LEAP 2025 Science 3–8
Technical Report: 2022–2023

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 2023 administration of the LEAP 2025 Science tests that 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 the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

The chapters of this technical report outline general information about the administration and scoring activities of the LEAP 2025 assessments, classical test theory (CTT) and item response theory (IRT) analysis results, 2023 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 2023 administration.

Table of Contents

1. Introduction

The Louisiana Department of Education (LDOE) has a long and distinguished history in the development and administration of assessments that support its state accountability system and are aligned to the Louisiana Student Standards. Per state law, the LDOE is to administer statewide science assessments in grades 3–8 and high school Biology annually. 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. 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 as part of the Louisiana Educational Assessment Program 2025 (LEAP 2025). This report outlines the testing procedures, including forms construction, administration, statistical analyses, scoring and analyses, and reporting of scores.

Summary of the 2022–2023 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 2023 Science grades 3–8 operational forms with embedded field test items (EFT). All tests were delivered in a computer-based format, with a paper-based option for grade 3. An accommodated paper-based format was available for students in grades 4–8 who are not physically able to test on a computer. Table 1.1 summarizes the key activities along with the months during which the activities were completed.

Table 1.1

Key Activities from August 2019 to May 2023

Date	Activity
August–December 2021	• Started item development planning for the spring 2023 test • The LDOE approved the item development plans, proposed bundles, and standalone item topics • WestEd updated the content development specifications, style guides, and training materials • WestEd developed outlines for the stimulus review committees and began standalone item development • The Technical Advisory Committee (TAC) meeting convened
January–February 2022	• The LDOE convened stimulus review committees • The LDOE provided feedback and approval to begin set/task development
March–June 2022	• WestEd led in item writing and development • LDOE staff reviewed proposed item sets, tasks, and standalones
July 2022	• WestEd and the LDOE convened Item Content/Bias Review Committees onsite in Baton Rouge • The LDOE and WestEd staff held reconciliation meetings
August–October 2022	• Content was finalized and the LDOE approved • Online content delivered to administration vendor • Conducted data review • Operational and field test forms were selected and the LDOE approved • The LDOE, WestEd, and DRC met for planning meeting
November–December 2022	• Fall 2022 test was administered • Frameworks were finalized and the LDOE approved • November TAC convened • Accommodated print/braille forms and alt text constructed, the LDOE approved, and delivered to administration vendor • The LDOE and WestEd staff reviewed proposed spring 2023 EFT selections in administration platform
February 2023	• The TAC convened
April 2023	• The LDOE, WestEd, and DRC met for planning meeting
May 2023	• Spring 2023 test was administered, including EFT

2. Assessment Frameworks

The assessment framework addresses:

- the test designs,
- test blueprints,
- range of standards to be covered,
- reporting categories,
- percentages of assessment items and score points by reporting category,
- projected testing times,
- the 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. 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 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 the 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. 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 address the same content.

The Assessment Frameworks were reviewed by the 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

Table 3.1 presents the acronyms used in item and test development. The test blueprints that guided item development projections for grades 3–8 is presented in Tables 3.2–3.7.

Table 3.1

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
ESS	Earth and Space Science
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
PS	Physical Science
Q/P	Asking Questions and Defining Problems

Table 3.1

Acronyms Used in Item and Test Development (continued)

Acronym	Meaning
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

Table 3.2

Test Blueprint for LEAP 2025: DCI Domain Coverage

Grade	Domain	LSSS PE #	LSSS Relative %	All Items % by Points
3	ESS	3	20%	15%–25%
	LS	8	53%	48%–58%
	PS	4	27%	22%–32%
	Total	15	100%	–
4	ESS	6	43%	38%–48%
	LS	2	14%	9%–19%
	PS	6	43%	38%–48%
	Total	14	100%	–
5	ESS	5	38%	33%–43%
	LS	2	15%	9%–20%
	PS	6	46%	41%–51%
	Total	13	100%	–
6	ESS	4	21%	15–26%
	LS	5	26%	21%–31%
	PS	10	53%	48%–58%
	Total	19	100%	–
7	ESS	4	25%	20%–35%
	LS	8	50%	45%–55%
	PS	4	25%	20%–35%
	Total	16	100%	–
8	ESS	7	37%	32%–42%
	LS	7	37%	32%–42%
	PS	5	26%	21%–31%
	Total	19	100%	–

Table 3.3

Test Blueprint for LEAP 2025: Minimal PE Coverage

Grade	PE*	SEP	CCC	Minimum Items
3	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
4	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 3.3

Test Blueprint for LEAP 2025: Minimal PE Coverage (continued)

Grade	PE*	SEP	CCC	Minimum Items
5	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
6	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

Table 3.3

Test Blueprint for LEAP 2025: Minimal PE Coverage (continued)

Grade	PE*	SEP	CCC	Minimum Items
6	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
7	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 3.3

Test Blueprint for LEAP 2025: Minimal PE Coverage (continued)

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

* *Note:* Every PE will be included at least one time in the test.

Table 3.4

Test Blueprint for LEAP 2025: CCC Coverage

Grade	CCC Overall	LSSS PE #	LSSS Relative %	CCC Items % by Points
3	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	–
4	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	–
5	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 3.4

Test Blueprint for LEAP 2025: CCC Coverage (continued)

Grade	CCC Overall	LSSS # in PEs	LSSS Relative %	CCC Items % by Points
6	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	–
7	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	–
8	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 3.5

Test Blueprint for LEAP 2025: SEP Coverage

Grade	SEP Overall	LSSS PE #	LSSS Relative%	SEP Items % by Points
3	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	–
4	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	–
5	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	0
	Total	13	100	–

Table 3.5

Test Blueprint for LEAP 2025: SEP Coverage (continued)

Grade	SEP Overall	LSSS PE #	LSSS Relative %	SEP Items % by Points
6	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	–
7	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	–
8	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 3.6

Test Blueprint for LEAP 2025: SEP Reporting Category Coverage

Grade	Reporting Category	LSSS PE #	LSSS Relative %	SEP Items % by Points	Min. Points
3	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	–	–
4	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	–	–
5	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	–	–
6	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	–	–
7	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	–	–
8	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	–	–

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 3.7

Test Blueprint for LEAP 2025 Grade 3–8: SEP Compared to CCC Ratio

Comparison	LSSS Relative Weight %	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 3.8 show the item development plans for the number of items developed by WestEd by reporting category for grades 3–8. There were no items developed for grades 3, 5 and 7.

Table 3.8

Number of New Items Developed for the Spring 2023 Field Test for Item Sets, Tasks, and Standalone Items

Grade	Development Type	Total Number of Sets or Tasks	1-pt SRs	1-pt TEs	2-pt TEs	TPD/ TPI	ER	CR	Total Number of Items (non-ER/CR)
4	Item Sets	5	16	15	3	11	0	5	50
	Tasks	0	0	0	0	0	0	0	0
	Standalone Items	n/a	7	3	1	1	0	0	12
6	Item Sets	4	15	9	15	5	0	4	34
	Tasks	0	0	0	0	0	0	0	0
	Standalone Items	n/a	4	5	2	5	0	0	16
8	Item Sets	1	18	7	14	5	0	4	48
	Tasks	0	0	0	0	0	0	0	0
	Standalone Items	n/a	6	2	6	2	0	0	16

The development plans also may include item sets and tasks that were revised and refield-tested. For spring 2023 field test, there were no revised and refield-tested item sets or tasks.

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 3.9 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 3.9
PE Bundling by Grade

Grade	Total Number of PE Bundles Approved	Number of Bundles Targeted for Development
3	18	3
4	21	3
5	22	0
6	19	4
7	23	0
8	21	4

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

Item sets were targeted for development for the 2022–2023 development cycle based on an analysis of the test bank for each grade. The development of item sets influenced the selection of phenomena. Like the tasks, the item sets are phenomena-based, but unlike the tasks, they are made up of independent items that do not 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. Table 3.10 shows the total number of CRs developed per grade.

Table 3.10

Constructed-Response Item Development by Grade

Grade	Number of CRs Developed
3	0
4	5
5	0
6	4
7	0
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 3.11 shows the number of phenomena submitted to the LDOE for item sets and tasks at grades 3–8.

Table 3.11

Phenomena Submitted by Grade

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

For the item sets, the LDOE identified four phenomena at grades 4, 5, 6, and 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 (2019–2023).

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 2022. 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; 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 content and bias review meetings. 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 3.12 provides the demographic characteristics of the review committee.

Table 3.12

Representation of Educators Participating in 2022–2023 Content and Bias Reviews

Grade Level	4	6	8
Classroom Teacher	3	4	4
Instructional Lead/Supervisor	2	1	2
School Administrator	0	1	0
Special Education Teacher	2	1	1
Visually or Hearing-Impaired Teacher	0	1	1
Other Staff	0	0	1
Black or African American	3	4	3
White	4	4	6
Male	1	1	1
Female	6	7	8
Total Participants	7	8	9

Before the committee members began the item review process, they received an orientation from the LDOE about the LEAP 2025 science assessments, and the WestEd content lead provided training on the criteria for evaluating items for content and bias considerations and the use of ABBI for item review. The committee members individually reviewed 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 bullet listed in the PE. An item was considered to have a lower degree of alignment if it aligned to another bullet listed in the learning progression for that SEP or CCC. Committee members also recorded whether the science for each item was accurate and whether each item was free

of bias. Areas of concern considered included opportunity and access, portrayal of groups represented, and protecting privacy and avoiding offensive content.

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

Following the individual reviewers’ votes, the group came together to view and discuss each stimulus and item as it was projected on-screen, with the goal of achieving consensus. The WestEd 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. Table 3.13 provides these results, based on the participants’ individual votes on each item following their initial review.

Table 3.12

Vote Totals Based on Individual Votes Following Initial Review

Grade	Item Type	*N* Items	Accept	Edits Accepted	No Vote	Reject	Total
4	CR	6	39	1	2	0	48
	ER	0	0	0	0	0	0
	MC	20	131	8	1	0	160
	MS	7	48	1	0	0	56
	TE	22	133	17	1	3	176
	TPD	8	50	6	0	0	64
	TPI	4	28	0	0	0	32
	All Grade 4	67	429	33	4	3	536
6	CR	4	27	0	5	0	36
	ER	0	0	0	0	0	0
	MC	15	111	8	0	0	134
	MS	4	32	0	0	0	36
	TE	31	232	12	4	0	279
	TPD	7	51	4	0	0	62
	TPI	3	24	0	0	0	27
	All Grade 6	64	477	24	9	0	574
8	CR	4	34	2	0	0	40
	ER	0	0	0	0	0	0
	MC	21	166	18	2	2	209
	MS	3	19	8	0	0	30
	TE	29	227	31	1	0	288
	TPD	4	32	4	0	0	40
	TPI	3	27	0	0	0	30
	All Grade 8	64	505	63	3	2	637

At the end of the meeting, consensus votes for each grade were compiled. There were no rejected items or item sets in any grade. All other items reviewed at each grade were either accepted as is or accepted with edits.

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 3.13 summarizes the disposition of field-tested items from data review.

Table 3.13

FT Item Dispositions by Item Type, 2023 Data Review

Grade	Item Type	Number of Items				
		Accept	Edits Accepted	Reject	Total	% of Total
3	CR	1	0	0	1	20
	MC	2	1	0	3	60
	MS	0	0	0	0	0
	TE	0	0	0	0	0
	TPI	0	1	0	1	20
	TPD	0	0	0	0	0
	Total	3	2	0	5	100
4	CR	6	1	0	7	13
	MC	6	7	0	13	24
	MS	7	0	0	7	13
	TE	10	3	0	13	24
	TPI	5	0	0	5	9
	TPD	5	4	0	9	17
	Total	39	15	0	54	100
5	CR	0	0	0	0	0
	ER	0	0	0	0	0
	MC	0	0	0	0	0
	MS	0	0	0	0	0
	TE	0	0	0	0	0
	TPI	0	0	0	0	0
	TPD	0	0	0	0	0
	Total	0	0	0	0	0

Table 3.13

FT Item Dispositions by Item Type, 2023 Data Review (continued)

Grade	Item Type	Number of Items				
		Accept	Edits Accepted	Reject	Total	% of Total
6	CR	2	2	0	4	7
	ER	0	0	0	0	0
	MC	10	3	2	15	25
	MS	1	3	0	4	7
	TE	15	13	0	28	46
	TPI	3	0	0	3	5
	TPD	3	2	2	7	11
	Total	34	23	4	61	100
7	CR	0	0	0	0	0
	ER	0	0	0	0	0
	MC	0	0	0	0	0
	MS	0	0	0	0	0
	TE	0	0	0	0	0
	TPI	0	0	0	0	0
	TPD	0	0	0	0	0
	Total	0	0	0	0	0
8	CR	2	0	0	2	3
	ER	0	0	0	0	0
	MC	12	6	4	22	37
	MS	2	0	0	2	3
	TE	16	12	0	28	47
	TPI	1	1	0	2	3
	TPD	2	2	0	4	7
	Total	35	21	14	60	100

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 grades 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 4.1 shows the Test Design for Science for grades 3–8.

Table 4.1
Test Design for Science

Grade	Session #	Test Session	Numbers of Items
3	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
	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

Table 4.1

Test Design for Science (continued)

Grade	Session #	Test Session	Numbers of Items
3	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
		OP Standalone Items	6 OP Standalone SR Items 1 OP Standalone TPD/TPI Items
	Total Items Field Tested Across Forms		1 FT Standalone SR Items 0 FT Standalone TPD/TPI Items 2 FT Item Set SR Items 1 FT Item Set TPD/TPI Items 1 Item Set CR Items
4	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
	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

Table 4.1

Test Design for Science (continued)

Grade	Session #	Test Session	Numbers of Items
4	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 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		6 FT Standalone SR items 6 FT Standalone TPD/TPI Items 11 FT Item Set SR Items 19 FT Item Set TPD/TPI Items 5 Item Set CR Items
5–8	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
	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

Table 4.1

Test Design for Science (continued)

Grade	Session #	Test Session	Numbers of Items
5–8	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
		OP Standalone Items	1 OP Standalone SR Item 2 OP Standalone TPD/TPI Items
	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
			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
5	Total Items Field Tested Across Forms		0 Items
6	Total Items Field Tested Across Forms		4 FT Standalone SR Items 7 FT Standalone TE Items 5 FT Standalone TPD/TPI Items 15 FT Item Set SR Items 22 FT Item Set TE Items 5 FT Item Set TPD/TPI Items 4 FT Item Set CR Items
7	Total Items Field Tested Across Forms		0 items
8	Total Items Field Tested Across Forms		7 FT Standalone SR Items 8 FT Standalone TE Items 2 FT Standalone TPD/TPI Items 17 FT Item Set SR Items 21 FT Item Set TE Items 5 FT Item Set TPD/TPI Items 4 FT Item Set CR Items

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

Initial Construction

The purpose of the spring 2023 forms construction activities was to create operational forms using the spring 2018, spring 2019, and spring 2022 field test items that were approved for operational use and to embed field test items in the spring 2023 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, 2019, or 2022 were available for use on the spring 2023 operational assessments.

For each of grades 3–8, WestEd completed item selection for one operational (OP) form for the spring 2023 administration. WestEd worked with the 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 the 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 4.2 and 4.3 provide the spring 2023 operational test composition for grades 3–8.

Table 4.2

LEAP 2025 Grades 3–4: Operational Test Composition

Grade	Item Sets / Item Types	Total Sets	Total Items per Set	Total Points per Set	SR	CR, 2-Part	Total Items	Total Points
3	4-Item Set	6	4	6	12	12	24	36
	Standalone Items	1	12	14	10	2	12	14
	Totals	–	–	–	22	14	36	50
4	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 4.3

LEAP 2025 Grades 5–8: Operational Test Composition

Grade	Item Sets / Item Types	Total Sets	Total Items per Set	Total Points per Set	SR, 1-pt TE	CR, 2-Pt TE, 2-part	ER	Total Items	Total Points
5	4-Item Set	5	4	6	10	10	0	20	30
	Standalone Items	1	12	16	0	0	0	12	16
	Task	1	5	15	2	2	1	5	15
	Totals	–	–	–	12	12	1	37	61
6	4-Item Set	5	4	6	10	10	0	20	30
	Standalone Items	1	12	16	0	0	0	12	16
	Task	1	5	15	2	2	1	5	15
	Totals	–	–	–	12	12	1	37	61
7	4-Item Set	5	4	6	10	10	0	20	30
	Standalone items	1	12	16	0	0	0	12	16
	Task	1	5	15	2	2	1	5	15
	Totals	–	–	–	12	12	1	37	61
8	4-Item Set	5	4	6	10	10	0	20	30
	Standalone Items	1	12	16	0	0	0	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 2023 varied by grade. These data are shown in Table 4.4.

Table 4.4
Spring 2023 Field Test Versions Administered by Grade

Grade	Number of Versions
3	1
4	12
5	0
6	12
7	0
8	12

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 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 grades 5-8, one item set and five standalone items were embedded in session 3.

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 the 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 on the operations forms including SR, CR, TPI, and TPD at grades 3 and 4; and SR, CR, TE, TPI, TPD, and ER at grades 5–8.

The answer keys for MC items were also 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. Criteria to flag items based on scoring point can be found in Table 4.4.

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 3–8 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.

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 4–8, 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 policies, 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 school systems. 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 systems. They disseminate information to each school, offer assistance with test administration, and serve as liaisons between the LDOE and their school systems. The LDOE also provides assistance with and 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, Data Recognition Corporation (DRC) produces two administration manuals: the *LEAP 2025 Grade 3 Paper-Based Test Administration Manual* (TAM) and the *LEAP 2025 Grades 3–8 Computer-Based Test Administration Manual* (TAM). The TAMs provide detailed instructions for administering the LEAP assessments. The manuals include information on test security, test administrator responsibilities, test preparation, administration of tests (computer-based or paper-based), and post-test procedures.

Table of Contents for *LEAP 2025 Paper-Based Test Administration Manual* (TAM)

- 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 Computer-Based Test Administration Manual* (TAM)

- 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 3–4 Social Studies Session 3 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

DRC also produces test coordinator manuals for paper- and computer-based test administrations. The TCMs 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 *LEAP 2025 Paper-Based Testing Test Coordinators Manual* (TCM)

- Key Dates
- 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
    - 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 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 DTC
- Void Notification
- Index

Table of Contents for *LEAP 2025 Computer-Based Testing Test Coordinators Manual* (TCM)

- Key Dates
- Resources Available in DRC INSIGHT Portal
- Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - DRC INSIGHT Portal 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 User Guide*
    - *LEAP 2025 Accommodations and Accessibility Features User Guide*
    - *INSIGHT Technology User Guide*
    - Online Tools Training (OTT)
    - Student Tutorials
- Void Notification

The LDOE assessment staff review, provide feedback, and give final approval for these manuals. The manuals are inclusive of grades 3–8 English Language Arts (ELA), Mathematics, Social Studies, and Science.

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

Standard 4.15. 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 TAMs 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 TAMs describe the following: general rules of paper and online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

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 TAMs. It should be noted that adhering to the test schedule is also a critical component. The TCMs included instructions for scheduling the test within the state testing window. The TAMs and TCMs also contained the schedule for timing each test session.

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

Department staff release annual test security reports that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; allowing students to correct or complete answers after tests have been submitted; splitting sessions into two parts; ignoring the standardized directions 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 Accommodation Plan/504 Plan (IAP), or English Learner Plan (EL plan); allowing accommodations for students who do not have an IEP, IAP, or EL plan; or defining terms on the test.

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

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

- Determine the layout of the classroom environment.
- Plan seating arrangements. Allow enough space between students to prevent the sharing of answers.
- Eliminate distractions such as bells or telephones.
- Use a Do Not Disturb sign on the door of the testing room.
- Make sure classroom maps, charts, and any other materials that relate to the content and processes of the test are covered or removed or are out of the students' view.

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

The TAMs 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 assessment, examiners are instructed to transcribe students' responses from the large-print or braille test book into the online testing system (INSIGHT) exactly as they responded in the large-print or braille test book.

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

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

Return Material Forms and Guidelines

The paper-based TCM instructs test coordinators regarding procedures for organizing and packing materials and returning them to DRC for secure inventory purposes. The LDOE assessment staff have opportunities to review, provide feedback, and give final approval of the guidelines. The purpose of the instructions is to ensure that secure test materials are properly accounted for and organized appropriately for the 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 barcodes on the tests match the packing list they received. The district test coordinator then packages the tests to be sent to schools. Upon returning 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 and administrators who receive the LEAP 2025 score reports. https://www.louisianabelieves.com/resources/search/assessment

Time

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

Online Forms Administration, Grades 3–8

The online forms are administered via DRC's INSIGHT online assessment system. School system and school personnel set up test sessions via DRC's INSIGHT portal and print test tickets. Students enter their ticket information to access the test in INSIGHT. In addition, students have access to the Online Tools Training (OTT) 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 at grade 3 had the option to participate in either paper-based or computer-based testing for the spring 2023 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 individual plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student information sheet at the time of test administration. AERA, APA, and NCME Standard 6.2 states:

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

In compliance with this standard, the 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 for online testers
- 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 testing window was available from April 25 through May 26, 2023. Paper-based testing occurred from April 26 through May 2, 2023.

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 TCMs and TAMs.

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 assessment and must account for all test materials and supervise the test administrators at all times.

Data Forensic Analyses

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

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

- Response Change Analysis

- Score Fluctuation Analysis
- Web Monitoring
- Plagiarism Detection
- Alerts for Disturbing Content

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

Score Fluctuation Analysis. It is anticipated that performance on the LEAP 2025 assessments will improve over time for legitimate reasons such as changes in the curriculum and improvement in instruction. However, large and unexpected score changes may be a sign of testing impropriety. The LDOE 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 are identified when the level of change is statistically unexpected.

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

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

Alerts for Disturbing Content. Scorers for the LEAP 2025 assessments also have the ability to apply an alert flag to student responses that may indicate disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to

themselves or others). All alerted responses are automatically routed to the scoring director, who reviews and forwards appropriate responses to senior project staff for review. If it is concluded that a response warrants an alert, project management will contact the LDOE to take the necessary action. At no point during this process do scorers or staff have access to demographic information for any students participating in the assessment.

6. Scoring Activities

Directory of Test Specification (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, subtype, session, alignment, 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 Social Studies 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 the 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. The scoring of MC and MS items is also evaluated at data review.

Scoring of Technology-Enhanced (TE) Items. All TE items are processed through DRC's autoscoring engine and scored according to the assigned scoring rules established 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 TE scoring process includes the following procedures:

- A scoring rubric is created for each TE item. The rubric describes the one and only correct answer for dichotomously scored items (i.e., items scored as either right or wrong). If partial credit is possible, the rubric describes in detail the type of response that could receive credit for each score point.
- The information from 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 TE and MS items. DRC provides a report listing the frequency distributions of TE item responses and MS items. Members of the LDOE and WestEd content staff examine the TE and MS response distributions and the auto-frequency reports to evaluate whether the items are scored appropriately. When scoring issues are identified, WestEd content staff and the LDOE recommend changes to the scoring algorithm. Any changes to the scoring algorithm are based on the LDOE's decisions. DRC, in turn, applies the approved scoring changes to any affected items.

Constructed-Response and Extended-Response Scoring

Constructed-response items are scored by human raters trained by DRC. Extended-response items are scored by Project Essay Grade (PEG), an Artificial Intelligence (AI) scoring engine. 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 2023 Handscoring/AI Documentation document.

Selection of Scoring Evaluators. Standard 4.20 states the following:

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

The following sections explain how scorers are selected and trained for the LEAP 2025 handscoring process and describe how the scorers are 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 2022–2023 LEAP 2025 test responses had at least a four-year college degree.

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

steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and 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 the handscoring process. When users log into DRC's secure, web-based scoring application, ScoreBoard, they are required to read and accept the security policy before they are allowed to access any project. For each project, scorers are also required to read and sign non-disclosure agreements, and during training emphasis is always given to what security means, the importance of maintaining security, and how this is accomplished.

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

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

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

Handscoring Training Process. Standard 6.9 specifies:

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

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

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

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

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

Any scorer who does not qualify by the end of the qualifying process for an item is not allowed to score live student responses.

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 review the statistics that are generated daily. DRC uses one team leader for every 10 to 12 readers. If scoring concerns are apparent among individual scorers or if a scorer needs clarification on the scoring rules, team leaders address those issues on an individual basis. DRC supervisors typically monitor one out of five of the scorer's readings, making adjustments to that ratio as needed. If a supervisor disagrees with a reader's scores during monitoring, the supervisor provides 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 provide to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducts validity scoring using the LDOE-approved validity responses identified by the DRC scoring supervisors during live scoring for newly operational items. Validity responses are inserted among the live student responses.

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

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

To monitor inter-rater reliability, DRC produces 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, and their definitions are included below.

- 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

Each reader is required to maintain a level of exact agreement on validity responses and on inter-rater reliability. Additionally, readers are required to maintain a low rate of nonadjacent agreement.

Calibration Sets. DRC pulls calibration responses for items. DRC uses these sets to perform calibration across the entire scorer population for an item if trends are detected (e.g., low agreement between certain score points if a certain type of response is missing from initial training). These calibrations are designed to help refocus scorers on how to properly use the scoring guidelines. They are selected to help illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After readers score a calibration set, the scoring director reviews it from the front

of the room, using rubric language and scoring concepts exemplified by the anchor responses to explain the reasoning behind each response's score.

Reports and Reader Feedback. Reader performance and intervention information are recorded in reader feedback logs. These logs track 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 for constructed- and extended-response items are scored independently by a second reader. This is the case regardless of whether the first reader is a human rater or AI. The statistics for inter-rater reliability are 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 is examined.

Tables 6.1–6.4 provide the inter-rater reliability and score point distributions by grade level for the constructed-response and extended-response items administered in the spring 2023 forms.

Table 6.1

Inter-Rater Reliability for Operational Constructed-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3**	Item 1	≥19,770	92	8	0
	Item 2	≥19,530	93	7	0
4	Item 1	≥18,310	90	10	0
	Item 2	≥17,500	92	8	0
5	Item 1	≥18,530	89	11	0
	Item 2	≥19,040	93	7	0
6	Item 1	≥16,050	88	12	0
	Item 2	≥21,260	91	9	0
7	Item 1	≥17,590	87	13	0
	Item 2	≥24,490	91	9	0
8	Item 1	≥16,380	87	13	0
	Item 2	≥17,410	85	15	0

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

** Grade 3 report combines both online and paper forms.

Table 6.2

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	≥61,350	51	21	6	6	17
	Item 2	≥61,280	40	25	13	7	16
4	Item 1	≥57,560	24	54	9	0	14
	Item 2	≥57,120	52	17	6	0	14
5	Item 1	≥57,070	46	33	4	0	15
	Item 2	≥57,510	38	37	8	0	16
6	Item 1	≥56,010	34	33	21	0	12
	Item 2	≥58,600	45	27	6	0	21
7	Item 1	≥57,270	27	39	20	0	14
	Item 2	≥60,770	28	35	11	0	26
8	Item 1	≥57,690	22	30	35	0	12
	Item 2	≥58,350	31	42	13	0	13

* 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.

*** Grade 3 report combines both online and paper forms.

Table 6.3

Inter-Rater Reliability for Operational Extended-Response Items

Grade	2x	Inter-Rater Reliability*			
		Dimension	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
5	≥44,850	Content	96	4	0
		Claim	96	4	0
6	≥46,480	Content	94	6	0
		Claim	93	7	0
7	≥43,980	Content	96	4	0
		Claim	96	4	0
8	≥60,720	Content	97	3	0
		Claim	97	3	0

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

Table 6.4

Score Point Distributions for Operational Extended-Response Items

Gr.	Score Point Distribution*								
	Total	Dimension	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	Blank (%)	Nonscore Codes (%)**
5	≥70,440	Content	43	28	7	1	0	0	20
		Claim	45	26	8	1	0	0	20
6	≥71,220	Content	36	32	9	2	0	0	21
		Claim	32	36	9	2	0	0	21
7	≥70,600	Content	29	37	10	2	0	0	20
		Claim	34	31	10	3	0	0	20
8	≥79,930	Content	14	33	24	8	1	0	18
		Claim	13	37	22	7	2	0	18

* 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 2023 test can be found in Appendix C: Item Analysis Summary Report. Tables C.1–C.4 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 also noted that a corrected point-biserial correlation indicates the correlation between an item score and the total test score, where the item score is not included in the total score. The results can be found in Tables C.2–C4. By the way, the statistical analysis results for operational and 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 (p <0.05) from 0.0 or is less than 1.0.
B (slight to moderate)	1. | *ΔMH DIF* | is significantly different (p <0.05) from 0.0 but not from 1.0, andis at least 1.0; OR 2. | *ΔMH DIF* | is significantly different (p <0.05) from 1.0 (p <0.05) 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 (p <0.05) 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 χP2P 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 (ERrmR) and the focal group (ERfmR). DIF is defined as DRmR = ERfmR – ERrmR, and SMD is a weighted average of DRmR using the weights wRmR = NRfmR (the number of students in the focal group with raw score m), which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

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

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

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

Table 7.2

DIF Categories for Polytomous Items

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

Three DIF analyses were conducted for 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.

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	38	[0],[1]	[0],[0]
7	36	[1],[1]	[0],[0]
8	39	[0],[0]	[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	36	[0],[0]	[0],[0]
5	36	[0],[1]	[0],[0]
6	39	[0],[0]	[0],[0]
7	37	[0],[1]	[0],[0]
8	38	[0],[1]	[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	37	[0],[0]	[0],[0]
6	38	[0],[1]	[0],[0]
7	37	[0],[1]	[0],[0]
8	39	[0],[0]	[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.

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 2023 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	2023 Spring Form C
99	791	787	791	790
95	775	773	777	773
90	765	762	765	765
85	760	755	759	757
80	755	750	751	753
75	750	745	748	748
70	745	740	743	743
65	742	734	737	738
60	737	731	734	733
55	734	725	731	730
50	731	722	725	727
45	728	719	721	721
40	722	715	718	718
35	719	712	714	714
30	715	703	709	710
25	712	698	705	705
20	703	693	700	700
15	698	687	694	694
10	693	679	687	687
5	679	669	679	663
1	650	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	2023 Spring Form C
99	798	798	803	809
95	782	779	782	789
90	774	770	771	776
85	766	762	764	770
80	764	756	759	765
75	758	751	754	757
70	753	748	749	754
65	751	742	747	749
60	748	739	741	744
55	743	734	739	739
50	740	731	733	737
45	737	725	730	732
40	734	721	727	729
35	728	718	723	723
30	725	712	720	720
25	722	707	716	717
20	716	703	711	714
15	708	695	701	706
10	704	690	695	701
5	690	678	687	690
1	668	651	664	672

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	2023 Spring Form C
99	807	807	804	813
95	788	785	785	791
90	776	773	774	779
85	768	765	766	771
80	762	760	761	763
75	757	752	756	758
70	752	747	750	752
65	747	742	745	745
60	745	737	739	739
55	740	735	733	737
50	735	729	730	731
45	732	723	724	725
40	726	717	718	719
35	723	714	714	715
30	717	707	706	708
25	714	703	702	700
20	707	694	693	695
15	698	689	688	690
10	689	677	676	677
5	677	671	660	670
1	654	650	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	2023 Spring Form C
99	797	794	800	793
95	779	776	778	773
90	769	767	766	763
85	763	758	758	756
80	758	753	753	749
75	753	749	747	745
70	749	744	741	740
65	744	739	736	735
60	742	734	730	730
55	736	731	727	725
50	734	725	721	722
45	728	722	717	716
40	725	719	714	713
35	722	716	706	709
30	719	709	702	706
25	712	704	697	698
20	709	700	692	693
15	704	695	687	687
10	695	683	680	681
5	683	676	665	664
1	657	650	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	2023 Spring Form C
99	809	805	812	802
95	786	783	784	783
90	775	770	773	774
85	767	762	765	765
80	759	754	757	760
75	754	748	751	754
70	751	743	746	750
65	746	740	743	745
60	743	735	737	740
55	737	732	735	735
50	735	726	729	730
45	729	723	726	728
40	726	717	723	722
35	723	714	717	716
30	717	711	713	713
25	714	707	710	706
20	707	699	702	702
15	703	695	698	694
10	695	690	688	689
5	685	679	681	677
1	662	651	653	650

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	2023 Spring Form C
99	803	799	802	802
95	784	778	781	785
90	773	768	773	774
85	766	761	765	767
80	761	756	758	759
75	756	750	754	755
70	752	745	749	750
65	747	743	744	745
60	743	738	740	740
55	741	733	735	737
50	736	729	730	732
45	731	726	728	727
40	729	721	723	724
35	723	718	717	717
30	721	712	711	714
25	715	708	708	710
20	708	701	701	706
15	705	697	697	698
10	697	687	687	693
5	682	675	682	680
1	658	650	658	662

Operational Item Parameters

Appendix C summarizes the distributions of item parameters and provides the 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 $ZQ_{1,crit}$ 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

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

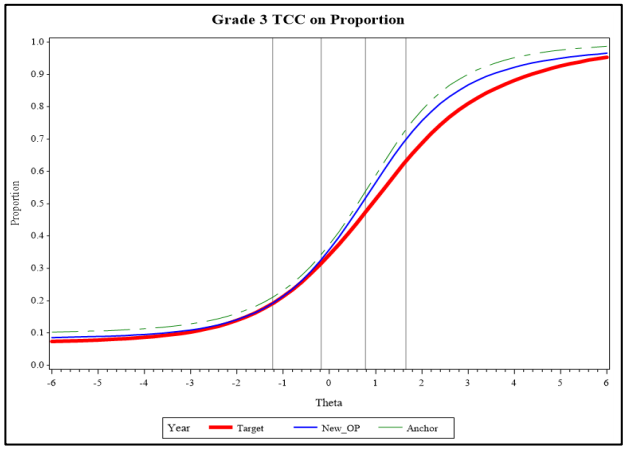
The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same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

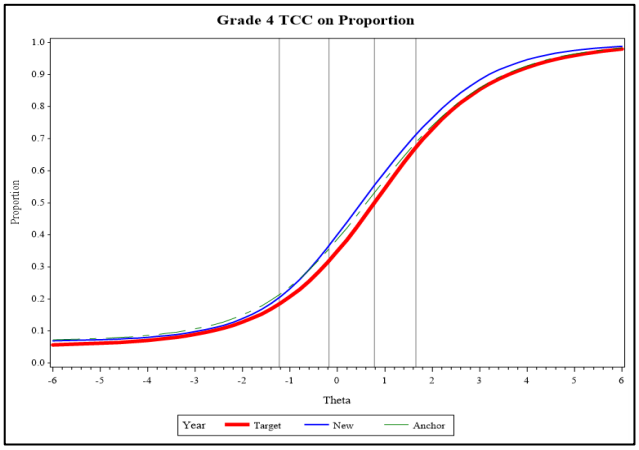
Test Characteristic Curve

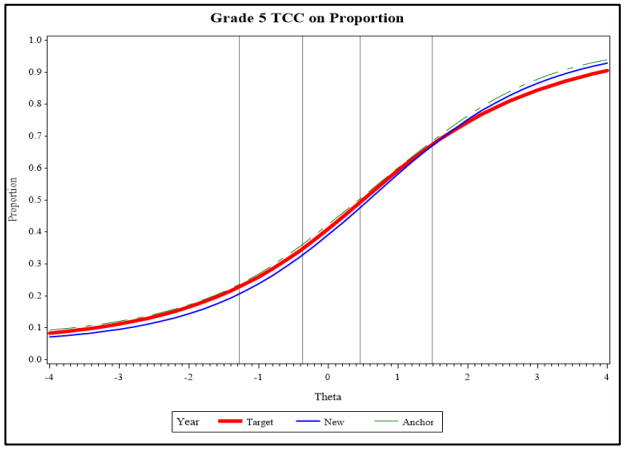
Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) across administrations of the LEAP 2025 Science assessments, as can be seen in the following figure. As seen from Plot 7.1 below, the TCCs between two years were similar across ability ranges. By the way, it should be noted that while the base form for all grades was the 2019 operational form, grade 4 used the 2022 operational test form. In addition, although the vertical lines are in theta scale, they indicate performance cuts. Each theta cut corresponding to the scale score of a performance-level cut (e.g., 704, 725, 750, and 778 for grade 4).

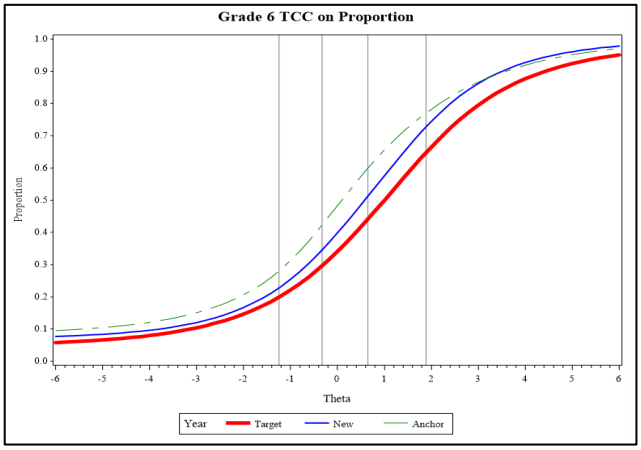
Plot 7.1

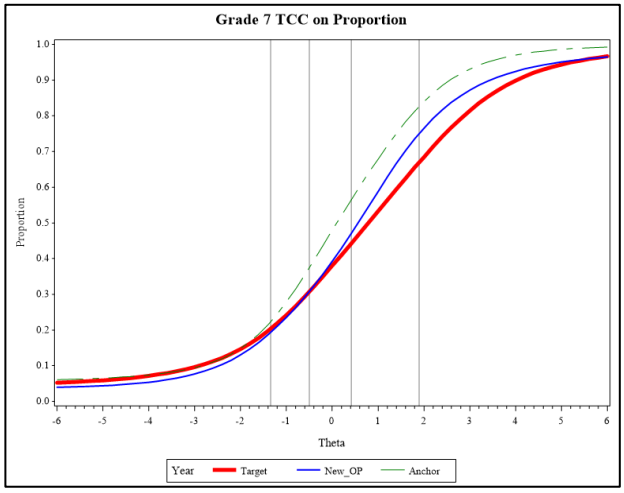
Test Characteristic Curve

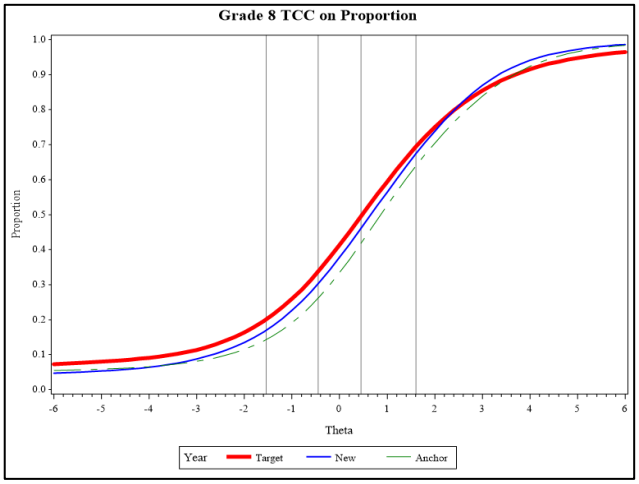












Test Information Curve, Score Distribution, and IRT Difficulty Distribution

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 2023 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 in Tables 7.5.1-7.5.6 and Plot 7.2. By the way, it should be noted that the base form of grade 4 was the 2022 operational test form. In addition, although the vertical lines are in theta scale, they indicate performance cuts. Each theta cut corresponding to the scale score of a performance-level cut (e.g., 704, 725, 750, and 778 for grade 4).

Table 7.5.1

SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 3

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
2.70	theta < -3.5	0
0.00	-3.5 <= theta < -3.0	0
2.34	-3.0 <= theta < -2.5	0
3.13	-2.5 <= theta < -2.0	0
4.08	-2.0 <= theta < -1.5	0
9.81	-1.5 <= theta < -1.0	0
15.16	-1.0 <= theta < -0.5	0
16.29	-0.5 <= theta < 0.0	3
15.85	0.0 <= theta < 0.5	8
14.60	0.5 <= theta < 1.0	8
8.70	1.0 <= theta < 1.5	8
4.51	1.5 <= theta < 2.0	6
2.19	2.0 <= theta < 2.5	1
0.46	2.5 <= theta < 3.0	0
0.09	3.0 <= theta < 3.5	1
0.10	3.5 <= theta	1
-6.00	Minimum	-0.50
5.57	Maximum	6.49
-0.21	Mean	1.08
1.33	SD	1.18
≥49,310	Total Number of	36

Table 7.5.2

SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 4

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.35	theta < -3.5	0
0.48	-3.5 <= theta < -3.0	0
0.00	-3.0 <= theta < -2.5	0
2.69	-2.5 <= theta < -2.0	0
5.05	-2.0 <= theta < -1.5	0
11.24	-1.5 <= theta < -1.0	0
15.43	-1.0 <= theta < -0.5	2
13.72	-0.5 <= theta < 0.0	6
15.38	0.0 <= theta < 0.5	8
15.52	0.5 <= theta < 1.0	7
8.49	1.0 <= theta < 1.5	3
6.26	1.5 <= theta < 2.0	5
3.77	2.0 <= theta < 2.5	4
0.93	2.5 <= theta < 3.0	1
0.45	3.0 <= theta < 3.5	0
0.21	3.5 <= theta	0
-6.00	Minimum	-0.95
5.32	Maximum	2.94
0.05	Mean	0.77
1.18	SD	0.99
≥48,870	Total Number of	36

Table 7.5.3

SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 5

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.61	theta < -3.5	0
0.81	-3.5 <= theta < -3.0	0
3.46	-3.0 <= theta < -2.5	0
2.61	-2.5 <= theta < -2.0	0
10.32	-2.0 <= theta < -1.5	0
10.73	-1.5 <= theta < -1.0	2
12.40	-1.0 <= theta < -0.5	2
14.02	-0.5 <= theta < 0.0	9
15.04	0.0 <= theta < 0.5	6
12.63	0.5 <= theta < 1.0	6
9.71	1.0 <= theta < 1.5	2
4.14	1.5 <= theta < 2.0	10
2.31	2.0 <= theta < 2.5	0
0.82	2.5 <= theta < 3.0	0
0.32	3.0 <= theta < 3.5	0
0.07	3.5 <= theta	0
-6.00	Minimum	-1.43
4.60	Maximum	1.90
-0.24	Mean	0.53
1.30	SD	0.94
≥48,320	Total Number of	37

Table 7.5.4

SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 6

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
1.67	theta < -3.5	0
1.37	-3.5 <= theta < -3.0	0
2.21	-3.0 <= theta < -2.5	0
6.64	-2.5 <= theta < -2.0	0
8.60	-2.0 <= theta < -1.5	0
13.40	-1.5 <= theta < -1.0	1
15.00	-1.0 <= theta < -0.5	4
14.98	-0.5 <= theta < 0.0	5
12.04	0.0 <= theta < 0.5	6
11.49	0.5 <= theta < 1.0	6
6.80	1.0 <= theta < 1.5	8
3.87	1.5 <= theta < 2.0	4
1.45	2.0 <= theta < 2.5	4
0.35	2.5 <= theta < 3.0	1
0.10	3.0 <= theta < 3.5	0
0.05	3.5 <= theta	0
-6.00	Minimum	-1.12
4.63	Maximum	2.85
-0.47	Mean	0.74
1.31	SD	0.99
≥48,300	Total Number of	39

Table 7.5.5

SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 7

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.60	theta < -3.5	0
0.77	-3.5 <= theta < -3.0	0
2.91	-3.0 <= theta < -2.5	0
5.02	-2.5 <= theta < -2.0	0
6.41	-2.0 <= theta < -1.5	2
13.30	-1.5 <= theta < -1.0	1
15.55	-1.0 <= theta < -0.5	1
13.93	-0.5 <= theta < 0.0	6
14.78	0.0 <= theta < 0.5	6
11.89	0.5 <= theta < 1.0	7
8.33	1.0 <= theta < 1.5	6
4.26	1.5 <= theta < 2.0	6
1.56	2.0 <= theta < 2.5	2
0.51	2.5 <= theta < 3.0	1
0.11	3.0 <= theta < 3.5	0
0.05	3.5 <= theta	0
-6.00	Minimum	-1.92
6.00	Maximum	2.64
-0.30	Mean	0.65
1.24	SD	1.05
≥48,900	Total Number of	38

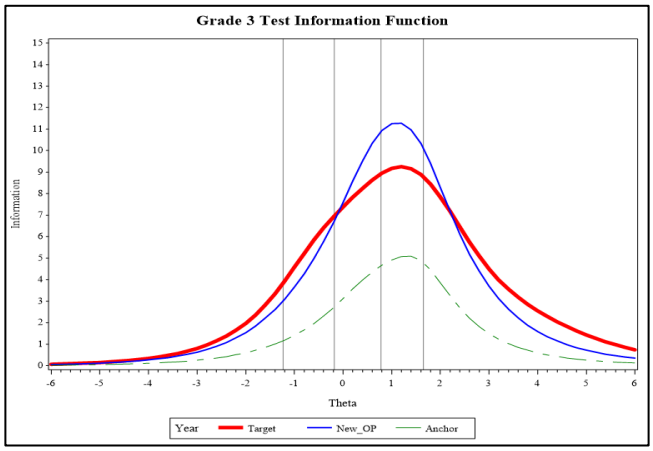
Table 7.5.6

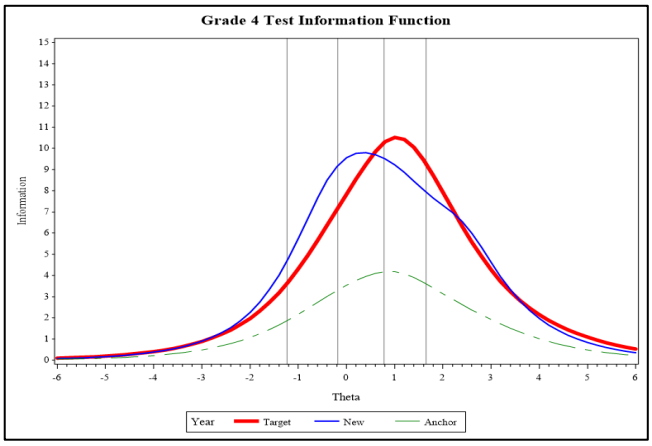
SPR 2023 Student's Score and IRT B-Parameter Distribution: Grade 8

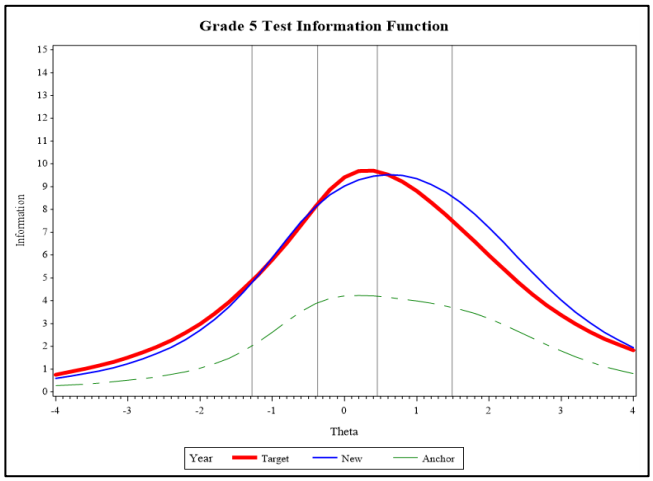
Percent of Students' Theta	Theta Range	Number of Items of IRT-B
0.28	theta < -3.5	0
0.41	-3.5 <= theta < -3.0	0
0.75	-3.0 <= theta < -2.5	0
3.65	-2.5 <= theta < -2.0	0
6.35	-2.0 <= theta < -1.5	1
12.02	-1.5 <= theta < -1.0	0
15.28	-1.0 <= theta < -0.5	5
19.02	-0.5 <= theta < 0.0	5
13.32	0.0 <= theta < 0.5	6
13.10	0.5 <= theta < 1.0	8
9.40	1.0 <= theta < 1.5	5
4.38	1.5 <= theta < 2.0	6
1.52	2.0 <= theta < 2.5	2
0.46	2.5 <= theta < 3.0	1
0.05	3.0 <= theta < 3.5	0
0.01	3.5 <= theta	0
-6.00	Minimum	-1.79
3.87	Maximum	2.75
-0.18	Mean	0.61
1.13	SD	0.99
≥50,160	Total Number of	39

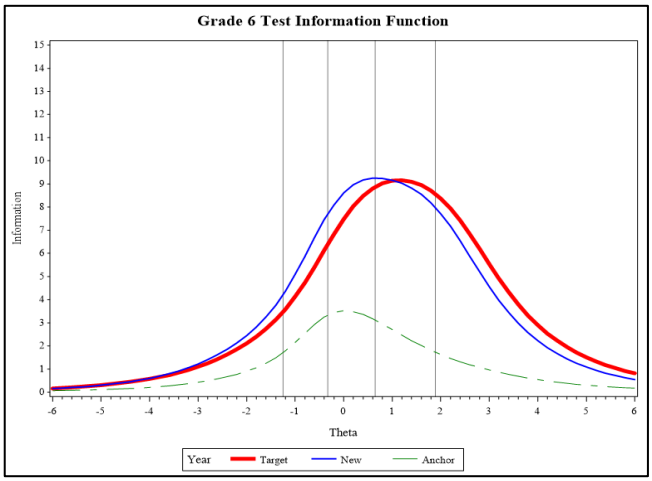
Plot 7.2

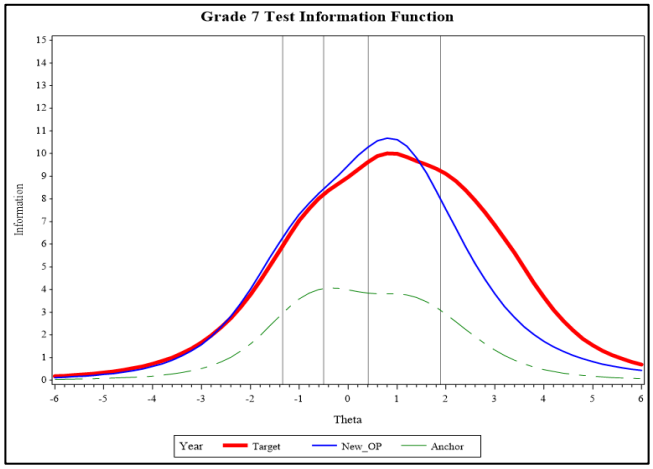
Test Information Curve

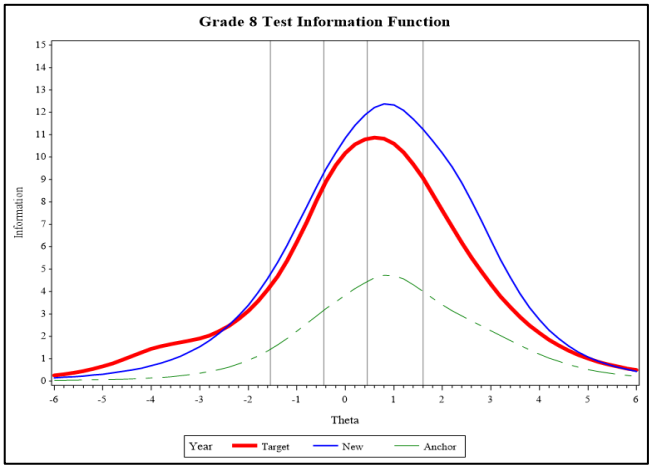












Field Test Data Review

The process used to complete the field test item equating is an anchored item equating process. In this process the item parameters from the 2023 operational items 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 Spring 2023 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 section provides the Spring LEAP 2025 Science test results including the scale score and performance levels. Presenting the results by performance level helps translate the numerical scale scores into descriptive categories reflecting student achievement levels (i.e., Level 1: Unsatisfactory, Level 2: Approaching Basic, Level 3: Basic, Level 4: Mastery, and Level 5: Advanced). Tables 8.1–8.6 present evidence of the score reliability and validity for the LEAP 2025 Science 3–8 tests.

Demographic Characteristics of Students

The operational Science tests were administered to all eligible students in the appropriate grade level during spring 2023. Grade 3 results combine both online and paper forms. Spring 2023 operational score results were based on the following student characteristics:

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

Test Results

For the spring 2023 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.

Table 8.1

LEAP 2025 State Test Results for Spring 2023: Grade 3

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
Total		≥49,310	725.29	30.80	17	33	27	17	6
Gender	Female	≥24,170	725.17	29.93	16	34	28	17	5
	Male	≥25,130	725.41	31.62	18	32	26	17	7
Race	African American	≥20,320	714.40	28.19	24	41	24	9	2
	AI/AN	≥270	728.48	28.79	12	32	31	19	5
	Asian	≥750	743.27	31.00	7	20	28	27	18
	Hispanic/Latino	≥5,260	719.06	29.28	22	37	25	13	3
	NHPI	≥50	729.87	28.67	11	30	33	24	2
	Two or More	≥1,950	729.51	29.07	13	30	31	20	6
	White	≥20,660	736.50	29.48	9	25	30	25	11
Economically Disadvantaged	No	≥14,550	740.06	29.63	8	21	30	28	13
	Yes	≥34,600	719.16	29.13	21	38	26	12	3
English Learner	No	≥46,400	726.41	30.74	16	32	28	18	6
	Yes	≥2,910	707.42	25.86	32	45	19	4	1
Education Classification	Regular	≥43,080	727.18	30.43	15	32	28	18	6
	Special	≥6,230	712.25	30.17	29	40	20	8	3
Section 504	No	≥45,650	725.75	30.94	17	32	27	17	6
	Yes	≥3,660	719.63	28.37	19	40	26	11	3
Migrant	No	≥49,210	725.31	30.81	17	33	27	17	6
	Yes	≥100	719.06	28.17	23	37	26	10	4
Homeless Status	No	≥48,150	725.64	30.78	17	33	27	17	6
	Yes	≥1,160	710.67	28.21	31	40	22	6	1
Military Affiliation	No	≥48,340	725.00	30.78	17	33	27	17	6
	Yes	≥970	739.56	28.26	8	20	33	28	11
Foster Care Status	No	≥49,140	725.32	30.81	17	33	27	17	6
	Yes	≥170	717.37	28.00	24	35	29	11	2

* Four students had invalid gender status. 24 students had missing ethnicity status. 165 students lacked economic status information; ** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander; *** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

Table 8.2

LEAP 2025 State Test Results for Spring 2023: Grade 4

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
Total		≥48,880	737.56	30.09	12	23	32	23	10
Gender	Female	≥23,870	736.16	29.13	12	24	33	22	9
	Male	≥25,000	738.90	30.91	12	22	31	24	11
Race	African American	≥20,120	725.90	26.07	18	32	33	14	3
	AI/AN	≥280	741.60	28.16	8	22	35	24	11
	Asian	≥820	757.10	31.14	4	13	25	32	27
	Hispanic/Latino	≥5,150	730.89	29.32	17	27	31	19	6
	NHPI	≥40	746.68	31.33	13	13	20	33	23
	Two or More	≥1,870	741.36	29.22	8	21	34	25	12
	White	≥20,540	749.45	28.97	6	14	31	32	17
Economically Disadvantaged	No	≥14,520	753.34	29.38	5	12	29	34	21
	Yes	≥34,050	730.98	27.79	15	28	33	18	5
English Learner	No	≥46,190	738.78	29.96	11	22	32	24	11
	Yes	≥2,680	716.55	23.86	28	39	25	7	1
Education Classification	Regular	≥42,840	739.85	29.56	10	22	33	24	11
	Special	≥6,030	721.29	28.73	26	35	24	11	4
Section 504	No	≥44,560	738.31	30.19	12	23	32	24	10
	Yes	≥4,310	729.83	27.84	16	29	33	16	5
Migrant	No	≥48,810	737.57	30.09	12	23	32	23	10
	Yes	≥60	730.82	29.33	18	26	28	21	7
Homeless Status	No	≥47,730	737.92	30.08	12	23	32	23	10
	Yes	≥1,140	722.50	26.20	23	34	29	12	2
Military Affiliation	No	≥47,960	737.28	30.07	12	23	32	23	10
	Yes	≥910	752.07	27.41	4	13	31	33	19
Foster Care Status	No	≥48,700	737.59	30.09	12	23	32	23	10
	Yes	≥170	727.37	29.43	23	22	30	19	5

* 20 students had missing ethnicity status. 304 students lacked economic status information.

** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

*** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

Table 8.3

LEAP 2025 State Test Results for Spring 2023: Grade 5

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
Total		≥48,330	729.44	37.83	22	22	24	23	9
Gender	Female	≥23,600	728.46	36.14	21	24	25	23	7
	Male	≥24,730	730.37	39.36	22	21	22	24	10
Race	African American	≥20,300	714.61	33.47	32	29	23	14	3
	AI/AN	≥250	736.23	35.02	15	19	27	28	10
	Asian	≥800	757.41	38.93	8	11	18	32	30
	Hispanic/Latino	≥5,170	722.17	37.48	28	23	23	20	6
	NHPI	≥30	735.50	35.25	12	26	21	35	6
	Two or More	≥1,680	736.03	36.46	15	19	27	28	10
	White	≥20,060	744.56	35.50	10	17	25	33	15
Economically Disadvantaged	No	≥14,660	748.80	35.76	9	14	23	35	19
	Yes	≥33,360	721.11	35.53	27	26	24	18	5
English Learner	No	≥46,090	730.96	37.55	20	22	24	24	9
	Yes	≥2,230	698.14	29.11	53	29	13	5	1
Education Classification	Regular	≥42,660	732.96	36.85	18	22	25	25	10
	Special	≥5,670	703.00	34.50	48	26	15	9	2
Section 504	No	≥43,590	730.48	37.88	21	22	24	24	9
	Yes	≥4,740	719.91	36.06	29	27	22	17	5
Migrant	No	≥48,260	729.45	37.83	22	22	24	23	9
	Yes	≥60	722.21	40.65	27	27	20	14	12
Homeless Status	No	≥47,220	729.86	37.81	21	22	24	24	9
	Yes	≥1,110	711.64	34.13	37	27	21	13	3
Military Affiliation	No	≥47,350	729.03	37.80	22	23	24	23	9
	Yes	≥970	749.25	34.13	8	14	25	34	19
Foster Care Status	No	≥48,180	729.48	37.84	22	22	24	23	9
	Yes	≥150	716.35	33.68	29	31	21	15	3

* 12 students had missing ethnicity status. 298 students lacked economic status information.

** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

*** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

Table 8.4

LEAP 2025 State Test Results for Spring 2023: Grade 6

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
Total		≥48,310	721.95	32.01	25	27	28	17	3
Gender	Female	≥23,580	720.92	30.58	25	29	28	16	2
	Male	≥24,720	722.94	33.28	25	25	27	19	3
Race	African American	≥20,380	709.69	28.22	37	33	23	7	1
	AI/AN	≥250	723.81	28.31	19	31	32	17	1
	Asian	≥730	747.94	32.20	7	15	26	38	14
	Hispanic/Latino	≥4,990	715.95	31.56	31	29	26	13	1
	NHPI	≥40	718.10	38.34	32	20	24	20	5
	Two or More	≥1,670	727.18	31.46	19	25	31	22	3
	White	≥20,210	734.41	30.39	13	22	33	28	4
Economically Disadvantaged	No	≥15,050	737.65	30.75	12	20	32	31	6
	Yes	≥32,990	714.93	29.95	31	31	26	12	1
English Learner	No	≥46,330	723.09	31.78	24	27	28	18	3
	Yes	≥1,970	695.16	24.89	57	31	10	2	NR
Education Classification	Regular	≥42,910	724.65	31.30	22	27	29	19	3
	Special	≥5,390	700.53	29.41	52	27	14	6	1
Section 504	No	≥43,290	722.88	32.05	24	27	28	18	3
	Yes	≥5,020	713.95	30.51	33	30	25	11	1
Migrant	No	≥48,240	721.96	32.01	25	27	28	17	3
	Yes	≥60	713.86	30.58	35	29	17	19	NR
Homeless Status	No	≥47,260	722.26	32.00	25	27	28	18	3
	Yes	≥1,040	708.15	29.09	40	31	22	7	1
Military Affiliation	No	≥47,380	721.63	31.97	25	27	28	17	2
	Yes	≥920	738.42	29.37	11	17	35	31	6
Foster Care Status	No	≥48,170	722.00	32.00	25	27	28	17	3
	Yes	≥130	704.10	29.81	48	27	19	7	NR

* 13 students had missing ethnicity status. 257 students lacked economic status information.

** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

*** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

Table 8.5

LEAP 2025 State Test Results for Spring 2023: Grade 7

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
Total		≥48,910	730.60	33.09	19	22	27	29	3
Gender	Female	≥23,690	731.69	31.82	17	23	29	29	3
	Male	≥25,210	729.59	34.20	21	22	26	28	3
Race	African American	≥20,470	719.22	29.63	27	29	27	16	1
	AI/AN	≥270	733.48	30.01	15	18	35	30	2
	Asian	≥790	756.14	34.99	8	9	18	51	14
	Hispanic/Latino	≥5,230	722.54	33.98	28	23	25	22	2
	NHPI	≥30	744.53	34.43	11	17	28	39	6
	Two or More	≥1,670	735.53	32.64	15	20	29	32	4
	White	≥20,400	742.65	31.27	10	16	28	41	5
Economically Disadvantaged	No	≥15,380	746.94	31.33	8	14	26	46	6
	Yes	≥33,250	723.18	31.11	24	26	27	21	1
English Learner	No	≥46,790	732.03	32.64	18	22	28	30	3
	Yes	≥2,120	699.14	26.46	54	28	14	4	NR
Education Classification	Regular	≥43,750	733.42	32.29	16	22	28	31	3
	Special	≥5,150	706.68	29.97	45	29	17	9	1
Section 504	No	≥43,680	731.88	33.12	18	22	27	30	3
	Yes	≥5,220	719.91	30.82	28	28	26	17	1
Migrant	No	≥48,840	730.62	33.08	19	22	27	29	3
	Yes	≥70	720.45	36.18	35	15	27	20	3
Homeless Status	No	≥47,880	730.95	33.05	19	22	27	29	3
	Yes	≥1,020	714.32	30.42	33	30	22	14	1
Military Affiliation	No	≥48,030	730.26	33.02	19	23	27	28	3
	Yes	≥870	749.46	30.90	7	12	25	49	7
Foster Care Status	No	≥48,760	730.65	33.08	19	22	27	29	3
	Yes	≥140	714.15	31.42	33	27	27	11	1

* 25 students had missing ethnicity status. 271 students lacked economic status information.

** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

*** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

Table 8.6

LEAP 2025 State Test Results for Spring 2023: Grade 8

Category*	Subgroup**	Scale Score			% at Performance Level***				
		N	Mean	SD	1	2	3	4	5
TOTAL		≥50,160	732.49	31.12	11	31	26	25	6
Gender	Female	≥24,810	732.61	29.59	10	31	29	25	5
	Male	≥25,350	732.36	32.54	13	30	24	25	7
Race	African American	≥21,430	720.02	27.07	17	42	25	14	2
	AI/AN	≥270	735.72	28.21	6	28	33	25	6
	Asian	≥750	757.80	31.08	3	13	18	40	26
	Hispanic/Latino	≥5,080	724.78	31.30	18	33	24	20	4
	NHPI	≥40	749.32	26.03	5	10	34	41	10
	Two or More	≥1,710	738.45	29.37	7	27	29	30	8
	White	≥20,860	745.69	29.06	5	19	28	37	11
Economically Disadvantaged	No	≥16,090	747.79	29.42	4	18	27	38	13
	Yes	≥33,750	725.36	29.21	15	37	26	19	3
English Learner	No	≥48,180	733.69	30.80	10	30	27	26	7
	Yes	≥1,980	703.19	23.64	38	45	12	4	NR
Education Classification	Regular	≥45,030	735.09	30.52	9	29	28	27	7
	Special	≥5,130	709.58	26.60	29	46	16	7	1
Section 504	No	≥44,790	733.58	31.14	11	30	27	26	7
	Yes	≥5,370	723.38	29.36	15	40	25	16	4
Migrant	No	≥50,100	732.50	31.11	11	31	26	25	6
	Yes	≥60	723.41	31.74	14	41	24	17	3
Homeless Status	No	≥49,160	732.78	31.10	11	31	26	25	7
	Yes	≥1,000	718.31	28.66	20	43	22	12	3
Military Affiliation	No	≥49,260	732.16	31.09	12	31	26	25	6
	Yes	≥890	750.29	27.48	3	17	26	41	13
Foster Care Status	No	≥50,010	732.52	31.12	11	31	26	25	6
	Yes	≥150	720.30	26.35	16	43	25	15	1

* Three students had missing ethnicity status. 320 students lacked economic status information.

** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

*** Level 1 = Unsatisfactory. Level 2 = Approaching Basic. Level 3 = Basic. Level 4 = Mastery. Level 5 = Advanced. The overall performance level may not add up to 100% due to rounding.

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 8.7. 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 8.7

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 0.861 and 0.897, the marginal alpha values were between 0.85 and 0.91 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

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

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

Additional reliabilities were calculated on various 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 2023 Science between 3.37 and 3.95, 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 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

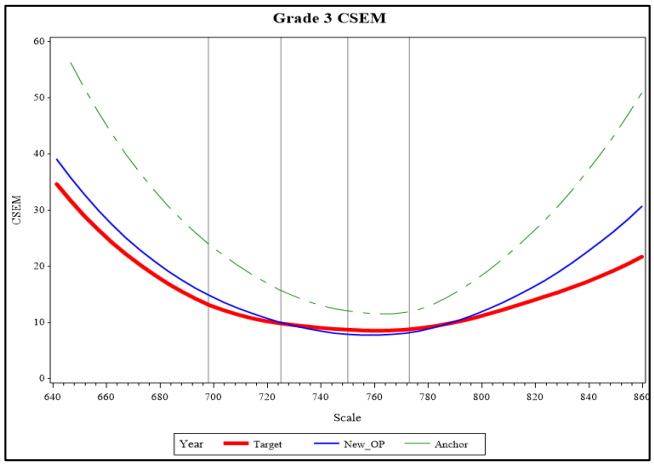
Conditional Standard Errors of Cut Scores: Spring 2023 LEAP

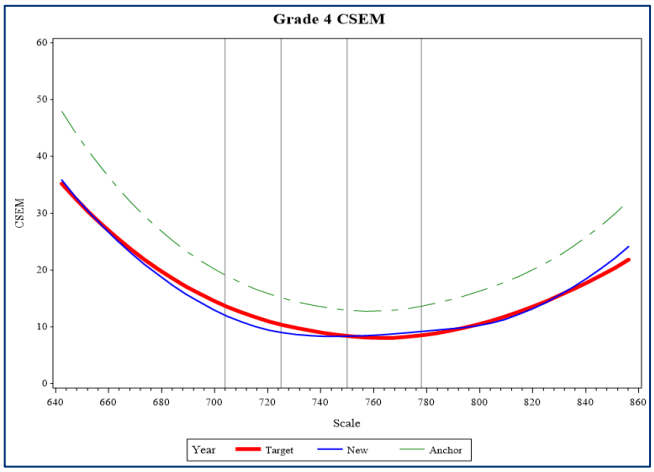
Grade	*Approaching Basic*		*Basic*		*Mastery*		*Advanced*	
	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
3	698	14	725	10	750	8	773	8
4	704	12	725	9	750	8	778	9
5	698	13	725	11	750	10	781	10
6	701	12	725	9	750	8	782	9
7	702	11	725	9	750	8	790	9
8	694	13	725	9	750	8	782	8

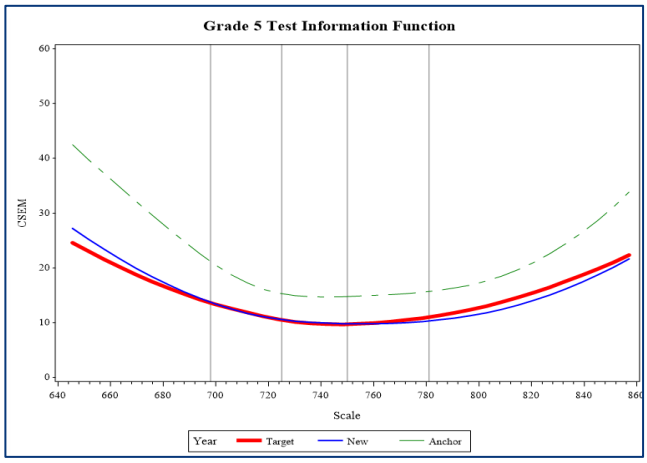
IRT methods are used for estimating CSEM and are presented in the following graph. With fixed-form assessments, the estimates of measurement error tend to be higher at the low and high ends of the scale-score range (i.e., theta-scale range), where few items measure the ability levels. Generally, there 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 below demonstrates that irrespective of grades, the tests are designed to minimize measurement error in the middle of the scale-score range, where the majority of students are located.

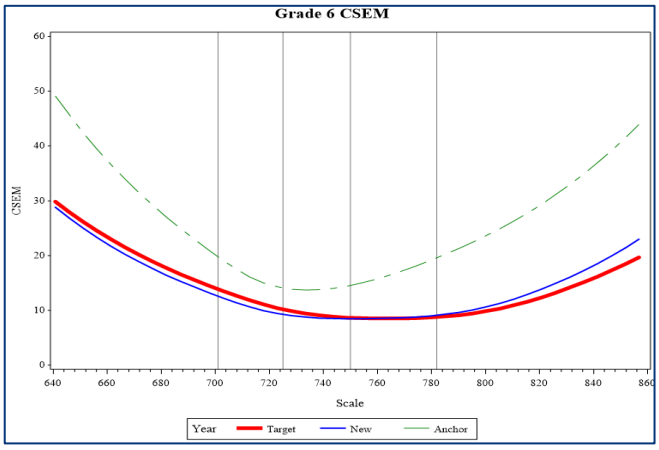
Plot 9.1

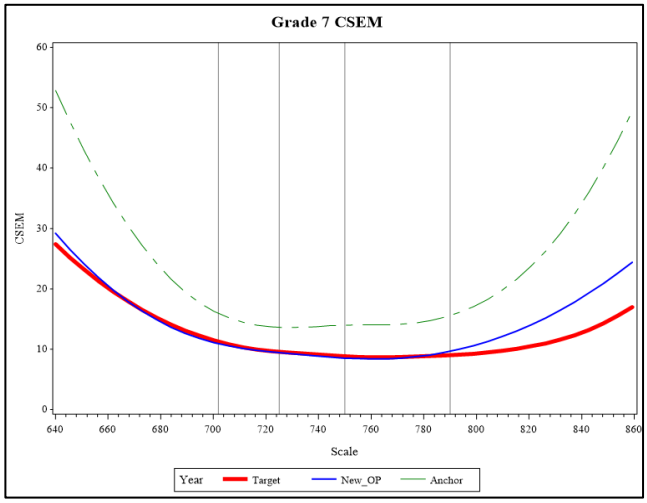
CSEM Curves

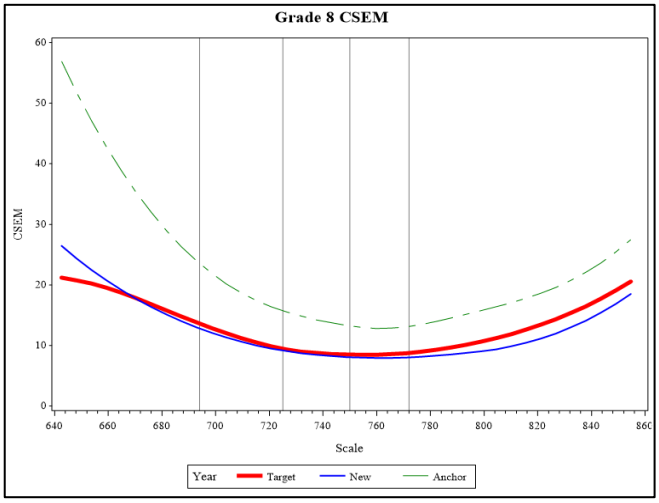












Student Classification Accuracy and Consistency

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 0.648 and 0.721, the overall consistency indices were 0.542 and 0.618 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 0.215 and 0.244 for the 2023 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 0.400 and 0.495 for the 2023 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 2022–2023 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 2023 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 2023 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 2023 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 2023 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 2023 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 2023 Science tests.

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 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.

Mode Effect Study

It is important to evaluate fairness in test administration in addition to evaluating fairness by examining performance among subgroups. Since two modes (i.e., paper-based tests and computer-based tests) were administered for grade 3, the following techniques (i.e., mode effect analysis and equating) were applied to operational test data to investigate the item mode effect. The mode effect analysis has been conducted, and the results indicate no items exhibiting C category DIF, suggesting no mode effect between online and paper tests; *all items* exhibited *A category DIF*.

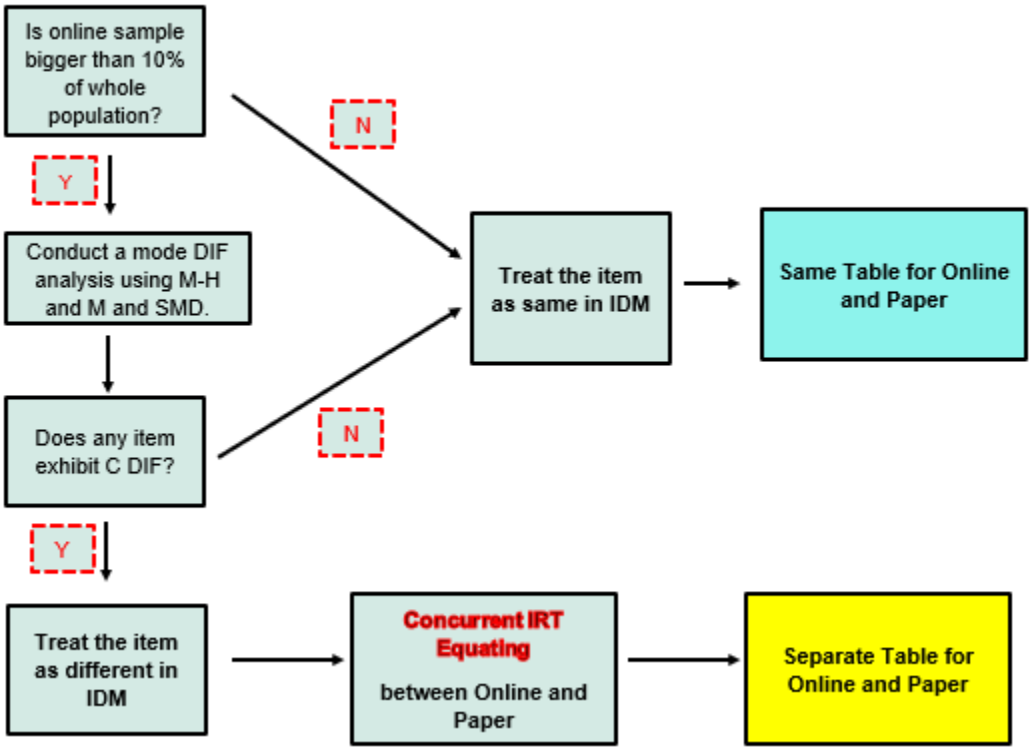


Figure 10.1 General overview of equating, including a mode effect analysis

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	Chapters 3, 7, & 9
	Data Review	LEAP 2025 Grades 3–8 Science Assessment Frameworks
	Classical Item Analysis	Chapter 7
	IRT Analysis	
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 the Consequences of Testing	Scale Score and Performance Level Information	Chapter 8
	Test Interpretive Guide	Chapter 8

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.

Andrich, A. (1988). *Rasch models for measurement*. 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*. ElsevierScience 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). Lawrence Erlbaum Associates.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. 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). Educational Testing Service.

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

Green, D. R. (1975). *Procedures for assessing bias in achievement tests*. Paper presented at the National Institute of Education (NIE) conference on Test Bias, Annapolis, MD.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. 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). Lawrence Erlbaum Associates.

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

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*. Lawrence Erlbaum Associates.

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*(4), 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). *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). Jossey-Bass.

Suen, H. K. (1990). *Principles of test theories*. Lawrence Erlbaum Associates.

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

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). 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*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

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

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement 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 à 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 > 0.95
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)
< 0.35	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 2023 Operational SC G3–8
Table B.2 Standard Coverage: Spring 2023 Operational SC G3–8
Table B.3 Item Type Summary: Spring 2023 Operational SC G3–8
Table B.4 Raw Score Summary: Spring 2023 Operational SC G3–8
Table B.5 Raw Score Summary by Reporting Category: Spring 2023 Operational SC G3–8
Table B.6 Scale Score and Raw Score Summary: Spring 2023 Operational SC G3–8

Table B.1.1

Percentage of Points by Reporting Category (includes Task Items): Spring 2023 Operational SC G3–8

Reporting Category	G3	G4	G5	G6	G7	G8
N/A	4.0%	11.5%	0.00%	1.6%	3.3%	9.8%
1 Investigate	28.0%	23.1%	16.4%	24.6%	13.1%	42.6%
2 Evaluate	46.0%	17.3%	34.4%	37.7%	18.0%	23.0%
3 Reason Scientifically	22.0%	48.1%	49.2%	36.1%	65.6%	24.6%

* N/A indicates no reporting category.

Table B.2
Standard Coverage: Spring 2023 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					2.78
	Sub-Total	**1**					**2.78**
1 Investigate	3-PS2-1		1		2		8.33
	3-PS2-2		1		1		5.56
	3-PS2-3				2		5.56
	3-PS2-4		1		1	1	8.33
	Sub-Total		**3**		**6**	**1**	**27.78**
2 Evaluate	3-ESS2-1			1	3		11.11
	3-ESS3-1		1		1		5.56
	3-LS2-1		1		3		11.11
	3-LS3-1				1		2.78
	3-LS4-1		1		1		5.56
	3-LS4-3				1		2.78
	3-LS4-4		2		2		11.11
	Sub-Total		**5**	**1**	**12**		**50.00**
3 Reason Scientifically	3-LS1-1	1	1			1	8.33
	3-LS3-2				2	1	8.33
	3-LS4-2				1		2.78
	Sub-Total	**1**	**1**		**3**	**2**	**19.44**
Total		**2**	**9**	**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-ESS3-1	1	1			1	8.33
	Sub-Total	**1**	**1**			**1**	**8.33**
1 Investigate	4-ESS2-1	1			1		5.56
	4-ESS2-3		2		2		11.11
	4-PS3-2				2		5.56
	4-PS3-3				1		2.78
	Sub-Total	**1**	**2**		**6**		**25.00**
2 Evaluate	4-ESS2-2			1	2		8.33
	4-LS1-1	1	1		2		11.11
	Sub-Total	**1**	**1**	**1**	**4**		**19.44**
3 Reason Scientifically	4-ESS1-1				3		8.33
	4-ESS3-2		2				5.56
	4-LS1-2				1		2.78
	4-PS3-1		2	1	2	1	16.67
	4-PS3-4		2				5.56
	4-PS4-1				1		2.78
	4-PS4-2				1	1	5.56
	Sub-Total		**6**	**1**	**8**	**2**	**47.22**
Total		**3**	**10**	**2**	**18**	**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							2.70
	5-PS1-3			2				1	8.11
	5-PS1-4					2			5.41
	Sub-Total	**1**		**2**		**2**		**1**	**16.22**
2 Evaluate	5-ESS1-1			3		1			10.81
	5-ESS1-2		2	1					8.11
	5-ESS2-2			1				1	5.41
	5-PS1-2	1		1	1				8.11
	5-PS2-1			2		1			8.11
	Sub-Total	**1**	**2**	**8**	**1**	**2**		**1**	**40.54**
3 Reason Scientifically	5-ESS2-1	1				1	1		8.11
	5-ESS3-1		1					1	5.41
	5-LS2-1			3		1			10.81
	5-PS1-1	2		1	1	1			13.51
	5-PS3-1			1		1			5.41
	Sub-Total	**3**	**1**	**5**	**1**	**4**	**1**	**1**	**43.24**
Total		**5**	**3**	**15**	**2**	**8**	**1**	**3**	**100.00**

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

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					2.70
	Sub-Total			**1**					**2.70**
1 Investigate	6-MS-LS1-1	1	1			1			8.11
	6-MS-PS2-2	1			1	1			8.11
	6-MS-PS2-3					2			5.41
	6-MS-PS2-5			1				1	5.41
	Sub-Total	**2**	**1**	**1**	**1**	**4**		**1**	**27.03**
2 Evaluate	6-MS-ESS1-3					1			2.70
	6-MS-ESS3-4				1				2.70
	6-MS-LS2-1		1						2.70
	6-MS-PS2-4	1				1			5.41
	6-MS-PS3-1			1		2			8.11
	6-MS-PS4-1			2			1		8.11
	Sub-Total	**1**	**1**	**3**	**1**	**4**	**1**		**29.73**
3 Reason Scientifically	6-MS-ESS1-1			1				1	5.41
	6-MS-ESS1-2					3			8.11
	6-MS-LS1-2			1					2.70
	6-MS-LS2-2					1			2.70
	6-MS-LS2-3			1					2.70
	6-MS-PS1-1			1					2.70
	6-MS-PS2-1					1		1	5.41
	6-MS-PS3-2			1					2.70
	6-MS-PS4-2		1			2			8.11
	Sub-Total		**1**	**5**		**7**		**2**	**40.54**
Total		**3**	**3**	**10**	**2**	**15**	**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-LS4-5			1					2.70
	Sub-Total			**1**					**2.70**
1 Investigate	7-MS-ESS2-5				1				2.70
	7-MS-ESS3-5	1							2.70
	7-MS-PS3-4		2	1					8.11
	Sub-Total	**1**	**2**	**1**	**1**				**13.51**
2 Evaluate	7-MS-LS1-3				1				2.70
	7-MS-LS2-4			3		2			13.51
	7-MS-PS1-2			1				1	5.41
	Sub-Total			**4**	**1**	**2**		**1**	**21.62**
3 Reason Scientifically	7-MS-ESS2-4			2		2	1	1	16.22
	7-MS-ESS2-6	1							2.70
	7-MS-LS1-6				1				2.70
	7-MS-LS1-7					1			2.70
	7-MS-LS2-5			2					5.41
	7-MS-LS3-2		1			2			8.11
	7-MS-LS4-4			2		1		1	10.81
	7-MS-PS1-4			2		1			8.11
	7-MS-PS1-5		1			1			5.41
	Sub-Total	**1**	**2**	**8**	**1**	**8**	**1**	**2**	**62.16**
Total		**2**	**4**	**14**	**3**	**10**	**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	
N/A	8-MS-ESS3-1			1					2.70
	8-MS-LS1-4		1	1					5.41
	Sub-Total		**1**	**2**					**8.11**
1 Investigate	8-MS-ESS3-2			2	1				8.11
	8-MS-ESS3-3				1				2.70
	8-MS-LS1-5			1		2			8.11
	8-MS-PS1-3			2	1		1		10.81
	8-MS-PS1-6				1	1			5.41
	8-MS-PS3-3				1	1			5.41
	Sub-Total			**5**	**5**	**4**	**1**		**40.54**
2 Evaluate	8-MS-ESS2-3					1			2.70
	8-MS-LS4-1					1			2.70
	8-MS-LS4-3			1	1				5.41
	8-MS-LS4-6		1	1		1			8.11
	8-MS-PS3-5		1					1	5.41
	Sub-Total		**2**	**2**	**1**	**3**		**1**	**24.32**
3 Reason Scientifically	8-MS-ESS1-4	1							2.70
	8-MS-ESS2-1			1				1	5.41
	8-MS-ESS2-2					1			2.70
	8-MS-ESS3-1			1		1			5.41
	8-MS-LS3-1					1			2.70
	8-MS-LS4-2			1				1	5.41
	8-MS-PS1-1			1					2.70
	Sub-Total	**1**		**4**		**3**		**2**	**27.03**
Total		**1**	**3**	**13**	**6**	**10**	**1**	**3**	**100.00**

* N/A indicates no reporting category.

Table B.3

Item Type Summary: Spring 2023 Operational SC G3–8

Grade	MC	MS	TEI	CR	ER	TPD	TPI
3	21	1	0	3	0	9	2
4	18	2	0	3	0	10	3
5	8	2	15	3	1	3	5
6	15	2	10	3	1	3	3
7	10	3	14	3	1	4	2
8	10	6	13	3	1	3	1

Note: 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 2023 Operational SC G3–8

Grade	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
3	≥49,300	18.33	9.03	0	48	0.38	0.41	0.86	3.37
4	≥48,860	22.06	9.94	0	51	0.44	0.44	0.88	3.39
5	≥48,310	23.21	11.51	0	58	0.42	0.45	0.88	3.95
6	≥48,270	21.89	10.37	0	58	0.38	0.42	0.88	3.64
7	≥48,870	22.88	11.63	0	60	0.39	0.45	0.89	3.82
8	≥50,100	22.56	10.51	0	57	0.40	0.45	0.90	3.37

* Reliability is Cronbach's alpha.

Table B.5

Raw Score Summary by Reporting Category: Spring 2023 Operational SC G3–8

Grade	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
3	Investigate	5.89	3.24	0	14	0.42	0.43	0.68	1.83
	Evaluate	8.20	4.39	0	23	0.38	0.39	0.73	2.28
	Reason Scientifically	3.62	2.24	0	11	0.33	0.40	0.54	1.52
4	Investigate	5.63	2.69	0	12	0.48	0.41	0.62	1.66
	Evaluate	3.82	2.04	0	9	0.42	0.42	0.56	1.35
	Reason Scientifically	10.81	5.20	0	25	0.45	0.45	0.79	2.38
5	Investigate	4.63	2.33	0	10	0.49	0.48	0.64	1.40
	Evaluate	7.33	4.48	0	21	0.36	0.47	0.79	2.05
	Reason Scientifically	11.24	5.80	0	30	0.44	0.43	0.72	3.07
6	Investigate	5.36	3.02	0	15	0.37	0.41	0.65	1.79
	Evaluate	6.90	4.02	0	22	0.34	0.42	0.70	2.20
	Reason Scientifically	9.29	4.34	0	22	0.44	0.42	0.73	2.26
7	Investigate	2.30	1.86	0	8	0.29	0.44	0.53	1.28
	Evaluate	4.16	2.69	0	11	0.39	0.50	0.71	1.45
	Reason Scientifically	15.89	7.89	0	40	0.41	0.44	0.83	3.25
8	Investigate	8.70	4.83	0	26	0.39	0.48	0.82	2.05
	Evaluate	5.61	3.11	0	14	0.42	0.47	0.67	1.79
	Reason Scientifically	4.69	2.62	0	15	0.35	0.38	0.60	1.66

Table B.6.1

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 3

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥49,310	100.00	725.29	30.80	18.33	9.03	–
Gender	Female	≥24,170	49.02	725.17	29.93	18.21	8.77	**0.03**
	Male	≥25,130	50.97	725.41	31.62	18.45	9.26	–
Race	African American	≥20,320	41.21	714.40	28.19	15.02	7.36	**0.80**
	AI/AN	≥270	0.56	728.48	28.79	19.09	8.61	**0.28**
	Asian	≥750	1.53	743.27	31.00	24.07	9.91	**-0.25**
	Hispanic/Latino	≥5,260	10.67	719.06	29.28	16.41	8.17	**0.58**
	NHPI	≥50	0.11	729.87	28.67	19.67	8.40	**0.22**
	Two or More	≥1,950	3.97	729.51	29.07	19.47	8.83	**0.24**
	White	≥20,660	41.90	736.50	29.48	21.74	9.34	–
Economically Disadvantaged	No	≥14,550	29.50	740.06	29.63	22.94	9.49	**-0.77**
	Yes	≥34,600	70.16	719.16	29.13	16.41	8.08	–
English Learner	No	≥46,400	94.09	726.42	30.74	18.66	9.08	**-0.63**
	Yes	≥2,910	5.91	707.42	25.86	13.05	6.12	–
Education Classification	Regular	≥43,080	87.36	727.18	30.43	18.86	9.05	**-0.47**
	Special	≥6,230	12.64	712.25	30.18	14.66	7.93	–
Section 504	No	≥45,650	92.58	725.75	30.94	18.48	9.09	**-0.23**
	Yes	≥3,660	7.42	719.63	28.37	16.42	7.93	–
Migrant	No	≥49,210	99.79	725.31	30.81	18.33	9.03	**-0.24**
	Yes	≥100	0.21	719.06	28.17	16.17	8.00	–
Homeless Status	No	≥48,150	97.64	725.65	30.78	18.43	9.04	**-0.49**
	Yes	≥1,160	2.36	710.67	28.21	14.06	7.05	–
Military Affiliation	No	≥48,340	98.02	725.00	30.78	18.24	9.00	**0.49**
	Yes	≥970	1.98	739.56	28.26	22.69	9.09	–
Foster Care Status	No	≥49,140	99.65	725.32	30.81	18.34	9.03	**-0.28**
	Yes	≥170	0.35	717.37	28.00	15.77	7.71	–

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table B.6.2

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 4

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥48,870	100.00	737.56	30.09	22.06	9.94	–
Gender	Female	≥23,860	48.83	736.16	29.13	21.55	9.64	**0.10**
	Male	≥25,000	51.17	738.90	30.91	22.55	10.20	–
Race	African American	≥20,120	41.18	725.90	26.06	18.11	8.25	**0.88**
	AI/AN	≥280	0.57	741.60	28.16	23.32	9.68	**0.28**
	Asian	≥820	1.69	757.10	31.14	28.62	10.44	**-0.26**
	Hispanic/Latino	≥5,150	10.55	730.89	29.32	19.88	9.44	**0.63**
	NHPI	≥40	0.08	746.68	31.33	25.20	10.74	**0.09**
	Two or More	≥1,870	3.84	741.36	29.22	23.30	9.79	**0.28**
	White	≥20,540	42.04	749.45	28.97	26.08	9.85	–
Economically Disadvantaged	No	≥14,510	29.70	753.34	29.37	27.43	9.95	**-0.82**
	Yes	≥34,050	69.67	730.98	27.79	19.82	9.03	–
English Learner	No	≥46,180	94.50	738.78	29.96	22.46	9.95	**-0.74**
	Yes	≥2,680	5.50	716.55	23.86	15.17	6.91	–
Education Classification	Regular	≥42,840	87.66	739.85	29.56	22.80	9.87	**-0.61**
	Special	≥6,030	12.34	721.30	28.73	16.83	8.81	–
Section 504	No	≥44,550	91.17	738.31	30.19	22.32	9.99	**-0.29**
	Yes	≥4,310	8.83	729.83	27.84	19.41	9.03	–
Migrant	No	≥48,810	99.88	737.57	30.09	22.06	9.95	**-0.21**
	Yes	≥60	0.12	730.82	29.33	19.98	9.34	–
Homeless Status	No	≥47,730	97.65	737.92	30.08	22.18	9.95	**-0.52**
	Yes	≥1,140	2.35	722.50	26.20	17.07	8.09	–
Military Affiliation	No	≥47,960	98.12	737.28	30.07	21.97	9.93	**0.50**
	Yes	≥910	1.88	752.07	27.41	26.92	9.57	–
Foster Care Status	No	≥48,700	99.64	737.60	30.08	22.07	9.95	**-0.32**
	Yes	≥170	0.36	727.37	29.43	18.88	9.32	–

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table B.6.3

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 5

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥48,320	100.00	729.44	37.83	23.21	11.51	–
Gender	Female	≥23,590	48.83	728.46	36.13	22.79	11.00	**0.07**
	Male	≥24,720	51.17	730.37	39.36	23.61	11.96	–
Race	African American	≥20,300	42.01	714.61	33.48	18.60	9.47	**0.89**
	AI/AN	≥250	0.53	736.23	35.02	25.18	11.06	**0.24**
	Asian	≥800	1.67	757.41	38.93	32.17	12.46	**-0.38**
	Hispanic/Latino	≥5,170	10.71	722.16	37.48	21.07	11.05	**0.60**
	NHPI	≥30	0.07	735.50	35.25	25.03	10.86	**0.25**
	Two or More	≥1,680	3.48	736.03	36.46	25.16	11.26	**0.24**
	White	≥20,050	41.50	744.56	35.50	27.87	11.40	–
Economically Disadvantaged	No	≥14,660	30.35	748.80	35.76	29.28	11.54	**-0.81**
	Yes	≥33,360	69.04	721.11	35.53	20.59	10.45	–
English Learner	No	≥46,090	95.38	730.96	37.55	23.65	11.49	**-0.84**
	Yes	≥2,230	4.62	698.11	29.07	14.10	7.27	–
Education Classification	Regular	≥42,660	88.28	732.96	36.85	24.21	11.41	**-0.76**
	Special	≥5,660	11.72	702.95	34.47	15.69	9.24	–
Section 504	No	≥43,580	90.19	730.47	37.88	23.53	11.56	**-0.29**
	Yes	≥4,740	9.81	719.91	36.06	20.26	10.62	–
Migrant	No	≥48,260	99.86	729.45	37.83	23.21	11.51	**-0.18**
	Yes	≥60	0.14	722.21	40.65	21.17	12.24	–
Homeless Status	No	≥47,210	97.70	729.86	37.81	23.33	11.52	**-0.48**
	Yes	≥1,110	2.30	711.64	34.13	17.86	9.50	–
Military Affiliation	No	≥47,340	97.98	729.03	37.80	23.08	11.48	**0.54**
	Yes	≥970	2.02	749.25	34.13	29.33	11.21	–
Foster Care Status	No	≥48,170	99.69	729.48	37.84	23.22	11.51	**-0.37**
	Yes	≥150	0.31	716.35	33.68	19.01	9.83	–

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table B.6.4

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 6

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥48,300	100.00	721.95	32.01	21.89	10.37	–
Gender	Female	≥23,580	48.83	720.92	30.58	21.41	9.90	**0.09**
	Male	≥24,710	51.17	722.94	33.28	22.34	10.79	–
Race	African American	≥20,370	42.18	709.70	28.22	17.81	8.28	**0.86**
	AI/AN	≥250	0.53	723.81	28.31	22.21	9.17	**0.36**
	Asian	≥730	1.52	747.98	32.20	30.96	11.67	**-0.47**
	Hispanic/Latino	≥4,990	10.33	715.93	31.55	19.98	9.75	**0.58**
	NHPI	≥40	0.08	718.10	38.34	21.41	11.56	**0.43**
	Two or More	≥1,670	3.47	727.18	31.46	23.55	10.41	**0.23**
	White	≥20,210	41.85	734.41	30.39	26.00	10.56	–
Economically Disadvantaged	No	≥15,050	31.17	737.65	30.75	27.18	10.82	**-0.79**
	Yes	≥32,990	68.30	714.92	29.95	19.51	9.21	–
English Learner	No	≥46,330	95.91	723.09	31.78	22.23	10.38	**-0.82**
	Yes	≥1,970	4.09	695.16	24.89	13.79	5.98	–
Education Classification	Regular	≥42,910	88.84	724.65	31.30	22.69	10.35	**-0.71**
	Special	≥5,390	11.16	700.50	29.39	15.51	8.07	–
Section 504	No	≥43,280	89.60	722.88	32.05	22.19	10.44	**-0.28**
	Yes	≥5,020	10.40	713.95	30.51	19.27	9.35	–
Migrant	No	≥48,230	99.86	721.96	32.01	21.89	10.37	**-0.25**
	Yes	≥60	0.14	713.86	30.58	19.29	9.15	–
Homeless Status	No	≥47,250	97.83	722.26	32.00	21.98	10.39	**-0.43**
	Yes	≥1,040	2.17	708.16	29.11	17.49	8.40	–
Military Affiliation	No	≥47,380	98.08	721.63	31.97	21.78	10.34	**0.54**
	Yes	≥920	1.92	738.42	29.37	27.40	10.40	–
Foster Care Status	No	≥48,170	99.72	722.00	32.00	21.90	10.37	**-0.52**
	Yes	≥130	0.28	704.10	29.81	16.52	8.30	–

***** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table B.6.5

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 7

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥48,900	100.00	730.60	33.09	22.88	11.63	–
Gender	Female	≥23,690	48.44	731.69	31.82	23.15	11.29	**-0.04**
	Male	≥25,210	51.56	729.59	34.20	22.63	11.93	–
Race	African American	≥20,470	41.86	719.22	29.63	18.73	9.78	**0.79**
	AI/AN	≥270	0.56	733.48	30.01	23.73	10.57	**0.30**
	Asian	≥780	1.61	756.16	35.01	32.50	12.87	**-0.46**
	Hispanic/Latino	≥5,230	10.70	722.54	33.98	20.27	11.29	**0.60**
	NHPI	≥30	0.07	744.53	34.43	27.61	12.58	**-0.04**
	Two or More	≥1,670	3.41	735.53	32.64	24.60	11.70	**0.22**
	White	≥20,400	41.73	742.65	31.27	27.19	11.59	–
Economically Disadvantaged	No	≥15,380	31.45	746.94	31.33	28.83	11.73	**-0.79**
	Yes	≥33,250	67.99	723.18	31.11	20.18	10.52	–
English Learner	No	≥46,780	95.67	732.03	32.64	23.35	11.58	**-0.93**
	Yes	≥2,120	4.33	699.14	26.46	12.69	7.10	–
Education Classification	Regular	≥43,750	89.46	733.42	32.29	23.82	11.53	**-0.78**
	Special	≥5,150	10.54	706.66	29.97	14.96	9.15	–
Section 504	No	≥43,670	89.31	731.89	33.12	23.34	11.70	**-0.37**
	Yes	≥5,220	10.69	719.91	30.82	19.04	10.27	–
Migrant	No	≥48,830	99.85	730.62	33.08	22.89	11.63	**-0.28**
	Yes	≥70	0.15	720.45	36.18	19.66	11.66	–
Homeless Status	No	≥47,880	97.91	730.95	33.05	23.00	11.64	**-0.49**
	Yes	≥1,020	2.09	714.37	30.39	17.29	9.72	–
Military Affiliation	No	≥48,020	98.21	730.26	33.02	22.76	11.59	**0.61**
	Yes	≥870	1.79	749.46	30.90	29.81	11.67	–
Foster Care Status	No	≥48,760	99.71	730.65	33.08	22.90	11.63	**-0.48**
	Yes	≥140	0.29	714.15	31.42	17.30	9.95	–

***** AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table B.6.6

Scale Score and Raw Score Summary: Spring 2023 Operational Science Grade 8

Category	Subgroup*	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total		≥50,160	100.00	732.49	31.12	22.56	10.51	–
Gender	Female	≥24,810	49.47	732.61	29.59	22.48	10.04	**0.01**
	Male	≥25,340	50.53	732.36	32.54	22.63	10.94	–
Race	African American	≥21,420	42.72	720.02	27.07	18.26	8.47	**0.93**
	AI/AN	≥270	0.56	735.72	28.21	23.47	9.79	**0.34**
	Asian	≥750	1.50	757.80	31.08	31.61	11.30	**-0.44**
	Hispanic/Latino	≥5,080	10.14	724.78	31.30	20.11	10.04	**0.67**
	NHPI	≥40	0.08	749.32	26.03	28.24	9.61	**-0.11**
	Two or More	≥1,710	3.41	738.47	29.38	24.48	10.26	**0.25**
	White	≥20,860	41.59	745.69	29.06	27.05	10.42	–
Economically Disadvantaged	No	≥16,090	32.08	747.79	29.42	27.84	10.58	**-0.79**
	Yes	≥33,750	67.28	725.36	29.21	20.09	9.50	–
English Learner	No	≥48,170	96.04	733.69	30.80	22.93	10.48	**-0.92**
	Yes	≥1,980	3.96	703.19	23.64	13.40	6.13	–
Education Classification	Regular	≥45,030	89.78	735.10	30.52	23.39	10.45	**-0.80**
	Special	≥5,120	10.22	709.57	26.60	15.24	7.80	–
Section 504	No	≥44,780	89.28	733.58	31.14	22.93	10.56	**-0.33**
	Yes	≥5,370	10.72	723.38	29.36	19.43	9.52	–
Migrant	No	≥50,090	99.87	732.50	31.12	22.56	10.51	**-0.27**
	Yes	≥60	0.13	723.41	31.74	19.75	9.96	–
Homeless Status	No	≥49,150	97.99	732.78	31.10	22.65	10.52	**-0.45**
	Yes	≥1,000	2.01	718.33	28.67	17.88	8.89	–
Military Affiliation	No	≥49,260	98.21	732.16	31.09	22.44	10.48	**0.60**
	Yes	≥890	1.79	750.29	27.48	28.68	10.12	–
Foster Care Status	No	≥50,000	99.69	732.53	31.12	22.57	10.51	**-0.41**
	Yes	≥150	0.31	720.30	26.35	18.25	8.38	–

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Appendix C: Item Analysis Summary Report

Summary Statistics Reports

Contents
Table C.1 P-Value Summary by Grade: Spring 2023 Operational SC G3–8 Plot C.1 P-Value Summary by Grade: Spring 2023 Operational SC G3–8
Table C.2 Item-Total Correlation Summary by Grade: Spring 2023 Operational SC G3–8 Plot C.2 Item-Total Correlation Summary by Grade: Spring 2023 Operational SC G3–8
Table C.3 Corrected Point-Biserial Correlation Summary by Grade: Spring 2023 Operational SC G3–8 Plot C.3 Corrected Point-Biserial Correlation Summary by Grade: Spring 2023 Operational SC G3–8
Table C.4 Item-Total Correlation Summary by Reporting Category and Grade: Spring 2023 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 2023 Operational SC G3–8 Plot C.5.1 IRT Parameter Summary: Spring 2023 Operational SC G3–8: A-Parameter
Table C.6 Statistically Flagged Items by Item Type: Spring 2023 Operational SC G3–8

Table C.1.1

P-Value Summary by Grade: Spring 2023 Operational SC G3–8

Grade*	No. of Items	0 ≤ p < 0.2	0.2 ≤ p < 0.4	0.4 ≤ p < 0.6	0.6 ≤ p < 0.8	0.8 ≤ p ≤ 1.0
3	36	3	15	18	0	0
4	36	2	12	18	3	1
5	37	3	15	12	6	1
6	39	5	14	17	3	0
7	38	3	21	10	3	1
8	39	6	11	19	2	1

* Classical analyses for Grades 6–8 were calculated and estimated separately for each dimension of the ER item,and the result summarize both dimensions.

Plot C.1.1

P-Value Summary by Grade: Spring 2023 Operational SC G3–8

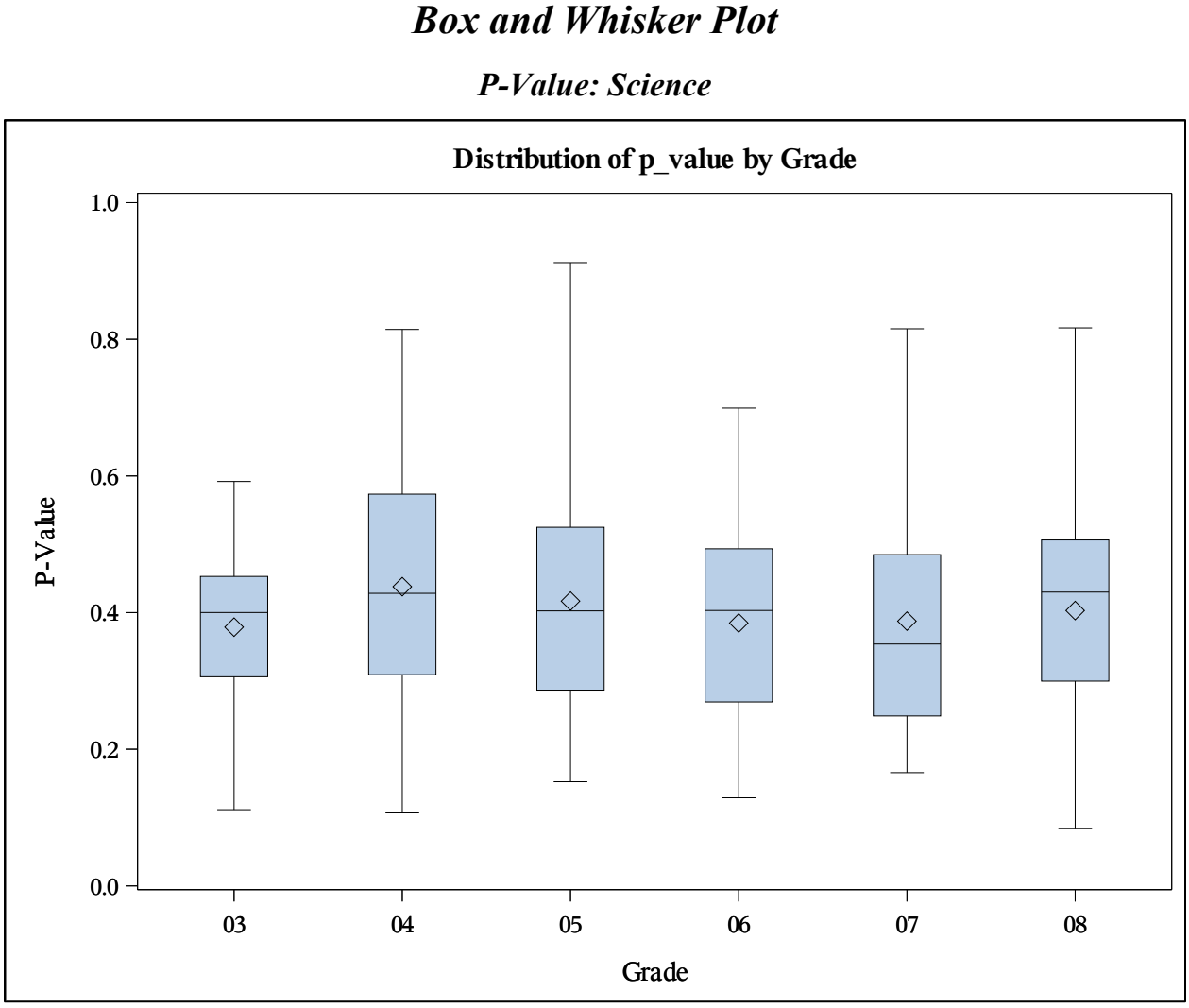


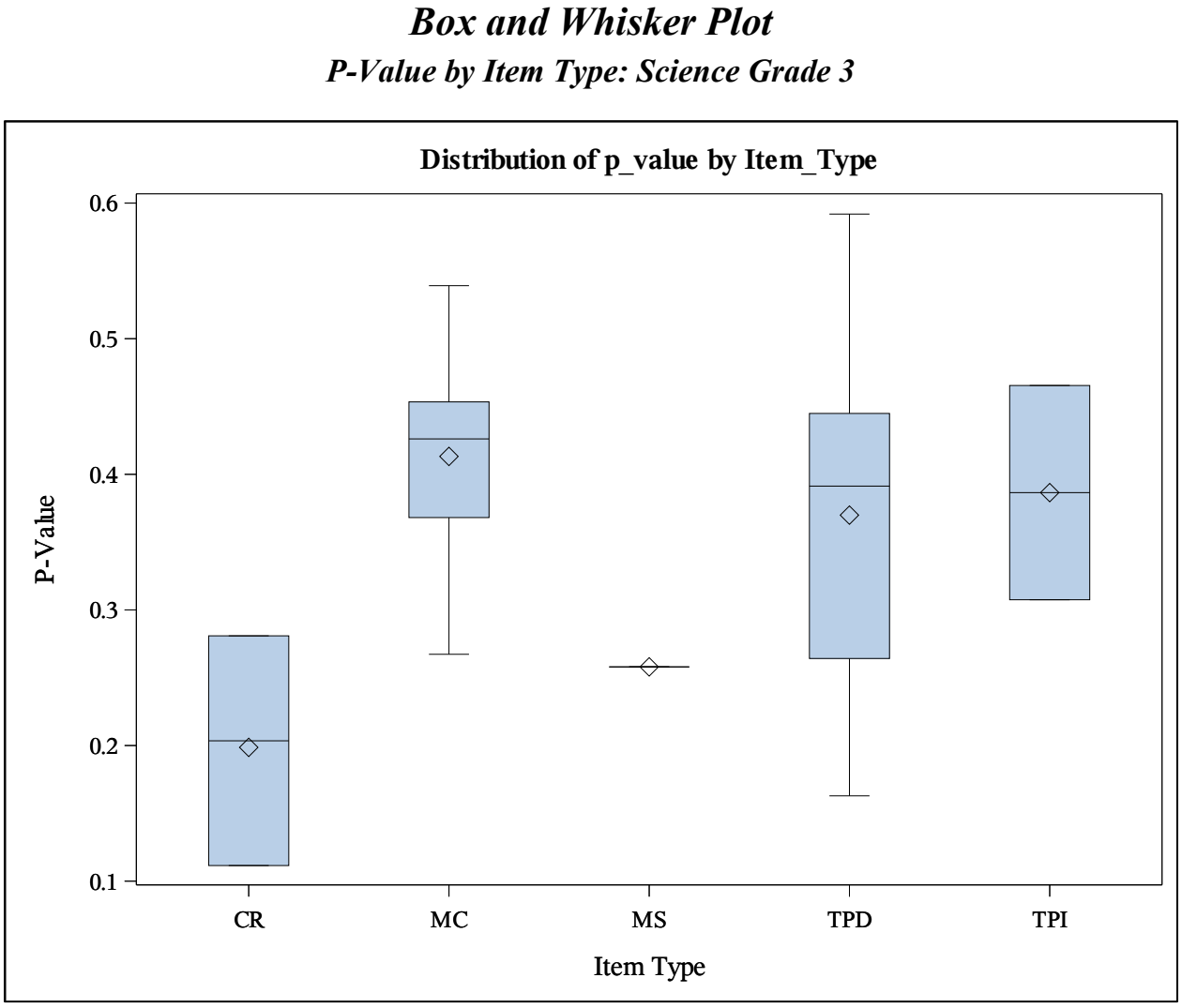
Table C.1.2

P-Value Summary by Item Type: Spring 2023 Operational SC G3–8

Grade	Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	CR	3	0.112	0.112	0.204	0.281	0.281
	MC	21	0.267	0.368	0.426	0.453	0.539
	MS	1	0.258	0.258	0.258	0.258	0.258
	TPD	9	0.163	0.264	0.391	0.445	0.592
	TPI	2	0.308	0.308	0.387	0.466	0.466
4	CR	3	0.107	0.107	0.185	0.255	0.255
	MC	18	0.230	0.367	0.511	0.589	0.814
	MS	2	0.201	0.201	0.250	0.298	0.298
	TPD	10	0.297	0.369	0.430	0.484	0.582
	TPI	3	0.393	0.393	0.405	0.591	0.591
5	CR	3	0.152	0.152	0.173	0.280	0.280
	ER	1	0.221	0.221	0.221	0.221	0.221
	MC	8	0.341	0.461	0.598	0.646	0.684
	MS	2	0.183	0.183	0.204	0.225	0.225
	TEI	15	0.206	0.290	0.359	0.583	0.912
	TPD	3	0.373	0.373	0.375	0.516	0.516
	TPI	5	0.227	0.428	0.442	0.450	0.488
6	CR	3	0.185	0.185	0.191	0.204	0.204
	ER	3	0.129	0.129	0.134	0.183	0.183
	MC	15	0.269	0.373	0.460	0.498	0.699
	MS	2	0.318	0.318	0.318	0.318	0.318
	TEI	10	0.217	0.272	0.504	0.564	0.630
	TPD	3	0.257	0.257	0.358	0.426	0.426
	TPI	3	0.310	0.310	0.403	0.448	0.448
7	CR	3	0.183	0.183	0.223	0.351	0.351
	ER	2	0.204	0.204	0.216	0.227	0.227
	MC	10	0.249	0.304	0.358	0.489	0.602
	MS	3	0.327	0.327	0.377	0.552	0.552
	TEI	14	0.166	0.269	0.369	0.587	0.815
	TPD	4	0.170	0.202	0.336	0.462	0.485
	TPI	2	0.338	0.338	0.387	0.437	0.437
8	CR	3	0.084	0.084	0.097	0.222	0.222
	ER	3	0.094	0.094	0.127	0.140	0.140
	MC	10	0.362	0.426	0.472	0.518	0.651
	MS	6	0.239	0.300	0.418	0.476	0.512
	TEI	13	0.151	0.390	0.466	0.565	0.817
	TPD	3	0.397	0.397	0.401	0.506	0.506
	TPI	1	0.273	0.273	0.273	0.273	0.273

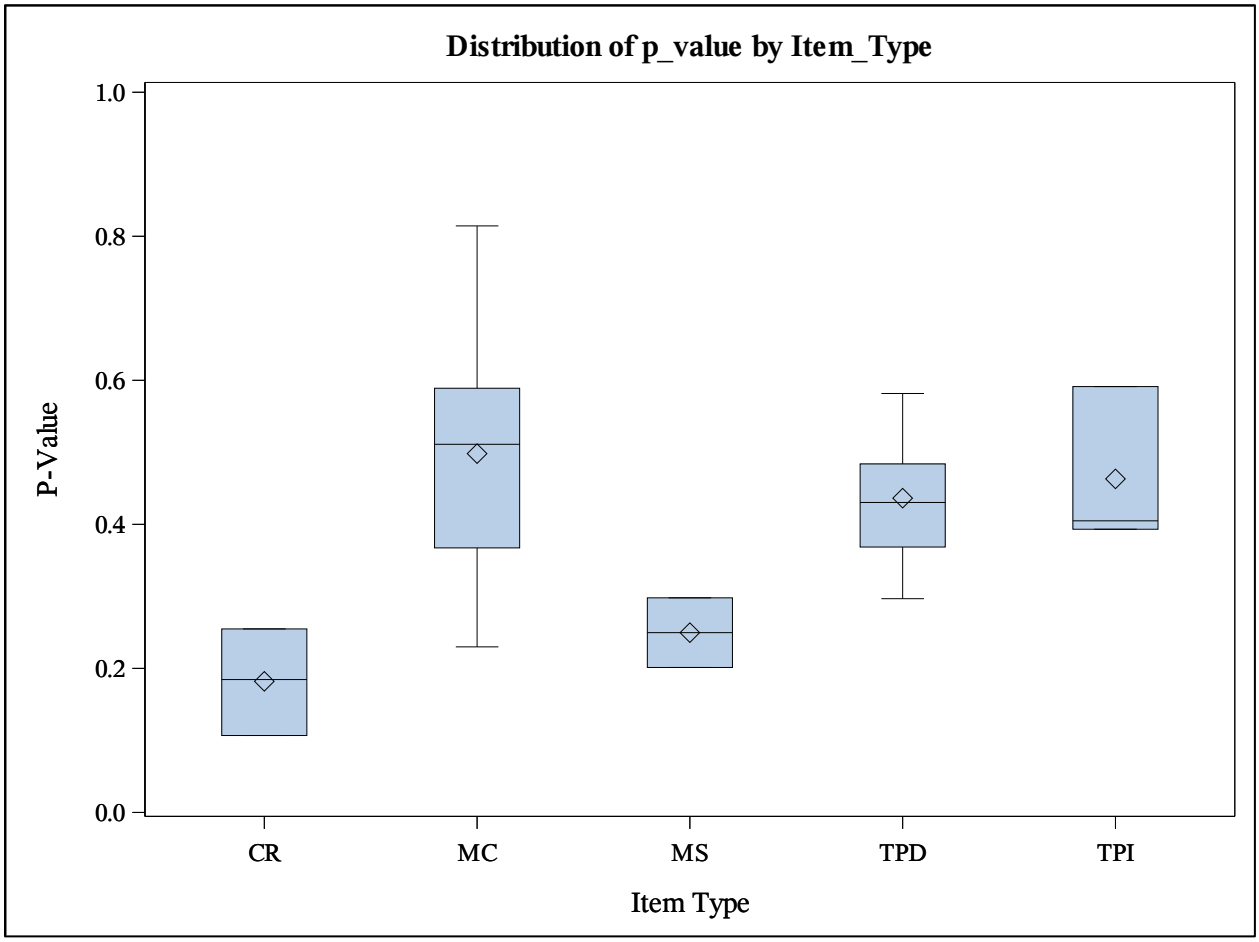
Plot C.1.2

P-Value Summary by Item Type: Spring 2023 Operational SC G3–8



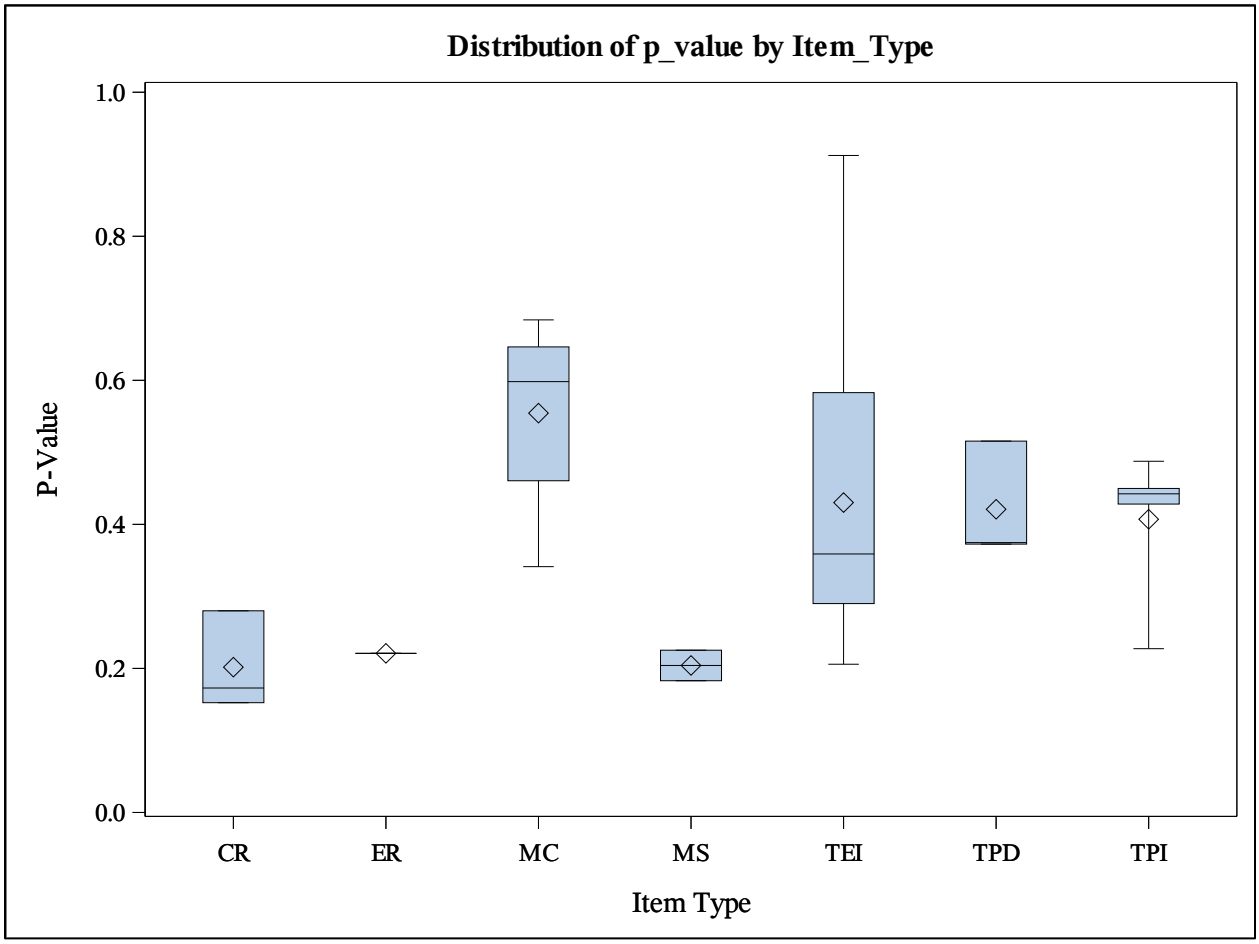
Box and Whisker Plot

P-Value by Item Type: Science Grade 4



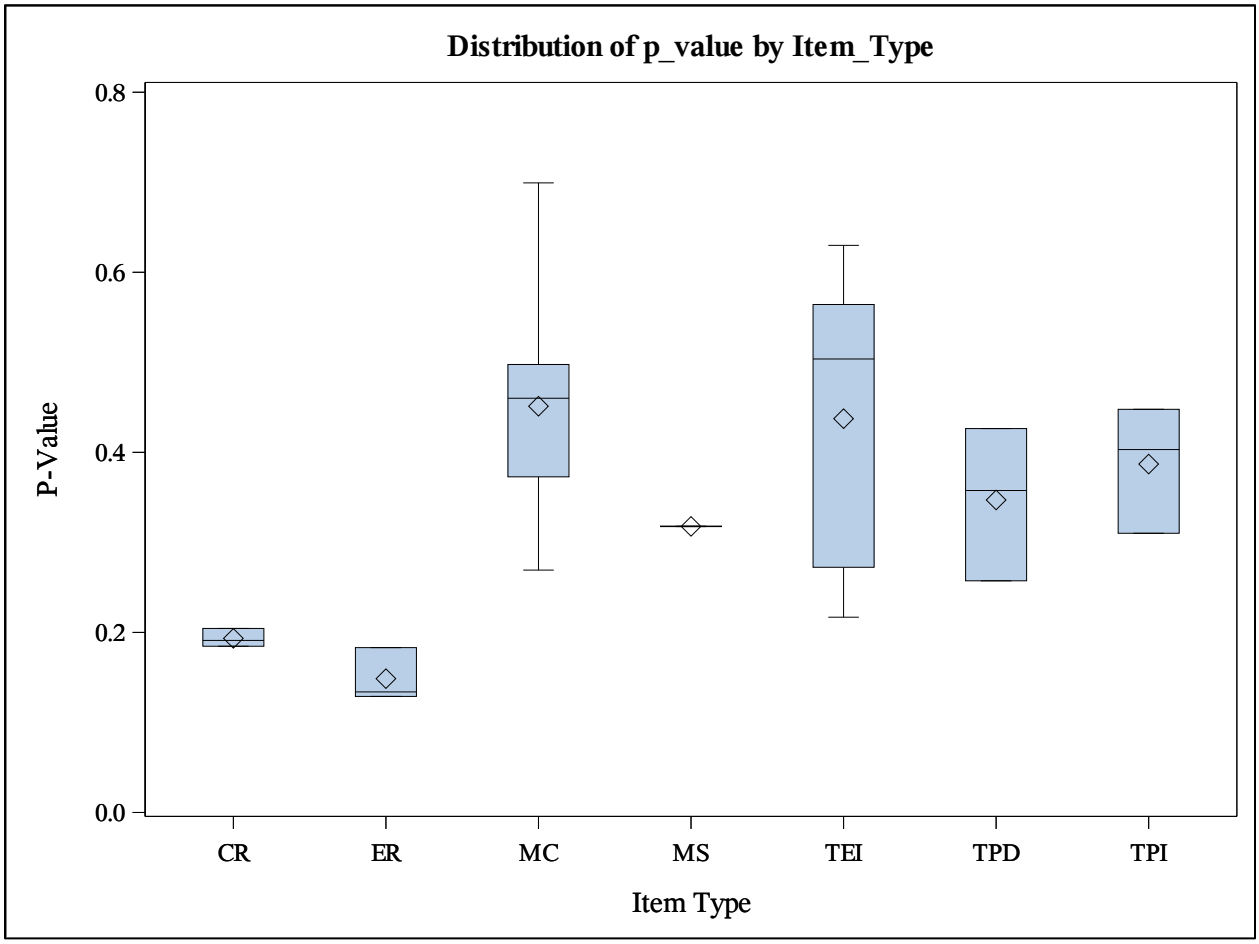
Box and Whisker Plot

P-Value by Item Type: Science Grade 5



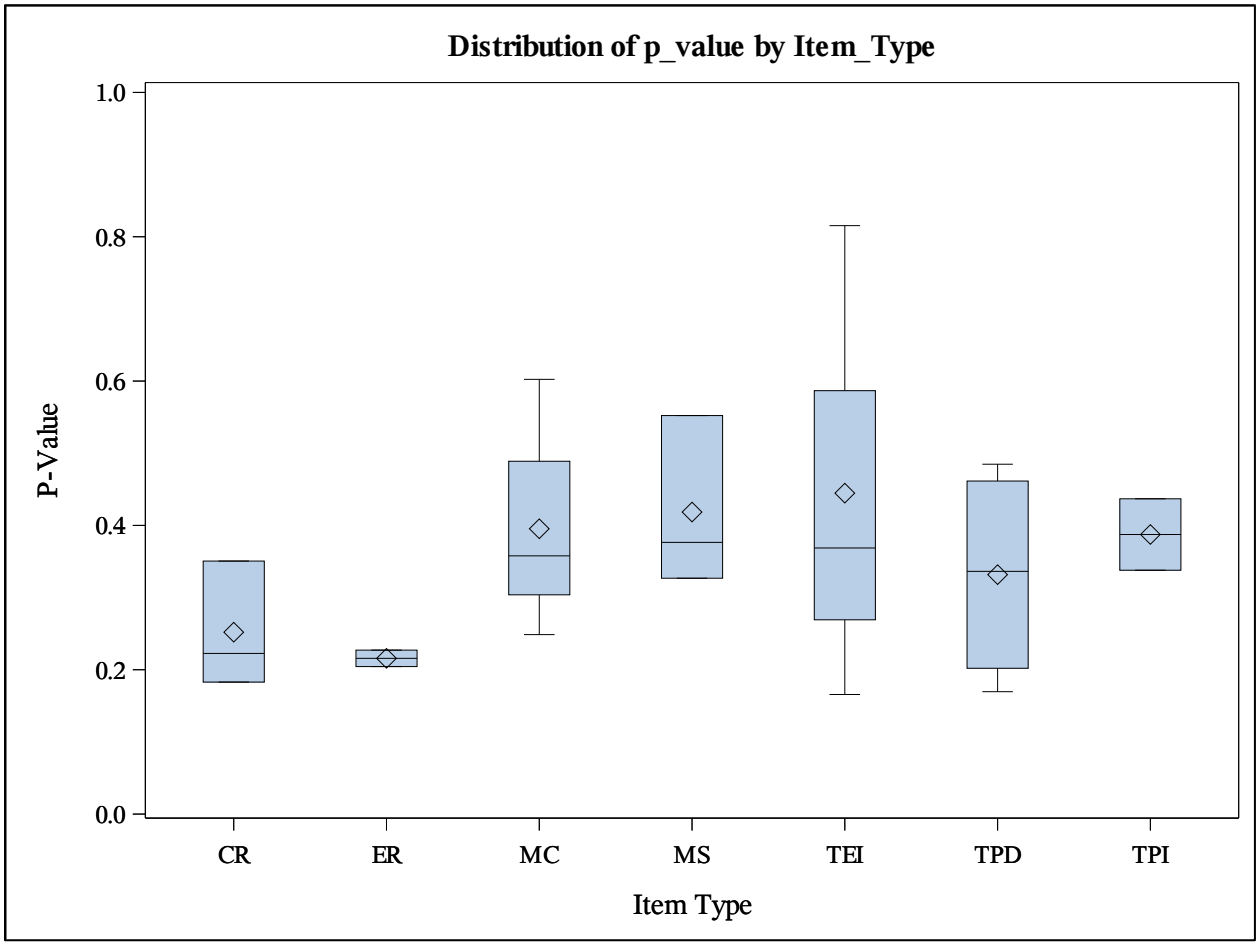
Box and Whisker Plot

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

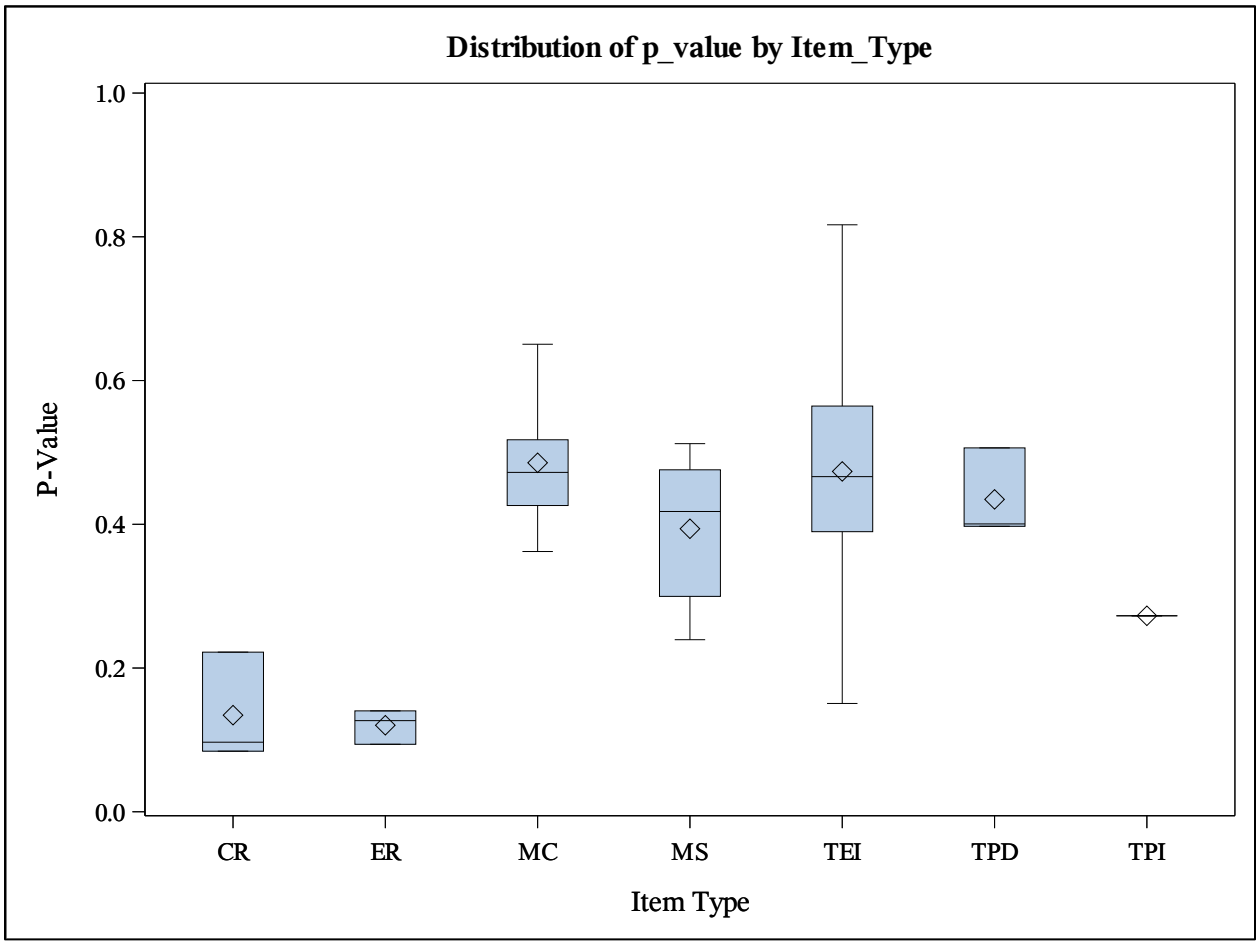


Table C.2.1

Item-Total Correlation by Grade: Spring 2023 Operational SC G3–8

Grade*	No. of Items	r < 0	0.0 ≤ r < 0.2	0.2 ≤ r < 0.3	0.3 ≤ r < 0.4	0.4 ≤ r < 0.5
3	36	0	1	7	10	9
4	36	0	2	3	4	14
5	37	0	0	4	6	17
6	39	0	2	5	8	11
7	38	0	0	4	8	12
8	39	0	0	3	8	16

* Classical analyses for Grades 6–8 were calculated and estimated separately for each dimension of the ER item,and the result summarize both dimensions.

Plot C.2.1

Item-Total Correlation by Grade: Spring 2023 Operational SC G3–8

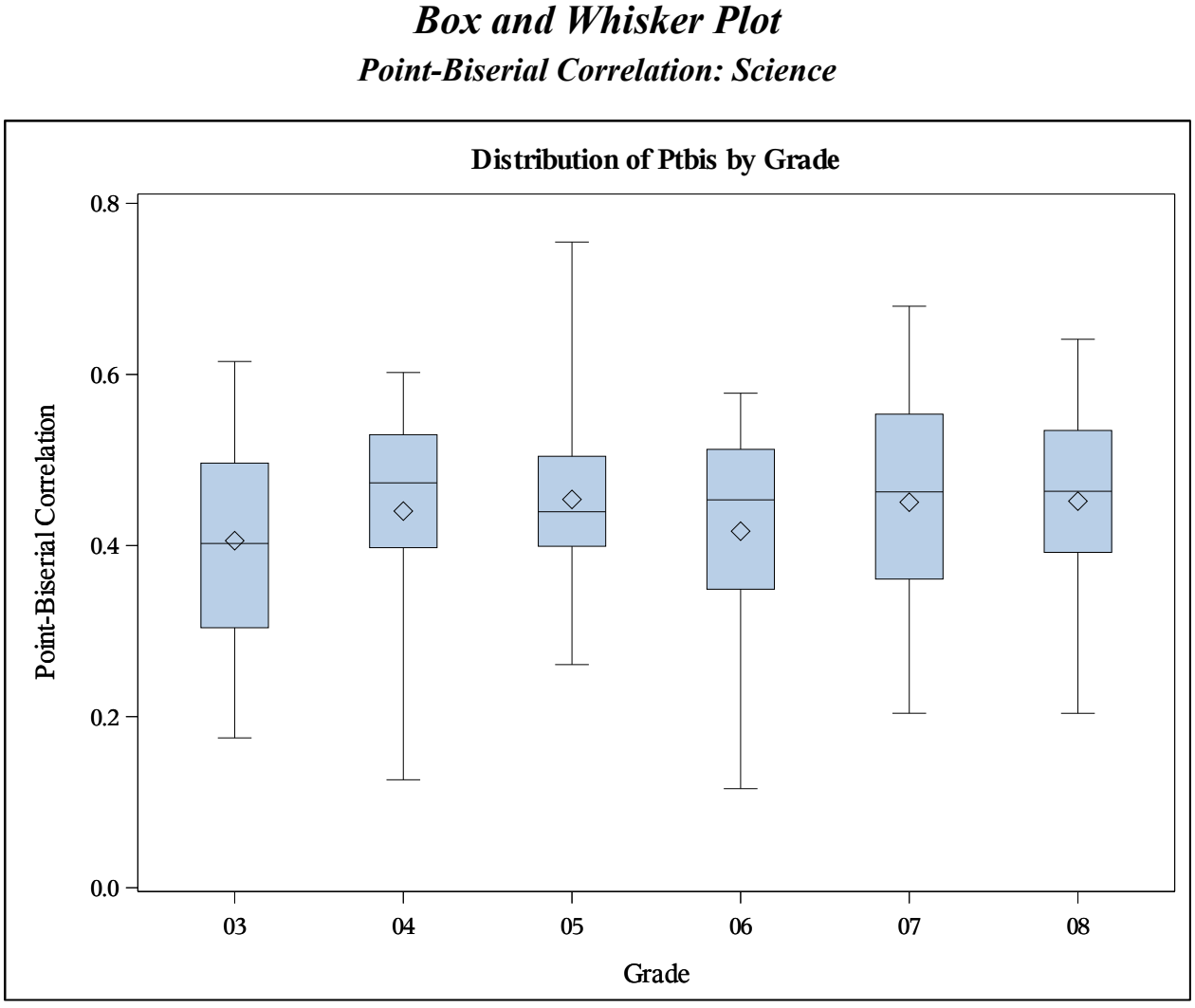


Table C.2.2

Item-Total Correlation Summary by Item Type: Spring 2023 Operational SC G3–8

Grade	Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	CR	3	0.322	0.322	0.529	0.603	0.603
	MC	21	0.201	0.290	0.368	0.411	0.478
	MS	1	0.306	0.306	0.306	0.306	0.306
	TPD	9	0.175	0.479	0.551	0.563	0.615
	TPI	2	0.490	0.490	0.496	0.502	0.502
4	CR	3	0.484	0.484	0.492	0.509	0.509
	MC	18	0.126	0.295	0.408	0.463	0.509
	MS	2	0.312	0.312	0.377	0.443	0.443
	TPD	10	0.393	0.505	0.566	0.576	0.602
	TPI	3	0.494	0.494	0.552	0.559	0.559
5	CR	3	0.478	0.478	0.496	0.613	0.613
	ER	1	0.755	0.755	0.755	0.755	0.755
	MC	8	0.280	0.355	0.400	0.450	0.567
	MS	2	0.324	0.324	0.369	0.415	0.415
	TEI	15	0.261	0.377	0.438	0.540	0.608
	TPD	3	0.493	0.493	0.544	0.623	0.623
	TPI	5	0.405	0.433	0.439	0.450	0.453
6	CR	3	0.373	0.373	0.512	0.512	0.512
	ER	3	0.491	0.491	0.550	0.578	0.578
	MC	15	0.116	0.237	0.349	0.438	0.497
	MS	2	0.487	0.487	0.503	0.520	0.520
	TEI	10	0.227	0.362	0.494	0.522	0.546
	TPD	3	0.389	0.389	0.453	0.523	0.523
	TPI	3	0.416	0.416	0.517	0.536	0.536
7	CR	3	0.462	0.462	0.491	0.528	0.528
	ER	2	0.554	0.554	0.617	0.680	0.680
	MC	10	0.204	0.303	0.338	0.406	0.536
	MS	3	0.378	0.378	0.512	0.563	0.563
	TEI	14	0.224	0.361	0.458	0.554	0.643
	TPD	4	0.378	0.428	0.519	0.567	0.575
	TPI	2	0.421	0.421	0.494	0.568	0.568

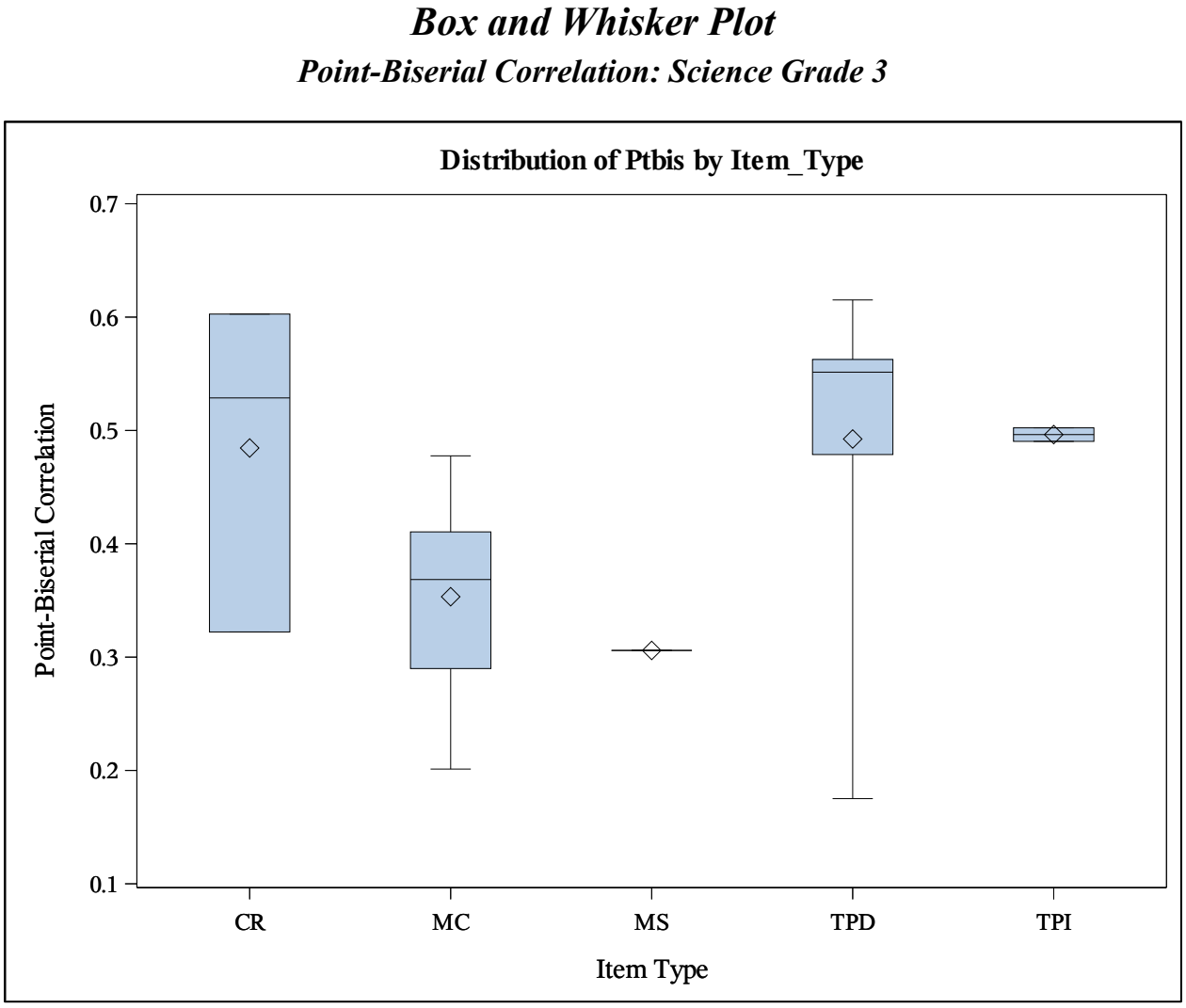
Table C.2.2

Item-Total Correlation Summary by Item Type: Spring 2023 Operational SC G3–8 (continued)

Grade	Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
8	CR	3	0.351	0.351	0.431	0.545	0.545
	ER	3	0.531	0.531	0.535	0.641	0.641
	MC	10	0.204	0.299	0.365	0.423	0.488
	MS	6	0.434	0.463	0.475	0.550	0.565
	TEI	13	0.311	0.415	0.467	0.526	0.641
	TPD	3	0.392	0.392	0.539	0.632	0.632
	TPI	1	0.432	0.432	0.432	0.432	0.432

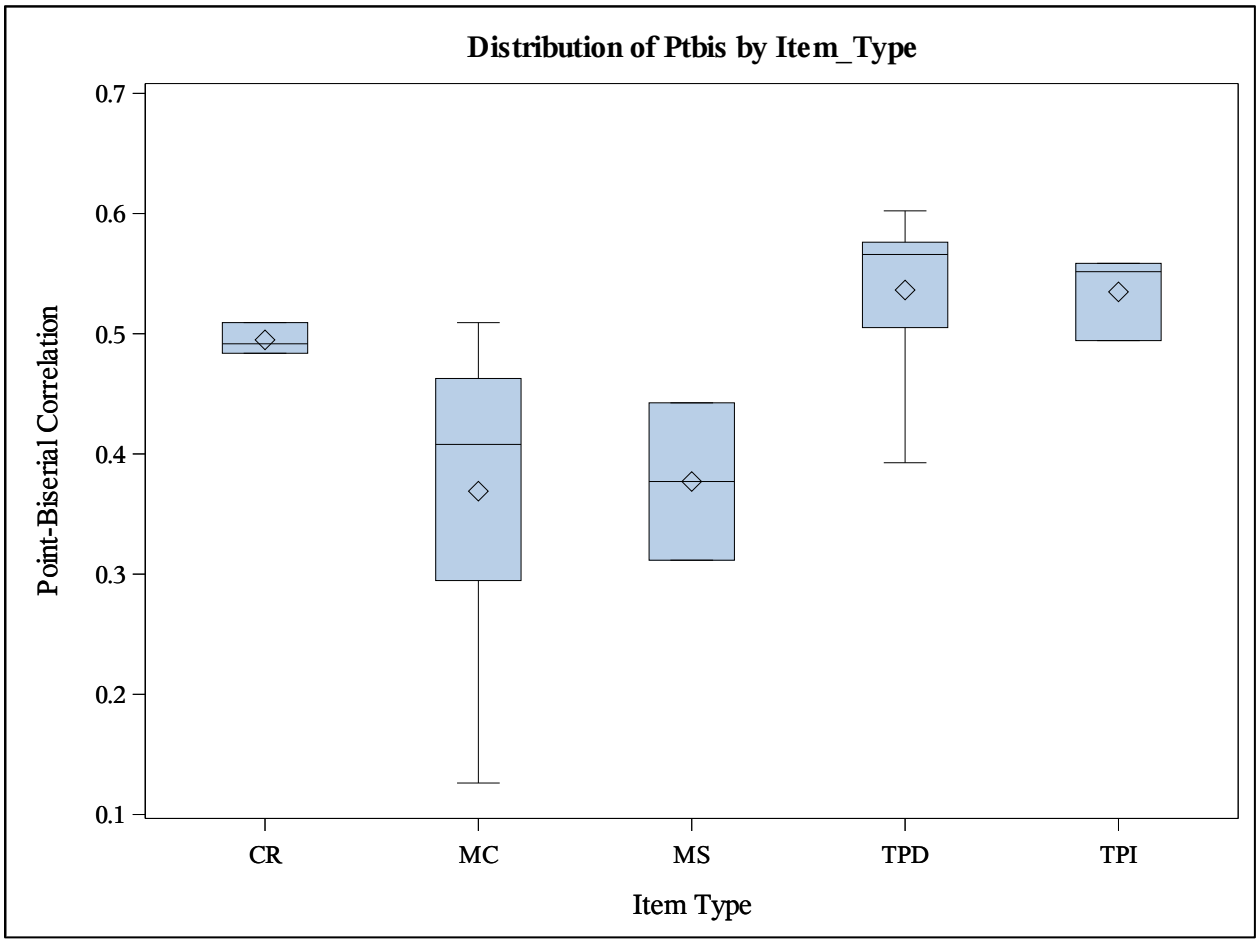
Plot C.2.2

Item-Total Correlation Summary by Item Type: Spring 2023 Operational SC G3–8



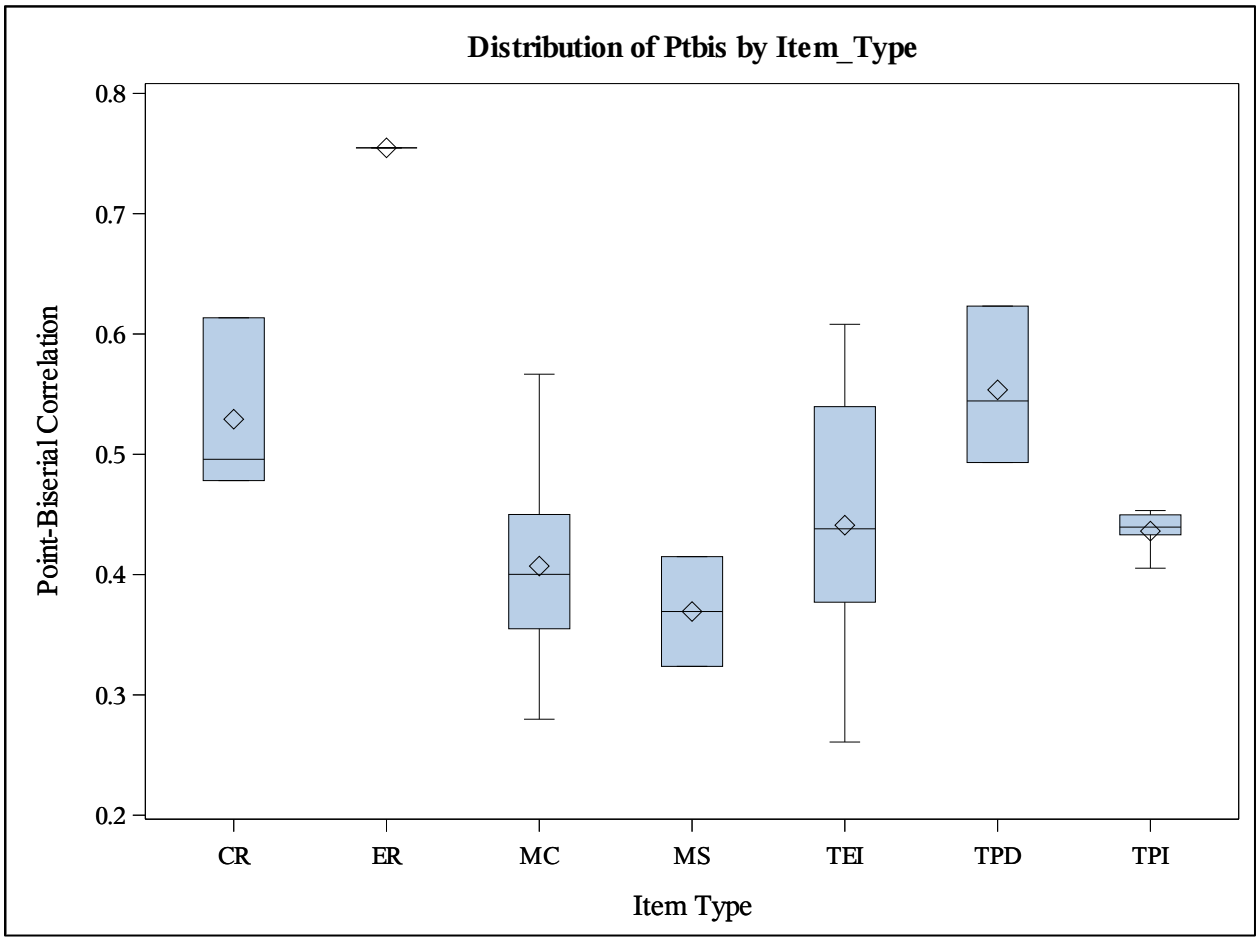
Box and Whisker Plot

Point-Biserial Correlation: Science Grade 4



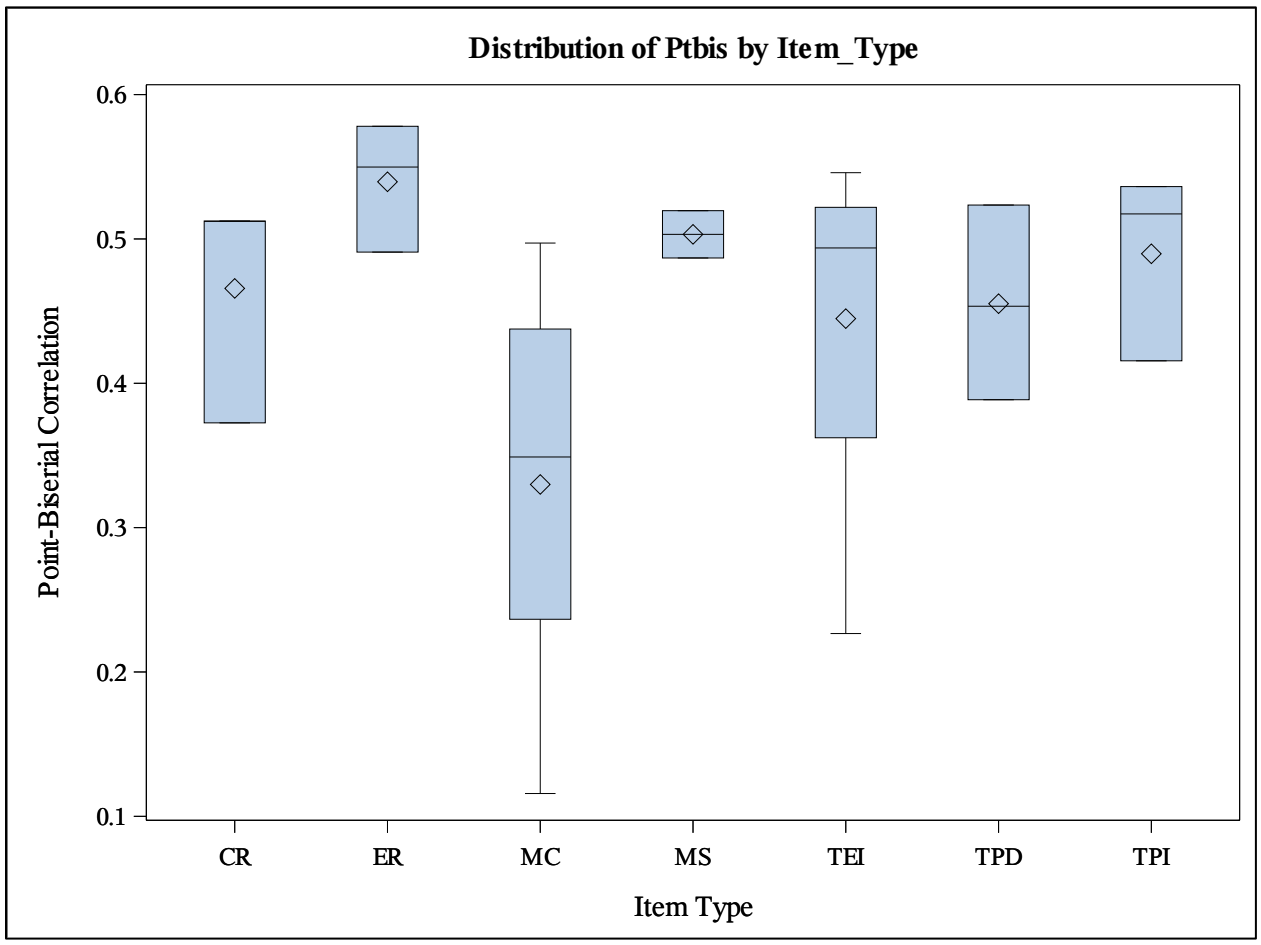
Box and Whisker Plot

Point-Biserial Correlation: Science Grade 5



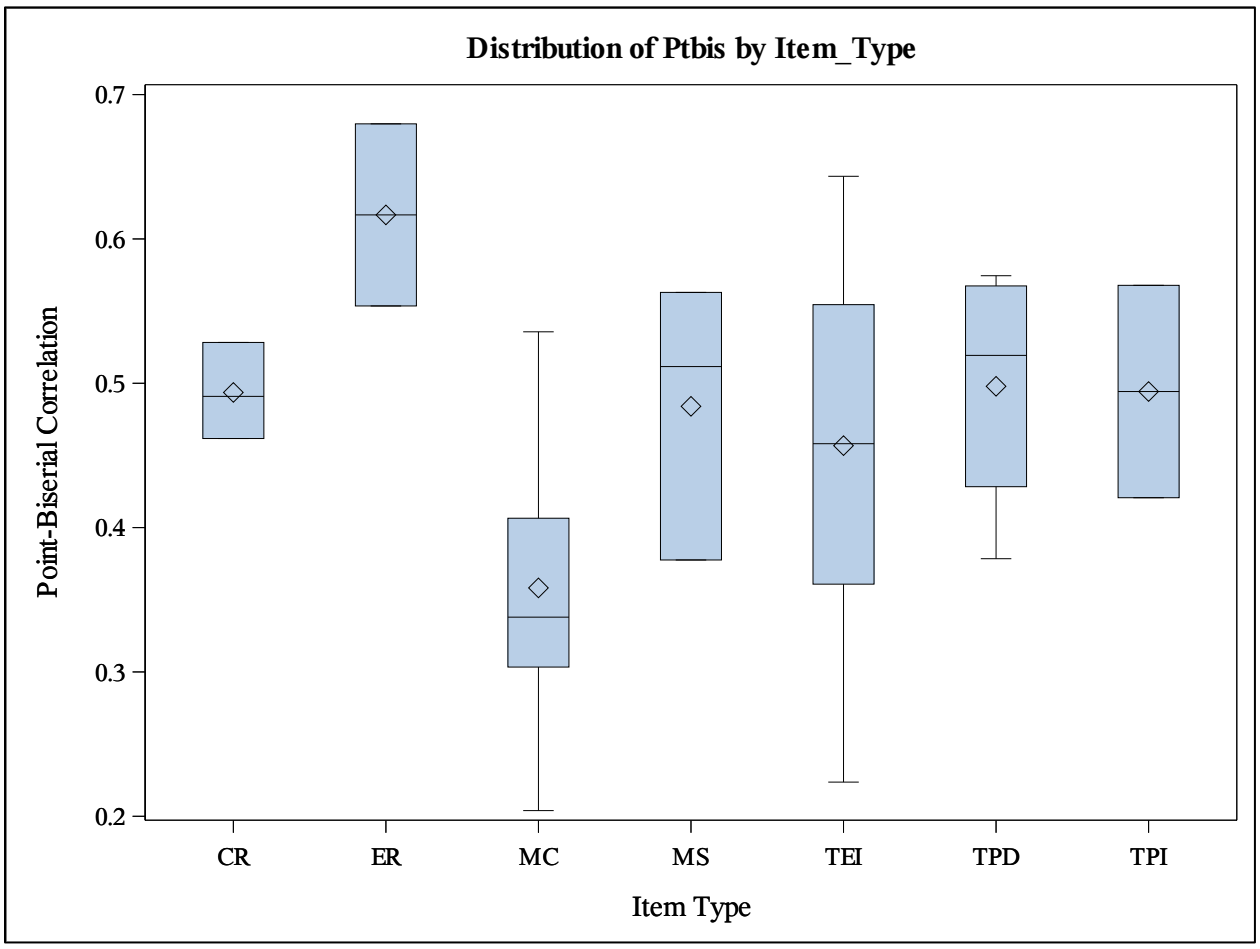
Box and Whisker Plot

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

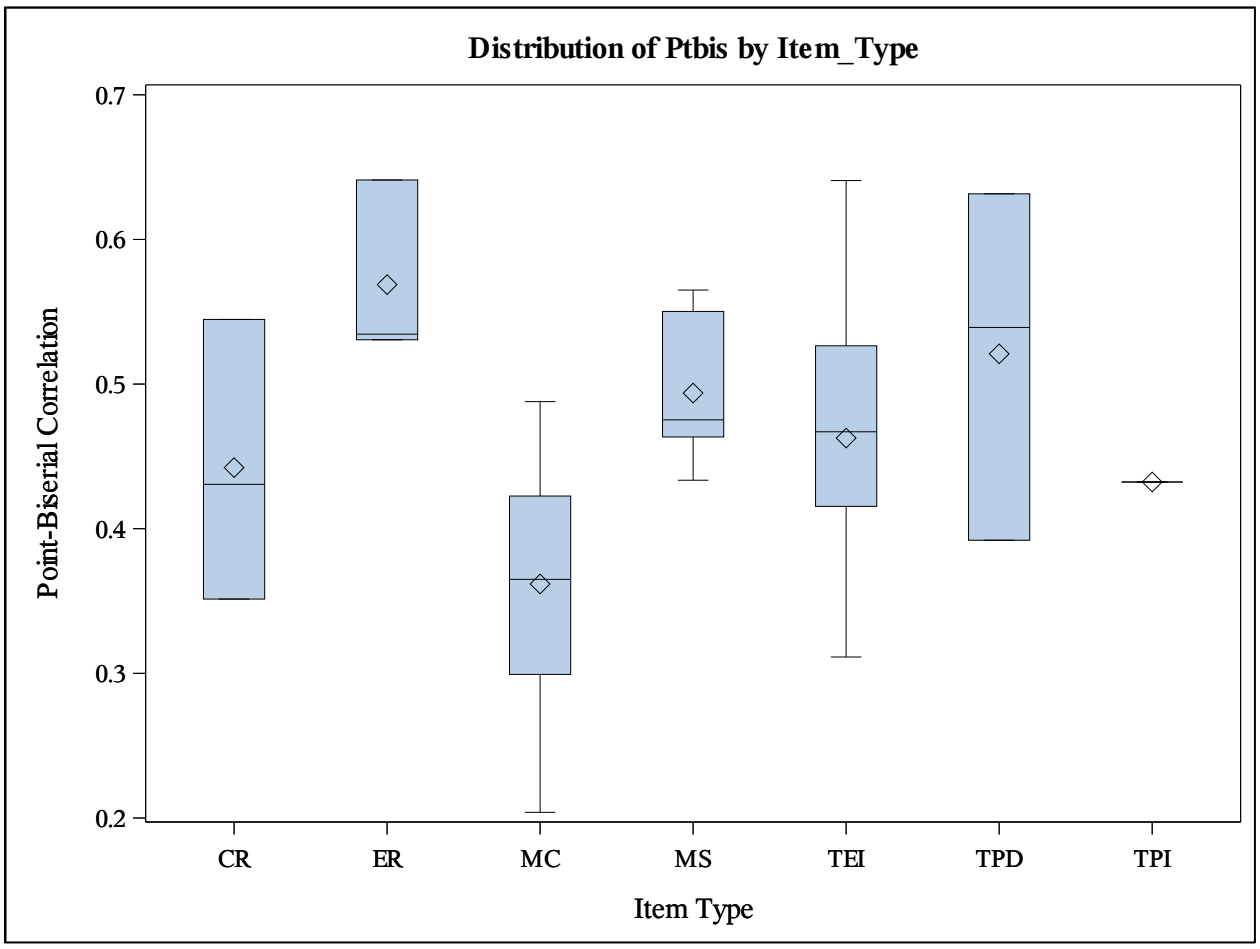


Table C.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2023 Operational SC G3–8*

Grade**	No. of Items	r < 0	0.0 ≤ r < 0.2	0.2 ≤ r < 0.3	0.3 ≤ r < 0.4	0.4 ≤ r < 0.5
3	36	0	3	9	12	9
4	36	0	4	4	8	14
5	37	0	0	6	13	11
6	39	0	5	3	12	18
7	38	0	2	6	8	13
8	39	0	1	5	13	13

* Corrected point-biserial correlation, which was 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 for Grades 6–8 were calculated and estimated separately for each dimension of the ER item,and the result summarize both dimensions.

Plot C.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2023 Operational SC G3–8

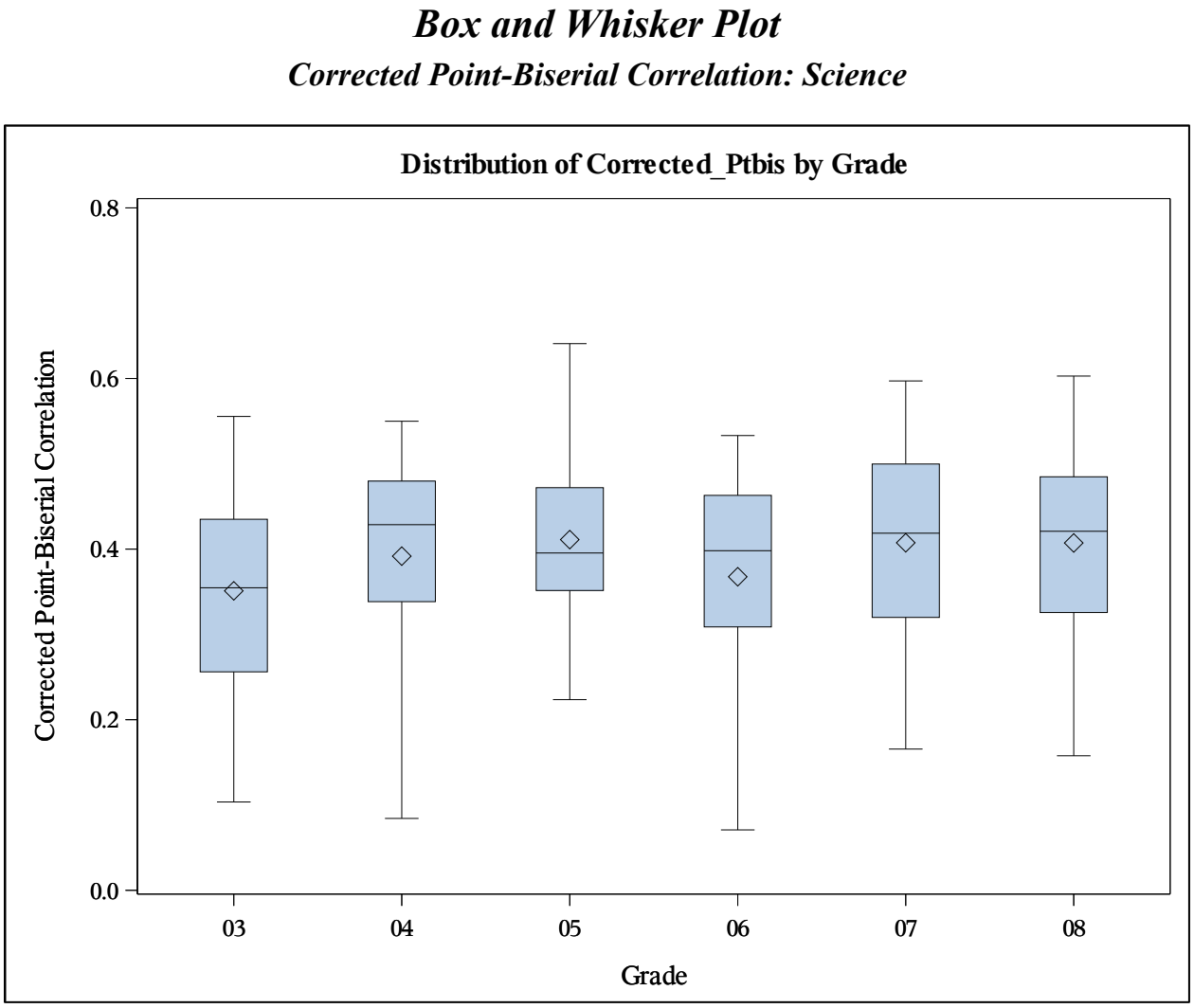


Table C.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2023 Operational SC G3–8*

Grade	Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	CR	3	0.275	0.275	0.474	0.556	0.556
	MC	21	0.150	0.244	0.321	0.363	0.433
	MS	1	0.261	0.261	0.261	0.261	0.261
	TPD	9	0.104	0.398	0.477	0.497	0.544
	TPI	2	0.424	0.424	0.433	0.442	0.442
4	CR	3	0.440	0.440	0.446	0.454	0.454
	MC	18	0.084	0.249	0.368	0.422	0.472
	MS	2	0.269	0.269	0.339	0.409	0.409
	TPD	10	0.318	0.440	0.502	0.516	0.550
	TPI	3	0.435	0.435	0.494	0.510	0.510
5	CR	3	0.431	0.431	0.456	0.566	0.566
	ER	1	0.641	0.641	0.641	0.641	0.641
	MC	8	0.241	0.317	0.364	0.416	0.537
	MS	2	0.291	0.291	0.339	0.387	0.387
	TEI	15	0.224	0.343	0.402	0.492	0.581
	TPD	3	0.436	0.436	0.493	0.571	0.571
	TPI	5	0.351	0.379	0.383	0.395	0.410
6	CR	3	0.326	0.326	0.465	0.469	0.469
	ER	3	0.421	0.421	0.490	0.533	0.533
	MC	15	0.071	0.196	0.309	0.398	0.460
	MS	2	0.451	0.451	0.469	0.486	0.486
	TEI	10	0.188	0.320	0.437	0.466	0.486
	TPD	3	0.315	0.315	0.384	0.463	0.463
	TPI	3	0.360	0.360	0.461	0.480	0.480
7	CR	3	0.419	0.419	0.444	0.481	0.481
	ER	2	0.473	0.473	0.532	0.591	0.591
	MC	10	0.166	0.264	0.300	0.375	0.505
	MS	3	0.342	0.342	0.480	0.533	0.533
	TEI	14	0.190	0.320	0.423	0.500	0.597
	TPD	4	0.329	0.373	0.464	0.514	0.516
	TPI	2	0.367	0.367	0.446	0.524	0.524

Table C.3.2

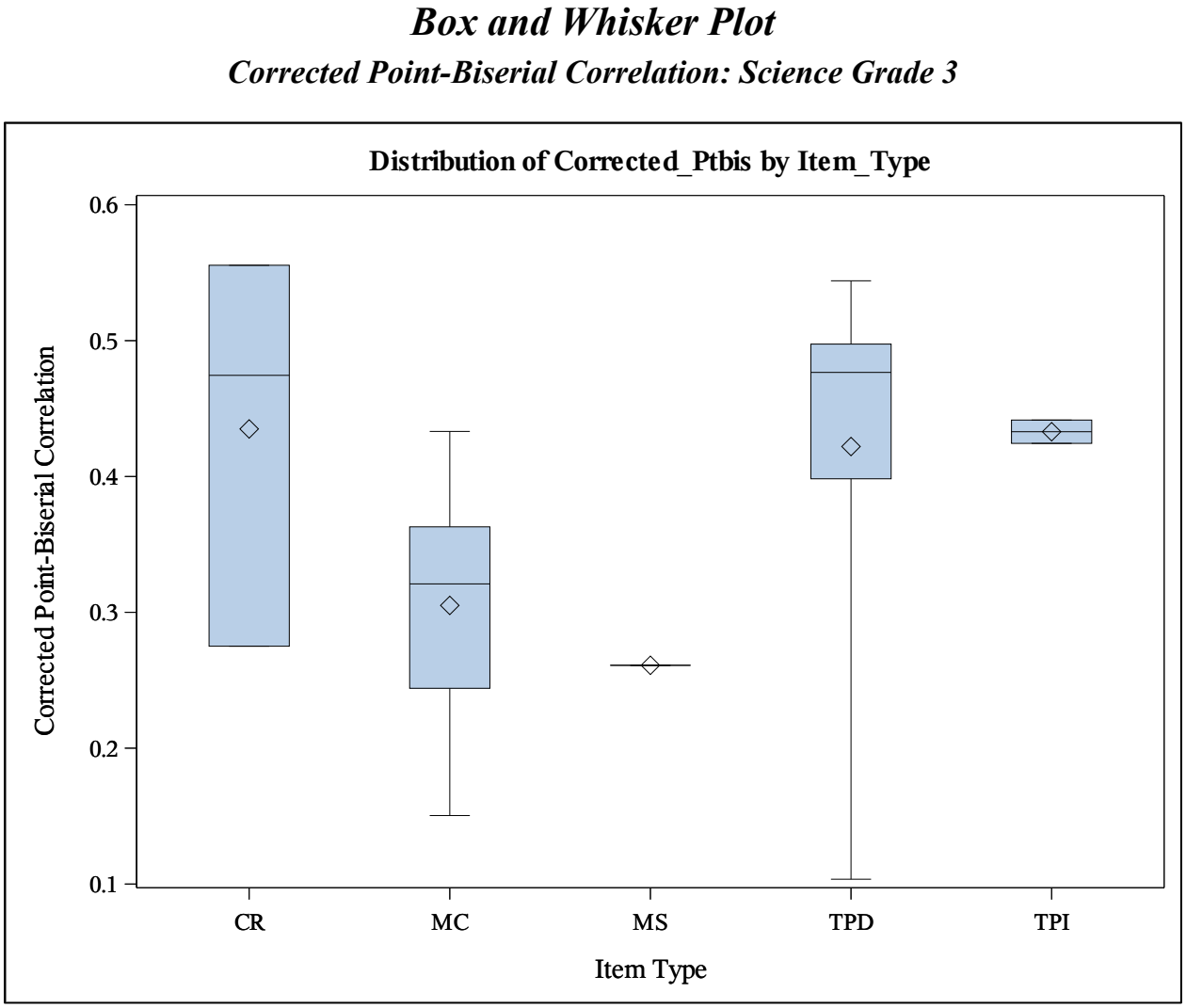
Corrected Point-Biserial Correlation Summary by Item Type: Spring 2023 Operational SC G3–8 (continued)*

Grade	Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
8	CR	3	0.315	0.315	0.395	0.507	0.507
	ER	3	0.485	0.485	0.495	0.603	0.603
	MC	10	0.158	0.257	0.323	0.384	0.450
	MS	6	0.397	0.426	0.440	0.516	0.531
	TEI	13	0.261	0.374	0.421	0.474	0.591
	TPD	3	0.316	0.316	0.476	0.577	0.577
	TPI	1	0.376	0.376	0.376	0.376	0.376

* Corrected point-biserial correlation, which was 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.

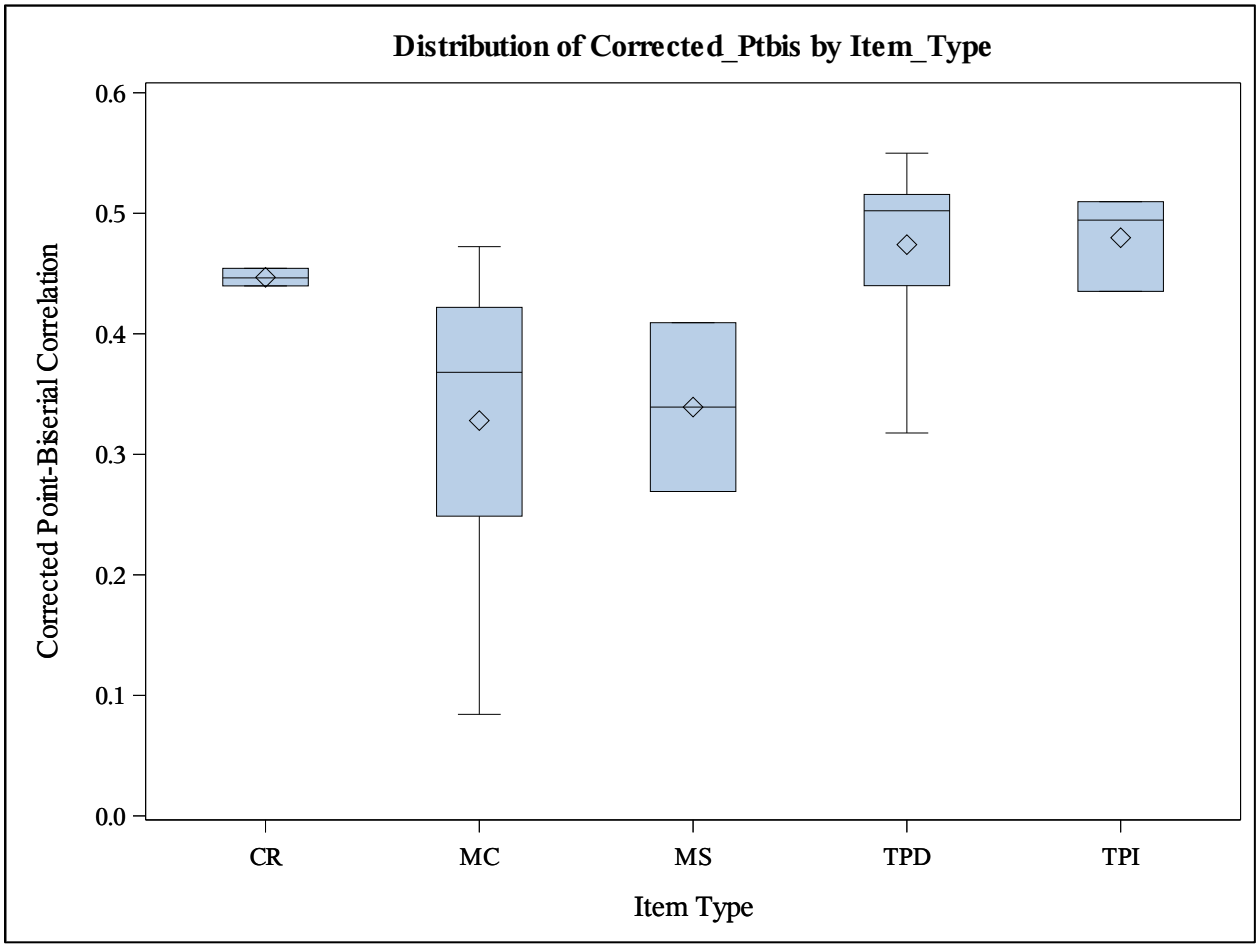
Plot C.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2023 Operational SC G3–8



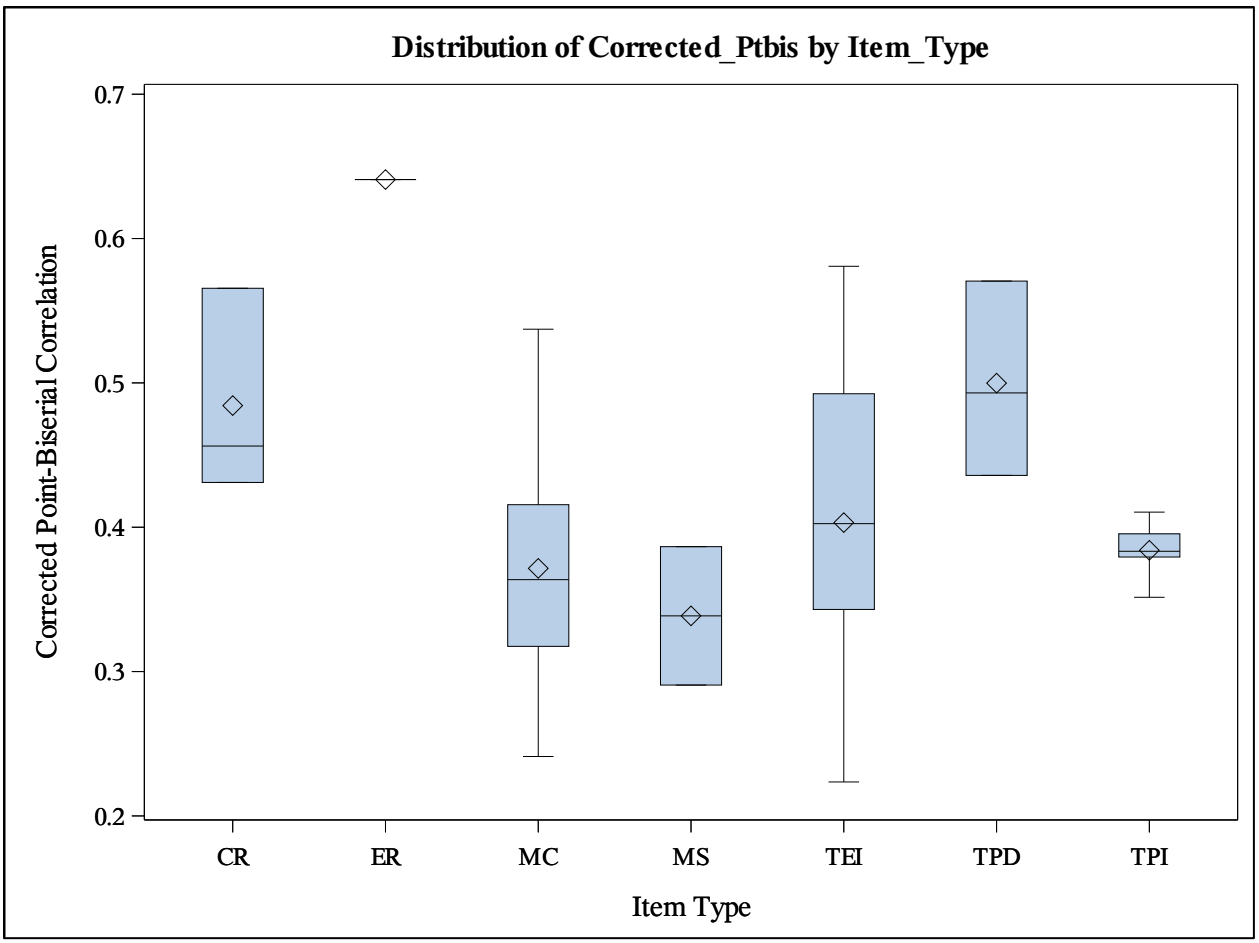
Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 4



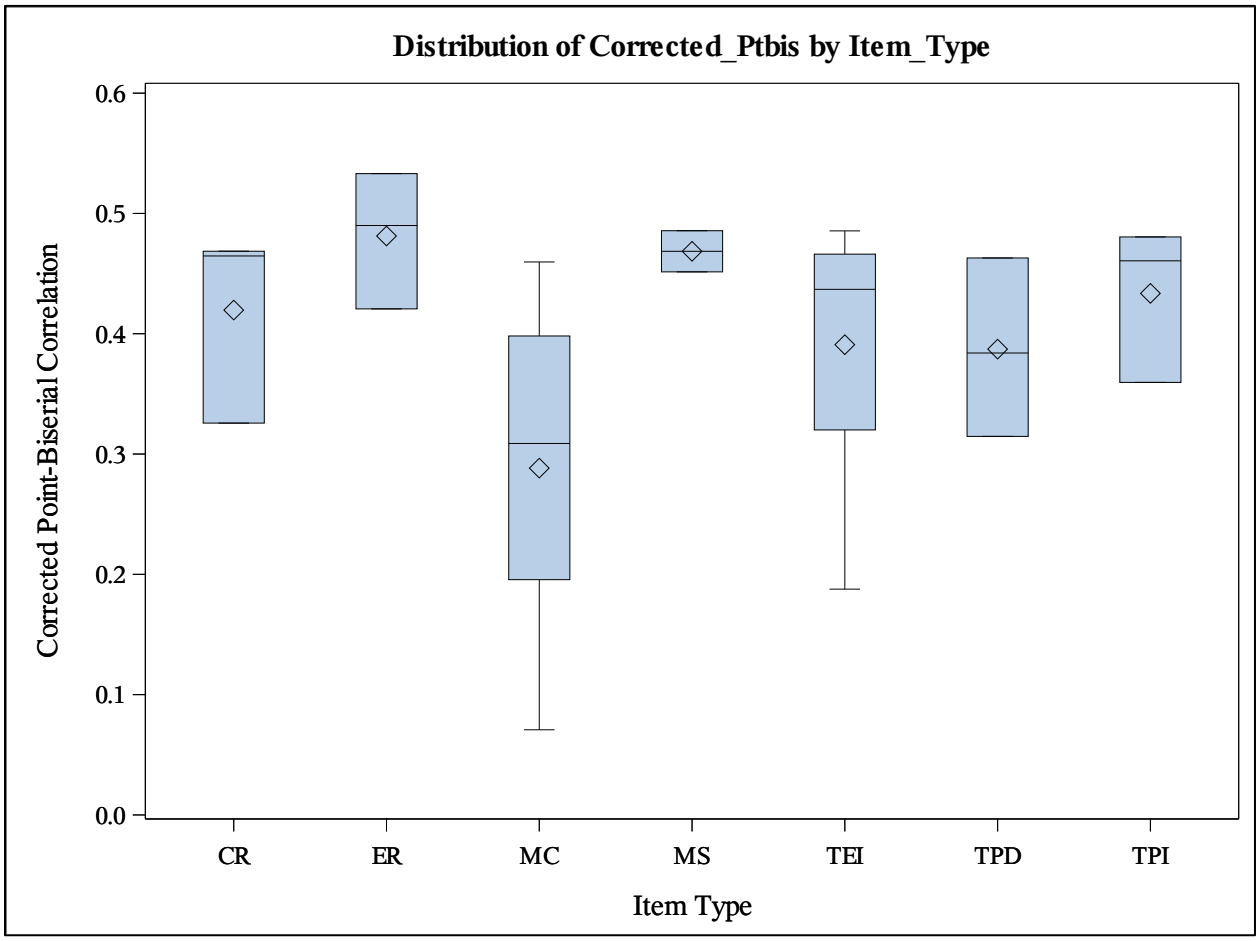
Box and Whisker Plot

Corrected Point-Biserial Correlation: Science Grade 5



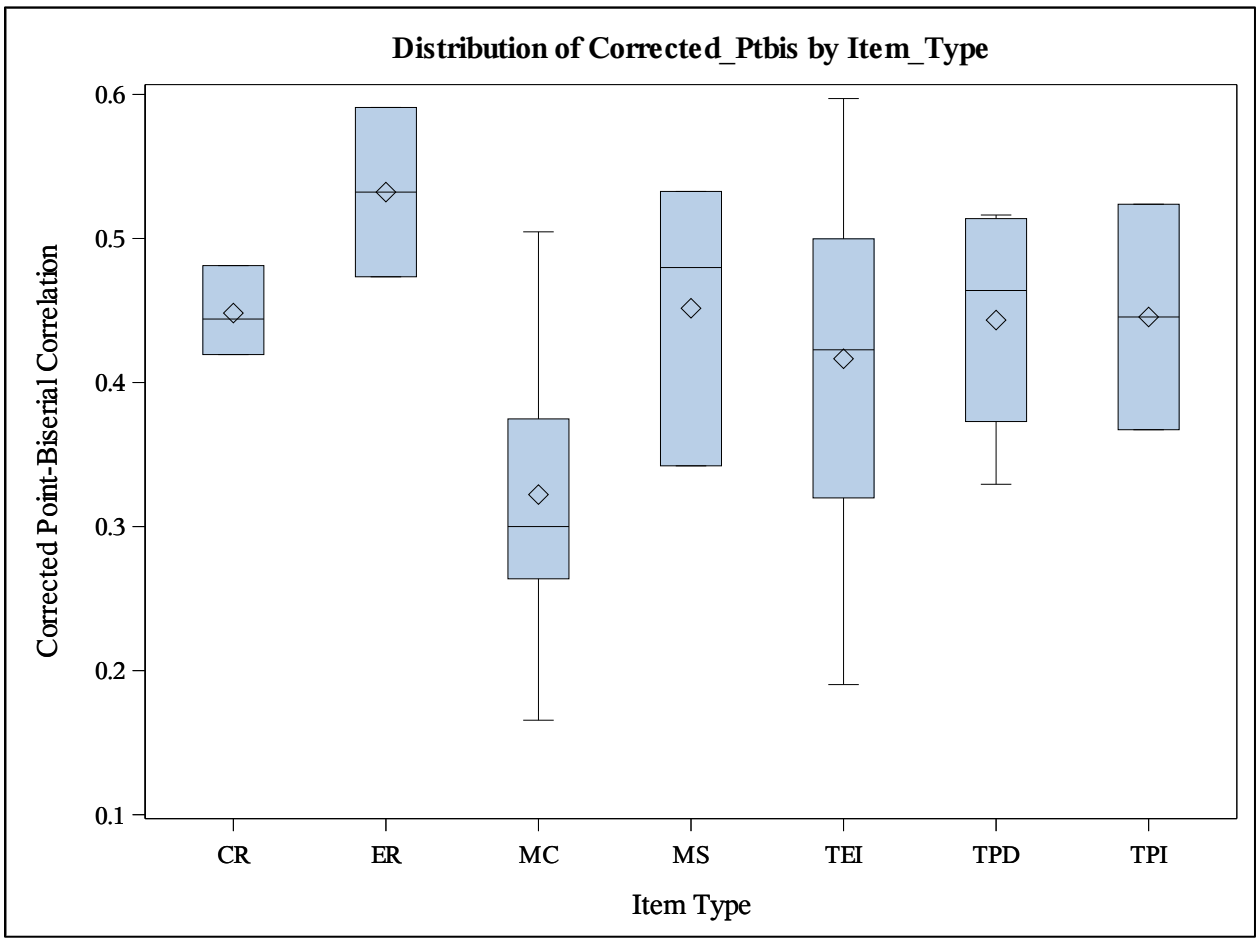
Box and Whisker Plot

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

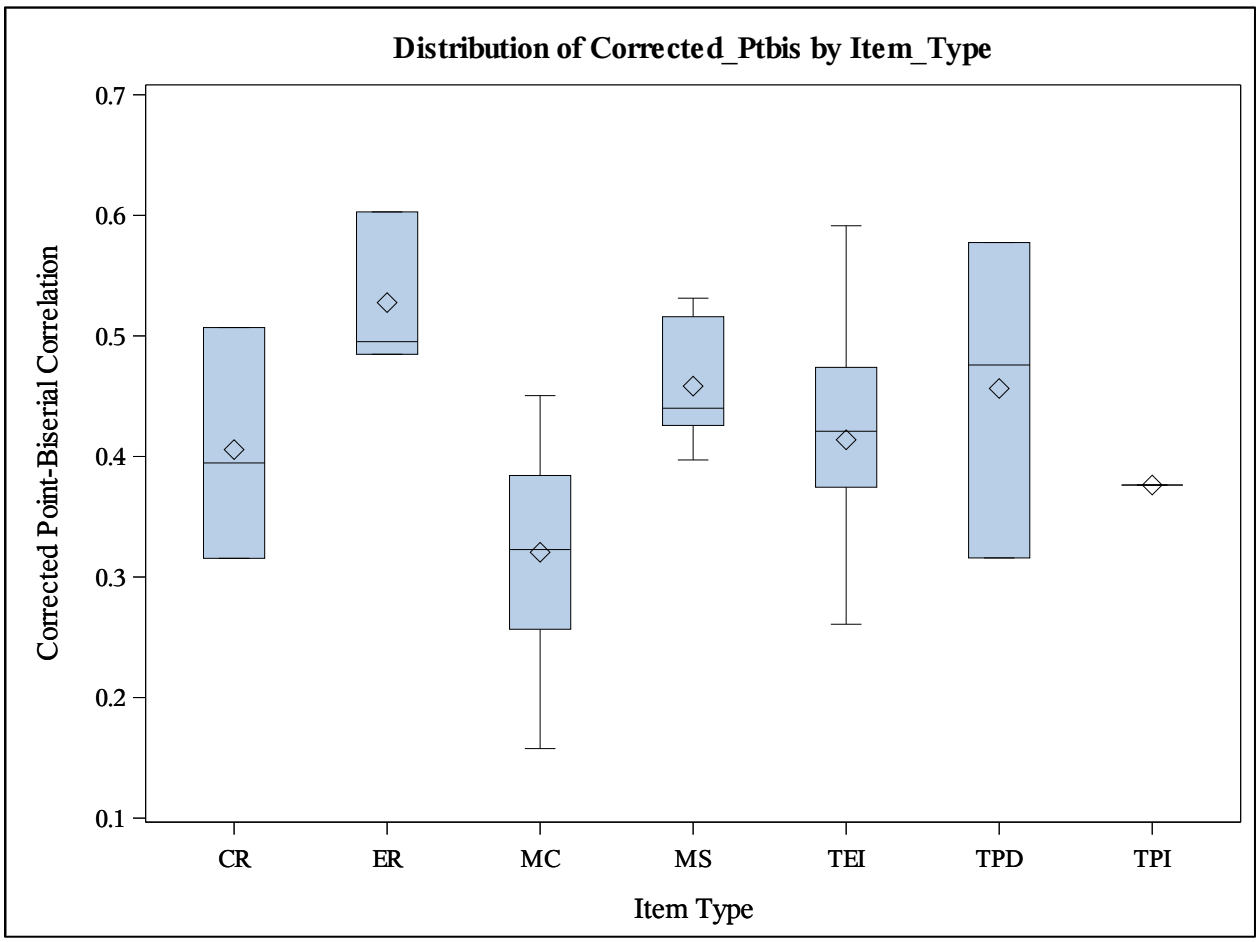


Table C.4.1

Item-Total Correlation Summary by Reporting Category: Spring 2023 Operational SC G3–8

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	1 Investigate	10	0.235	0.342	0.424	0.556	0.584
	2 Evaluate	18	0.175	0.294	0.393	0.447	0.615
	3 Reason Scientifically	7	0.201	0.29	0.388	0.519	0.603
4	1 Investigate	9	0.126	0.402	0.419	0.494	0.592
	2 Evaluate	7	0.229	0.295	0.445	0.506	0.552
	3 Reason Scientifically	17	0.154	0.393	0.492	0.509	0.602
5	1 Investigate	6	0.385	0.399	0.468	0.558	0.596
	2 Evaluate	15	0.285	0.413	0.475	0.544	0.613
	3 Reason Scientifically	16	0.261	0.352	0.407	0.484	0.755
6	1 Investigate	10	0.237	0.285	0.451	0.522	0.536
	2 Evaluate	13	0.116	0.354	0.453	0.517	0.578
	3 Reason Scientifically	15	0.117	0.362	0.468	0.502	0.546
7	1 Investigate	5	0.324	0.378	0.421	0.512	0.560
	2 Evaluate	8	0.405	0.434	0.471	0.581	0.643
	3 Reason Scientifically	24	0.204	0.326	0.463	0.545	0.680
8	1 Investigate	17	0.368	0.434	0.48	0.531	0.641
	2 Evaluate	9	0.299	0.392	0.526	0.544	0.565
	3 Reason Scientifically	10	0.204	0.319	0.377	0.431	0.641

Table C.4.2.1

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2023 SC G3–4

Grade	Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	CR	1 Investigate	1	0.529	0.529	0.529	0.529	0.529
		3 Reason Scientifically	2	0.322	0.322	0.462	0.603	0.603
	MC	1 Investigate	6	0.235	0.302	0.354	0.409	0.438
		2 Evaluate	12	0.273	0.285	0.390	0.424	0.478
		3 Reason Scientifically	3	0.201	0.201	0.290	0.388	0.388
	MS	2 Evaluate	1	0.306	0.306	0.306	0.306	0.306
	TPD	1 Investigate	3	0.556	0.556	0.563	0.584	0.584
		2 Evaluate	5	0.175	0.390	0.479	0.551	0.615
		3 Reason Scientifically	1	0.519	0.519	0.519	0.519	0.519
	TPI	3 Reason Scientifically	1	0.490	0.490	0.490	0.490	0.490
4	CR	3 Reason Scientifically	2	0.492	0.492	0.500	0.509	0.509
	MC	1 Investigate	6	0.126	0.313	0.403	0.419	0.424
		2 Evaluate	4	0.229	0.262	0.379	0.484	0.506
		3 Reason Scientifically	8	0.154	0.286	0.419	0.495	0.509
	MS	2 Evaluate	1	0.443	0.443	0.443	0.443	0.443
		3 Reason Scientifically	1	0.312	0.312	0.312	0.312	0.312
	TPD	1 Investigate	2	0.550	0.550	0.571	0.592	0.592
		2 Evaluate	1	0.445	0.445	0.445	0.445	0.445
		3 Reason Scientifically	6	0.393	0.505	0.566	0.576	0.602
	TPI	1 Investigate	1	0.494	0.494	0.494	0.494	0.494
		2 Evaluate	1	0.552	0.552	0.552	0.552	0.552

Table C.4.2.2

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2023 SC G5–6

Grade	Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
5	CR	1 Investigate	1	0.496	0.496	0.496	0.496	0.496
		2 Evaluate	1	0.613	0.613	0.613	0.613	0.613
		3 Reason Scientifically	1	0.478	0.478	0.478	0.478	0.478
	ER	3 Reason Scientifically	1	0.755	0.755	0.755	0.755	0.755
	MC	1 Investigate	2	0.385	0.385	0.392	0.399	0.399
		2 Evaluate	2	0.325	0.325	0.446	0.567	0.567
		3 Reason Scientifically	4	0.280	0.341	0.405	0.450	0.491
	MS	2 Evaluate	1	0.415	0.415	0.415	0.415	0.415
		3 Reason Scientifically	1	0.324	0.324	0.324	0.324	0.324
	TEI	1 Investigate	2	0.558	0.558	0.577	0.596	0.596
		2 Evaluate	8	0.285	0.395	0.457	0.522	0.608
		3 Reason Scientifically	5	0.261	0.284	0.381	0.403	0.492
	TPD	2 Evaluate	2	0.493	0.493	0.519	0.544	0.544
		3 Reason Scientifically	1	0.623	0.623	0.623	0.623	0.623
	TPI	1 Investigate	1	0.439	0.439	0.439	0.439	0.439
		2 Evaluate	1	0.453	0.453	0.453	0.453	0.453
		3 Reason Scientifically	3	0.405	0.405	0.433	0.450	0.450
6	CR	1 Investigate	1	0.512	0.512	0.512	0.512	0.512
		3 Reason Scientifically	2	0.373	0.373	0.443	0.512	0.512
	ER	2 Evaluate	3	0.491	0.491	0.550	0.578	0.578
	MC	1 Investigate	4	0.237	0.256	0.280	0.317	0.349
		2 Evaluate	4	0.116	0.235	0.355	0.394	0.433
		3 Reason Scientifically	7	0.117	0.201	0.438	0.475	0.497
	MS	1 Investigate	1	0.487	0.487	0.487	0.487	0.487
		2 Evaluate	1	0.520	0.520	0.520	0.520	0.520
	TEI	1 Investigate	1	0.522	0.522	0.522	0.522	0.522
		2 Evaluate	3	0.227	0.227	0.322	0.486	0.486
		3 Reason Scientifically	5	0.362	0.501	0.502	0.546	0.546
	TPD	1 Investigate	1	0.523	0.523	0.523	0.523	0.523
		2 Evaluate	1	0.453	0.453	0.453	0.453	0.453
	TPI	3 Reason Scientifically	1	0.389	0.389	0.389	0.389	0.389
		1 Investigate	2	0.416	0.416	0.476	0.536	0.536
		2 Evaluate	1	0.517	0.517	0.517	0.517	0.517

Table C.4.2.3

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2023 SC G7–8

Grade	Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
7	CR	2 Evaluate	1	0.462	0.462	0.462	0.462	0.462
		3 Reason Scientifically	2	0.491	0.491	0.510	0.528	0.528
	ER	3 Reason Scientifically	2	0.554	0.554	0.617	0.680	0.680
	MC	2 Evaluate	2	0.405	0.405	0.406	0.406	0.406
		3 Reason Scientifically	8	0.204	0.280	0.326	0.409	0.536
	MS	1 Investigate	1	0.512	0.512	0.512	0.512	0.512
		2 Evaluate	1	0.563	0.563	0.563	0.563	0.563
		3 Reason Scientifically	1	0.378	0.378	0.378	0.378	0.378
	TEI	1 Investigate	1	0.324	0.324	0.324	0.324	0.324
		2 Evaluate	4	0.464	0.471	0.539	0.621	0.643
		3 Reason Scientifically	8	0.224	0.351	0.447	0.541	0.624
	TPD	1 Investigate	2	0.378	0.378	0.469	0.560	0.560
		3 Reason Scientifically	2	0.478	0.478	0.526	0.575	0.575
	TPI	1 Investigate	1	0.421	0.421	0.421	0.421	0.421
		3 Reason Scientifically	1	0.568	0.568	0.568	0.568	0.568
8	CR	2 Evaluate	1	0.545	0.545	0.545	0.545	0.545
		3 Reason Scientifically	2	0.351	0.351	0.391	0.431	0.431
	ER	1 Investigate	3	0.531	0.531	0.535	0.641	0.641
	MC	1 Investigate	4	0.368	0.389	0.416	0.451	0.480
		2 Evaluate	3	0.299	0.299	0.308	0.488	0.488
		3 Reason Scientifically	3	0.204	0.204	0.276	0.362	0.362
	MS	1 Investigate	5	0.434	0.463	0.465	0.485	0.550
		2 Evaluate	1	0.565	0.565	0.565	0.565	0.565
	TEI	1 Investigate	5	0.415	0.439	0.493	0.499	0.540
		2 Evaluate	2	0.526	0.526	0.535	0.544	0.544
		3 Reason Scientifically	4	0.319	0.356	0.410	0.534	0.641
	TPD	2 Evaluate	2	0.392	0.392	0.466	0.539	0.539
	TPI	3 Reason Scientifically	1	0.432	0.432	0.432	0.432	0.432

Table C.5.1.1

IRT-A Parameter Summary by Reporting Category: SC G3

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
3	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	1	0	1
	0.2 <= a < 0.4	0	4	2	6
	0.4 <= a < 0.6	5	2	1	9
	0.6 <= a < 0.8	2	7	1	10
	0.8 <= a < 1.0	3	0	2	5
	1.0 <= a < 1.2	0	2	0	2
	1.2 <= a < 1.4	0	2	0	2
	1.4 <= a < 1.6	0	0	1	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	0.42	0.09	0.37	0.09
	Maximum	0.95	1.33	1.53	1.53
	Mean	0.64	0.68	0.73	0.67
	SD	0.19	0.33	0.41	0.31
	Number of Items	10	18	7	36

Table C.5.1.2

IRT-A Parameter Summary by Reporting Category: SC G4

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
4	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	0	0	0
	0.2 <= a < 0.4	0	1	1	2
	0.4 <= a < 0.6	3	3	9	16
	0.6 <= a < 0.8	2	1	4	9
	0.8 <= a < 1.0	1	1	1	3
	1.0 <= a < 1.2	2	1	1	4
	1.2 <= a < 1.4	0	0	1	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	0.42	0.39	0.24	0.24
	Maximum	1.45	1.10	1.34	1.45
	Mean	0.81	0.64	0.61	0.66
	SD	0.35	0.25	0.27	0.28
	Number of Items	9	7	17	36

Table C.5.1.3

IRT-A Parameter Summary by Reporting Category: SC G5

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
5	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	0	0	0
	0.2 <= a < 0.4	1	2	5	8
	0.4 <= a < 0.6	4	5	7	16
	0.6 <= a < 0.8	1	2	2	5
	0.8 <= a < 1.0	0	4	2	6
	1.0 <= a < 1.2	0	1	0	1
	1.2 <= a < 1.4	0	1	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	0.32	0.35	0.26	0.26
	Maximum	0.60	1.28	0.93	1.28
	Mean	0.51	0.68	0.52	0.58
	SD	0.10	0.29	0.21	0.24
	Number of Items	6	15	16	37

Table C.5.1.4

IRT-A Parameter Summary by Reporting Category: SC G6

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
6	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	1	0	1
	0.2 <= a < 0.4	2	2	3	7
	0.4 <= a < 0.6	4	5	6	15
	0.6 <= a < 0.8	3	2	2	8
	0.8 <= a < 1.0	0	2	1	3
	1.0 <= a < 1.2	1	1	2	4
	1.2 <= a < 1.4	0	0	1	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	0.32	0.19	0.22	0.19
	Maximum	1.17	1.16	1.31	1.31
	Mean	0.61	0.56	0.63	0.60
	SD	0.25	0.28	0.32	0.28
	Number of Items	10	13	15	39

Table C.5.1.5

IRT-A Parameter Summary by Reporting Category: SC G7

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
7	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	0	0	0
	0.2 <= a < 0.4	2	1	4	8
	0.4 <= a < 0.6	1	2	10	13
	0.6 <= a < 0.8	1	1	4	6
	0.8 <= a < 1.0	1	1	4	6
	1.0 <= a < 1.2	0	3	1	4
	1.2 <= a < 1.4	0	0	1	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	0.30	0.40	0.28	0.28
	Maximum	0.81	1.16	1.35	1.35
	Mean	0.53	0.77	0.62	0.63
	SD	0.23	0.30	0.27	0.27
	Number of Items	5	8	24	38

Table C.5.1.6

IRT-A Parameter Summary by Reporting Category: SC G8

Grade	IRT-a Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
8	a < 0.0	0	0	0	0
	0.0 <= a < 0.2	0	0	0	0
	0.2 <= a < 0.4	2	2	2	7
	0.4 <= a < 0.6	2	4	3	9
	0.6 <= a < 0.8	6	1	4	13
	0.8 <= a < 1.0	6	2	0	8
	1.0 <= a < 1.2	1	0	0	1
	1.2 <= a < 1.4	0	0	1	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	0.39	0.23	0.37	0.23
	Maximum	1.19	0.92	1.22	1.22
	Mean	0.75	0.56	0.61	0.65
	SD	0.22	0.24	0.24	0.23
	Number of Items	17	9	10	39

Table C.5.2.1

IRT-B Parameter Summary by Reporting Category: SC G3

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
3	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	1	1	1	3
	0.0 <= b < 0.5	4	3	1	8
	0.5 <= b < 1.0	2	5	1	8
	1.0 <= b < 1.5	2	4	1	8
	1.5 <= b < 2.0	1	4	1	6
	2.0 <= b < 2.5	0	0	1	1
	2.5 <= b < 3.0	0	0	0	0
	3.0 <= b < 3.5	0	0	1	1
	3.5 <= b	0	1	0	1
	Minimum	-0.50	-0.05	-0.21	-0.50
	Maximum	1.95	6.49	3.18	6.49
	Mean	0.67	1.25	1.24	1.08
	SD	0.67	1.41	1.17	1.18
	Number of Items	10	18	7	36

Table C.5.2.2

IRT-B Parameter Summary by Reporting Category: SC G4

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
4	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	2	0	0	2
	-0.5 <= b < 0.0	1	1	4	6
	0.0 <= b < 0.5	1	2	5	8
	0.5 <= b < 1.0	3	0	2	7
	1.0 <= b < 1.5	0	1	2	3
	1.5 <= b < 2.0	1	2	2	5
	2.0 <= b < 2.5	1	1	1	4
	2.5 <= b < 3.0	0	0	1	1
	3.0 <= b < 3.5	0	0	0	0
	3.5 <= b	0	0	0	0
	Minimum	-0.95	-0.41	-0.39	-0.95
	Maximum	2.41	2.34	2.94	2.94
	Mean	0.62	0.93	0.72	0.77
	SD	1.11	1.00	0.99	0.99
	Number of Items	9	7	17	36

Table C.5.2.3

IRT-B Parameter Summary by Reporting Category: SC G5

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
5	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	1	2
	-1.0 <= b < -0.5	1	0	1	2
	-0.5 <= b < 0.0	2	3	4	9
	0.0 <= b < 0.5	1	1	4	6
	0.5 <= b < 1.0	0	6	0	6
	1.0 <= b < 1.5	0	1	1	2
	1.5 <= b < 2.0	1	4	5	10
	2.0 <= b < 2.5	0	0	0	0
	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	-1.07	-0.48	-1.43	-1.43
	Maximum	1.79	1.86	1.90	1.90
	Mean	-0.05	0.78	0.51	0.53
	SD	1.03	0.77	1.01	0.94
	Number of Items	6	15	16	37

Table C.5.2.4

IRT-B Parameter Summary by Reporting Category: SC G6

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
6	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	1	1	2	4
	-0.5 <= b < 0.0	0	2	3	5
	0.0 <= b < 0.5	2	3	1	6
	0.5 <= b < 1.0	2	1	2	6
	1.0 <= b < 1.5	4	1	3	8
	1.5 <= b < 2.0	1	2	1	4
	2.0 <= b < 2.5	0	3	1	4
	2.5 <= b < 3.0	0	0	1	1
	3.0 <= b < 3.5	0	0	0	0
	3.5 <= b	0	0	0	0
	Minimum	-0.54	-0.53	-1.12	-1.12
	Maximum	1.68	2.26	2.85	2.85
	Mean	0.80	0.86	0.60	0.74
	SD	0.69	1.04	1.18	0.99
	Number of Items	10	13	15	39

Table C.5.2.5

IRT-B Parameter Summary by Reporting Category: SC G7

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
7	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	2	2
	-1.5 <= b < -1.0	0	0	1	1
	-1.0 <= b < -0.5	0	0	1	1
	-0.5 <= b < 0.0	0	3	3	6
	0.0 <= b < 0.5	2	1	3	6
	0.5 <= b < 1.0	1	2	4	7
	1.0 <= b < 1.5	0	1	5	6
	1.5 <= b < 2.0	1	1	4	6
	2.0 <= b < 2.5	1	0	1	2
	3.5 <= b	0	0	0	0
	Minimum	0.25	-0.45	-1.92	-1.92
	Maximum	2.11	1.67	2.14	2.64
	Mean	1.07	0.43	0.54	0.65
	SD	0.78	0.79	1.11	1.05
	Number of Items	5	8	24	38

Table C.5.2.6

IRT-B Parameter Summary by Reporting Category: SC G8

Grade	IRT-b Range	Investigate	Evaluate	Reason Scientifically	Total Number of Items
8	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	1
	-1.5 <= b < -1.0	0	0	0	0
	-1.0 <= b < -0.5	3	1	1	5
	-0.5 <= b < 0.0	2	1	1	5
	0.0 <= b < 0.5	2	3	0	6
	0.5 <= b < 1.0	5	2	1	8
	1.0 <= b < 1.5	2	1	2	5
	1.5 <= b < 2.0	2	1	3	6
	2.0 <= b < 2.5	1	0	1	2
	2.5 <= b < 3.0	0	0	1	1
	3.0 <= b < 3.5	0	0	0	0
	3.5 <= b	0	0	0	0
	Minimum	-0.70	-1.00	-0.60	-1.79
	Maximum	2.21	1.65	2.75	2.75
	Mean	0.55	0.41	1.23	0.61
	SD	0.87	0.77	1.05	0.99
	Number of Items	17	9	10	39

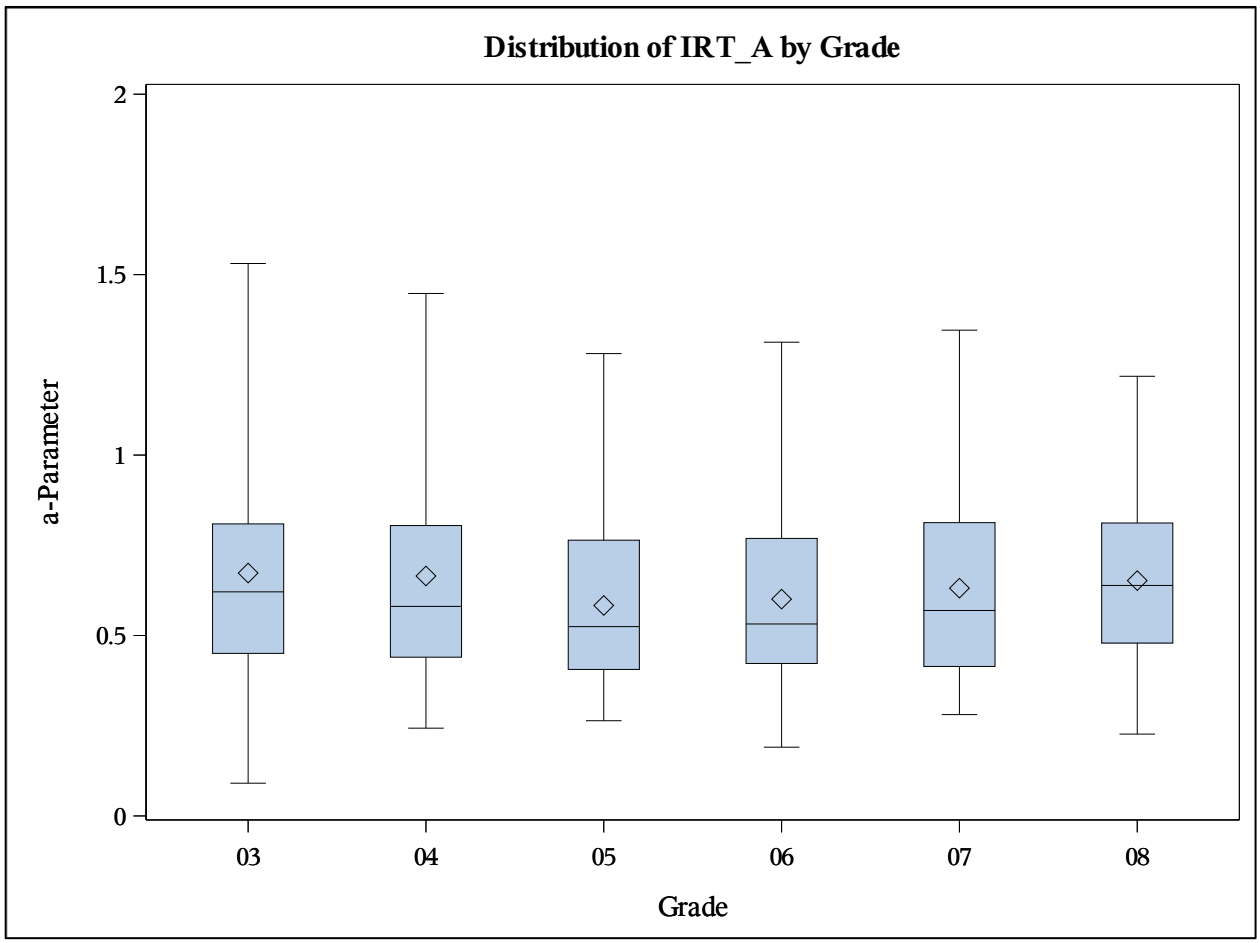
Table C.5.3

IRT Parameter Summary: Spring 2023 Operational SC G3–8

Grade	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	a	36	0.091	0.45	0.621	0.809	1.531
	b	36	-0.496	0.451	0.825	1.451	6.491
	c	22	0	0.151	0.205	0.234	0.281
4	a	36	0.243	0.44	0.581	0.805	1.448
	b	36	-0.953	0.088	0.594	1.537	2.94
	c	20	0.027	0.122	0.164	0.23	0.304
5	a	37	0.264	0.406	0.525	0.764	1.281
	b	37	-1.429	-0.277	0.421	1.515	1.903
	c	20	0.003	0.042	0.118	0.186	0.614
6	a	39	0.191	0.422	0.532	0.769	1.313
	b	39	-1.124	-0.07	0.688	1.494	2.852
	c	20	0	0.141	0.216	0.316	0.371
7	a	38	0.281	0.414	0.569	0.813	1.346
	b	38	-1.922	-0.051	0.703	1.479	2.643
	c	20	0.001	0.036	0.106	0.133	0.33
8	a	39	0.227	0.479	0.639	0.812	1.218
	b	39	-1.792	-0.181	0.6	1.39	2.754
	c	20	0.001	0.026	0.116	0.179	0.421

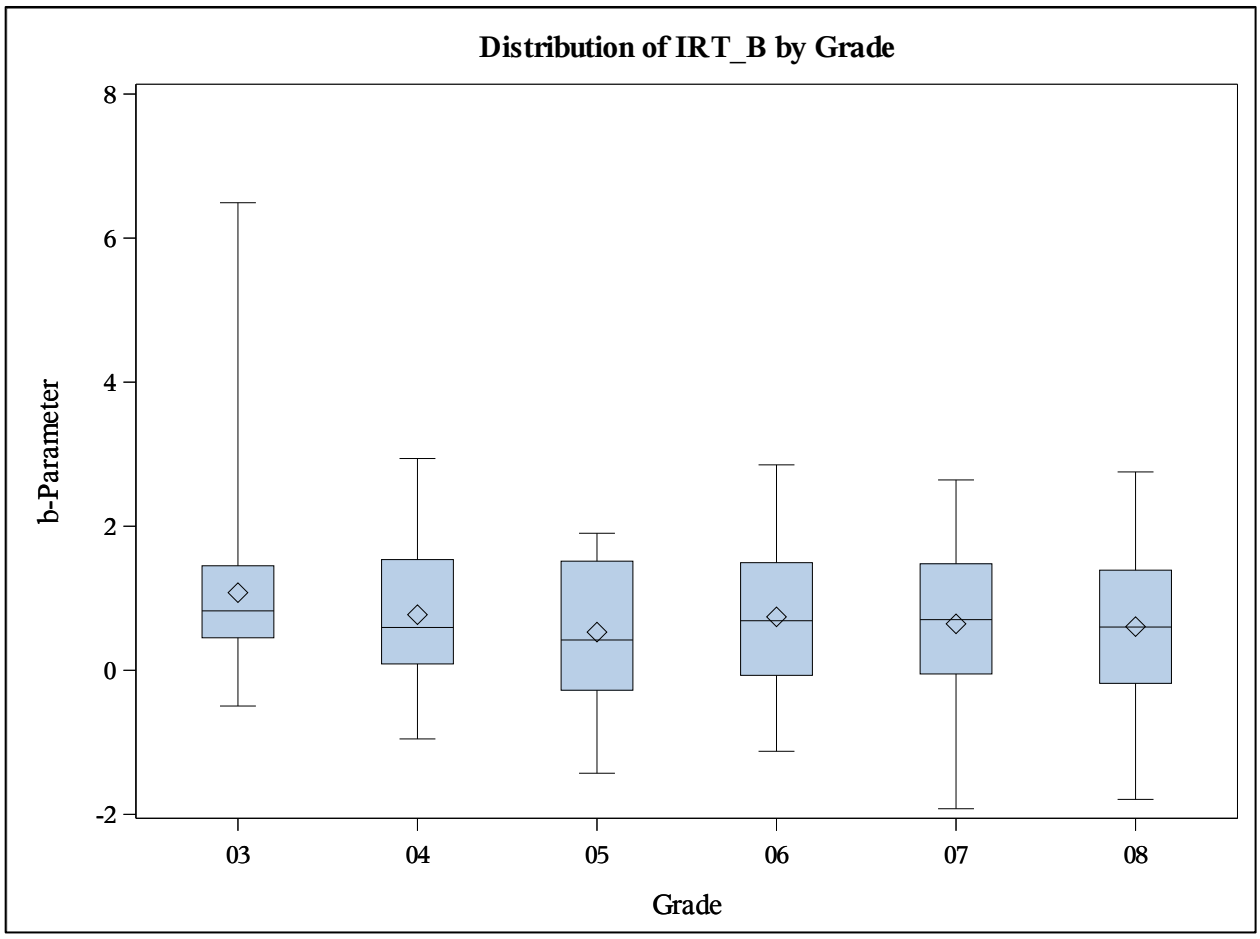
Plot C.5.1

IRT Item Parameter Summary for Spring 2023 Operational SC G3–8: A-Parameter



Plot C.5.2

IRT Item Parameter Summary for Spring 2023 Operational SC G3–8: B-Parameter



Plot C.5.3

IRT Item Parameter Summary for Spring 2023 Operational SC G3–8: C-Parameter

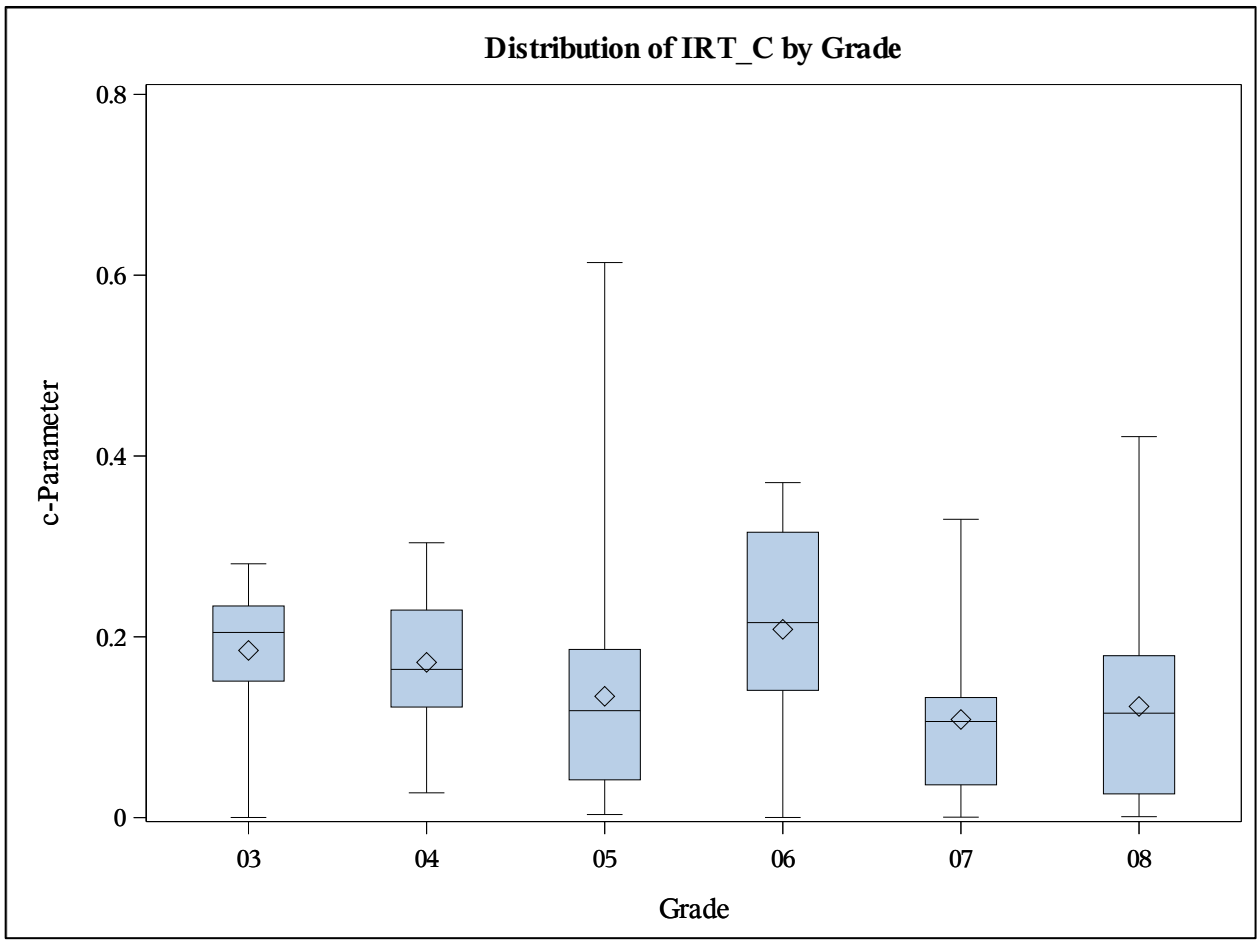


Table C.5.4

IRT Parameter Summary by Item Type: Spring 2023 Operational SC G3–8

Grade	Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	CR	a	3	0.384	0.384	0.604	0.807	0.807
		b	3	1.139	1.139	1.249	3.179	3.179
	MC	a	21	0.387	0.644	0.781	0.946	1.531
		b	21	-0.047	0.592	0.822	1.542	2.076
		c	21	0.029	0.158	0.208	0.234	0.281
	MS*	a	1	0.379	0.379	0.379	0.379	0.379
		b	1	1.328	1.328	1.328	1.328	1.328
		c	1	0	0	0	0	0
	TPD	a	9	0.091	0.343	0.452	0.517	0.63
		b	9	-0.496	0.095	0.404	1.36	6.491
	TPI	a	2	0.449	0.449	0.457	0.465	0.465
		b	2	0.093	0.093	0.572	1.05	1.05
4	CR	a	3	0.479	0.479	0.493	0.724	0.724
		b	3	1.421	1.421	1.762	2.289	2.289
	MC	a	18	0.409	0.662	0.763	1.027	1.448
		b	18	-0.953	-0.130	0.328	1.939	2.940
		c	18	0.067	0.135	0.179	0.241	0.304
	MS	a	2	0.411	0.411	0.616	0.822	0.822
		b	2	1.568	1.568	1.626	1.685	1.685
		c	2	0.027	0.027	0.035	0.043	0.043
	TPD	a	10	0.243	0.42	0.44	0.527	0.623
		b	10	-0.382	0.139	0.383	0.875	1.507
	TPI	a	3	0.423	0.423	0.536	0.613	0.613
		b	3	-0.411	-0.411	0.661	0.690	0.690
5	CR	a	3	0.427	0.427	0.539	0.549	0.549
		b	3	0.781	0.781	1.627	1.788	1.788
	ER	a	1	0.264	0.264	0.264	0.264	0.264
		b	1	1.642	1.642	1.642	1.642	1.642
	MC	a	8	0.368	0.502	0.563	0.764	1.281
		b	8	-1.067	-0.547	-0.265	0.36	1.701
		c	8	0.026	0.073	0.163	0.197	0.206

* The value of c parameter is 0.00046.

Table C.5.4

IRT Parameter Summary by Item Type: Spring 2023 Operational SC G3–8 (continued)

Grade	Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
5	MS	a	2	0.406	0.406	0.611	0.816	0.816
		b	2	1.559	1.559	1.702	1.845	1.845
		c	2	0.003	0.003	0.02	0.036	0.036
	TEI	a	15	0.383	0.47	0.724	0.876	1.094
		b	15	-1.429	-0.357	0.421	1.481	1.903
		c	10	0.014	0.048	0.118	0.184	0.614
	TPD	a	3	0.345	0.345	0.428	0.496	0.496
		b	3	-0.277	-0.277	0.508	0.567	0.567
	TPI	a	5	0.284	0.312	0.32	0.329	0.436
		b	5	-0.114	0.174	0.322	0.382	1.856
6	CR	a	3	0.359	0.359	0.508	0.559	0.559
		b	3	1.353	1.353	1.494	2.487	2.487
	ER	a	3	0.320	0.320	0.427	0.629	0.629
		b	3	1.535	1.535	1.931	2.002	2.002
	MC	a	15	0.374	0.534	0.784	1.072	1.313
		b	15	-0.550	0.069	0.925	1.683	2.852
		c	15	0.084	0.181	0.242	0.340	0.371
	MS	a	2	0.720	0.720	0.773	0.825	0.825
		b	2	0.049	0.049	0.055	0.060	0.060
		c	2	0	0	0	0.001	0.001
	TEI	a	10	0.191	0.422	0.434	0.549	0.639
		b	10	-1.124	-0.54	-0.313	1.004	2.219
		c	3	0.061	0.061	0.149	0.224	0.224
	TPD	a	3	0.222	0.222	0.282	0.391	0.391
		b	3	0.094	0.094	0.686	0.854	0.854
	TPI	a	3	0.317	0.317	0.425	0.444	0.444
		b	3	-0.070	-0.070	0.177	1.228	1.228
7	CR	a	3	0.453	0.453	0.465	0.493	0.493
		b	3	0.736	0.736	1.245	1.666	1.666
	ER	a	2	0.281	0.281	0.309	0.337	0.337
		b	2	1.017	1.017	1.248	1.479	1.479

Table C.5.4

IRT Parameter Summary by Item Type: Spring 2023 Operational SC G3–8 (continued)

Grade	Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
7	MC	a	10	0.396	0.499	0.741	1.07	1.346
		b	10	-0.469	0.114	1.126	1.549	1.984
		c	10	0.001	0.105	0.113	0.240	0.330
	MS	a	3	0.524	0.524	0.813	0.889	0.889
		b	3	-0.440	-0.440	0.427	0.956	0.956
		c	3	0.019	0.019	0.054	0.054	0.054
	TEI	a	14	0.373	0.429	0.633	0.805	1.065
		b	14	-1.922	-0.621	0.422	1.598	2.643
		c	7	0.008	0.008	0.074	0.122	0.144
	TPD	a	4	0.304	0.304	0.359	0.457	0.501
		b	4	-0.194	-0.043	0.532	1.534	2.113
	TPI	a	2	0.314	0.314	0.44	0.566	0.566
		b	2	0.252	0.252	0.438	0.623	0.623
8	CR	a	3	0.503	0.503	0.639	0.760	0.760
		b	3	1.651	1.651	2.226	2.754	2.754
	ER	a	3	0.600	0.600	0.812	0.956	0.956
		b	3	1.697	1.697	1.812	2.206	2.206
	MC	a	10	0.367	0.505	0.68	0.755	1.218
		b	10	-0.998	0.001	0.715	1.016	1.644
		c	10	0.039	0.128	0.146	0.182	0.421
	MS	a	6	0.759	0.764	0.874	0.921	0.989
		b	6	-0.507	-0.196	0.319	1.044	1.085
		c	6	0.001	0.004	0.026	0.083	0.103
	TEI	a	13	0.292	0.4	0.522	0.651	1.187
		b	13	-1.792	-0.385	0.031	0.904	1.801
		c	4	0.009	0.011	0.115	0.217	0.219
	TPD	a	3	0.227	0.227	0.402	0.628	0.628
		b	3	-0.181	-0.181	0.35	0.544	0.544
	TPI	a	1	0.367	0.367	0.367	0.367	0.367
		b	1	1.390	1.390	1.390	1.390	1.390

Table C.6

Statistically Flagged Operational Items: Spring 2023 Operational SC G3–8

Grade	Type	No. of 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
3	CR	3	2	0	0	1
	MC	21	0	0	0	0
	MS	1	0	0	0	0
	TEI	9	2	1	0	0
	TPD	2	0	0	0	0
	TPI	3	2	0	0	1
4	CR	3	2	0	0	0
	MC	18	3	2	0	0
	MS	2	1	0	0	0
	TEI	10	0	0	0	0
	TPD	3	0	0	0	0
	TPI	3	2	0	0	0
5	CR	3	2	0	0	0
	ER	1	1	0	0	0
	MC	8	0	0	0	0
	MS	2	2	0	0	0
	TEI	15	2	0	1	0
	TPD	3	0	0	0	0
	TPI	5	1	0	0	0
6	CR	3	3	0	0	0
	ER**	1	1	0	0	0
	MC	15	0	2	0	0
	MS	2	0	0	0	0
	TEI	10	2	0	1	0
	TPD	3	0	0	0	0
	TPI	3	0	0	0	0

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

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

Table C.6

Statistically Flagged Operational Items: Spring 2023 Operational SC G3–8 (continued)

Grade	Type	No. of 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
7	CR	3	2	0	1	0
	ER**	1	1	0	0	0
	MC	10	1	0	1	0
	MS	3	0	0	0	0
	TEI	14	3	0	0	0
	TPD	4	2	0	0	0
	TPI	2	0	0	0	0
8	CR	3	3	0	0	0
	ER**	1	1	0	0	0
	MC	10	0	0	0	0
	MS	6	1	0	0	0
	TEI	13	1	0	0	0
	TPD	3	0	0	0	0
	TPI	1	0	0	0	0

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

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

Appendix D: Dimensionality

Dimensionality Reports: Science

Contents
Table D.1 Zq1 Statistics and Summary Data: Spring 2023 Operational SC G3–8
Table D.2 Q3 Statistics and Summary Data: Spring 2023 Operational SC G3–8
Table D.3 Reporting Category Intercorrelation Coefficients: Spring 2023 Operational SC G3–8
Table D.4 First and Second Eigenvalues: Spring 2023 Operational SC G3–8 Plot D.1 Principal Component Analysis: Spring 2023 Operational SC G3–8

Table D.1

Zq1 Statistics and Summary Data: Spring 2023 Operational SC G3–8

Grade	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	No. of Items with Poor Fit
3	CR	41.97	41.97	118.29	188.77	188.77	1
	MC	9.63	21.00	25.04	30.93	83.68	0
	MS	175.79	175.79	175.79	175.79	175.79	1
	TPD	100.37	130.61	238.66	351.32	475.06	6
	TPI	100.55	100.55	200.91	301.27	301.27	1
4	CR	33.23	33.23	70.33	85.84	85.84	0
	MC	3.03	16.14	26.93	45.06	97.71	0
	MS	10.42	10.42	16.86	23.30	23.30	0
	TPD	61.64	84.79	148.01	222.50	284.13	5
	TPI	43.08	43.08	60.86	243.13	243.13	1
5	CR	41.78	41.78	62.64	69.65	69.65	0
	ER	153.47	153.47	153.47	153.47	153.47	1
	MC	7.50	18.50	28.51	41.09	86.50	0
	MS	17.93	17.93	31.87	45.82	45.82	0
	TEI	10.15	19.22	38.99	55.80	131.11	1
	TPD	20.10	20.10	78.61	89.32	89.32	0
	TPI	25.72	35.35	112.49	130.71	135.03	2
6	CR	48.67	48.67	56.49	88.04	88.04	0
	ER	69.30	69.30	128.97	279.79	279.79	2
	MC	5.50	11.60	25.85	32.59	91.17	0
	MS	464.48	464.48	481.35	498.23	498.23	2
	TEI	9.79	32.41	82.65	133.28	298.90	3
	TPD	45.26	45.26	73.46	269.25	269.25	1
	TPI	58.05	58.05	115.52	229.03	229.03	1
7	CR	20.55	20.55	63.91	183.74	183.74	1
	ER	56.33	56.33	132.32	208.30	208.30	1
	MC	6.96	16.42	27.25	30.59	292.81	1
	MS	12.14	12.14	33.03	50.77	50.77	0
	TEI	11.97	26.10	35.27	129.01	262.77	3
	TPD	43.37	63.85	158.71	303.22	373.34	2
	TPI	54.34	54.34	64.45	74.56	74.56	0

Table D.1

Zq1 Statistics and Summary Data: Spring 2023 Operational SC G3–8 (continued)

Grade	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	No. of Items with Poor Fit
8	CR	30.76	30.76	38.26	75.28	75.28	0
	ER	30.41	30.41	31.55	61.87	61.87	0
	MC	1.73	11.39	27.83	51.75	130.46	0
	MS	16.36	21.98	49.55	114.51	1065.51	1
	TEI	7.25	67.25	111.88	159.16	563.35	4
	TPD	92.66	92.66	258.12	350.85	350.85	2
	TPI	110.24	110.24	110.24	110.24	110.24	0

Table D.2

Q3 Statistics and Summary Data: Spring 2023 Operational SC G3–8

Grade	Average Zero Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.142	-0.158	-0.086	-0.017	0.083	0.238
4	0.172	-0.213	-0.109	-0.015	0.109	0.199
5	0.188	-0.368	-0.121	-0.003	0.121	0.296
6	0.153	-0.200	-0.098	-0.008	0.102	0.303
7	0.184	-0.275	-0.130	-0.002	0.137	0.290
8	0.184	-0.208	-0.086	-0.013	0.111	0.212

Table D.3

Reporting Category Intercorrelation Coefficients: Spring 2023 Operational SC G3–8

Grade	Reporting Category	1 Investigate	2 Evaluate	3 Reason Scientifically
3	1 Investigate	1.00	–	–
	2 Evaluate	0.69	1.00	–
	3 Reason Scientifically	0.60	0.64	1.00
4	1 Investigate	1.00	–	–
	2 Evaluate	0.61	1.00	–
	3 Reason Scientifically	0.71	0.67	1.00
5	1 Investigate	1.00	–	–
	2 Evaluate	0.68	1.00	–
	3 Reason Scientifically	0.68	0.79	1.00
6	1 Investigate	1.00	–	–
	2 Evaluate	0.67	1.00	–
	3 Reason Scientifically	0.70	0.71	1.00
7	1 Investigate	1.00	–	–
	2 Evaluate	0.59	1.00	–
	3 Reason Scientifically	0.64	0.78	1.00
8	1 Investigate	1.00	–	–
	2 Evaluate	0.75	1.00	–
	3 Reason Scientifically	0.72	0.66	1.00

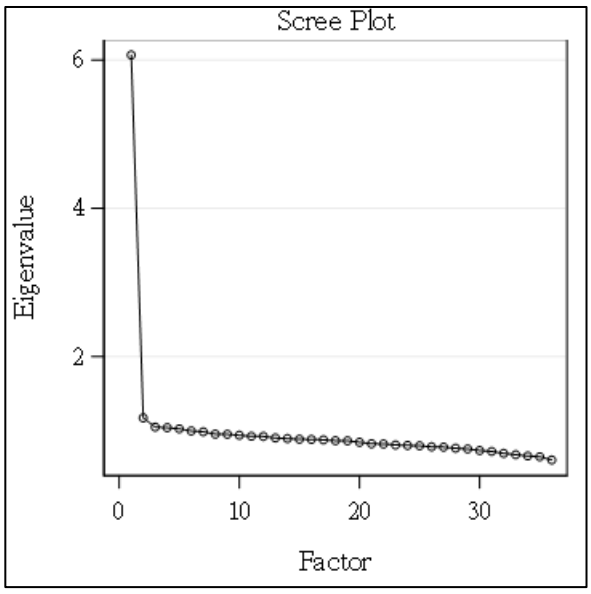
Table D.4

First and Second Eigenvalue: Spring 2023 Operational SC G3–8

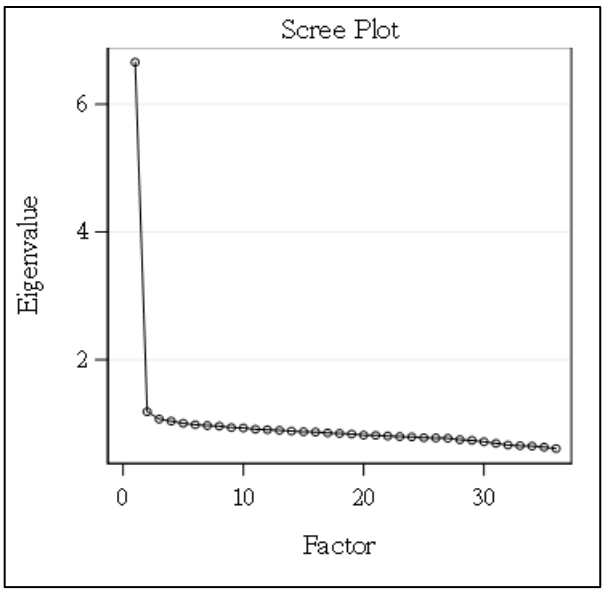
Grade	Mode	First Eigenvalue	Second Eigenvalue	Ratio
3	Online	6.066	1.175	5.163
	Paper	6.661	1.184	5.624
4	Online	7.554	1.239	6.097
5	Online	8.152	1.253	6.507
6	Online	7.405	1.134	6.531
7	Online	8.317	1.354	6.141
8	Online	8.413	1.238	6.795

Plot D.1

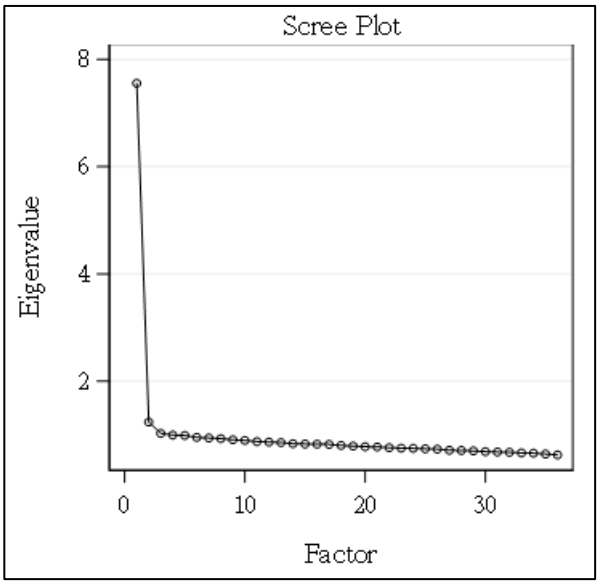
Principal Component Analysis Plot: Spring 2023 Operational SC G3-8



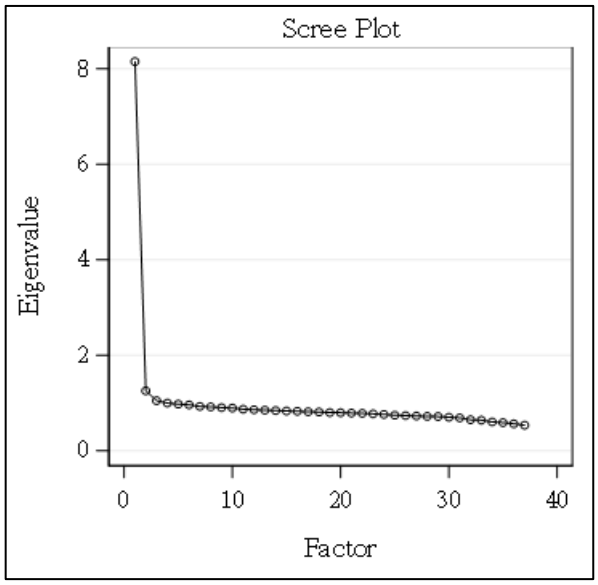
Grade 3: Online



Grade 3: Paper



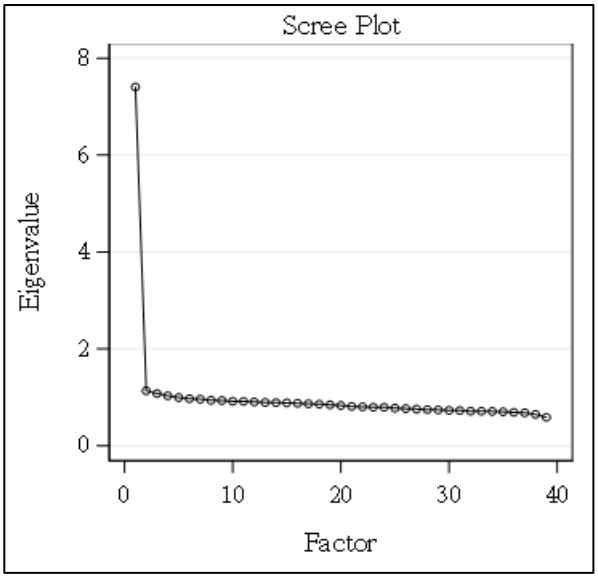
Grade 4



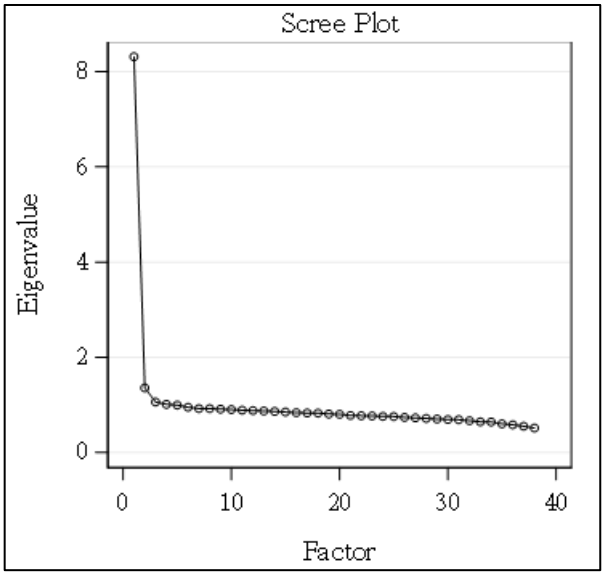
Grade 5

Plot D.1

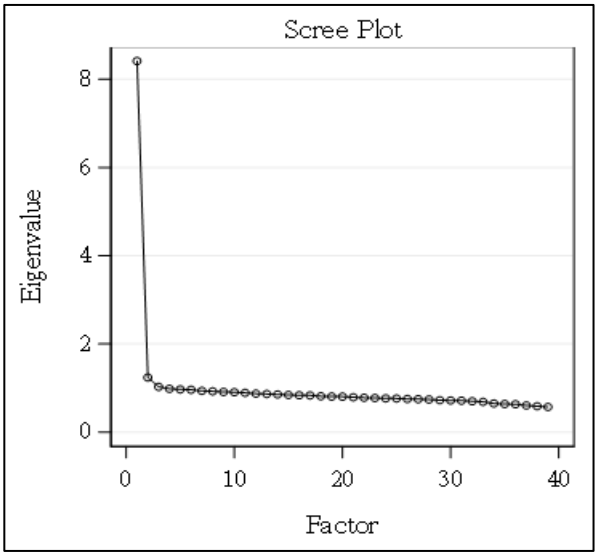
Principal Component Analysis Plot: Spring 2023 Operational SC G3-8 (continued)



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 2023 Operational Science Grade 3 Table E.1.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 3
Table E.2.1 Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science Grade 4 Table E.2.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 4
Table E.3.1 Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science Grade 5 Table E.3.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 5
Table E.4.1 Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science Grade 6 Table E.4.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 6
Table E.5.1 Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science Grade 7 Table E.5.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 7
Table E.6.1 Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science Grade 8 Table E.6.2 Frequency Distribution of Scale Scores: Spring 2023 Operational Science Grade 8

Table E.1.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 3

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

N	≥49310		
Mean	725.29	Median	727.00
Std deviation	30.80	Variance	948.77
Skewness	-0.1837	Kurtosis	-0.0378
Mode	710.00	Std Error Mean	0.1387
Range	200.00	Interquartile Range	43.00

Quantile	Estimate
100% Max	850
99%	790
95%	773
90%	765
75% Q3	748
50% Median	727
25% Q1	705
10%	687
5%	663
1%	650
0% Min	650

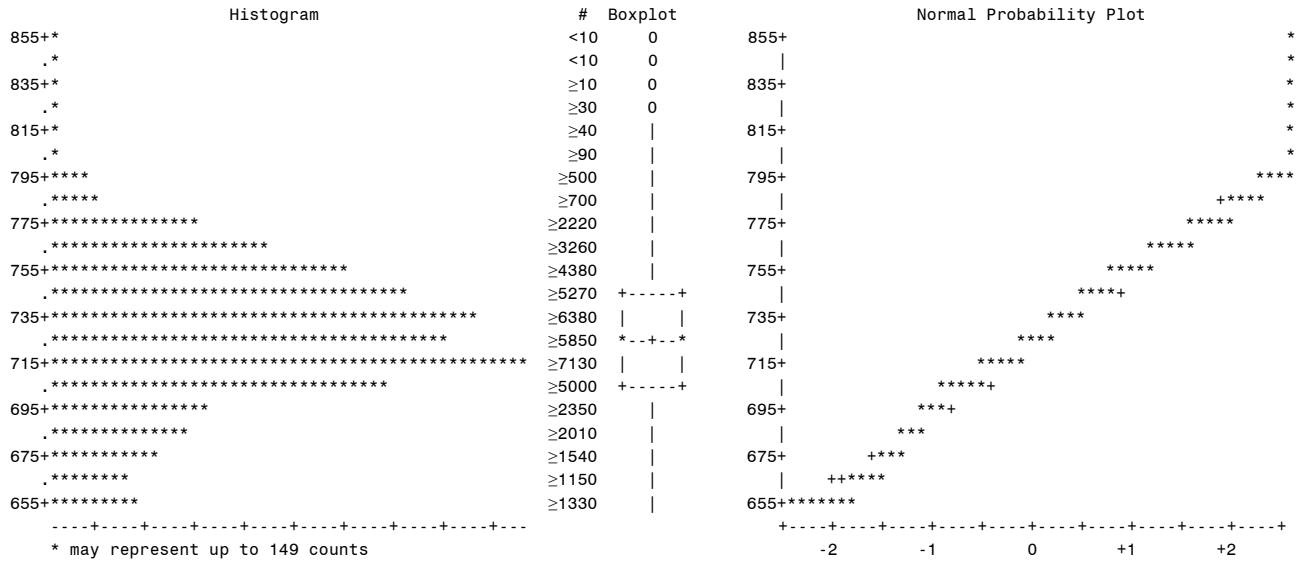


Table E.1.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 3

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	*************************	≥1330	≥1330	2.70	2.70
663	**********************	≥1150	≥2480	2.34	5.05
677	******************************	≥1540	≥4030	3.13	8.18
687	************************************	≥2010	≥6040	4.08	12.25
694	**	≥2350	≥8400	4.78	17.03
700	**	≥2470	≥10870	5.03	22.06
705	**	≥2520	≥13400	5.12	27.18
710	**	≥2560	≥15960	5.20	32.38
714	**	≥2380	≥18350	4.84	37.21
718	**	≥2180	≥20530	4.43	41.64
721	**************************************	≥2060	≥22600	4.19	45.83
724	************************************	≥1970	≥24580	4.01	49.84
727	**********************************	≥1800	≥26380	3.66	53.50
730	**********************************	≥1730	≥28110	3.51	57.02
733	******************************	≥1610	≥29730	3.27	60.28
735	******************************	≥1580	≥31310	3.21	63.50
738	****************************	≥1450	≥32760	2.95	66.44
741	*****************************	≥1430	≥34200	2.92	69.36
743	**************************	≥1360	≥35570	2.78	72.14
745	************************	≥1260	≥36840	2.57	74.70
748	**********************	≥1200	≥38040	2.44	77.14
750	**********************	≥1200	≥39250	2.45	79.59
753	*********************	≥1100	≥40350	2.23	81.82
755	*********************	≥1050	≥41400	2.14	83.96
757	********************	≥1020	≥42430	2.08	86.04
760	******************	≥890	≥43320	1.82	87.86
762	*****************	≥850	≥44180	1.74	89.60
765	***************	≥770	≥44960	1.57	91.17
767	***************	≥730	≥45690	1.49	92.66
770	**************	≥670	≥46370	1.37	94.03
773	**********	≥550	≥46930	1.13	95.17
776	**********	≥530	≥47470	1.09	96.26
779	*********	≥440	≥47920	0.91	97.17
782	********	≥380	≥48300	0.78	97.95
786	******	≥310	≥48620	0.64	98.59
790	*****	≥230	≥48850	0.48	99.07
794	***	≥140	≥48990	0.29	99.36
799	***	≥120	≥49120	0.26	99.62
805	**	≥90	≥49220	0.20	99.82
812	*	≥40	≥49260	0.09	99.90
821	*	≥30	≥49300	0.06	99.97
832		≥10	≥49310	0.03	99.99
848		<10	≥49310	0.00	100.00
850		<10	≥49310	0.00	100.00

400 800 1200 1600 2000 2400

Frequency

Table E.2.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 4

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

N	≥48870		
Mean	737.56	Median	737.00
Std deviation	30.09	Variance	905.23
Skewness	0.0960	Kurtosis	-0.0544
Mode	717.00	Std Error Mean	0.1361
Range	200.00	Interquartile Range	40.00

Quantile	Estimate
100% Max	850
99%	809
95%	789
90%	776
75% Q3	757
50% Median	737
25% Q1	717
10%	701
5%	690
1%	672
0% Min	650

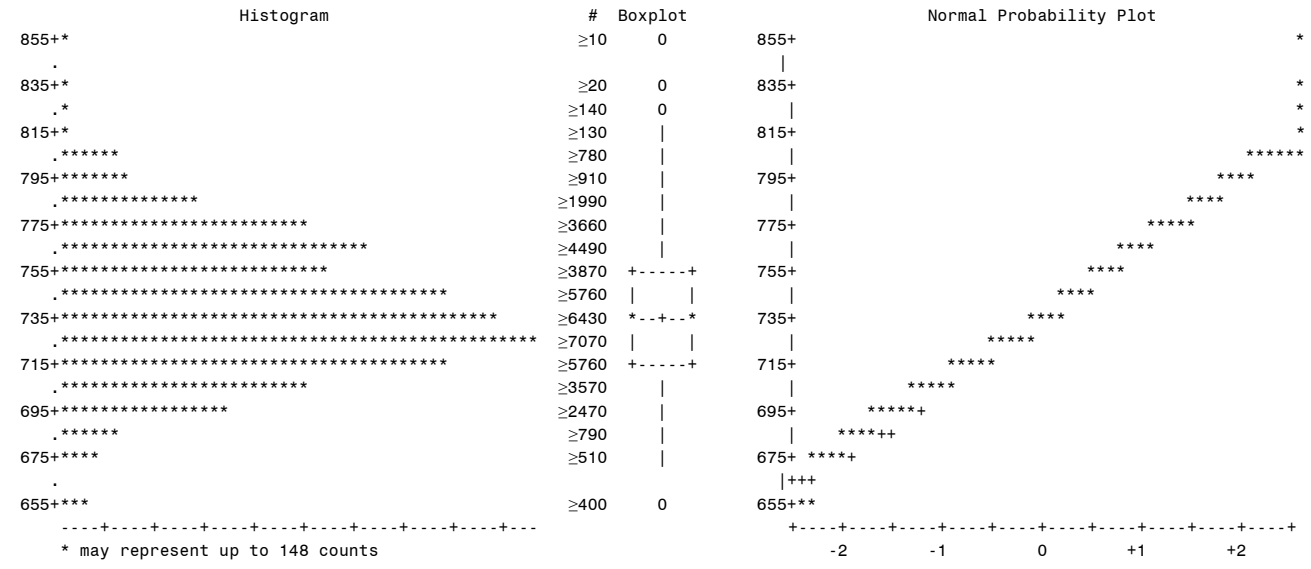


Table E.2.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 4

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

Scale_Score	Freq	Cum. Freq	Percent	Cum. Percent
650	≥170	≥170	0.35	0.35
656	≥230	≥400	0.48	0.84
672	≥510	≥920	1.06	1.89
682	≥790	≥1720	1.63	3.53
690	≥1090	≥2820	2.24	5.77
696	≥1370	≥4190	2.81	8.58
701	≥1720	≥5910	3.52	12.10
706	≥1850	≥7770	3.80	15.90
710	≥1910	≥9690	3.92	19.83
714	≥1900	≥11590	3.90	23.73
717	≥1940	≥13540	3.98	27.70
720	≥1890	≥15430	3.87	31.58
723	≥1800	≥17230	3.68	35.26
726	≥1760	≥18990	3.60	38.86
729	≥1620	≥20620	3.32	42.19
732	≥1670	≥22290	3.42	45.61
734	≥1640	≥23940	3.37	48.98
737	≥1580	≥25520	3.25	52.23
739	≥1530	≥27050	3.13	55.36
742	≥1500	≥28560	3.08	58.44
744	≥1450	≥30020	2.98	61.42
747	≥1430	≥31450	2.94	64.36
749	≥1360	≥32820	2.79	67.15
752	≥1330	≥34150	2.73	69.88
754	≥1280	≥35430	2.62	72.50
757	≥1250	≥36690	2.58	75.07
760	≥1170	≥37860	2.41	77.48
762	≥1170	≥39040	2.41	79.89
765	≥1060	≥40110	2.18	82.07
768	≥1070	≥41180	2.19	84.26
770	≥1030	≥42210	2.12	86.38
773	≥970	≥43190	2.00	88.38
776	≥790	≥43990	1.62	90.00
779	≥850	≥44850	1.76	91.76
782	≥730	≥45580	1.50	93.26
786	≥670	≥46250	1.39	94.64
789	≥590	≥46840	1.21	95.85
793	≥480	≥47330	0.99	96.85
796	≥430	≥47760	0.88	97.73
800	≥330	≥48100	0.68	98.41
805	≥270	≥48370	0.55	98.97
809	≥180	≥48550	0.38	99.34
815	≥130	≥48690	0.27	99.62
821	≥80	≥48770	0.18	99.79
828	≥60	≥48830	0.13	99.92
838	≥20	≥48860	0.05	99.97
850	≥10	≥48870	0.03	100.00

200 400 600 800 1000 1200 1400 1600 1800

Frequency

Table E.3.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 5

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 05

N	≥48320		
Mean	729.44	Median	731.00
Std deviation	37.83	Variance	1431.25
Skewness	0.0060	Kurtosis	-0.4662
Mode	695.00	Std Error Mean	0.1721
Range	200.00	Interquartile Range	58.00

Quantile	Estimate
100% Max	850
99%	813
95%	791
90%	779
75% Q3	758
50% Median	731
25% Q1	700
10%	677
5%	670
1%	650
0% Min	650

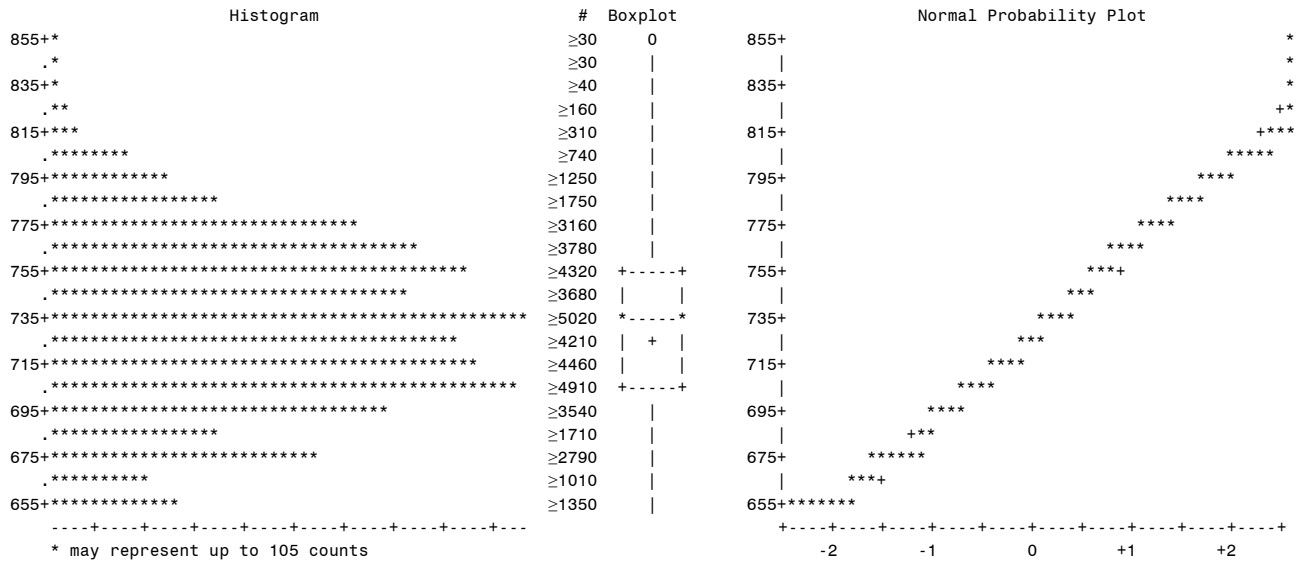


Table E.3.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 5

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 05

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**	≥1350	≥1350	2.80	2.80
660	*************************************	≥1010	≥2360	2.09	4.89
670	**	≥1260	≥3620	2.61	7.49
677	**	≥1530	≥5150	3.17	10.67
684	**	≥1710	≥6860	3.54	14.21
690	**	≥1740	≥8610	3.61	17.82
695	**	≥1790	≥10400	3.72	21.54
700	***	≥1720	≥12120	3.56	25.10
704	***	≥1660	≥13790	3.45	28.54
708	**	≥1520	≥15320	3.16	31.71
712	**	≥1570	≥16890	3.26	34.96
715	***	≥1470	≥18370	3.06	38.03
719	***	≥1400	≥19780	2.92	40.94
722	**	≥1480	≥21260	3.07	44.01
725	***	≥1380	≥22650	2.87	46.88
728	***	≥1340	≥24000	2.79	49.67
731	**	≥1280	≥25280	2.65	52.32
734	***	≥1270	≥26550	2.63	54.96
737	**	≥1220	≥27770	2.52	57.48
739	***	≥1240	≥29020	2.58	60.06
742	**	≥1220	≥30250	2.54	62.60
745	**	≥1220	≥31470	2.54	65.13
747	**	≥1230	≥32710	2.56	67.69
750	**	≥1110	≥33820	2.30	69.99
752	***	≥1110	≥34940	2.31	72.30
755	**	≥1070	≥36010	2.22	74.52
758	**************************************	≥1020	≥37030	2.12	76.64
760	**************************************	≥1000	≥38040	2.08	78.72
763	**********************************	≥920	≥38960	1.92	80.63
765	**********************************	≥960	≥39920	1.99	82.62
768	**********************************	≥880	≥40810	1.84	84.46
771	**********************************	≥890	≥41710	1.85	86.31
773	*********************************	≥820	≥42540	1.72	88.03
776	******************************	≥750	≥43290	1.56	89.58
779	**************************	≥690	≥43980	1.43	91.02
782	************************	≥630	≥44620	1.32	92.34
784	**********************	≥580	≥45200	1.21	93.55
787	*********************	≥530	≥45730	1.10	94.65
791	********************	≥490	≥46230	1.02	95.66
794	**********	≥390	≥46620	0.81	96.47
797	**********	≥360	≥46990	0.76	97.24
801	**********	≥290	≥47280	0.60	97.84
804	**********	≥240	≥47520	0.50	98.34
808	*********	≥210	≥47740	0.45	98.79
813	*******	≥170	≥47910	0.35	99.14
817	******	≥140	≥48050	0.29	99.43
822	****	≥80	≥48130	0.18	99.61
828	***	≥70	≥48210	0.16	99.77
834	**	≥40	≥48260	0.10	99.86
842	*	≥30	≥48290	0.06	99.93
850	*	≥30	≥48320	0.07	100.00

200 400 600 800 1000 1200 1400 1600 1800

Frequency

Table E.4.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 6

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

N	≥48300		
Mean	721.95	Median	722.00
Std deviation	32.01	Variance	1024.49
Skewness	-0.0220	Kurtosis	-0.3488
Mode	698.00	Std Error Mean	0.1456
Range	200.00	Interquartile Range	47.00

Quantile	Estimate
100% Max	850
99%	793
95%	773
90%	763
75% Q3	745
50% Median	722
25% Q1	698
10%	681
5%	664
1%	650
0% Min	650

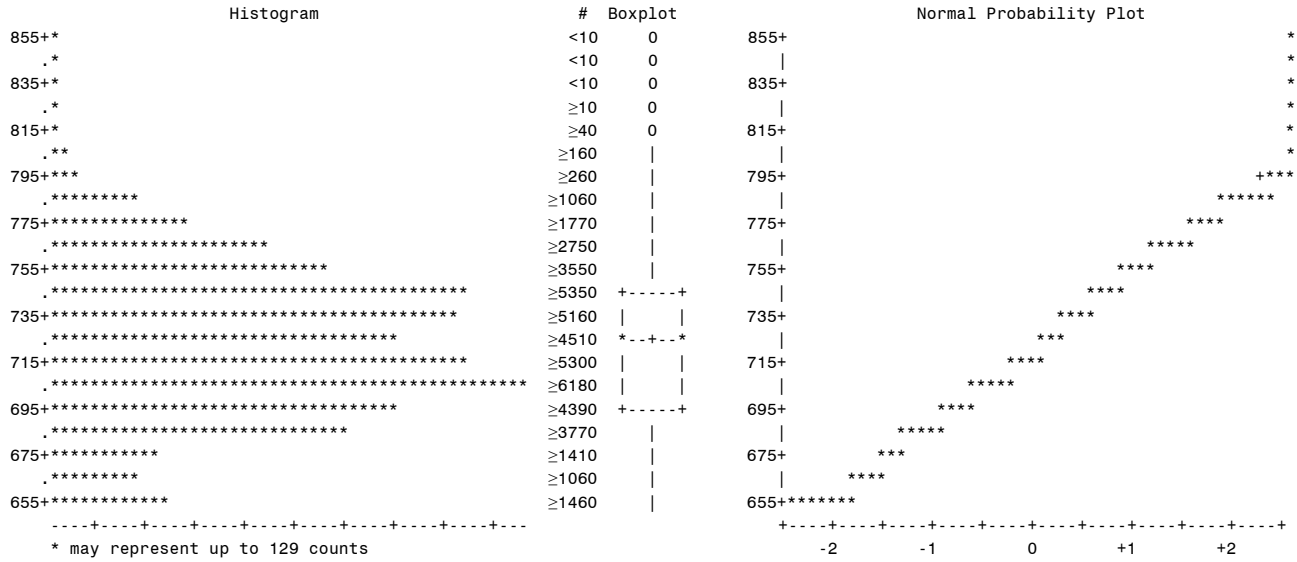


Table E.4.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 6

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥800	≥800	1.67	1.67
652	*********	≥660	≥1460	1.37	3.04
664	**********	≥1060	≥2530	2.21	5.25
673	**********	≥1410	≥3950	2.93	8.18
681	**********	≥1790	≥5740	3.71	11.89
687	**********	≥1980	≥7720	4.10	15.99
693	**********	≥2170	≥9890	4.50	20.49
698	**********	≥2220	≥12120	4.60	25.09
702	**********	≥2120	≥14240	4.40	29.49
706	**********	≥2120	≥16360	4.39	33.89
709	**********	≥1930	≥18300	4.01	37.90
713	**********	≥1850	≥20160	3.84	41.74
716	**********	≥1780	≥21940	3.69	45.43
719	**********	≥1670	≥23610	3.46	48.89
722	**********	≥1650	≥25260	3.42	52.30
725	**********	≥1480	≥26750	3.08	55.38
728	**********	≥1370	≥28130	2.85	58.23
730	**********	≥1360	≥29490	2.83	61.06
733	**********	≥1350	≥30850	2.81	63.86
735	**********	≥1270	≥32120	2.64	66.51
738	**********	≥1170	≥33290	2.43	68.93
740	**********	≥1140	≥34440	2.37	71.30
742	**********	≥1150	≥35590	2.39	73.69
745	**********	≥1060	≥36660	2.21	75.90
747	**********	≥1010	≥37680	2.10	78.00
749	**********	≥970	≥38650	2.02	80.02
752	**********	≥900	≥39550	1.87	81.88
754	**********	≥930	≥40480	1.93	83.81
756	**********	≥860	≥41340	1.78	85.59
759	**********	≥860	≥42210	1.79	87.38
761	**********	≥780	≥42990	1.63	89.01
763	**********	≥710	≥43710	1.49	90.50
766	**********	≥650	≥44370	1.35	91.85
768	**********	≥590	≥44960	1.24	93.09
771	**********	≥520	≥45490	1.09	94.18
773	*********	≥450	≥45940	0.93	95.11
776	********	≥410	≥46360	0.87	95.98
778	********	≥380	≥46740	0.79	96.77
781	*******	≥330	≥47080	0.70	97.47
784	******	≥280	≥47360	0.59	98.06
786	*****	≥230	≥47590	0.48	98.54
789	****	≥200	≥47800	0.43	98.96
793	***	≥150	≥47960	0.32	99.29
796	**	≥100	≥48060	0.22	99.51
800	**	≥80	≥48140	0.17	99.67
803	*	≥50	≥48200	0.11	99.78
808	*	≥30	≥48230	0.08	99.86
812	*	≥20	≥48260	0.06	99.92
818		≥10	≥48280	0.04	99.95
824		≥10	≥48290	0.03	99.98
831		<10	≥48300	0.01	99.99
841		<10	≥48300	0.01	100.00
850		<10	≥48300	0.00	100.00

400 800 1200 1600 2000

Frequency

Table E.5.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 7

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 07

N	≥48900		
Mean	730.60	Median	730.00
Std deviation	33.09	Variance	1094.69
Skewness	-0.0488	Kurtosis	-0.3169
Mode	698.00	Std Error Mean	0.1496
Range	200.00	Interquartile Range	48.00

Quantile	Estimate
100% Max	850
99%	802
95%	783
90%	774
75% Q3	754
50% Median	730
25% Q1	706
10%	689
5%	677
1%	650
0% Min	650

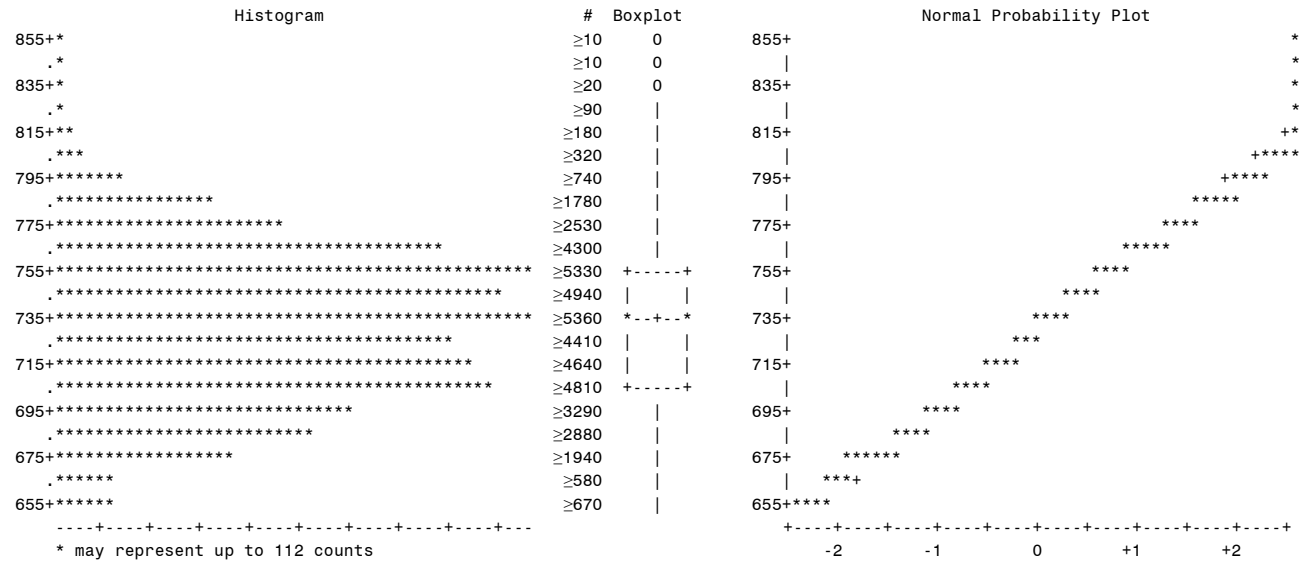


Table E.5.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 7

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 07

Scale_Score		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥670	≥670	1.37	1.37
660	**********	≥580	≥1250	1.19	2.56
670	**********	≥840	≥2090	1.73	4.28
677	**********	≥1100	≥3190	2.25	6.54
684	**********	≥1350	≥4550	2.77	9.31
689	**********	≥1520	≥6070	3.12	12.43
694	**********	≥1610	≥7680	3.29	15.72
698	**********	≥1680	≥9370	3.45	19.17
702	**********	≥1580	≥10950	3.24	22.41
706	**********	≥1600	≥12560	3.29	25.69
709	**********	≥1620	≥14190	3.32	29.02
713	**********	≥1530	≥15720	3.13	32.15
716	**********	≥1570	≥17290	3.23	35.37
719	**********	≥1530	≥18830	3.14	38.51
722	**********	≥1490	≥20330	3.06	41.57
725	**********	≥1460	≥21790	3.00	44.57
728	**********	≥1450	≥23240	2.97	47.54
730	**********	≥1450	≥24700	2.98	50.51
733	**********	≥1320	≥26030	2.72	53.23
735	**********	≥1320	≥27360	2.72	55.95
738	**********	≥1240	≥28600	2.55	58.50
740	**********	≥1290	≥29900	2.64	61.14
743	**********	≥1290	≥31190	2.64	63.78
745	**********	≥1160	≥32350	2.38	66.16
747	**********	≥1200	≥33550	2.45	68.61
750	**********	≥1190	≥34750	2.45	71.06
752	**********	≥1080	≥35830	2.22	73.28
754	**********	≥1060	≥36900	2.19	75.47
756	**********	≥1000	≥37910	2.05	77.52
758	**********	≥980	≥38890	2.01	79.53
760	**********	≥960	≥39850	1.96	81.49
763	**********	≥960	≥40810	1.96	83.45
765	**********	≥840	≥41650	1.72	85.18
767	**********	≥780	≥42440	1.61	86.79
769	**********	≥750	≥43190	1.54	88.33
771	**********	≥740	≥43930	1.52	89.84
774	**********	≥640	≥44580	1.31	91.16
776	**********	≥600	≥45180	1.25	92.40
778	**********	≥540	≥45720	1.10	93.51
781	**********	≥550	≥46280	1.13	94.64
783	**********	≥440	≥46730	0.91	95.55
786	**********	≥420	≥47150	0.87	96.43
789	**********	≥350	≥47510	0.73	97.16
792	**********	≥300	≥47810	0.61	97.77
795	**********	≥220	≥48040	0.46	98.23
798	**********	≥210	≥48250	0.44	98.67
802	**********	≥160	≥48420	0.34	99.01
805	**********	≥150	≥48570	0.32	99.33
810	**********	≥100	≥48680	0.22	99.55
814	**********	≥70	≥48760	0.16	99.71
820	**********	≥60	≥48820	0.13	99.84
826	**********	≥20	≥48850	0.06	99.90
834	**********	≥20	≥48880	0.05	99.95
844	**********	≥10	≥48890	0.02	99.98
850	**********	≥10	≥48900	0.02	100.00

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

Frequency

Table E.6.1

Scale Score Descriptive Statistics and Plots: Spring 2023 Operational Science: Grade 8

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 08

N	≥50160		
Mean	732.49	Median	732.00
Std deviation	31.12	Variance	968.33
Skewness	0.0408	Kurtosis	-0.3657
Mode	706.00	Std Error Mean	0.1389
Range	195.00	Interquartile Range	45.00

Quantile	Estimate
100% Max	845
99%	802
95%	785
90%	774
75% Q3	755
50% Median	732
25% Q1	710
10%	693
5%	680
1%	662
0% Min	650

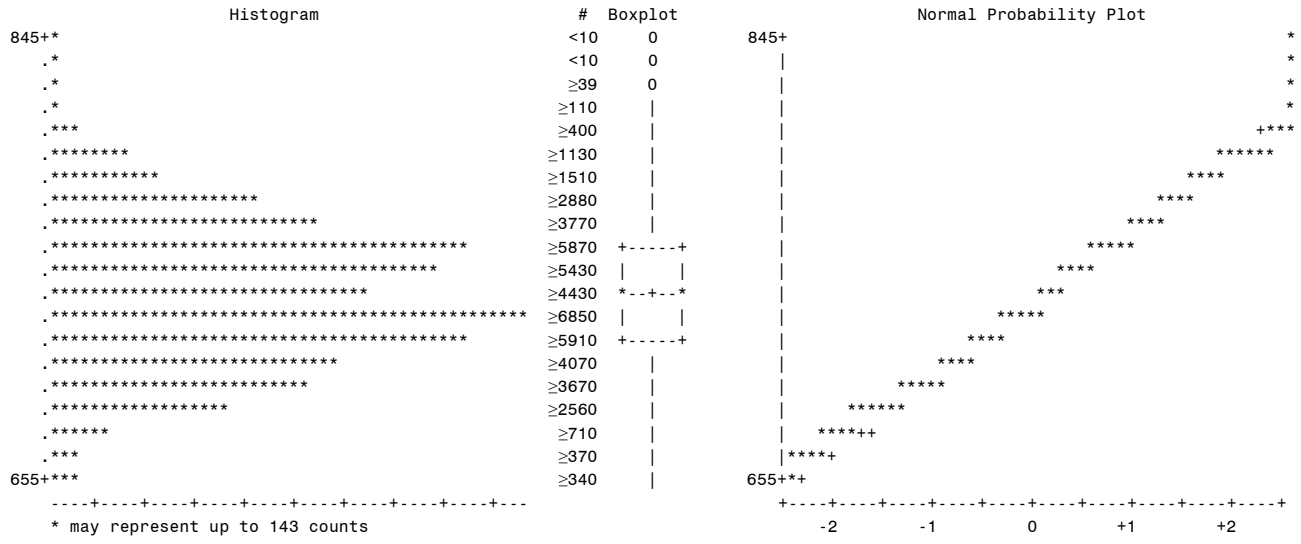


Table E.6.2

Frequency Distribution of Scale Scores: Spring 2023 Operational Science: Grade 8

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 08

Scale_Score	Freq	Cum. Freq	Percent	Cum. Percent
650	≥340	≥340	0.69	0.69
662	≥370	≥720	0.75	1.44
672	≥710	≥1440	1.43	2.88
680	≥1110	≥2550	2.21	5.09
687	≥1450	≥4000	2.90	7.99
693	≥1720	≥5730	3.44	11.43
698	≥1950	≥7680	3.89	15.32
702	≥2010	≥9690	4.01	19.33
706	≥2060	≥11760	4.12	23.45
710	≥2060	≥13820	4.11	27.56
714	≥2000	≥15830	4.01	31.57
717	≥1840	≥17680	3.68	35.25
721	≥1740	≥19430	3.49	38.73
724	≥1730	≥21160	3.47	42.20
727	≥1670	≥22840	3.35	45.55
729	≥1680	≥24530	3.37	48.91
732	≥1540	≥26080	3.08	52.00
735	≥1410	≥27500	2.83	54.82
737	≥1470	≥28970	2.93	57.75
740	≥1440	≥30410	2.87	60.63
742	≥1370	≥31780	2.74	63.36
745	≥1350	≥33130	2.69	66.06
747	≥1270	≥34410	2.54	68.60
750	≥1240	≥35650	2.48	71.08
752	≥1230	≥36890	2.47	73.54
755	≥1140	≥38030	2.28	75.82
757	≥1140	≥39180	2.29	78.11
759	≥1100	≥40280	2.19	80.31
762	≥1010	≥41290	2.02	82.32
764	≥930	≥42220	1.86	84.18
767	≥910	≥43140	1.82	86.00
769	≥910	≥44050	1.83	87.83
772	≥830	≥44890	1.66	89.49
774	≥770	≥45660	1.54	91.03
777	≥690	≥46350	1.39	92.42
779	≥580	≥46930	1.16	93.57
782	≥570	≥47500	1.14	94.71
785	≥500	≥48010	1.00	95.72
787	≥440	≥48450	0.89	96.60
790	≥370	≥48830	0.75	97.35
793	≥300	≥49130	0.61	97.96
796	≥260	≥49400	0.52	98.48
799	≥190	≥49590	0.38	98.86
802	≥170	≥49760	0.34	99.20
806	≥130	≥49890	0.27	99.48
809	≥100	≥49990	0.20	99.68
813	≥80	≥50080	0.16	99.84
817	≥30	≥50110	0.07	99.90
821	≥10	≥50130	0.04	99.94
826	≥20	≥50150	0.04	99.98
831	<10	≥50150	0.01	99.99
838	<10	≥50160	0.00	100.00
845	<10	≥50160	0.00	100.00

200 400 600 800 1000 1200 1400 1600 1800 2000

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 2023 Operational SC G3-8
Table F.2 Cronbach's Alpha and Marginal Reliability: Spring 2023 Operational SC G3-8
Table F.3.1–F.3.9 Classification Accuracy and Decision Consistency Matrices: Spring 2023 Operational SC G3-8

Table F.1.1

Reliability for Overall and Subgroups: Spring 2023 Operational Science

Category	Subgroup*	Grade					
		3	4	5	6	7	8
All Students		0.861	0.884	0.882	0.877	0.892	0.897
Gender	Female	0.852	0.876	0.869	0.864	0.885	0.887
	Male	0.869	0.891	0.892	0.888	0.899	0.905
Race	African American	0.802	0.837	0.849	0.824	0.862	0.852
	AI/AN	0.844	0.877	0.867	0.839	0.867	0.879
	Asian	0.882	0.896	0.891	0.891	0.903	0.904
	Hispanic/Latino	0.834	0.875	0.878	0.866	0.891	0.891
	NHPI	0.831	0.906	0.859	0.900	0.905	0.864
	Two or More	0.850	0.880	0.872	0.875	0.890	0.887
	White	0.865	0.881	0.870	0.873	0.884	0.889
Economically Disadvantaged	No	0.869	0.884	0.872	0.878	0.884	0.891
	Yes	0.830	0.861	0.867	0.851	0.876	0.878
English Learner	No	0.862	0.884	0.880	0.876	0.890	0.895
	Yes	0.732	0.784	0.801	0.710	0.795	0.762
Education Classification	Regular	0.860	0.882	0.877	0.874	0.888	0.894
	Special	0.835	0.863	0.864	0.833	0.866	0.844
Section 504	No	0.862	0.885	0.882	0.878	0.892	0.897
	Yes	0.825	0.863	0.874	0.858	0.875	0.881
Migrant	No	0.861	0.884	0.882	0.877	0.892	0.897
	Yes	0.827	0.868	0.897	0.848	0.905	0.890
Homeless Status	No	0.861	0.885	0.882	0.877	0.892	0.897
	Yes	0.792	0.834	0.852	0.831	0.868	0.869
Military Affiliation	No	0.860	0.884	0.881	0.877	0.892	0.896
	Yes	0.857	0.874	0.867	0.867	0.884	0.879
Foster Care Status	No	0.861	0.884	0.882	0.877	0.892	0.897
	Yes	0.818	0.875	0.867	0.832	0.873	0.845

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table F.1.2

SEM for Overall and Subgroups: Spring 2023 Operational Science

Category	Subgroup*	Grade					
		3	4	5	6	7	8
All Students		3.37	3.39	3.95	3.64	3.82	3.37
Gender	Female	3.37	3.39	3.98	3.65	3.83	3.37
	Male	3.35	3.37	3.93	3.61	3.79	3.37
Race	African American	3.27	3.33	3.68	3.47	3.63	3.26
	AI/AN	3.40	3.39	4.03	3.68	3.85	3.41
	Asian	3.40	3.37	4.11	3.85	4.01	3.50
	Hispanic/Latino	3.33	3.34	3.86	3.57	3.73	3.31
	NHPI	3.45	3.29	4.08	3.66	3.88	3.54
	Two or More	3.42	3.39	4.03	3.68	3.88	3.45
	White	3.43	3.40	4.11	3.76	3.95	3.47
Economically Disadvantaged	No	3.43	3.39	4.13	3.78	4.00	3.49
	Yes	3.33	3.37	3.81	3.56	3.70	3.32
English Learner	No	3.37	3.39	3.98	3.66	3.84	3.40
	Yes	3.17	3.21	3.24	3.22	3.21	2.99
Education Classification	Regular	3.39	3.39	4.00	3.67	3.86	3.40
	Special	3.22	3.26	3.41	3.30	3.35	3.08
Section 504	No	3.38	3.39	3.97	3.65	3.85	3.39
	Yes	3.32	3.34	3.77	3.52	3.63	3.28
Migrant	No	3.37	3.39	3.95	3.64	3.82	3.37
	Yes	3.33	3.39	3.93	3.57	3.59	3.30
Homeless Status	No	3.37	3.37	3.96	3.64	3.83	3.38
	Yes	3.22	3.30	3.65	3.45	3.53	3.22
Military Affiliation	No	3.37	3.38	3.96	3.63	3.81	3.38
	Yes	3.44	3.40	4.09	3.79	3.97	3.52
Foster Care Status	No	3.37	3.39	3.95	3.64	3.82	3.37
	Yes	3.29	3.30	3.58	3.40	3.55	3.30

* AI/AN = American Indian or Alaska Native. NHPI = Native Hawaiian or Other Pacific Islander.

Table F.2

Cronbach's Alpha and Marginal Reliability: Spring 2023 Operational SC G3–8

Grade	Cronbach's Alpha	Marginal Reliability
3	0.861	0.85
4	0.884	0.89
5	0.882	0.89
6	0.877	0.89
7	0.892	0.91
8	0.897	0.91

Table F.3.1

Classification Accuracy Matrices: Spring 2023 Operational SC G3–8

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
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
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
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
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
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

Table F.3.2

Decision Consistency Matrices: Spring 2023 Operational SC G3–8

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
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
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
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
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
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.3.3

Estimates of Accuracy and Consistency of Achievement Level Classification

Grade	Accuracy	Consistency	PChance	Kappa
3	0.648	0.542	0.237	0.400
4	0.689	0.579	0.231	0.453
5	0.664	0.557	0.215	0.436
6	0.695	0.593	0.247	0.460
7	0.717	0.618	0.244	0.495
8	0.721	0.616	0.240	0.494

Table F.3.4

Accuracy of Classification at Each Achievement Level

Grade	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
3	0.794	0.723	0.583	0.519	0.615
4	0.765	0.652	0.666	0.676	0.812
5	0.836	0.611	0.574	0.644	0.706
6	0.838	0.656	0.632	0.66	N/A
7	0.843	0.64	0.64	0.771	0.655
8	0.792	0.756	0.63	0.737	0.773

Table F.3.5

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	0.931	0.876	0.893	0.942
4	0.941	0.892	0.902	0.952
5	0.926	0.894	0.898	0.939
6	0.918	0.888	0.911	0.975
7	0.935	0.902	0.904	0.974
8	0.95	0.899	0.908	0.963

Table F.3.6

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	0.899	0.828	0.847	0.926
4	0.913	0.850	0.862	0.933
5	0.894	0.852	0.856	0.915
6	0.882	0.844	0.874	0.967
7	0.907	0.863	0.865	0.965
8	0.926	0.858	0.870	0.948

Table F.3.7

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	0.676	0.657	0.581	0.186
4	0.631	0.669	0.688	0.616
5	0.697	0.701	0.674	0.484
6	0.699	0.688	0.612	0.031
7	0.707	0.717	0.688	0.252
8	0.664	0.709	0.699	0.556

Table F.3.8

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	0.036	0.052	0.052	0.056
4	0.027	0.047	0.053	0.032
5	0.034	0.050	0.052	0.039
6	0.040	0.050	0.049	0.025
7	0.029	0.046	0.050	0.023
8	0.023	0.047	0.051	0.026

Table F.3.9

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	0.033	0.072	0.055	0.002
4	0.032	0.061	0.045	0.016
5	0.039	0.056	0.051	0.021
6	0.042	0.062	0.040	0.000
7	0.036	0.052	0.046	0.003
8	0.027	0.054	0.042	0.011

Appendix G: Accommodated Print and Braille Creation

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 creates accommodated versions of 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 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 letter or number of 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.
- Match interaction 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 or select the correct answer in each row.
- 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 and lettered or numbered. The directions are then modified to ask the student to select the letter/number of 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 forms, a box is provided for the student to write the response. In the braille forms, students are asked to answer on the paper provided.
- 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 how 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 alongside the online form, and braille forms are reviewed by an outside third-party braille expert against the AP form. Throughout the braille creation process, the braille vendor relies on the AP form and consults with the content experts at LDOE for additional clarification or modifications

for specific items as needed. Students' responses to the accommodated print or braille test are captured in 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. Louisiana's sample sizes are too small for traditional studies of comparability for both AP and braille forms.

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. Table H.1. outlines where information about monitoring, maintaining, and improving quality can be found within this report.

Table H.1

On-Going Quality Control

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

Table H.1

On-Going Quality Control (continued)

Related Information			Related Chapter/Source
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	Pearson and the LDOE Internal documentation
	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