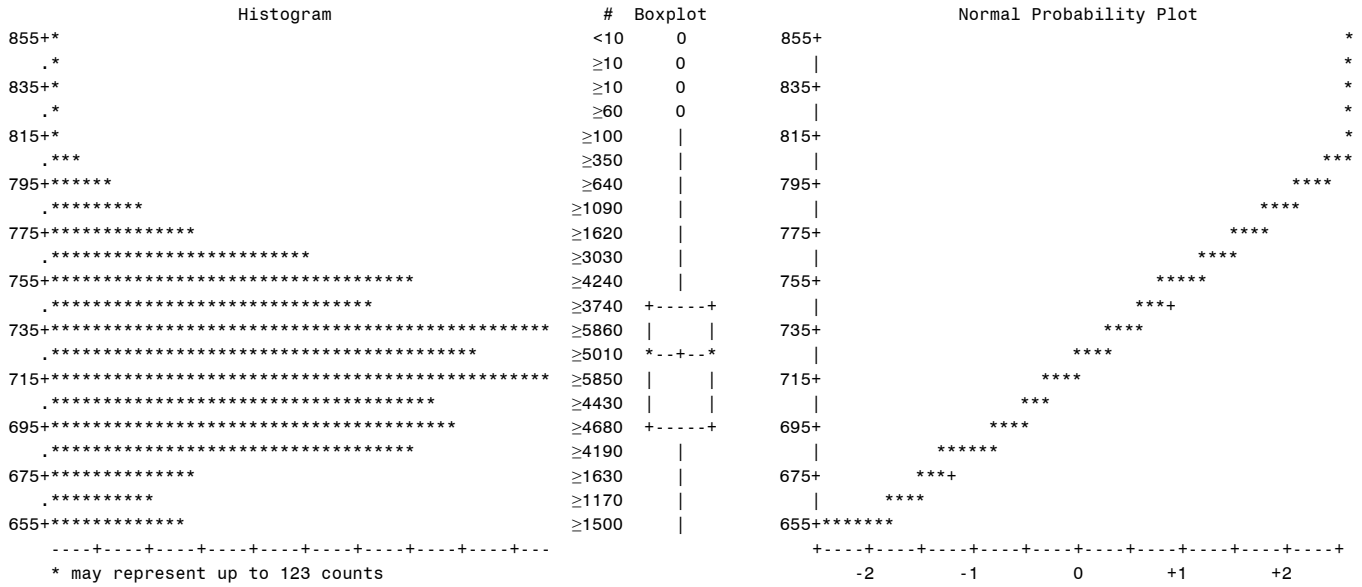


Table E.4.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 6

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥690	≥690	1.42	1.42
654	**********	≥800	≥1500	1.63	3.04
665	**********	≥1170	≥2670	2.38	5.43
673	**********	≥1630	≥4300	3.31	8.73
680	**********	≥1950	≥6250	3.96	12.69
687	**********	≥2230	≥8490	4.54	17.23
692	**********	≥2300	≥10800	4.68	21.92
697	**********	≥2370	≥13170	4.81	26.73
702	**********	≥2250	≥15430	4.57	31.30
706	**********	≥2170	≥17610	4.42	35.71
710	**********	≥2080	≥19690	4.23	39.94
714	**********	≥1910	≥21610	3.89	43.83
717	**********	≥1850	≥23460	3.76	47.59
721	**********	≥1680	≥25150	3.42	51.01
724	**********	≥1670	≥26820	3.40	54.41
727	**********	≥1650	≥28480	3.36	57.77
730	**********	≥1530	≥30020	3.12	60.89
733	**********	≥1480	≥31500	3.01	63.90
736	**********	≥1480	≥32990	3.01	66.91
739	**********	≥1360	≥34350	2.77	69.67
741	**********	≥1330	≥35680	2.70	72.37
744	**********	≥1220	≥36910	2.48	74.85
747	**********	≥1190	≥38100	2.41	77.27
750	**********	≥1140	≥39240	2.33	79.60
753	**********	≥1060	≥40310	2.16	81.75
755	**********	≥1060	≥41370	2.16	83.92
758	**********	≥960	≥42340	1.95	85.87
761	**********	≥850	≥43190	1.74	87.61
763	**********	≥810	≥44000	1.64	89.25
766	**********	≥700	≥44710	1.42	90.67
769	**********	≥660	≥45370	1.35	92.03
772	**********	≥620	≥46000	1.27	93.29
775	**********	≥520	≥46520	1.05	94.35
778	**********	≥480	≥47000	0.98	95.33
781	**********	≥410	≥47420	0.84	96.17
784	**********	≥350	≥47770	0.72	96.89
787	**********	≥320	≥48100	0.66	97.55
790	**********	≥250	≥48350	0.51	98.06
793	**********	≥210	≥48560	0.44	98.50
797	**********	≥170	≥48740	0.36	98.86
800	**********	≥150	≥48900	0.32	99.18
804	**********	≥100	≥49010	0.22	99.40
808	**********	≥80	≥49100	0.17	99.58
813	**********	≥60	≥49160	0.13	99.71
817	**********	≥30	≥49200	0.08	99.78
822	**********	≥40	≥49240	0.08	99.86
828	**********	≥20	≥49260	0.05	99.92
834	**********	≥10	≥49280	0.04	99.96
842	**********	≥10	≥49300	0.03	99.98
850	**********	<10	≥49300	0.02	100.00

400 800 1200 1600 2000
Frequency

Table E.5.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 7

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 07

N	≥50990		
Mean	730.41	Median	729.00
Std deviation	32.50	Variance	1056.40
Skewness	0.1434	Kurtosis	0.1210
Mode	720.00	Std Error Mean	0.1439
Range	200.00	Interquartile Range	41.00

Quantile	Estimate
100% Max	850
99%	812
95%	784
90%	773
75% Q3	751
50% Median	729
25% Q1	710
10%	688
5%	681
1%	653
0% Min	650

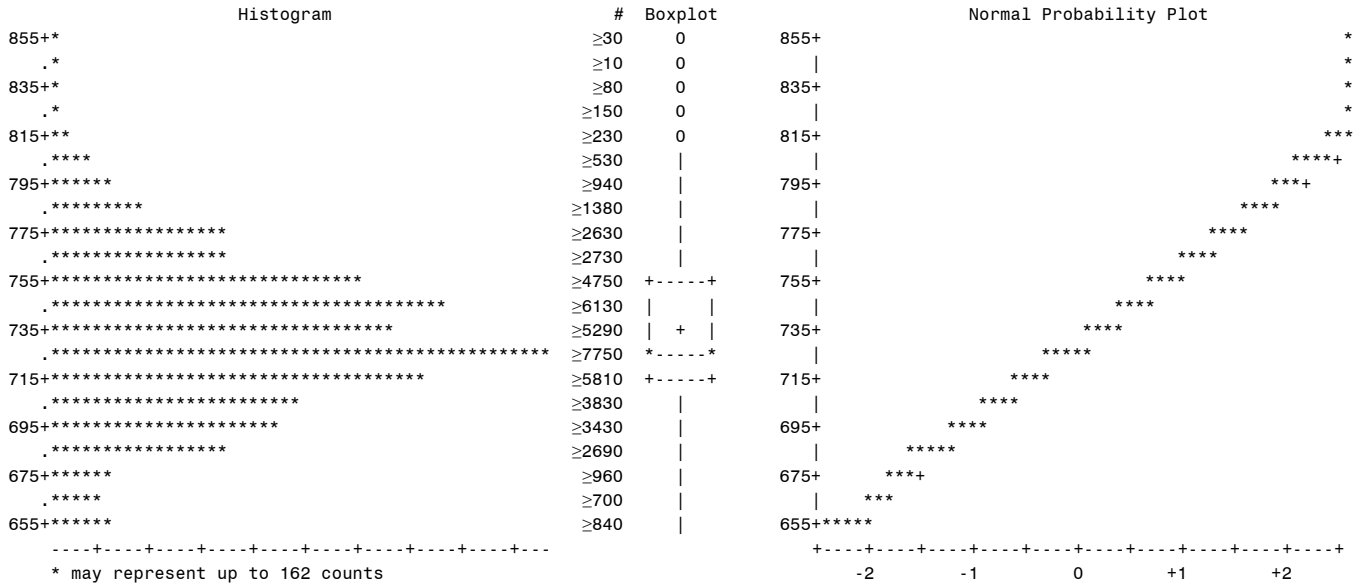


Table E.5.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 7

FREQUENCY DISTRIBUTION - SCALE SCORES
Science

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥400	≥400	0.79	0.79
653	**********	≥430	≥840	0.86	1.65
665	**********	≥700	≥1540	1.38	3.03
674	**********	≥960	≥2500	1.88	4.91
681	**********	≥1230	≥3730	2.42	7.33
688	**********	≥1450	≥5190	2.86	10.19
693	**********	≥1680	≥6880	3.30	13.49
698	**********	≥1750	≥8630	3.44	16.93
702	**********	≥1880	≥10510	3.69	20.63
706	**********	≥1950	≥12470	3.83	24.46
710	**********	≥1950	≥14420	3.82	28.28
713	**********	≥1940	≥16370	3.82	32.11
717	**********	≥1910	≥18280	3.76	35.86
720	**********	≥1980	≥20270	3.89	39.75
723	**********	≥1940	≥22210	3.81	43.57
726	**********	≥1970	≥24190	3.88	47.44
729	**********	≥1840	≥26040	3.62	51.07
732	**********	≥1830	≥27870	3.60	54.67
735	**********	≥1720	≥29600	3.39	58.06
737	**********	≥1730	≥31330	3.39	61.45
740	**********	≥1670	≥33000	3.28	64.73
743	**********	≥1590	≥34590	3.12	67.85
746	**********	≥1520	≥36120	2.99	70.84
749	**********	≥1340	≥37470	2.64	73.48
751	**********	≥1260	≥38740	2.49	75.97
754	**********	≥1280	≥40020	2.52	78.50
757	**********	≥1180	≥41210	2.33	80.82
759	**********	≥1010	≥42220	1.99	82.81
762	**********	≥990	≥43220	1.96	84.77
765	**********	≥910	≥44130	1.79	86.55
768	**********	≥820	≥44960	1.62	88.18
770	**********	≥750	≥45720	1.49	89.66
773	**********	≥690	≥46410	1.36	91.02
776	**********	≥610	≥47020	1.20	92.22
779	**********	≥570	≥47590	1.12	93.35
782	**********	≥510	≥48100	1.00	94.35
784	**********	≥470	≥48570	0.92	95.27
787	**********	≥400	≥48980	0.80	96.07
791	**********	≥330	≥49320	0.66	96.72
794	**********	≥320	≥49640	0.63	97.36
797	**********	≥280	≥49930	0.56	97.92
800	**********	≥200	≥50130	0.40	98.32
804	**********	≥170	≥50310	0.35	98.67
808	**********	≥150	≥50460	0.31	98.97
812	**********	≥130	≥50600	0.27	99.24
817	**********	≥100	≥50700	0.20	99.44
821	**********	≥90	≥50790	0.18	99.62
827	**********	≥60	≥50850	0.12	99.74
833	**********	≥40	≥50900	0.09	99.83
839	**********	≥30	≥50930	0.07	99.89
847	**********	≥10	≥50950	0.04	99.93
850	**********	≥30	≥50990	0.07	100.00

200 400 600 800 1000 1200 1400 1600 1800

Frequency

Table E.6.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational Science: Grade 8

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 08

N	≥50720		
Mean	730.81	Median	730.00
Std deviation	32.05	Variance	1027.45
Skewness	0.0122	Kurtosis	-0.2914
Mode	697.00	Std Error Mean	0.1423
Range	200.00	Interquartile Range	46.00

Quantile	Estimate
100% Max	850
99%	802
95%	781
90%	773
75% Q3	754
50% Median	730
25% Q1	708
10%	687
5%	682
1%	658
0% Min	650

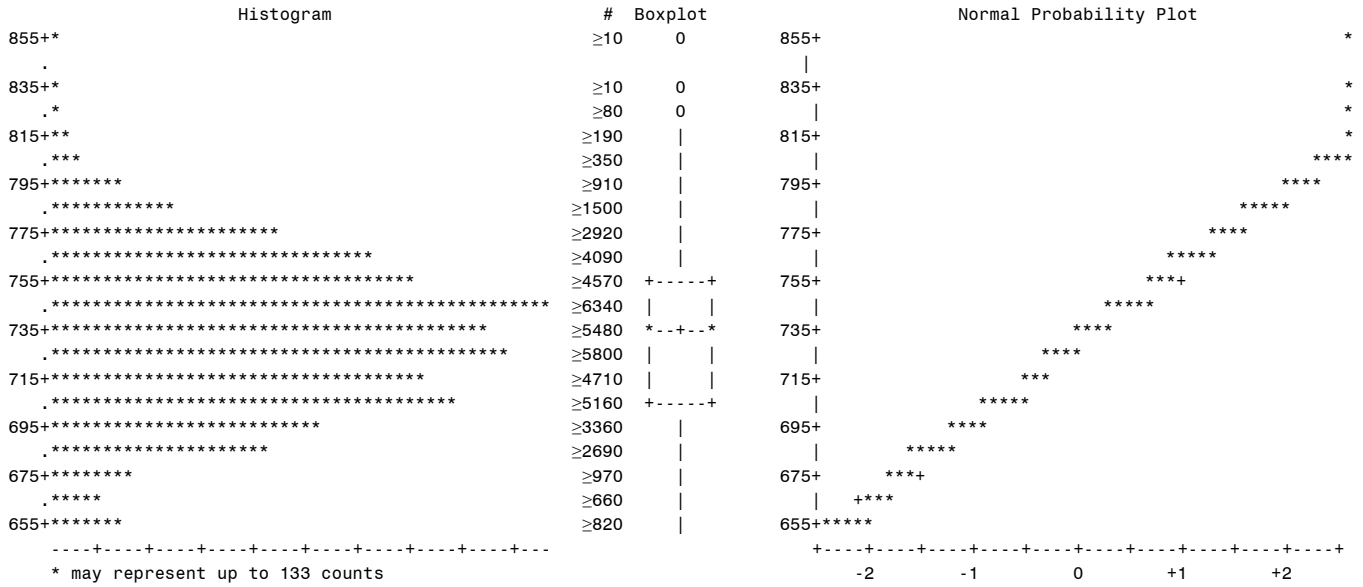


Table E.6.2

Frequency Distribution of Scale Scores: Spring 2022 Operational Science: Grade 8

FREQUENCY DISTRIBUTION - SCALE SCORES
Science

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥400	≥400	0.81	0.81
658	**********	≥410	≥820	0.83	1.63
668	**********	≥660	≥1480	1.30	2.94
675	**********	≥970	≥2460	1.93	4.87
682	**********	≥1160	≥3630	2.30	7.17
687	**********	≥1520	≥5160	3.01	10.18
692	**********	≥1610	≥6770	3.17	13.35
697	**********	≥1750	≥8530	3.47	16.82
701	**********	≥1720	≥10250	3.41	20.22
704	**********	≥1750	≥12010	3.45	23.68
708	**********	≥1680	≥13690	3.33	27.01
711	**********	≥1600	≥15290	3.15	30.16
714	**********	≥1570	≥16870	3.10	33.26
717	**********	≥1540	≥18410	3.04	36.31
720	**********	≥1480	≥19900	2.93	39.24
723	**********	≥1480	≥21380	2.92	42.16
725	**********	≥1430	≥22820	2.84	44.99
728	**********	≥1390	≥24210	2.75	47.74
730	**********	≥1430	≥25650	2.83	50.58
733	**********	≥1360	≥27020	2.70	53.27
735	**********	≥1320	≥28340	2.62	55.89
737	**********	≥1350	≥29700	2.68	58.57
740	**********	≥1260	≥30970	2.50	61.06
742	**********	≥1310	≥32290	2.60	63.66
744	**********	≥1290	≥33580	2.56	66.22
747	**********	≥1230	≥34820	2.43	68.65
749	**********	≥1220	≥36040	2.42	71.07
751	**********	≥1170	≥37210	2.31	73.38
754	**********	≥1130	≥38350	2.24	75.62
756	**********	≥1130	≥39490	2.24	77.86
758	**********	≥1130	≥40620	2.23	80.09
760	**********	≥1110	≥41730	2.21	82.29
763	**********	≥1010	≥42750	2.01	84.30
765	**********	≥1010	≥43770	2.00	86.30
768	**********	≥940	≥44710	1.85	88.15
770	**********	≥810	≥45520	1.61	89.76
773	**********	≥760	≥46280	1.50	91.26
775	**********	≥700	≥46990	1.40	92.66
778	**********	≥640	≥47630	1.26	93.92
781	**********	≥580	≥48220	1.15	95.08
784	**********	≥470	≥48690	0.93	96.01
787	**********	≥450	≥49140	0.89	96.89
790	**********	≥380	≥49520	0.75	97.64
794	**********	≥280	≥49810	0.56	98.20
798	**********	≥240	≥50050	0.49	98.69
802	**********	≥180	≥50240	0.36	99.06
806	**********	≥160	≥50400	0.33	99.38
811	**********	≥110	≥50520	0.23	99.61
816	**********	≥70	≥50590	0.14	99.76
822	**********	≥60	≥50650	0.12	99.88
829	**********	≥20	≥50680	0.06	99.93
838	**********	≥10	≥50700	0.04	99.97
850	**********	≥10	≥50720	0.03	100.00

200 400 600 800 1000 1200 1400 1600
Frequency

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports Science

Contents
Tables F.1.1-F.1.2 Reliability and SEM for Overall and Subgroups: Spring 2022 Operational SC G3-8
Table F.2 Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational SC G3-8
Table F.3 Classification Accuracy and Decision Consistency: Spring 2022 Operational SC G3-8

- Because the spring 2022 test was administered during the 2022 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table F.1.1

Reliability for Overall and Subgroups: Spring 2022 Operational Science

Grade	3	4	5	6	7	8
All Students	**0.853**	**0.868**	**0.888**	**0.841**	**0.851**	**0.891**
Female	0.844	0.858	0.878	0.826	0.837	0.884
Male	0.861	0.877	0.896	0.854	0.862	0.898
African American	0.785	0.802	0.862	0.783	0.79	0.849
American Indian or Alaska Native	0.839	0.852	0.863	0.807	0.811	0.889
Asian	0.869	0.888	0.878	0.876	0.88	0.91
Hispanic/Latino	0.832	0.858	0.881	0.827	0.847	0.89
Multi-Racial	0.843	0.861	0.877	0.823	0.849	0.884
Native Hawaiian or Other Pacific Islander	0.838	0.823	0.889	0.85	0.839	0.916
White	0.854	0.866	0.873	0.837	0.851	0.88
Economically Disadvantaged: No	0.857	0.868	0.871	0.835	0.856	0.882
Economically Disadvantaged: Yes	0.821	0.841	0.874	0.813	0.823	0.874
English Learner: No	0.854	0.868	0.887	0.84	0.85	0.89
English Learner: Yes	0.732	0.77	0.813	0.713	0.692	0.788
Regular Education	0.852	0.866	0.882	0.837	0.847	0.888
Special Education	0.825	0.836	0.878	0.797	0.789	0.838
Section 504: No	0.854	0.869	0.888	0.842	0.852	0.892
Section 504: Yes	0.824	0.855	0.876	0.814	0.823	0.869
Migrant: No	0.853	0.868	0.888	0.841	0.851	0.891
Migrant: Yes	0.826	0.875	0.864	0.836	0.785	0.88
Homeless: No	0.853	0.869	0.888	0.841	0.851	0.891
Homeless: Yes	0.812	0.836	0.87	0.81	0.83	0.885
Military Affiliation: No	0.852	0.868	0.887	0.84	0.85	0.89
Military Affiliation: Yes	0.856	0.869	0.874	0.844	0.846	0.89
Foster Care: No	0.853	0.868	0.888	0.841	0.851	0.891
Foster Care: Yes	0.786	0.848	0.86	0.804	0.793	0.866

Table F.1.2

SEM for Overall and Subgroups: Spring 2022 Operational Science

Grade	3	4	5	6	7	8
All Students	**3.23**	**3.31**	**3.66**	**3.86**	**3.88**	**3.76**
Female	3.23	3.30	3.71	3.89	3.87	3.76
Male	3.23	3.30	3.62	3.83	3.90	3.74
African American	3.15	3.19	3.42	3.64	3.77	3.66
American Indian or Alaska Native	3.27	3.39	3.73	3.88	3.92	3.73
Asian	3.29	3.36	4.00	4.08	3.93	3.69
Hispanic/Latino	3.19	3.27	3.59	3.79	3.86	3.72
Multi-Racial	3.27	3.35	3.78	3.92	3.90	3.78
Native Hawaiian or Other Pacific Islander	3.23	3.35	4.01	3.96	3.93	3.67
White	3.30	3.38	3.83	4.01	3.93	3.78
Economically Disadvantaged: No	3.31	3.39	3.90	4.07	3.92	3.77
Economically Disadvantaged: Yes	3.19	3.25	3.53	3.74	3.83	3.71
English Learner: No	3.24	3.32	3.67	3.88	3.88	3.76
English Learner: Yes	3.08	3.09	3.20	3.25	3.51	3.36
Regular Education	3.25	3.33	3.71	3.91	3.89	3.77
Special Education	3.12	3.12	3.24	3.33	3.56	3.46
Section 504: No	3.24	3.31	3.67	3.88	3.89	3.76
Section 504: Yes	3.18	3.24	3.51	3.64	3.78	3.67
Migrant: No	3.23	3.31	3.66	3.86	3.88	3.76
Migrant: Yes	3.13	3.24	3.36	3.82	3.81	3.68
Homeless: No	3.24	3.30	3.66	3.87	3.88	3.76
Homeless: Yes	3.13	3.18	3.44	3.66	3.79	3.67
Military Affiliation: No	3.24	3.30	3.67	3.86	3.88	3.76
Military Affiliation: Yes	3.30	3.37	3.85	4.03	3.89	3.75
Foster Care: No	3.24	3.31	3.66	3.86	3.88	3.76
Foster Care: Yes	3.22	3.26	3.37	3.74	3.74	3.61

Table F.2

Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational SC G3–8

Grade	Cronbach's Alpha	Marginal Reliability
3	0.85	0.86
4	0.87	0.86
5	0.89	0.90
6	0.84	0.87
7	0.85	0.89
8	0.89	0.90

Table F.3

Classification Accuracy and Decision Consistency: Spring 2022 Operational SC G3–8

Accuracy Matrix: Grade 3

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
3	1	0.13	0.04	0.00	0.00	0.00	0.17
	2	0.04	0.20	0.06	0.00	0.00	0.29
	3	0.00	0.07	0.20	0.05	0.00	0.32
	4	0.00	0.00	0.05	0.10	0.04	0.19
	5	0.00	0.00	0.00	0.01	0.02	0.03
	Total	0.17	0.30	0.31	0.16	0.06	1.00

Accuracy Matrix: Grade 4

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
4	1	0.12	0.03	0.00	0.00	0.00	0.15
	2	0.03	0.15	0.05	0.00	0.00	0.23
	3	0.00	0.06	0.22	0.06	0.00	0.33
	4	0.00	0.00	0.05	0.16	0.03	0.24
	5	0.00	0.00	0.00	0.01	0.03	0.05
	Total	0.15	0.24	0.32	0.23	0.07	1.00

Accuracy Matrix: Grade 5

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
5	1	0.17	0.03	0.00	0.00	0.00	0.20
	2	0.03	0.18	0.05	0.00	0.00	0.26
	3	0.00	0.05	0.13	0.05	0.00	0.23
	4	0.00	0.00	0.05	0.17	0.04	0.27
	5	0.00	0.00	0.00	0.02	0.03	0.04
	Total	0.20	0.26	0.23	0.24	0.07	1.00

Accuracy Matrix: Grade 6

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
6	1	0.21	0.04	0.00	0.00	0.00	0.26
	2	0.05	0.17	0.06	0.00	0.00	0.28
	3	0.00	0.06	0.13	0.05	0.00	0.24
	4	0.00	0.00	0.04	0.13	0.03	0.20
	5	0.00	0.00	0.00	0.01	0.01	0.02
	Total	0.27	0.28	0.23	0.19	0.04	1.00

Accuracy Matrix: Grade 7

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
7	1	0.13	0.03	0.00	0.00	0.00	0.16
	2	0.04	0.17	0.06	0.00	0.00	0.27
	3	0.00	0.06	0.19	0.05	0.00	0.31
	4	0.00	0.00	0.05	0.16	0.02	0.24
	5	0.00	0.00	0.00	0.01	0.02	0.02
	Total	0.17	0.27	0.30	0.23	0.04	1.00

Accuracy Matrix: Grade 8

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
8	1	0.10	0.02	0.00	0.00	0.00	0.13
	2	0.03	0.21	0.05	0.00	0.00	0.29
	3	0.00	0.05	0.20	0.05	0.00	0.30
	4	0.00	0.00	0.04	0.18	0.02	0.25
	5	0.00	0.00	0.00	0.01	0.02	0.03
	Total	0.13	0.29	0.29	0.24	0.05	1.00

Consistency Matrix: Grade 3

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
3	1	0.12	0.06	0.01	0.00	0.00	0.19
	2	0.04	0.15	0.07	0.01	0.00	0.27
	3	0.00	0.08	0.16	0.05	0.01	0.30
	4	0.00	0.01	0.06	0.08	0.03	0.18
	5	0.00	0.00	0.01	0.02	0.03	0.06
	Total	0.17	0.30	0.31	0.16	0.06	1.00

Consistency Matrix: Grade 4

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
4	1	0.11	0.05	0.01	0.00	0.00	0.17
	2	0.04	0.12	0.07	0.01	0.00	0.22
	3	0.00	0.07	0.17	0.06	0.00	0.31
	4	0.00	0.01	0.07	0.13	0.03	0.23
	5	0.00	0.00	0.00	0.03	0.03	0.07
	Total	0.15	0.24	0.32	0.23	0.07	1.00

Consistency Matrix: Grade 5

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
5	1	0.16	0.05	0.00	0.00	0.00	0.22
	2	0.04	0.14	0.06	0.01	0.00	0.25
	3	0.00	0.06	0.10	0.05	0.00	0.22
	4	0.00	0.01	0.06	0.14	0.04	0.25
	5	0.00	0.00	0.00	0.04	0.03	0.07
	Total	0.20	0.26	0.23	0.24	0.07	1.00

Consistency Matrix: Grade 6

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
6	1	0.20	0.07	0.01	0.00	0.00	0.28
	2	0.06	0.13	0.06	0.01	0.00	0.26
	3	0.01	0.07	0.09	0.05	0.00	0.23
	4	0.00	0.01	0.06	0.10	0.02	0.20
	5	0.00	0.00	0.00	0.02	0.01	0.04
	Total	0.27	0.28	0.23	0.19	0.04	1.00

Consistency Matrix: Grade 7

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
7	1	0.12	0.06	0.01	0.00	0.00	0.19
	2	0.04	0.13	0.08	0.01	0.00	0.26
	3	0.01	0.07	0.14	0.06	0.00	0.28
	4	0.00	0.01	0.07	0.14	0.02	0.24
	5	0.00	0.00	0.00	0.02	0.02	0.04
	Total	0.17	0.27	0.30	0.23	0.04	1.00

Consistency Matrix: Grade 8

Grade	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
8	1	0.10	0.04	0.00	0.00	0.00	0.14
	2	0.03	0.17	0.07	0.00	0.00	0.28
	3	0.00	0.07	0.16	0.06	0.00	0.29
	4	0.00	0.00	0.06	0.15	0.02	0.24
	5	0.00	0.00	0.00	0.02	0.02	0.05
	Total	0.13	0.29	0.29	0.24	0.05	1.00

Table F.3.1

Estimates of Accuracy and Consistency of Achievement Level Classification

Grade	Accuracy	Consistency	PChance	Kappa
3	0.650	0.538	0.237	0.395
4	0.672	0.560	0.233	0.425
5	0.674	0.569	0.223	0.445
6	0.647	0.543	0.238	0.400
7	0.662	0.551	0.239	0.410
8	0.716	0.610	0.242	0.485

Table F.3.2

Accuracy of Classification at Each Achievement Level

Grade	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
3	0.781	0.677	0.624	0.532	0.689
4	0.795	0.632	0.649	0.659	0.721
5	0.841	0.679	0.565	0.643	0.637
6	0.821	0.598	0.525	0.639	0.665
7	0.780	0.611	0.612	0.696	0.758
8	0.807	0.729	0.657	0.729	0.718

Table F.3.3

Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.926	0.873	0.898	0.947
4	0.935	0.886	0.893	0.954
5	0.933	0.896	0.898	0.941
6	0.900	0.871	0.898	0.967
7	0.920	0.868	0.894	0.972
8	0.945	0.897	0.907	0.966

Table F.3.4

Consistency of Dichotomous Categorizations by Form (PAC Metric)

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.891	0.824	0.855	0.929
4	0.905	0.842	0.850	0.934
5	0.903	0.855	0.857	0.920
6	0.857	0.822	0.856	0.954
7	0.884	0.818	0.851	0.962
8	0.920	0.857	0.869	0.951

Table F.3.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.647	0.647	0.595	0.335
4	0.655	0.668	0.643	0.468
5	0.715	0.709	0.668	0.368
6	0.646	0.641	0.595	0.321
7	0.615	0.632	0.627	0.442
8	0.674	0.706	0.683	0.476

Table F.3.6

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

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.037	0.057	0.054	0.043
4	0.030	0.052	0.057	0.033
5	0.032	0.049	0.049	0.043
6	0.046	0.063	0.058	0.028
7	0.035	0.065	0.056	0.022
8	0.025	0.047	0.050	0.024

Table F.3.7

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

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.037	0.070	0.048	0.010
4	0.035	0.062	0.050	0.013
5	0.035	0.055	0.053	0.016
6	0.054	0.065	0.045	0.005
7	0.044	0.067	0.050	0.006
8	0.030	0.055	0.043	0.010

Appendix G: Accommodated Print and Braille Creation

Guidelines for Accommodated Print and Braille

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

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

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

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

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

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

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

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

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

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

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

Appendix H: On-Going Quality Control

A system for monitoring, maintaining, and increasing the quality of its assessment system, including precise and technically sound criteria for the analyses of all of the assessments in its assessment system, is crucial and critical for keeping a high quality of assessments.

The places where information about monitoring, maintaining, and improving quality is incorporated are included in the following table.

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