LEAP 2025 Science Grades 3–8 Technical Report: 2020–2021

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 addendum provides information on the operational test administrations, scoring activities, analyses, and results of the spring 2021 administration of the LEAP 2025 science tests, which used intact forms based on previously administered operational forms. For information on the development and construction processes for these forms, see the 2019 LEAP 2025 Science Grades 3-8 Technical Report.

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

The chapters of this technical report outline general information about the administration and scoring activities of the LEAP 2025 assessments, CTT (Classical Test Theory) and IRT (Item Response Theory) analysis results, 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 2021 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 its state content standards. Per state law, the LDOE is to administer statewide summative science assessments in grades 3–8 and in science. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. Further, the LDOE and the BESE are committed to the development of rigorous assessments as one component of their comprehensive plan—Louisiana Believes—designed to ensure that every Louisiana student is on track to be successful in postsecondary education and the workforce.

The purpose of this technical addendum is to describe the processes for the spring 2021 administration of LEAP 2025 science tests. This report outlines the testing administrations, scoring activities, and psychometric analyses.

2. Test Administration

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

Training of School Systems

To ensure that the LEAP 2025 assessments are administered and scored in accordance with the department's mandates, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE provides train-the-trainer opportunities for the district test coordinators, who in turn convey test administration training to schools within their system. The LDOE conducts quality-assurance visits during testing to ensure adherence to the standardized administration of the tests.

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

Ancillary Materials

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

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

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

DRC also produced Test Coordinators Manuals for paper-based test administrations and for computer-based test administration. LDOE assessment staff review, provide feedback, and give final approval for these manuals. The Test Coordinators Manuals are inclusive of grades 3–8 English Language Arts (ELA), Mathematics, Social Studies, and Science. They provide detailed instructions for district and school test coordinators' responsibilities for distributing, collecting, and returning test materials to DRC for scoring.

Table of Contents for Paper-Based Testing Test Coordinators Manual

- Key Dates
- Spring 2021 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 Grades 3 and 4
- District Test Coordinator
    - Conduct Training Session
    - Receive Test Materials
    - Spanish Mathematics
    - Large-Print and Braille Test Materials and Communication Assistance Scripts (CAS)
    - Accommodated Materials
    - Verify and Distribute Test Materials to School Test Coordinators
    - Request Additional Test Materials and Bar-Code Labels
    - Collect Materials from Schools After Testing
    - Used and Unused Consumable Test Booklets (Defined)
    - Unscorable Documents and Unscorable Document Labels
- Directions for Returning Test Materials to DRC in May
    - Pickup 1: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials
    - Pickup 3: Nonscorable Test Materials
    - Final Checklist for Returning Test Materials to DRC
- School Test Coordinator
    - Receive and Verify Test Materials
    - Conduct Test Administration and Security Training Session
    - Supervise Application of Bar-Code Labels and Coding of Consumable Test Booklets
    - Soiled, Damaged, and Other Unscorable Consumable Test Booklets

    - Verify and Distribute Materials to Test Administrators
    - Supervise Test Administration
    - Collect Test Materials
    - Used and Unused Consumable Test Booklets (Defined)
    - Coding Responsibilities of Principals—Before Testing
    - Coding Responsibilities of Principals—Before or After Testing
    - Coding Responsibilities of Principals—After Testing
- Directions for Returning Test Materials to the DTC
    - Pickup 1: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials
    - Pickup 3: Nonscorable Test Materials
    - Final Checklist for Returning Test Materials to the District Test Coordinator
- Void Notification—Spring 2021
- Index

Table of Contents for Computer-Based Testing Test Coordinators Manual

- Key Dates Spring 2021
- Resources Available in DRC INSIGHT Portal (eDIRECT) Spring 2021
- Spring 2021 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - DRC INSIGHT Portal (eDIRECT) and INSIGHT
- Test Security
    - Key Definitions
    - Violations of Test Security
- Testing Guidelines
    - Testing Eligibility

    - Testing Conditions
    - Testing Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Accommodations
    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grades 3 through 8
- Roles and Responsibilities
    - District Test Coordinator
    - School Test Coordinator
    - Technology Coordinator
- Managing Test Tickets
    - Student Transfers
    - Locked Test Tickets
    - Technical Issues
    - Invalidating Test Tickets
- Resources for Online Testing
    - Test Administration Manuals
    - *DRC INSIGHT Portal (eDIRECT) User Guides*
    - *LEAP 2025 Accommodations and Accessibility Features User Guide*
    - *INSIGHT Technology User Guide*
    - Online Tools Training (OTT)
    - Student Tutorials
- Void Notification—Spring 2021

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

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

- Testing Times for Grades 3 and 4
    - Makeup Testing
    - Testing Conditions
- Special Populations and Accommodations
    - IDEA Special Education Students
    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard of Hearing Students
    - English Learners (ELs)
- Hand-Coded Consumable Test Booklets
- Students Absent from Testing
- Consumable Test Booklet Coding
    - Coding the Demographic Section
- Sample Grade 3 English Language Arts Consumable Test Booklet
- General Instructions for LEAP 2025
    - Student Marking/Erasing on Consumable Test Booklet
    - Reading Directions to Students
    - Special Instructions
- Directions for Administering LEAP 2025 Tests
- Post-Test Procedures
    - Test Administrator Oath of Security and Confidentiality Statement
    - Used and Unused Consumable Test Booklets (Defined)
    - Transferring Student Responses
    - Returning Test Materials to the School Test Coordinator
- Index

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

- Student Tutorials
    - Student Tutorials
- Special Populations and Accommodations
    - IDEA Special Education Students
    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard-of-Hearing Students
    - English Learners (ELs)
- Test Materials
    - Receipt of Test Materials
- 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 1–3)
- LEAP 2025: Grades 3–8 Social Studies (Grades 3–4 Sessions 1–2, Grades 5–8 Sessions 1–3)
- Post-Test Procedures
    - Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    - Returning Test Materials to the School Test Coordinator
- Index

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

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

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

The LEAP 2025 test administration manuals provide instructions for activities that happen before, during, and after testing with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the test administration manuals describe the following: general rules of paper and online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

Furthermore, the standardized procedures addressed in the test administration manual need to be followed, as the *Standards* state in Standard 6.1: "Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user" (114). To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the LEAP 2025 was administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The test coordinator's manual included instructions for scheduling the test within the state testing window. The test coordinator's manual and test administration manual also contained the schedule for timing each test session.

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

Department staff administer reports on testing concerns that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; allowing students to correct or complete answers after tests have

been submitted; splitting sessions into two parts; ignoring the standardized directions in the online assessment; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the Individualized Education Program (IEP), Individual Accommodations Plan/504 Plan (IAP), or English Learner Plan (EL plan); allowing accommodations for students who do not have an IEP/IAP/EL plan; or defining terms on the test.

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

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

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

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

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

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

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

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

Security Checklists. As soon as printed test materials are received by a school system, the district test coordinator ensures that the first and last security bar codes on the tests match the packing list they received. The district test coordinator then packages the tests to be sent to schools. Upon returning the test books to DRC, school and district test coordinators are required to complete and submit an accountability form that details the number of test books or printed test forms returned. This form also requires that systems/schools document nonstandard situations, including lost, damaged, destroyed, extra, or missing test books.

Interpretive Guides

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

Time

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

Online Forms Administration, Grades 3–8

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

Paper-Based Forms Administration, Grades 3 and 4

Schools with testers in grades 3 and 4 had the option to participate in either paper-based or computer-based testing for the spring 2021 test. DRC prints and ships paper materials to the sites that choose paper-based testing. These materials are returned to DRC after testing, for processing and scoring with the online tests.

Accessibility and Accommodations

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

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

Accommodations may be used with students who qualify under the Individuals with

Disabilities Education Act (IDEA) and have an IEP or Section 504 of the Americans with Disabilities Act and have an IAP, or who are identified as English Learners (ELs) and have an EL plan.

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

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

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

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

The following additional access and accommodation features were also available:

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

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

Testing Windows

The computer-based test window was available from April 26 through May 28, 2021. Paper-based testing occurred from April 28 through May 4, 2021.

Test Security Procedures

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

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

Data Forensic Analyses

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

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

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

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

Item Exposure Monitoring. Due to the re-use of the 2019 operational forms for the spring of 2021 administration, item performance was examined to ensure that item content had not been exposed. Frequently during the testing window, every item's moving *p*-value and point-biserial averages were produced. If an item's moving average *p*-value was larger than expected compared to the previous administrations, the item was flagged. Additionally, plots were produced for a visual inspection of the day-to-day patterns of item performance.

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

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

3. Scoring Activities

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

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

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

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

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

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

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

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

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

Constructed- and Extended-Response Item Scoring Process

Constructed- and extended-response items are scored by human raters trained by DRC. Ten percent of the responses are scored twice to monitor and maintain inter-rater reliability. Scoring supervisors also conduct read-behinds and review all nonscores and alerts. Handscoring processing rules are detailed in the *LEAP 2025 Spring 2021 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 were selected and trained for the LEAP 2025 handscoring process and describe how the scorers were monitored throughout the handscoring process.

Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2020–2021 LEAP 2025 test responses have at least a four-year college degree.

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

and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and does not demonstrate the ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

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

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

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

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

Handscoring Training Process. Standard 6.9 specifies:

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

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

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

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

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

Table 3.1
Science Training Set Composition

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

* Examples of responses at the top score points or for all score point combinations were not present in some anchor, training, and qualifying sets, as there were few or no examples found during rangefinding or subsequent field test scoring. DRC scoring directors identified examples of these scores during live scoring to supplement reader training.
** For grades 3 and 4, ER is 6-point item.

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

Any scorer who did not qualify by the end of the qualifying process for an item was not allowed to score live student responses. The qualifying standards for the science constructed- and extended-response items are shown in Table 3.2.

Table 3.2

Science Qualifying Standards

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

* Qualifying sets are made up of 10 responses comparable to the anchor set responses. For multi-part ERs, the appropriate qualifying standard should be achieved on each part of the item. For example, if an item has Part A with a top score of 6 and Part B with a top score of 3, a scorer would need to achieve 60% perfect agreement on Part A and 70% perfect agreement on Part B on one or more of the qualifying sets. A scorer may qualify on one part in the first qualifying set and the other part in the second qualifying set.
** For grades 3 and 4, ER is 6-point item.

Monitoring the Scoring Process. Standard 6.8 states:

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

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

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

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

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

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

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

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

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

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

Table 3.3
Agreement Rate Requirements for Validity and Inter-Rater Reliability

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

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

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

readers scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and the anchor responses to explain the reasoning behind each response's score.

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

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

Tables 3.4–3.7 provide the inter-rater reliability and score point distributions by grade level for the constructed-response and extended-response items administered in the spring 2021 forms.

Table 3.4

Inter-Rater Reliability for Operational Constructed-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3	Item 1	≥14,620	94	6	0
	Item 2	≥14,830	94	6	0
	Item 3	≥16,140	94	6	0
4	Item 1	≥12,170	96	4	0
	Item 2	≥16,200	95	5	0
	Item 3	≥12,630	94	6	0
5	Item 1	≥12,270	94	6	0
	Item 2	≥13,770	91	9	0
	Item 3	≥11,810	95	5	0
6	Item 1	≥13,220	91	9	1
	Item 2	≥12,730	87	13	1
	Item 3	≥11,580	81	19	0
7	Item 1	≥14,800	88	11	0
	Item 2	≥9,570	98	2	0
	Item 3	≥13,070	93	7	0
8	Item 1	≥15,870	92	7	0
	Item 2	≥15,060	93	6	0
	Item 3	≥16,460	94	6	0

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

Table 3.5

Score Point Distributions for Operational Constructed-Response Items

Grade	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
3	Item 1	≥59,980	55	26	4	8	7
	Item 2	≥60,080	35	47	3	8	7
	Item 3	≥60,720	39	32	12	9	8
4	Item 1	≥58,240	72	14	4	7	3
	Item 2	≥60,240	62	17	4	7	9
	Item 3	≥58,390	75	11	1	9	3
5	Item 1	≥55,800	63	14	18	0	5
	Item 2	≥56,370	29	55	7	0	7
	Item 3	≥55,420	65	7	24	0	3
6	Item 1	≥57,560	80	13	2	0	5
	Item 2	≥57,100	73	17	3	0	6
	Item 3	≥56,730	61	32	2	0	4
7	Item 1	≥59,050	32	47	9	0	11
	Item 2	≥55,800	90	2	0	0	6
	Item 3	≥57,760	41	39	12	0	8
8	Item 1	≥59,120	69	17	3	0	10
	Item 2	≥58,620	75	15	3	0	7
	Item 3	≥58,940	46	41	1	0	11

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

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

Table 3.6

Inter-Rater Reliability for Operational-Extended Response Items

Grade	Inter-Rater Reliability*				
	2x	Part	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3	≥17,400	N/A	93	5	2
4	≥13,170	N/A	85	15	0
5	≥15,450	N/A	81	15	3
6	≥14,070	Part A	93	4	3
		Part B	89	10	1
		Part C	93	6	1
7	≥14,400	Part A	93	7	0
		Part B	93	7	0
		Part C	98	2	0
8	≥17,230	Part A	86	13	1
		Part B	81	15	4

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

Table 3.7

Score Point Distributions for Operational-Extended Response Items

Grade	Score Point Distribution*													
	Total	Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	"6" (%)	"7" (%)	"8" (%)	"9" (%)	Blank (%)	Nonscore Codes (%)**
3	≥61,390	N/A	59	8	8	2	3	0	0				8	11
4	≥58,860	N/A	14	18	37	20	1	0	0				6	4
5	≥57,170	N/A	40	9	8	8	8	7	5	3	1	0	0	11
6	≥57,830	A	70	12	0	11							0	5
		B	61	24	8	1							0	5
		C	63	13	16	1							0	5
7	≥57,900	A	78	8	4	2							1	8
		B	72	11	5	2	1						1	8
		C	79	9	3								1	8
8	≥59,460	A	31	28	21	8							0	12
		B	12	16	21	19	13	6	2				0	12

* 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.

4. 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 2021 test can be found in Appendix B: Item Analysis Summary Report. Tables B.1.1–B.5.1 show 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. In general, statistical analysis results for field-test (FT) items are stored in Pearson's Assessment Banking and Building solutions for Interoperable assessment (ABBI) system. Placeholder (PH) items included on test forms are not part of any statistical analyses because the purpose of PH items is to maintain a consistent testing length and experience by occupying FT-item positions for administrations when no field testing takes place; therefore, these items do not require any statistical analysis.

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/Hispanic) 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 4.1 defines the DIF categories for dichotomous items.

Table 4.1
DIF Categories for Dichotomous Items

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

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

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

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

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

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

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

The *p*-value associated with the Mantel χ^2 statistic and the *SMD* (on an effect-size metric) are used to determine DIF classifications. Table 4.2 defines the DIF categories for polytomous items.

Table 4.2
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 Table 4.3 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. Tables 4.3.1 to 4.3.3 summarize the operational test DIF statistics for the operational items appearing on the spring 2021 test form. Because the spring 2021 tests were administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 4.3.1
Summary of Female – Male DIF Flags for Spring 2021 Science Operational Items

Grade	A	[B+],[B-]	[C+],[C-]
3	39	[0],[0]	[0],[0]
4	40	[1],[0]	[0],[0]
5	39	[0],[1]	[0],[0]
6	40	[0],[1]	[0],[0]
7	39	[0],[2]	[0],[0]
8	38	[0],[2]	[0],[1]

Table 4.3.2

Summary of African American – White DIF Flags for Spring 2021 Science Operational Items

Grade	A	[B+],[B-]	[C+],[C-]
3	39	[0],[0]	[0],[0]
4	41	[0],[0]	[0],[0]
5	40	[0],[0]	[0],[0]
6	41	[0],[0]	[0],[0]
7	38	[0],[3]	[0],[0]
8	40	[0],[1]	[0],[0]

Table 4.3.3

Summary of Hispanic – White DIF Flags for Spring 2021 Science Operational Items

Grade	A	[B+],[B-]	[C+],[C-]
3	38	[0],[1]	[0],[0]
4	41	[0],[0]	[0],[0]
5	40	[0],[0]	[0],[0]
6	40	[0],[1]	[0],[0]
7	40	[0],[1]	[0],[0]
8	41	[0],[0]	[0],[0]

Pre-Equating for Intact Forms

In general, the LEAP 2025 science assessment utilizes a statistical procedure called the post-equating method based on Item Response Theory (IRT) models to place the new forms administered on the same scale. For the spring 2021 administration, however, one of the 2019 operational forms was used; therefore, the pre-equating method was applied. That is, existing scoring tables were used for score reports and performance classifications.

Unidimensionality and Principal Component Analysis

Appendix C: 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 2021 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 C.2.1 and Figures C.1.1 and C.1.2 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 spring 2021 tests. Because the spring tests were administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Scaling

Although the spring 2021 tests used the preexisting scoring tables, general procedures for the scaling method are described here since scaling is directly associated with performance-level cuts. Based on the Standard Setting 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 are subsequently interpolated and vary by grades and subjects. 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 D: Scale Distribution and Statistical Report.

5. Reliability and Validity

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 ith 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 4. Data Analysis."

The reliability and classification accuracy reports in Appendix E: Reliability and Classification Accuracy provide coefficient alpha for the total test. First, the coefficient alpha values for the spring 2021 assessments were between 0.851 and 0.873. Because the spring tests were administered during the COVID-19 pandemic, any statistical inferences should be cautiously drawn from the result. Additional reliabilities were calculated on various demographic subgroups using the population of students (Appendix E: Reliability and Classification Accuracy). The subgroups are male/female, white/Black/Hispanic/Asian/American Indian or Alaska Native/Native Hawaiian or Other

Pacific Islander/multi-racial, Economically Disadvantaged Status, English Learners, Section 504, and Education Classification.

Cronbach's 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.

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, but 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 E: 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 values of the spring 2021 tests were between 0.653 and 0.694, the overall consistency values were between 0.550 and 0.587. Because the spring 2021 tests were administered during the COVID-19 pandemic, however, great caution should be applied when any statistical inference is drawn.

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 0.222 and 0.246 for the spring 2021 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 0.407 and 0.452 for the spring 2021 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.

Validity

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating 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 spring 2021 tests were designed and developed to provide fair and accurate scores that support appropriate, meaningful, and useful educational decisions. As the technical addendum/report progresses, it reflects the phases of the testing cycle. Each part of the technical addendum details the procedures and processes applied in the creation of the LEAP 2025 science tests and their results. Validity evidence may be found in the following portions: Chapter 2 (Test Administration), Chapter 3 (Scoring Activities), Chapter 4 (Data Analysis), Chapter 5 (Reliability and Validity), and Chapter 6 (Statistical Summaries). For validity evidence related to the development and construction of the test forms used in the spring 2021 administration, please refer to the 2019 LEAP 2025 Science Grades 3-8 Technical Report. Because the spring 2021 tests were administered during the COVID-19 pandemic, any validity evidence associated with the spring test should be carefully interpreted and argued.

The knowledge, expertise, and professional judgment offered by Louisiana educators ultimately ensure that the content for 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. Participation by Louisiana educators throughout the process—from source selection, item development, and content and bias review to rangefinding and standard setting—reinforces confidence in the content and design of the LEAP 2025 science tests to derive valid inferences about Louisiana student performance.

Chapter 2 of the technical addendum 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 3 describes scoring processes and activities for the LEAP 2025 science assessments.

Although the spring 2021 tests are based on a pre-equating method, Chapter 4 briefly describes classical data analysis, IRT, and scaling of the science tests, which derive scale scores from students' raw scores. In addition, Chapter 4 describes an analysis of DIF and includes gender and ethnicity DIF results. A summary of classical analysis and DIF results for the operational items is presented in Appendix B: Item Analysis Summary Report.

Chapter 5 addresses Cronbach's alpha as a measure of internal consistency and also describes analysis procedures for classification consistency and classification accuracy.

Chapter 6 reports the statistical summaries of the spring 2021 science tests.

6. Statistical Summaries

For the spring 2021 science tests, the lowest obtainable scale score (LOSS) on the tests is 650 and the highest obtainable scale score (HOSS) is 850. Test results are presented in Tables 6.1 through 6.6. 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 Tables 6.1 through 6.6, scale score frequency distributions are presented in Appendix D: Scale Distribution and Statistical Report. Finally, because the spring 2021 tests were administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 6.1

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 3

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥49,560	721.62	31.23	20	32	28	15	5
Gender								
Female	≥24,260	721.44	30.47	20	33	28	14	5
Male	≥25,280	721.80	31.95	21	31	27	16	5
Ethnicity								
African American	≥20,940	709.62	28.51	30	39	22	7	1
American Indian or Alaska Native	≥310	724.97	29.08	15	34	28	19	4
Asian	≥820	740.24	30.91	9	19	30	27	15
Hispanic/Latino	≥4,830	717.48	30.06	23	35	27	12	3
Multi-Racial	≥1,650	726.25	29.60	15	30	33	17	6
Native Hawaiian or Other Pacific Islander	≥40	731.36	24.11	5	32	45	11	7
White	≥20,950	733.40	29.39	10	25	33	23	9
Economically Disadvantaged*								
No	≥12,610	740.17	28.67	7	19	32	28	13
Yes	≥36,720	715.32	29.48	25	36	26	11	2
English Learner								
No	≥47,080	722.50	31.25	19	32	28	16	5
Yes	≥2,470	704.89	25.75	34	43	19	3	NR
Education Classification								
Regular	≥43,400	723.43	30.92	18	31	29	16	5
Special	≥6,150	708.86	30.45	33	38	19	8	2
Section 504								
No	≥46,230	722.01	31.32	20	32	28	16	5
Yes	≥3,320	716.18	29.50	23	38	26	9	4

* ≥220 students with no record of either No or Yes.

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

Table 6.2

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 4

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥49,540	729.25	31.28	22	23	28	22	6
Gender								
Female	≥24,120	728.98	29.80	20	24	30	21	5
Male	≥25,410	729.51	32.62	23	22	27	22	7
Ethnicity								
African American	≥21,130	715.60	27.84	34	29	26	10	1
American Indian or Alaska Native	≥280	732.07	27.46	15	22	37	23	3
Asian	≥760	749.51	31.51	8	13	25	36	19
Hispanic/Latino	≥4,590	726.87	29.78	22	24	30	20	4
Multi-Racial	≥1,630	734.57	30.24	16	21	31	25	7
Native Hawaiian or Other Pacific Islander	≥30	737.49	33.87	18	15	23	36	8
White	≥21,090	742.25	28.90	10	16	31	33	10
Economically Disadvantaged*								
No	≥12,990	747.94	28.46	7	13	28	38	14
Yes	≥36,310	722.64	29.48	27	26	28	16	3
English Learner								
No	≥47,410	730.07	31.28	21	22	29	22	6
Yes	≥2,130	711.17	25.13	38	33	23	6	NR
Education Classification								
Regular	≥43,430	731.65	30.68	19	22	30	23	6
Special	≥6,110	712.18	30.15	42	27	20	10	2
Section 504								
No	≥45,100	729.93	31.37	21	22	28	22	6
Yes	≥4,430	722.33	29.41	27	27	27	15	3

* ≥230 students with no record of either No or Yes.

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

Table 6.3

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 5

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥49,860	727.87	35.88	20	26	24	23	7
Gender								
Female	≥24,200	728.92	34.46	18	28	25	23	7
Male	≥25,660	726.89	37.15	22	25	23	23	7
Ethnicity								
African American	≥21,110	712.83	31.55	31	35	21	12	2
American Indian or Alaska Native	≥320	729.44	30.39	15	26	30	25	3
Asian	≥800	752.65	36.88	8	13	23	33	23
Hispanic/Latino	≥4,770	722.40	35.00	24	27	25	20	4
Multi-Racial	≥1,600	732.65	33.20	15	24	28	26	7
Native Hawaiian or Other Pacific Islander	≥20	739.07	37.73	14	18	18	39	11
White	≥21,210	742.73	33.64	9	19	26	33	12
Economically Disadvantaged*								
No	≥12,780	750.68	32.54	5	15	25	38	17
Yes	≥36,760	720.03	33.54	25	30	24	17	3
English Learner								
No	≥47,650	729.06	35.69	19	26	24	23	7
Yes	≥2,200	702.12	29.79	44	34	16	6	NR
Education Classification								
Regular	≥43,890	731.30	34.87	16	26	25	25	8
Special	≥5,960	702.61	33.06	47	29	14	8	2
Section 504								
No	≥44,760	729.12	35.96	19	26	24	24	7
Yes	≥5,100	716.92	33.22	28	33	22	14	3

* ≥310 students with no record of either No or Yes.

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

Table 6.4

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 6

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥51,540	726.19	30.92	22	25	30	20	3
Gender								
Female	≥25,340	726.46	29.25	20	26	32	20	2
Male	≥26,200	725.93	32.44	23	24	28	21	3
Ethnicity								
African American	≥22,290	713.90	27.90	33	31	26	9	1
American Indian or Alaska Native	≥310	729.42	28.53	16	23	36	22	3
Asian	≥760	749.39	30.48	7	13	28	40	12
Hispanic/Latino	≥4,600	721.43	30.74	26	27	29	16	2
Multi-Racial	≥1,670	731.85	29.13	15	24	33	25	3
Native Hawaiian or Other Pacific Islander	≥40	729.56	28.52	19	17	31	33	NR
White	≥21,830	738.45	28.65	10	19	34	32	5
Economically Disadvantaged*								
No	≥13,210	744.13	28.07	7	15	33	38	7
Yes	≥38,010	720.02	29.40	26	29	29	14	1
English Learner								
No	≥49,570	727.08	30.75	21	25	31	21	3
Yes	≥1,970	703.83	26.32	46	33	17	4	NR
Education Classification								
Regular	≥45,670	729.09	29.97	18	25	32	22	3
Special	≥5,860	703.66	28.83	49	29	16	6	1
Section 504								
No	≥46,070	727.38	30.90	20	24	31	22	3
Yes	≥5,460	716.16	29.16	31	31	26	11	1

* ≥310 students with no record of either No or Yes.

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

Table 6.5

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 7

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥52,330	727.85	31.85	20	28	28	21	3
Gender								
Female	≥25,670	728.46	30.53	18	29	29	21	3
Male	≥26,660	727.27	33.07	22	27	26	22	3
Ethnicity								
African American	≥22,470	715.23	27.55	31	35	24	10	1
American Indian or Alaska Native	≥300	733.59	29.61	12	30	29	25	4
Asian	≥830	756.41	35.67	6	13	24	41	16
Hispanic/Latino	≥4,640	725.41	31.97	23	27	27	21	2
Multi-Racial	≥1,590	731.38	31.13	16	27	30	24	3
Native Hawaiian or Other Pacific Islander	≥40	742.84	36.24	16	22	16	39	6
White	≥22,420	739.59	30.55	10	21	32	32	5
Economically Disadvantaged*								
No	≥14,090	746.43	30.31	7	17	31	38	7
Yes	≥37,890	721.05	29.58	25	31	26	15	1
English Learner								
No	≥50,410	728.71	31.74	19	27	28	22	3
Yes	≥1,920	705.48	26.14	44	34	16	5	NR
Education Classification								
Regular	≥46,720	730.65	31.14	17	27	29	23	3
Special	≥5,600	704.57	28.00	49	31	13	6	1
Section 504								
No	≥46,800	729.18	31.89	19	27	28	23	3
Yes	≥5,520	716.66	29.25	30	34	23	12	1

* ≥340 students with no record of either No or Yes.

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

Table 6.6

LEAP 2025 State Test Results: Spring 2021 Operational Science Grade 8

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥51,850	728.78	31.51	14	30	29	23	4
Gender								
Female	≥25,540	729.24	30.22	13	30	30	24	3
Male	≥26,300	728.34	32.72	16	29	27	23	5
Ethnicity								
African American	≥22,160	715.21	27.95	23	40	25	11	1
American Indian or Alaska Native	≥310	730.68	30.58	13	26	33	24	4
Asian	≥800	754.11	32.37	5	13	21	44	17
Hispanic/Latino	≥4,210	722.88	32.02	20	30	27	20	3
Multi-Racial	≥1,520	733.77	29.70	10	28	31	27	5
Native Hawaiian or Other Pacific Islander	≥40	737.41	29.78	9	18	34	32	7
White	≥22,770	741.82	28.63	6	20	32	36	7
Economically Disadvantaged*								
No	≥14,560	746.94	28.26	4	16	31	40	9
Yes	≥36,940	721.71	29.84	18	35	28	17	2
English Learner								
No	≥49,960	729.81	31.23	13	29	29	24	4
Yes	≥1,880	701.71	26.65	41	40	14	4	NR
Education Classification								
Regular	≥46,650	731.58	30.65	12	29	30	25	4
Special	≥5,190	703.65	27.82	40	39	15	6	1
Section 504								
No	≥46,550	730.00	31.46	14	29	29	25	4
Yes	≥5,290	718.13	29.95	21	39	25	13	2

* ≥340 students with no record of either No or Yes.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A: Test Summary

Science

Contents
Table A.1 Item Type Summary: Spring 2021 Operational Science Tests
Table A.2 Raw Score Summary: Spring 2021 Operational Science Tests
Table A.3 Raw Score Summary by Reporting Category: Spring 2021 Operational Science Tests
Table A.4 Scale Score and Raw Score Summary: Spring 2021 Operational Science Tests

- Because the spring 2021 tests were administered during the 2021 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table A.1

Item Type Summary: Spring 2021 Operational Science Tests

Grade	MC	MS	TEI	CR	ER	TPD	TPI
3	21	4	0	3	1	7	3
4	22	3	0	3	1	6	6
5	14	1	13	3	1	4	4
6	16	3	10	3	1*	6	2
7	11	6	14	3	1*	4	2
8	14	2	14	3	1*	4	3

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

Table A.2

Raw Score Summary: Spring 2021 Operational Science Tests

Grade	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
3	≥49,560	18.62	8.95	0	50	0.38	0.39	0.86	3.35
4	≥49,540	23.26	9.46	0	56	0.42	0.39	0.87	3.41
5	≥49,860	25.17	11.22	1	62	0.45	0.40	0.86	4.20
6	≥51,540	22.11	9.95	1	61	0.37	0.37	0.85	3.85
7	≥52,330	22.66	10.21	0	65	0.36	0.39	0.87	3.68
8	≥51,850	25.62	11.03	0	65	0.41	0.40	0.87	3.98

***** Reliability is Cronbach's alpha.

Table A.3

Raw Score Summary by Grade and Reporting Category: Spring 2021 Operational Science Tests

Grade	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
3	Investigate	4.31	2.51	0	14	0.33	0.33	0.57	1.65
	Evaluate	8.77	4.70	0	26	0.39	0.42	0.75	2.35
	Reason Scientifically	4.12	1.94	0	9	0.48	0.37	0.51	1.36
4	Investigate	5.62	2.70	0	15	0.43	0.41	0.64	1.62
	Evaluate	4.01	2.13	0	10	0.41	0.35	0.49	1.52
	Reason Scientifically	11.15	4.61	0	25	0.44	0.38	0.73	2.40
5	Investigate	4.45	2.37	0	10	0.49	0.40	0.61	1.48
	Evaluate	10.67	4.85	0	26	0.44	0.40	0.76	2.38
	Reason Scientifically	10.06	5.24	0	28	0.44	0.42	0.65	3.10
6	Investigate	3.26	2.01	0	12	0.31	0.36	0.51	1.41
	Evaluate	8.79	5.01	0	30	0.34	0.41	0.74	2.55
	Reason Scientifically	10.07	4.26	0	25	0.42	0.34	0.66	2.48
7	Investigate	2.20	1.51	0	10	0.24	0.27	0.27	1.29
	Evaluate	4.21	2.54	0	11	0.40	0.40	0.58	1.65
	Reason Scientifically	13.44	6.63	0	40	0.37	0.41	0.82	2.81
8	Investigate	8.06	3.95	0	20	0.42	0.42	0.73	2.05
	Evaluate	8.20	3.56	0	18	0.49	0.40	0.67	2.05
	Reason Scientifically	9.37	4.86	0	29	0.35	0.37	0.70	2.66

Table A.4.1

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 3 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥49,560	100.00	721.62	31.23	18.62	8.95
Female	≥24,260	48.96	721.44	30.47	18.49	8.75
Male	≥25,280	51.02	721.80	31.95	18.74	9.14
African American	≥20,940	42.26	709.62	28.51	15.14	7.29
American Indian or Alaska Native	≥310	0.64	724.97	29.08	19.43	8.62
Asian	≥820	1.66	740.24	30.91	24.33	9.91
Hispanic/Latino	≥4,830	9.75	717.48	30.06	17.35	8.31
Multi-Racial	≥1,650	3.33	726.25	29.60	19.83	8.79
Native Hawaiian or Other Pacific Islander	≥40	0.09	731.36	24.11	20.86	7.95
White	≥20,950	42.28	733.40	29.39	22.05	9.15
Economically Disadvantaged: No*	≥12,610	25.44	740.17	28.67	24.23	9.26
Economically Disadvantaged: Yes*	≥36,720	74.11	715.32	29.48	16.71	7.99
EL: No	≥47,080	95.01	722.50	31.25	18.87	9.01
EL: Yes	≥2,470	4.99	704.89	25.75	13.74	6.03
Regular Education	≥43,400	87.58	723.43	30.92	19.11	8.98
Special Education	≥6,150	12.42	708.86	30.45	15.12	7.90
Section 504: No	≥46,230	93.28	722.01	31.32	18.74	8.99
Section 504: Yes	≥3,320	6.72	716.18	29.50	16.89	8.19

* Economic Status was not available for all students.

Table A.4.2

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 4 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥49,540	100.00	729.25	31.28	23.26	9.46
Female	≥24,120	48.69	728.98	29.80	23.09	9.03
Male	≥25,410	51.30	729.51	32.62	23.42	9.84
African American	≥21,130	42.66	715.60	27.84	19.09	7.76
American Indian or Alaska Native	≥280	0.58	732.07	27.46	23.99	8.40
Asian	≥760	1.55	749.51	31.51	29.66	10.10
Hispanic/Latino	≥4,590	9.27	726.87	29.78	22.30	8.95
Multi-Racial	≥1,630	3.30	734.57	30.24	24.84	9.36
Native Hawaiian or Other Pacific Islander	≥30	0.08	737.49	33.87	26.10	10.33
White	≥21,090	42.57	742.25	28.90	27.27	9.23
Economically Disadvantaged: No*	≥12,990	26.23	747.94	28.46	29.13	9.23
Economically Disadvantaged: Yes*	≥36,310	73.30	722.64	29.48	21.18	8.62
EL: No	≥47,410	95.69	730.07	31.28	23.52	9.48
EL: Yes	≥2,130	4.31	711.17	25.13	17.49	6.74
Regular Education	≥43,430	87.67	731.65	30.68	23.96	9.39
Special Education	≥6,110	12.33	712.18	30.15	18.27	8.37
Section 504: No	≥45,100	91.04	729.93	31.37	23.47	9.51
Section 504: Yes	≥4,430	8.96	722.33	29.41	21.12	8.65

* Economic Status was not available for all students.

Table A.4.3

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 5 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥49,860	100.00	727.87	35.88	25.17	11.22
Female	≥24,200	48.53	728.92	34.46	25.40	10.90
Male	≥25,660	51.47	726.89	37.15	24.97	11.51
African American	≥21,110	42.34	712.83	31.55	20.39	9.20
American Indian or Alaska Native	≥320	0.64	729.44	30.39	25.40	9.73
Asian	≥800	1.62	752.65	36.88	33.23	12.01
Hispanic/Latino	≥4,770	9.57	722.40	35.00	23.50	10.61
Multi-Racial	≥1,600	3.22	732.65	33.20	26.56	10.70
Native Hawaiian or Other Pacific Islander	≥20	0.06	739.07	37.73	28.89	12.03
White	≥21,210	42.55	742.73	33.64	29.89	11.05
Economically Disadvantaged: No*	≥12,780	25.64	750.68	32.54	32.52	10.87
Economically Disadvantaged: Yes*	≥36,760	73.74	720.03	33.54	22.65	10.17
EL: No	≥47,650	95.57	729.06	35.69	25.53	11.22
EL: Yes	≥2,200	4.43	702.12	29.79	17.40	7.95
Regular Education	≥43,890	88.03	731.30	34.87	26.18	11.09
Special Education	≥5,960	11.97	702.61	33.06	17.77	9.20
Section 504: No	≥44,760	89.77	729.12	35.96	25.57	11.28
Section 504: Yes	≥5,100	10.23	716.92	33.22	21.67	9.97

* Economic Status was not available for all students.

Table A.4.4

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 6 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥51,540	100.00	726.19	30.92	22.11	9.95
Female	≥25,340	49.17	726.46	29.25	22.03	9.50
Male	≥26,200	50.83	725.93	32.44	22.19	10.37
African American	≥22,290	43.26	713.90	27.90	18.10	8.00
American Indian or Alaska Native	≥310	0.61	729.42	28.53	22.95	9.35
Asian	≥760	1.48	749.39	30.48	30.23	11.11
Hispanic/Latino	≥4,600	8.94	721.43	30.74	20.61	9.50
Multi-Racial	≥1,670	3.26	731.85	29.13	23.82	9.83
Native Hawaiian or Other Pacific Islander	≥40	0.09	729.56	28.52	23.10	9.23
White	≥21,830	42.37	738.45	28.65	26.11	10.04
Economically Disadvantaged: No*	≥13,210	25.63	744.13	28.07	28.15	10.15
Economically Disadvantaged: Yes*	≥38,010	73.75	720.02	29.40	20.04	8.99
EL: No	≥49,570	96.18	727.08	30.75	22.39	9.96
EL: Yes	≥1,970	3.82	703.83	26.32	15.31	6.71
Regular Education	≥45,670	88.62	729.09	29.97	22.96	9.89
Special Education	≥5,860	11.38	703.66	28.83	15.50	7.70
Section 504: No	≥46,070	89.40	727.38	30.90	22.50	10.02
Section 504: Yes	≥5,460	10.60	716.16	29.16	18.85	8.71

* Economic Status was not available for all students.

Table A.4.5

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 7 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥52,330	100.00	727.85	31.85	22.66	10.21
Female	≥25,670	49.05	728.46	30.53	22.77	9.87
Male	≥26,660	50.95	727.27	33.07	22.56	10.53
African American	≥22,470	42.95	715.23	27.55	18.58	8.21
American Indian or Alaska Native	≥300	0.59	733.59	29.61	24.29	9.90
Asian	≥830	1.60	756.41	35.67	32.28	12.26
Hispanic/Latino	≥4,640	8.88	725.41	31.97	21.96	10.01
Multi-Racial	≥1,590	3.05	731.38	31.13	23.75	10.09
Native Hawaiian or Other Pacific Islander	≥40	0.09	742.84	36.24	27.69	12.36
White	≥22,420	42.84	739.59	30.55	26.43	10.29
Economically Disadvantaged: No*	≥14,090	26.93	746.43	30.31	28.74	10.42
Economically Disadvantaged: Yes*	≥37,890	72.41	721.05	29.58	20.43	9.16
EL: No	≥50,410	96.33	728.71	31.74	22.92	10.22
EL: Yes	≥1,920	3.67	705.48	26.14	15.82	7.15
Regular Education	≥46,720	89.28	730.65	31.14	23.50	10.14
Special Education	≥5,600	10.72	704.57	28.00	15.68	7.87
Section 504: No	≥46,800	89.44	729.18	31.89	23.08	10.27
Section 504: Yes	≥5,520	10.56	716.66	29.25	19.09	8.91

* Economic Status was not available for all students.

Table A.4.6

Scale Score and Raw Score Summary: Spring 2021 Operational Science Grade 8 Test

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥51,850	100.00	728.78	31.51	25.62	11.03
Female	≥25,540	49.27	729.24	30.22	25.70	10.63
Male	≥26,300	50.73	728.34	32.72	25.55	11.39
African American	≥22,160	42.75	715.21	27.95	20.78	8.96
American Indian or Alaska Native	≥310	0.61	730.68	30.58	26.30	10.69
Asian	≥800	1.56	754.11	32.37	35.07	11.92
Hispanic/Latino	≥4,210	8.14	722.88	32.02	23.70	10.75
Multi-Racial	≥1,520	2.95	733.77	29.70	27.28	10.77
Native Hawaiian or Other Pacific Islander	≥40	0.08	737.41	29.78	28.77	10.78
White	≥22,770	43.92	741.82	28.63	30.23	10.70
Economically Disadvantaged: No*	≥14,560	28.08	746.94	28.26	32.17	10.69
Economically Disadvantaged: Yes*	≥36,940	71.25	721.71	29.84	23.07	10.06
EL: No	≥49,960	96.36	729.81	31.23	25.96	11.00
EL: Yes	≥1,880	3.64	701.71	26.65	16.73	7.55
Regular Education	≥46,650	89.98	731.58	30.65	26.55	10.91
Special Education	≥5,190	10.02	703.65	27.82	17.35	8.29
Section 504: No	≥46,550	89.78	730.00	31.46	26.05	11.06
Section 504: Yes	≥5,290	10.22	718.13	29.95	21.85	9.92

* Economic Status was not available for all students.

Appendix B: Item Analysis Summary Report

Summary Statistics Reports

Contents
Table B.1.1 P-Value Summary by Item Type: Spring 2021 Operational Science Tests
Plot B.1.1 P-Value Summary by Item Type: Spring 2021 Operational Science Tests
Table B.2.1 Item-Total Correlation Summary by Item Type: Spring 2021 Operational Science Tests
Plot B.2.1 Item-Total Correlation Summary by Item Type: Spring 2021 Operational Science Tests
Table B.3.1 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2021 Operational Science Tests
Plot B.3.1 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2021 Operational Science Tests
Table B.4.1 Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2021 Operational Science Tests
Table B.5.1 Statistically Flagged Items by Item Type: Spring 2021 Operational Science Tests

- Because the spring 2021 tests were administered during the 2021 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table B.1.1

P-Value Summary by Grade: Spring 2021 Operational Science Tests

Grade	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	39	0.059	0.289	0.344	0.467	0.709
4	41	0.077	0.307	0.426	0.532	0.801
5	40	0.104	0.327	0.446	0.579	0.820
6	41	0.089	0.255	0.389	0.478	0.708
7	41	0.014	0.267	0.348	0.466	0.763
8	41	0.116	0.327	0.427	0.490	0.762

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

Plot B.1.1

P-Value Summary by Grade: Spring 2021 Operational Science Tests

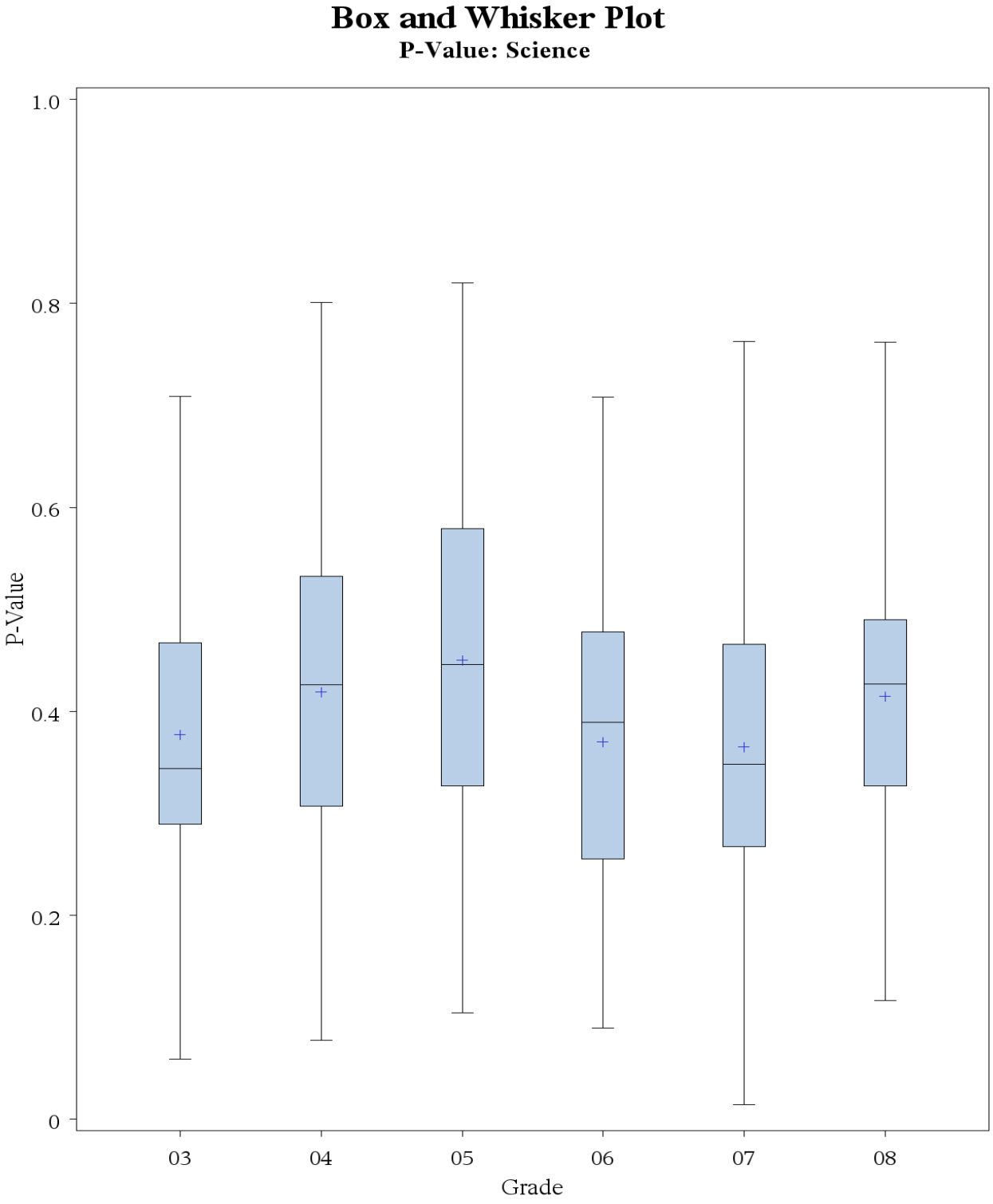


Table B.2.1

Item-Total Correlation Summary by Grade: Spring 2021 Operational Science Tests

Grade	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	39	0.084	0.290	0.406	0.491	0.585
4	41	0.120	0.326	0.401	0.477	0.588
5	40	0.223	0.322	0.379	0.480	0.740
6	41	0.171	0.311	0.372	0.455	0.572
7	41	0.040	0.292	0.358	0.498	0.618
8	41	0.062	0.319	0.388	0.457	0.664

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

Plot B.2.1

Item-Total Correlation Summary by Grade: Spring 2021 Operational Science Tests

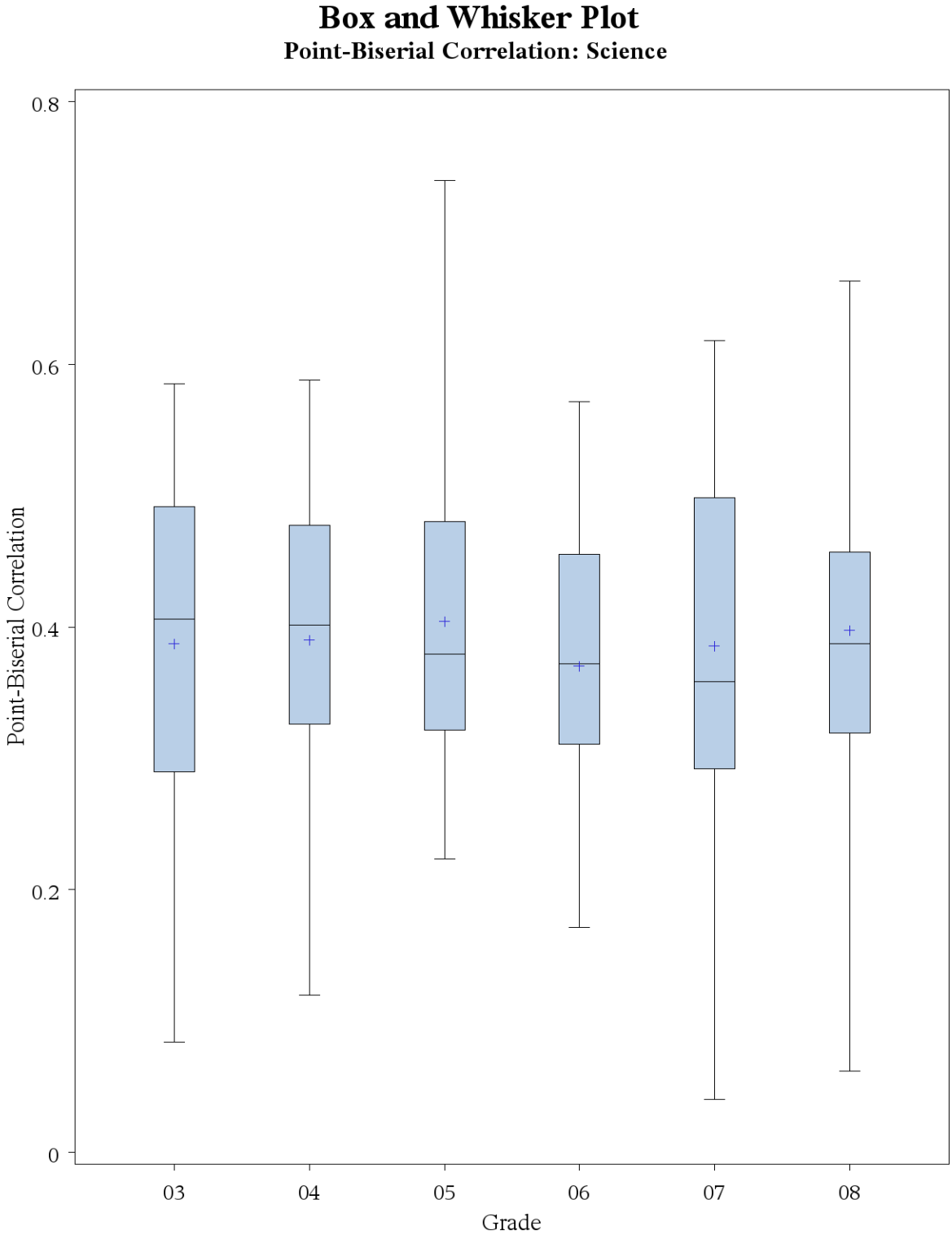


Table B.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2021 Operational Science Tests*

Grade	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	39	0.052	0.243	0.340	0.439	0.524
4	41	0.076	0.269	0.355	0.417	0.529
5	40	0.166	0.278	0.340	0.432	0.617
6	41	0.121	0.249	0.306	0.401	0.521
7	41	-0.006	0.240	0.313	0.457	0.589
8	41	0.032	0.279	0.338	0.420	0.570

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

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

Plot B.3.1

Corrected Point-Biserial Correlation Summary by Grade: Spring 2021 Operational Science Tests

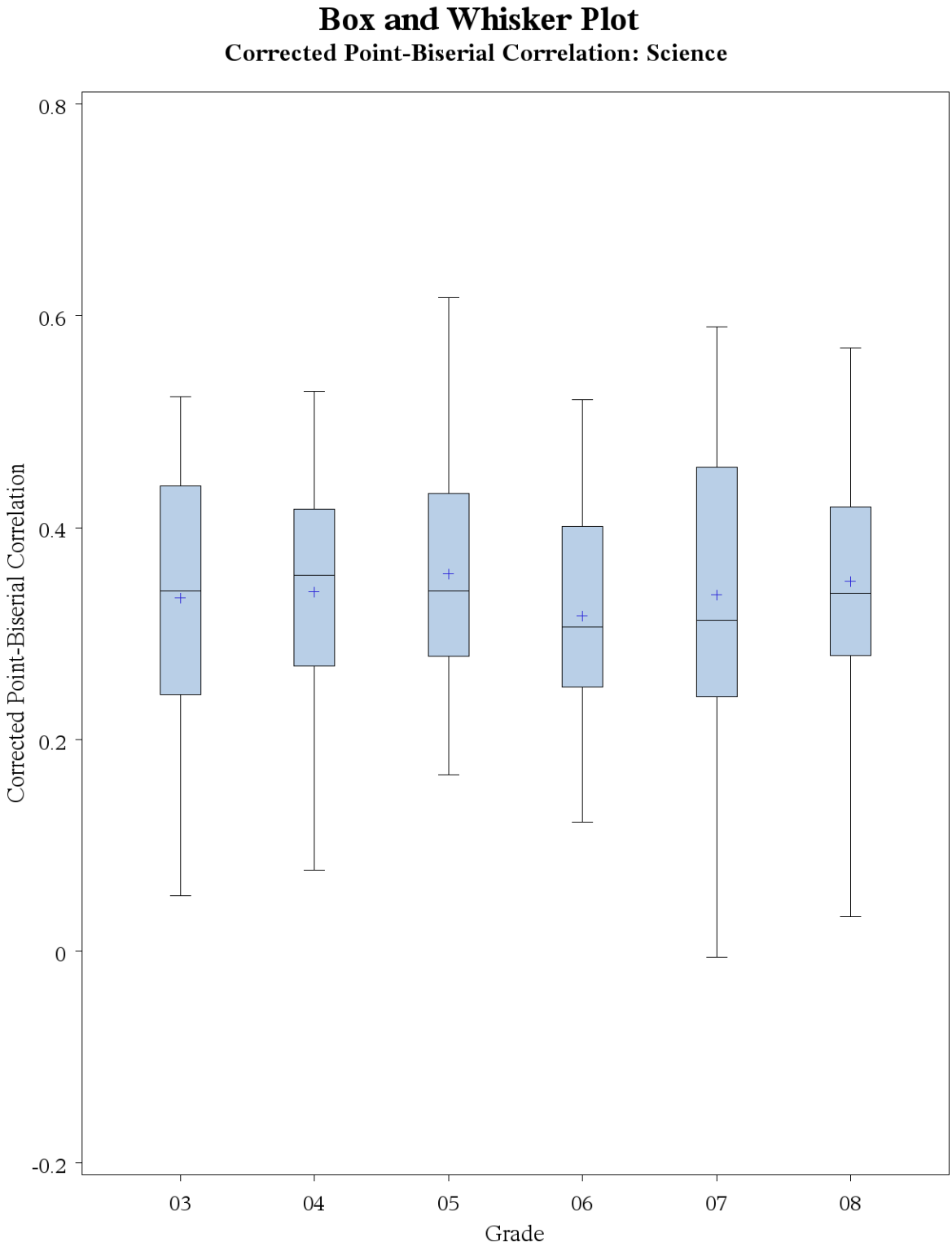


Table B.4.1

Item-Total Correlation Summary by Reporting Category: Spring 2021 Operational Science Tests

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	Investigate	11	0.084	0.185	0.334	0.466	0.569
	Evaluate	16	0.238	0.334	0.438	0.517	0.585
	Reason Scientifically	8	0.193	0.289	0.382	0.454	0.490
4	Investigate	10	0.264	0.353	0.401	0.465	0.588
	Evaluate	8	0.120	0.295	0.350	0.427	0.529
	Reason Scientifically	18	0.153	0.318	0.400	0.477	0.569
5	Investigate	8	0.303	0.337	0.366	0.453	0.590
	Evaluate	19	0.234	0.311	0.392	0.499	0.606
	Reason Scientifically	13	0.223	0.362	0.379	0.465	0.740
6	Investigate	8	0.219	0.310	0.333	0.430	0.515
	Evaluate	15	0.229	0.343	0.379	0.469	0.572
	Reason Scientifically	18	0.171	0.253	0.361	0.427	0.497
7	Investigate	7	0.040	0.139	0.265	0.294	0.618
	Evaluate	8	0.231	0.302	0.411	0.497	0.520
	Reason Scientifically	22	0.177	0.310	0.435	0.506	0.598
8	Investigate	14	0.283	0.382	0.415	0.502	0.545
	Evaluate	12	0.298	0.325	0.374	0.471	0.538
	Reason Scientifically	15	0.062	0.295	0.348	0.436	0.664

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

Table B.5.1

Statistically Flagged Operational Items: Spring 2021 Operational Science Tests

Grade	Item Type	*N* of OP Items	*N* of Items Flagged for *P*-Value	*N* of Items Flagged for Point-Biserial Correlation	*N* of Items Flagged for DIF*	*N* of Items Flagged for Omitting
3	CR	3	1	0	0	0
	ER	1	1	0	0	0
	MC	21	0	2	1	0
	MS	4	1	1	0	0
	TPD	7	3	1	0	0
	TPI	3	0	0	0	0
4	CR	3	3	0	0	1
	ER	1	0	0	1	0
	MC	22	1	4	0	0
	MS	3	1	0	0	0
	TPD	6	0	0	0	0
	TPI	6	1	0	0	0
5	CR	3	0	0	0	0
	ER	1	1	0	0	0
	MC	14	0	0	0	0
	MS	1	1	0	0	0
	TEI	13	2	0	1	0
	TPD	4	0	0	0	0
	TPI	4	0	0	0	0

Table B.5.1 (Continued)

Grade	Item Type	N of OP Items	N of Items Flagged for P-Value	N of Items Flagged for Point-Biserial Correlation	N of Items Flagged for DIF*	N of Items Flagged for Omitting
6	CR	3	3	0	0	0
	ER**	1	1	0	0	0
	MC	16	0	2	0	0
	MS	3	0	0	0	0
	TEI	10	3	0	2	0
	TPD	6	1	0	0	0
	TPI	2	0	0	0	0
7	CR	3	1	0	0	0
	ER**	1	1	0	0	0
	MC	11	0	2	1	0
	MS	6	2	0	1	0
	TEI	14	2	1	2	0
	TPD	4	0	0	0	0
	TPI	2	0	0	0	0
8	CR	3	3	0	0	0
	ER**	1	0	0	0	0
	MC	14	1	1	0	0
	MS	2	0	0	0	0
	TEI	14	1	0	3	0
	TPD	4	0	0	0	0
	TPI	3	0	0	0	0

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

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

Appendix C: Dimensionality

Dimensionality Reports

Science

Contents
Table C.1.1 Reporting Category Intercorrelation Coefficients for Spring 2021 Operational Science Tests
Table C.2.1 First and Second Eigenvalues: Spring 2021 Operational Science Tests
Figure C.1.1 Principal Component Analysis Plot: Spring 2021 Operational Science Grades 3 and 4 Tests
Figure C.1.2 Principal Component Analysis Plot: Spring 2021 Operational Science Grades 5 through 8 Tests

- Because the spring 2021 tests were administered during the 2021 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table C.1.1

Reporting Category Intercorrelation Coefficients for Spring 2021 Operational Science Tests

Grade	Reporting Category	Investigate	Evaluate	Reason Scientifically
3	Investigate	1.00		
	Evaluate	0.65	1.00	
	Reason Scientifically	0.52	0.62	1.00
4	Investigate	1.00		
	Evaluate	0.57	1.00	
	Reason Scientifically	0.69	0.60	1.00
5	Investigate	1.00		
	Evaluate	0.67	1.00	
	Reason Scientifically	0.67	0.73	1.00
6	Investigate	1.00		
	Evaluate	0.61	1.00	
	Reason Scientifically	0.59	0.69	1.00
7	Investigate	1.00		
	Evaluate	0.35	1.00	
	Reason Scientifically	0.45	0.69	1.00
8	Investigate	1.00		
	Evaluate	0.68	1.00	
	Reason Scientifically	0.71	0.68	1.00

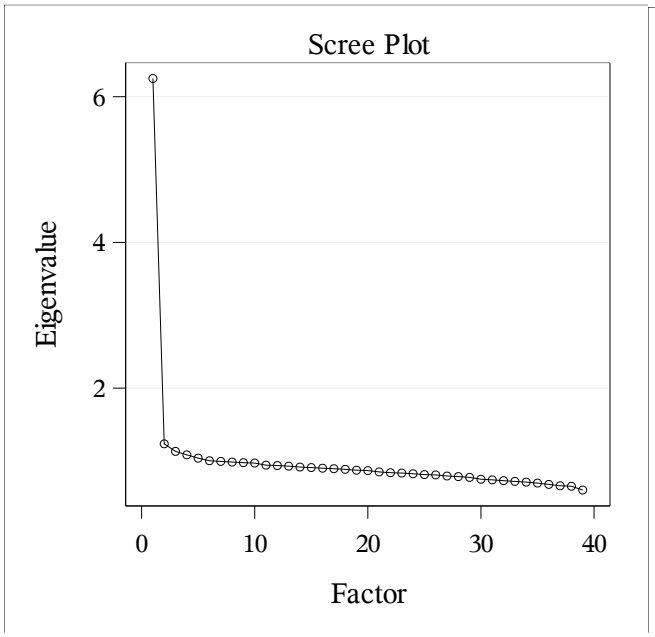
Table C.2.1

First and Second Eigenvalue by Grade: Spring 2021 Operational Science Tests

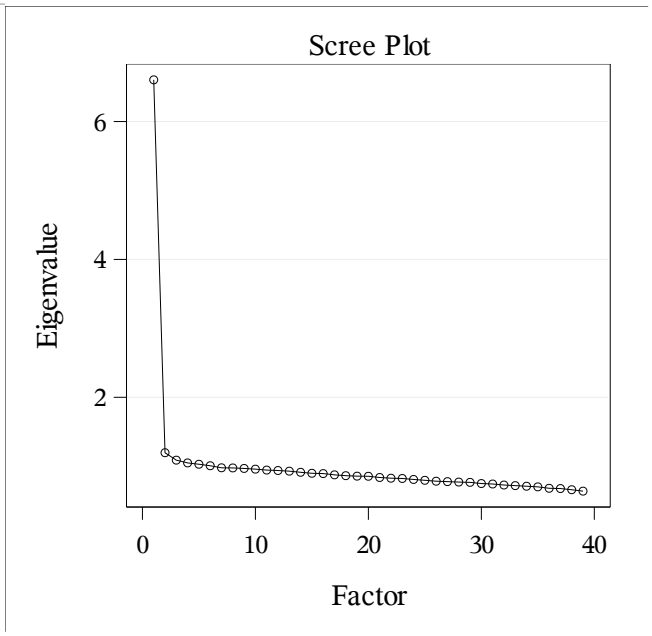
Grade	Form	First Eigenvalue	Second Eigenvalue	Ratio
3	Online	6.250	1.236	5.057
	Paper	6.605	1.196	5.523
4	Online	6.846	1.287	5.319
	Paper	7.009	1.300	5.392
5	Online	7.418	1.221	6.075
6	Online	6.607	1.205	5.483
7	Online	7.461	1.580	4.722
8	Online	7.465	1.231	6.064

Figure C.1.1

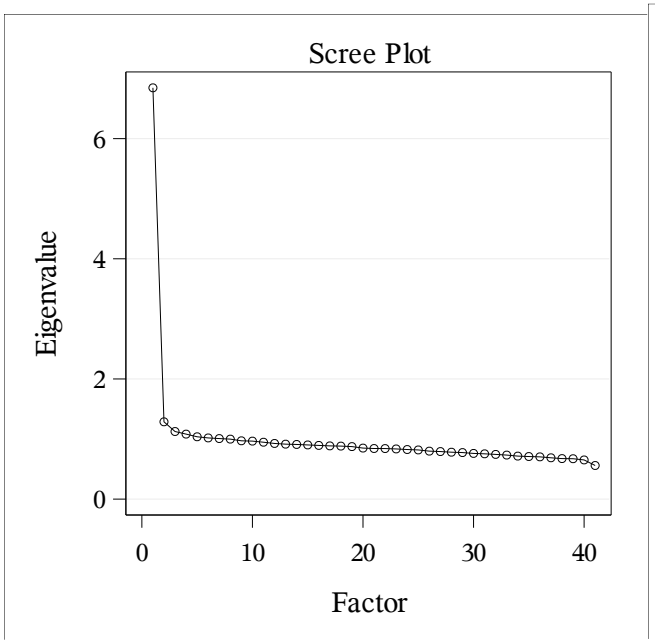
Principal Component Analysis Plot: Spring 2021 Operational Science Grades 3 and 4 Tests



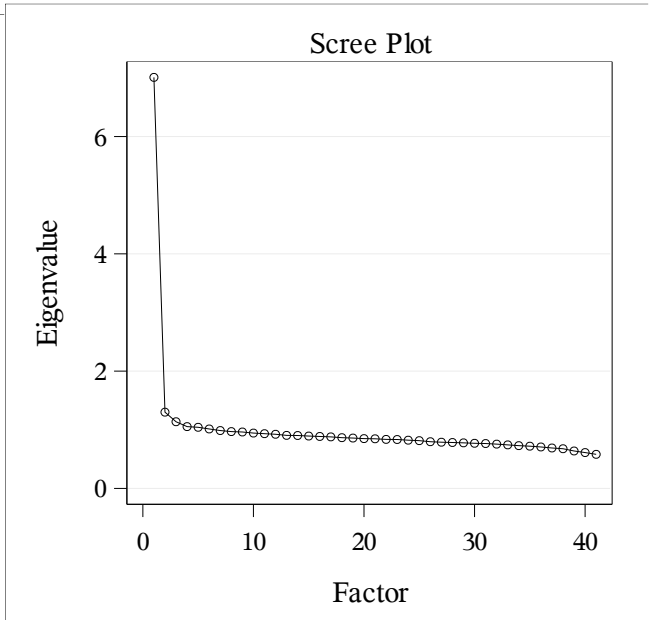
LEAP Science Online: Grade 3



LEAP Science Paper: Grade 3



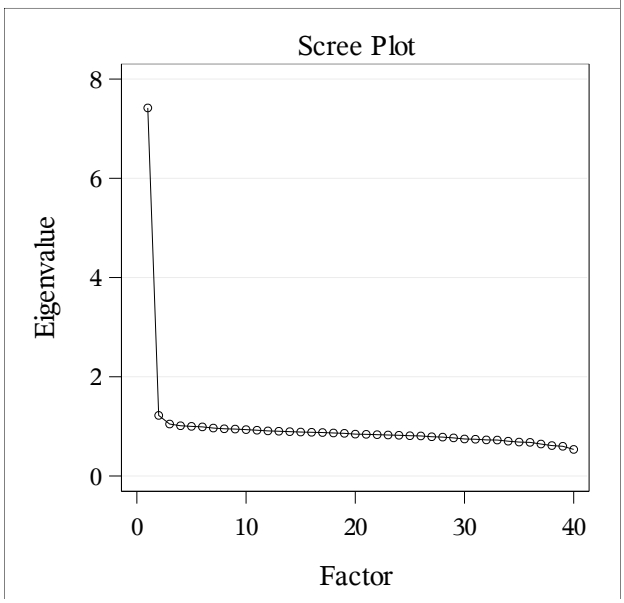
LEAP Science Online: Grade 4



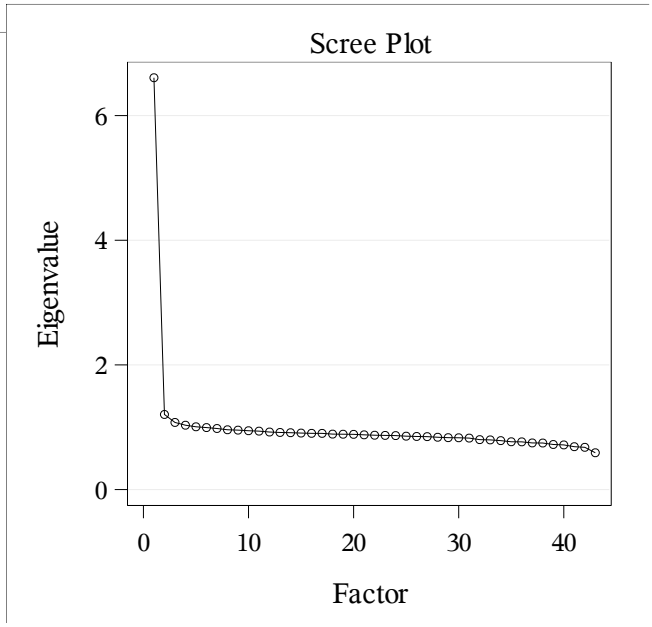
LEAP Science Paper: Grade 4

Figure C.1.2

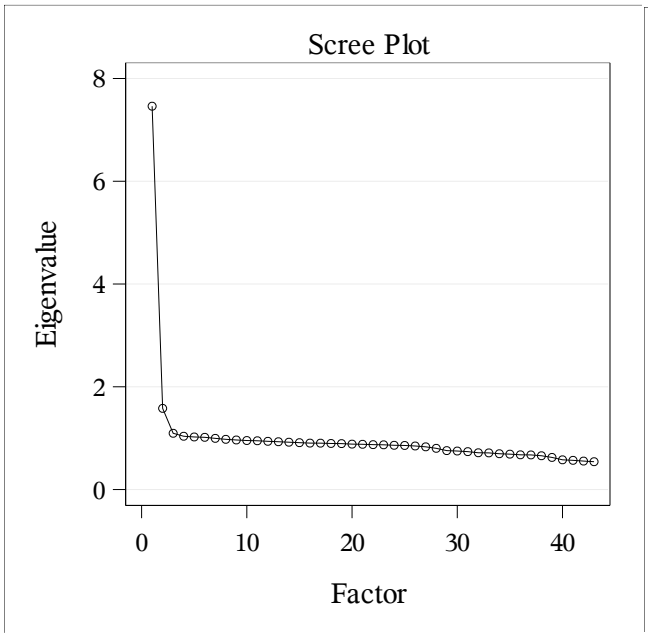
Principal Component Analysis Plot: Spring 2021 Operational Science Grades 5 through 8 Tests



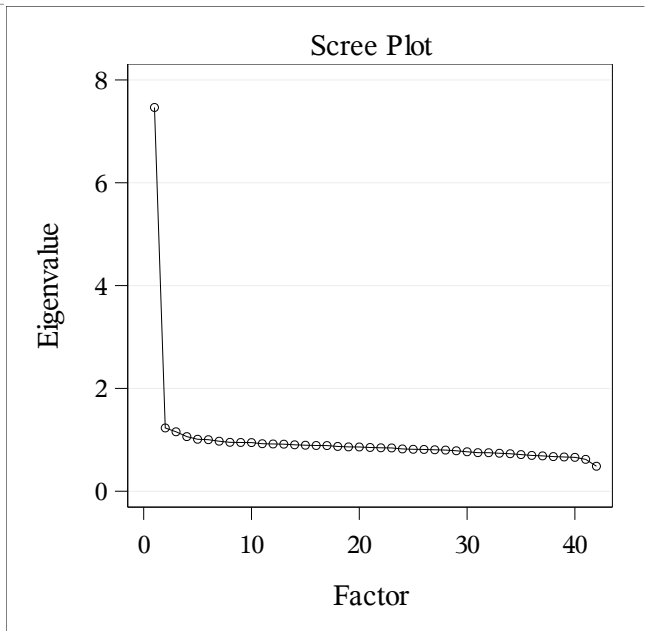
LEAP Social Science Online: Grade 5



LEAP Social Science Online: Grade 6



LEAP Social Science Online: Grade 7



LEAP Social Science Online: Grade 8

Appendix D: Scale Distribution and Statistical Report

Contents
Figure D.1 Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Tests
Figure D.2 Frequency Distribution of Scale Scores for Spring 2021 Operational Science Tests

- Because the spring 2021 tests were administered during the 2021 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table D.1.1 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 3 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

N	≥49560		
Mean	721.62	Median	722.00
Std deviation	31.23	Variance	975.51
Skewness	-0.1197	Kurtosis	-0.2634
Mode	693.00	Std Error Mean	0.1403
Range	182.00	Interquartile Range	47.00

Quantile	Estimate
100% Max	832
99%	787
95%	773
90%	762
75% Q3	745
50% Median	722
25% Q1	698
10%	679
5%	669
1%	650
0% Min	650

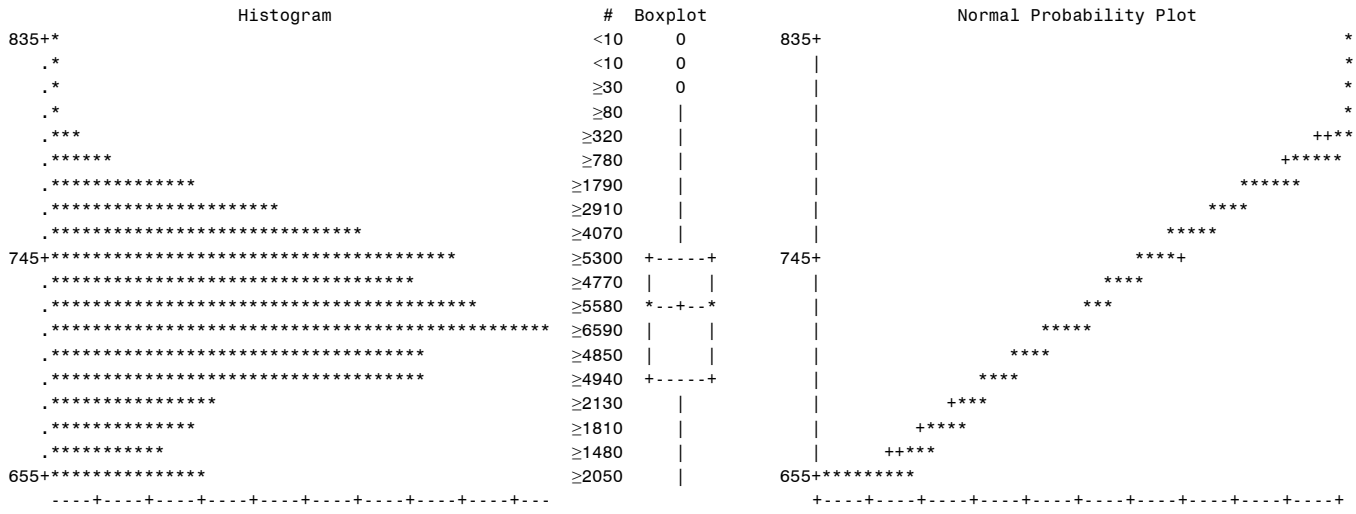


Table D.2.1 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 3 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 03

SCALE_SCORE	Freq	Cum. Freq	Percent	Cum. Percent
650	≥1050	≥1050	2.14	2.14
654	≥1000	≥2050	2.02	4.15
669	≥1480	≥3540	2.99	7.14
679	≥1810	≥5350	3.66	10.80
687	≥2130	≥7480	4.30	15.10
693	≥2470	≥9960	4.99	20.10
698	≥2470	≥12430	4.98	25.08
703	≥2470	≥14900	4.99	30.07
708	≥2380	≥17280	4.80	34.87
712	≥2240	≥19520	4.52	39.40
715	≥2240	≥21770	4.53	43.93
719	≥2110	≥23880	4.26	48.18
722	≥1980	≥25860	4.01	52.19
725	≥1830	≥27690	3.70	55.89
728	≥1760	≥29460	3.55	59.44
731	≥1690	≥31150	3.41	62.86
734	≥1570	≥32720	3.18	66.04
737	≥1500	≥34230	3.03	69.07
740	≥1450	≥35680	2.93	72.00
742	≥1370	≥37060	2.78	74.78
745	≥1200	≥38270	2.44	77.22
747	≥1260	≥39530	2.55	79.77
750	≥1050	≥40580	2.12	81.89
753	≥1080	≥41660	2.18	84.07
755	≥1010	≥42680	2.05	86.12
758	≥920	≥43600	1.87	87.99
760	≥810	≥44420	1.65	89.64
762	≥720	≥45150	1.47	91.11
765	≥720	≥45870	1.46	92.56
768	≥640	≥46510	1.30	93.86
770	≥550	≥47070	1.11	94.98
773	≥470	≥47540	0.96	95.93
775	≥410	≥47960	0.84	96.78
778	≥350	≥48310	0.71	97.49
781	≥310	≥48620	0.64	98.12
784	≥280	≥48910	0.58	98.70
787	≥180	≥49100	0.37	99.08
791	≥120	≥49230	0.26	99.34
794	≥90	≥49330	0.20	99.54
798	≥90	≥49420	0.20	99.73
802	≥40	≥49470	0.09	99.82
807	≥40	≥49510	0.08	99.91
812	≥20	≥49540	0.06	99.96
818	<10	≥49550	0.02	99.98
824	<10	≥49550	0.01	100.00
832	<10	≥49560	0.00	100.00

400 800 1200 1600 2000 2400

Frequency

Table D.1.2 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 4 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

N	≥49540		
Mean	729.25	Median	731.00
Std deviation	31.28	Variance	978.32
Skewness	-0.1015	Kurtosis	-0.2400
Mode	734.00	Std Error Mean	0.1405
Range	200.00	Interquartile Range	44.00

Quantile	Estimate
100% Max	850
99%	798
95%	779
90%	770
75% Q3	751
50% Median	731
25% Q1	707
10%	690
5%	678
1%	651
0% Min	650

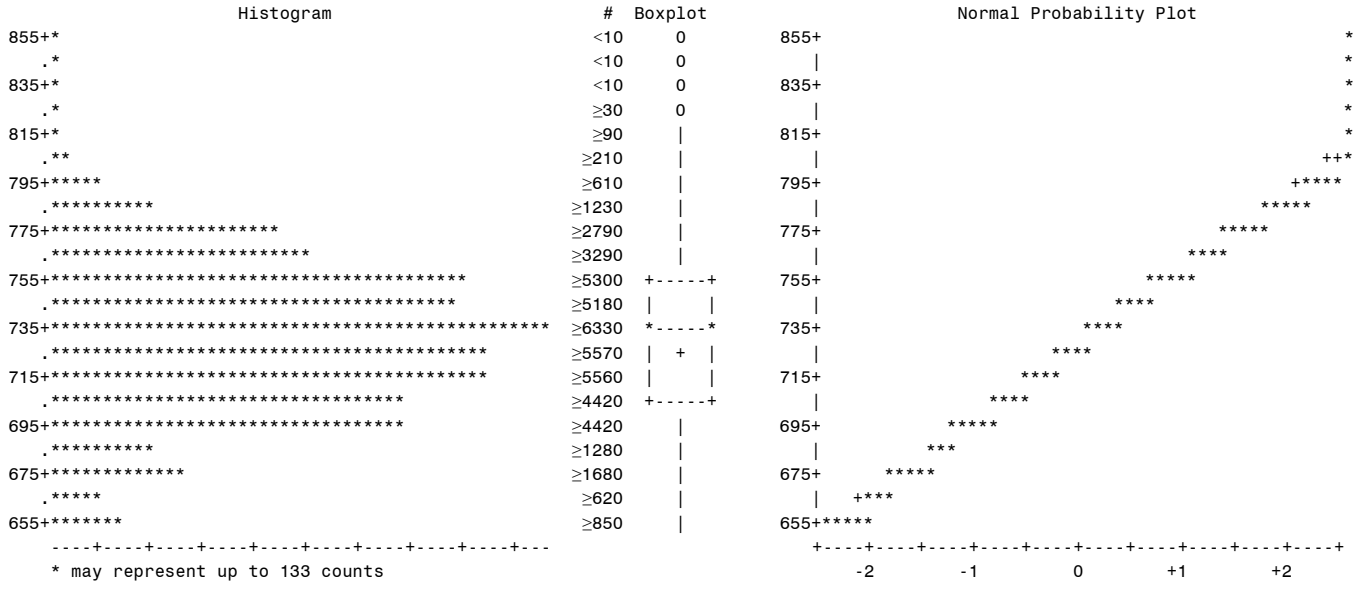


Table D.2.2 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 4 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 04

SCALE_SCORE Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥690	≥690	1.40	1.40
654		<10	≥690	0.00	1.40
658	***	≥150	≥850	0.32	1.72
662	********	≥380	≥1240	0.78	2.51
666		<10	≥1240	0.00	2.51
670	**********	≥790	≥2030	1.61	4.12
674		<10	≥2030	0.00	4.12
678	**********	≥1120	≥3160	2.27	6.38
682	*********	≥440	≥3600	0.90	7.28
686	**********	≥830	≥4440	1.68	8.97
690	**********	≥1540	≥5990	3.12	12.09
694	**********	≥1720	≥7710	3.48	15.57
698	**********	≥1140	≥8860	2.32	17.89
702	**********	≥1790	≥10660	3.63	21.52
706	**********	≥1920	≥12580	3.88	25.40
710	**********	≥1860	≥14440	3.77	29.16
714	**********	≥1860	≥16310	3.76	32.92
718	**********	≥2530	≥18850	5.12	38.05
722	**********	≥1830	≥20680	3.71	41.76
726	**********	≥3080	≥23770	6.23	47.99
730	**********	≥2390	≥26170	4.84	52.83
734	**********	≥1770	≥27950	3.58	56.42
738	**********	≥2810	≥30760	5.67	62.09
742	**********	≥2190	≥32950	4.43	66.52
746	**********	≥1500	≥34460	3.05	69.56
750	**********	≥2920	≥37390	5.91	75.47
754	**********	≥1310	≥38700	2.65	78.12
758	**********	≥2540	≥41250	5.14	83.26
762	**********	≥1100	≥42350	2.23	85.49
766	**********	≥1890	≥44250	3.83	89.32
770	**********	≥1150	≥45410	2.33	91.65
774	**********	≥670	≥46080	1.35	93.00
778	**********	≥1250	≥47330	2.54	95.55
782	**********	≥500	≥47840	1.01	96.56
786	*********	≥420	≥48260	0.86	97.42
790	*******	≥370	≥48640	0.75	98.17
794	*****	≥240	≥48880	0.50	98.67
798	******	≥300	≥49190	0.61	99.29
802	**	≥120	≥49310	0.24	99.53
806	*	≥20	≥49340	0.06	99.58
810	**	≥90	≥49430	0.18	99.77
814	*	≥50	≥49480	0.11	99.88
818		≥20	≥49500	0.04	99.92
822		≥20	≥49520	0.04	99.96
826		<10	≥49530	0.01	99.97
830		<10	≥49540	0.02	99.99
834		<10	≥49540	0.00	99.99
838		<10	≥49540	0.00	100.00
842		<10	≥49540	0.00	100.00
846		<10	≥49540	0.00	100.00
850		<10	≥49540	0.00	100.00

400 800 1200 1600 2000 2400 2800

Frequency

Table D.1.3 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 5 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 05

N	≥49860		
Mean	727.87	Median	729.00
Std deviation	35.88	Variance	1287.65
Skewness	0.0258	Kurtosis	-0.3195
Mode	698.00	Std Error Mean	0.1607
Range	200.00	Interquartile Range	49.00

Quantile	Estimate
100% Max	850
99%	807
95%	785
90%	773
75% Q3	752
50% Median	729
25% Q1	703
10%	677
5%	671
1%	650
0% Min	650

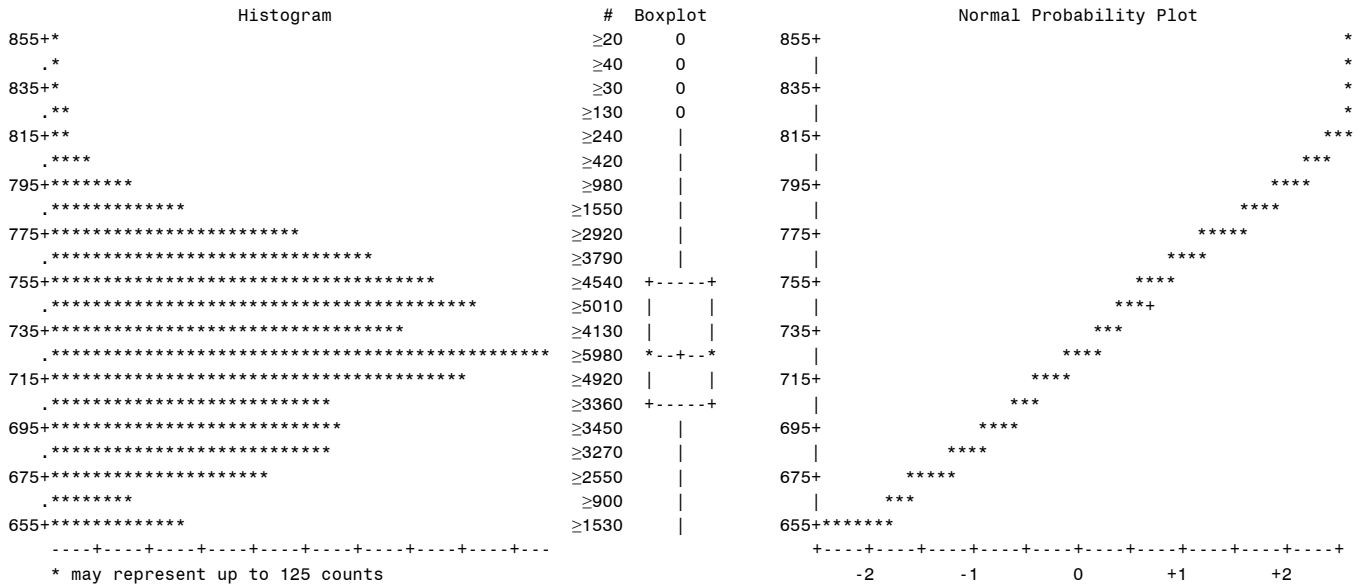


Table D.2.3 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 5 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 05

SCALE_SCORE	Freq	Cum. Freq	Percent	Cum. Percent
650	≥870	≥870	1.75	1.75
654	≥660	≥1530	1.33	3.08
663	≥900	≥2440	1.81	4.90
671	≥1140	≥3590	2.30	7.20
677	≥1400	≥4990	2.82	10.02
684	≥1580	≥6570	3.18	13.19
689	≥1680	≥8260	3.38	16.58
694	≥1720	≥9980	3.45	20.03
698	≥1730	≥11720	3.48	23.51
703	≥1690	≥13410	3.39	26.90
707	≥1670	≥15080	3.36	30.25
710	≥1660	≥16750	3.34	33.60
714	≥1620	≥18370	3.26	36.86
717	≥1630	≥20010	3.28	40.14
720	≥1660	≥21670	3.33	43.46
723	≥1480	≥23150	2.97	46.44
726	≥1420	≥24580	2.86	49.30
729	≥1410	≥25990	2.84	52.14
732	≥1370	≥27370	2.76	54.90
735	≥1400	≥28770	2.81	57.71
737	≥1350	≥30130	2.73	60.44
740	≥1260	≥31400	2.54	62.98
742	≥1270	≥32670	2.55	65.53
745	≥1250	≥33920	2.52	68.05
747	≥1220	≥35150	2.46	70.50
750	≥1200	≥36360	2.42	72.93
752	≥1160	≥37520	2.33	75.26
755	≥1060	≥38590	2.14	77.40
757	≥1100	≥39690	2.21	79.61
760	≥1020	≥40720	2.06	81.67
762	≥970	≥41690	1.96	83.63
765	≥930	≥42630	1.87	85.50
768	≥860	≥43490	1.72	87.22
770	≥840	≥44330	1.68	88.91
773	≥760	≥45090	1.53	90.44
776	≥700	≥45790	1.41	91.84
779	≥620	≥46410	1.25	93.09
782	≥590	≥47000	1.19	94.28
785	≥490	≥47500	0.99	95.27
788	≥460	≥47960	0.93	96.20
792	≥360	≥48330	0.73	96.93
795	≥340	≥48670	0.69	97.62
799	≥280	≥48950	0.56	98.18
803	≥230	≥49180	0.46	98.65
807	≥190	≥49380	0.39	99.04
812	≥130	≥49510	0.26	99.30
817	≥110	≥49620	0.22	99.52
822	≥70	≥49690	0.15	99.68
828	≥60	≥49760	0.12	99.80
834	≥30	≥49790	0.07	99.87
841	≥20	≥49810	0.04	99.91
849	≥10	≥49830	0.04	99.95
850	≥20	≥49860	0.05	100.00

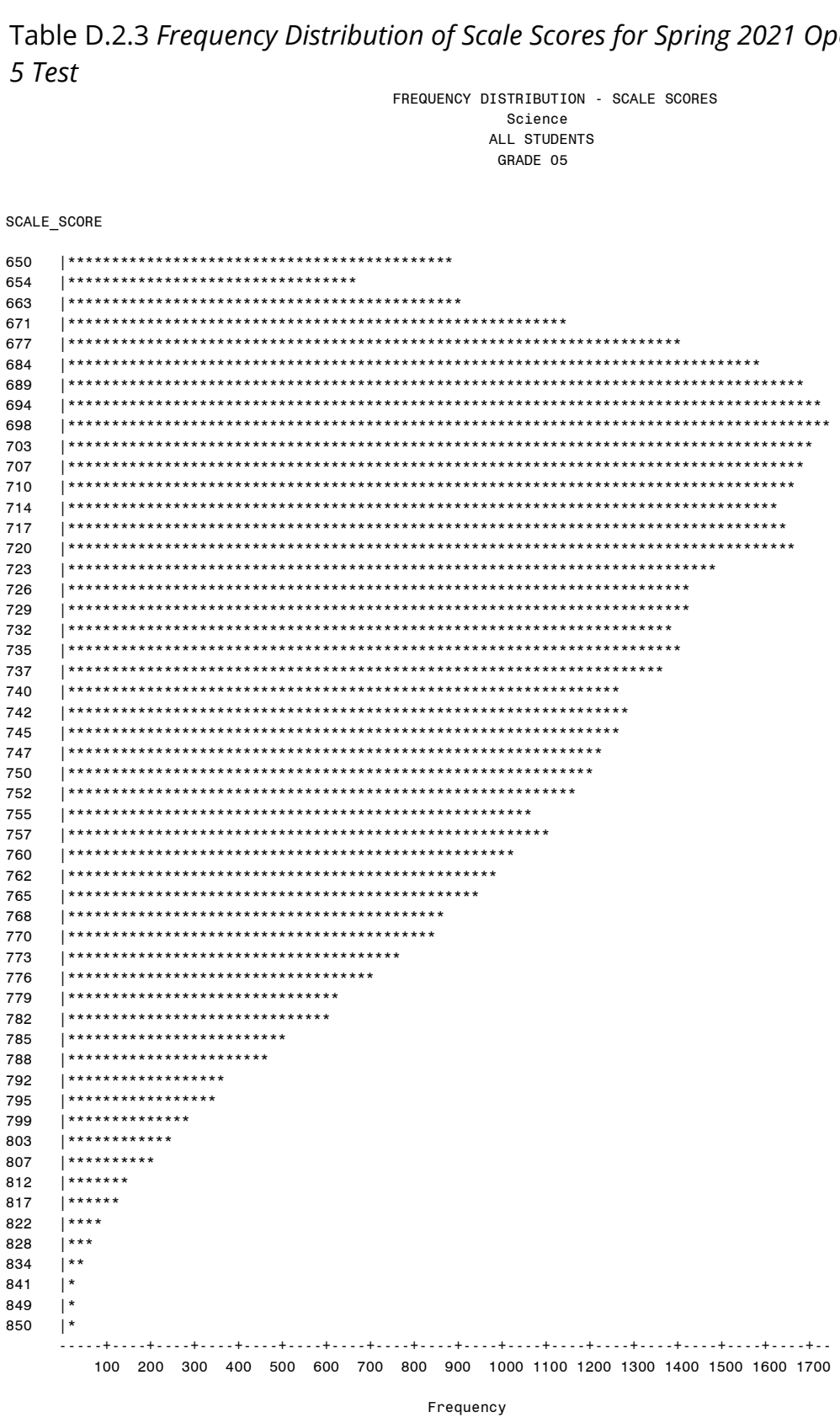


Table D.1.4 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 6 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

N	≥51540		
Mean	726.19	Median	725.00
Std deviation	30.92	Variance	955.84
Skewness	-0.1184	Kurtosis	-0.2308
Mode	704.00	Std Error Mean	0.1362
Range	200.00	Interquartile Range	45.00

Quantile	Estimate
100% Max	850
99%	794
95%	776
90%	767
75% Q3	749
50% Median	725
25% Q1	704
10%	683
5%	676
1%	650
0% Min	650

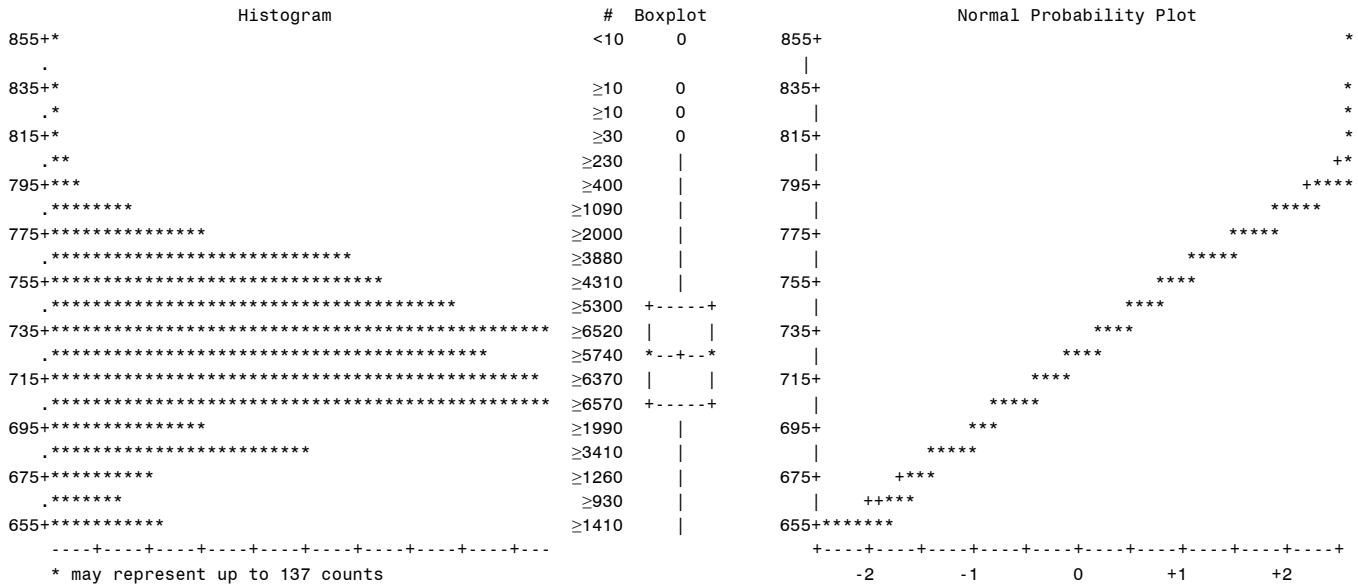


Table D.2.4 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 6 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 06

SCALE_SCORE		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥750	≥750	1.46	1.46
657	**********	≥660	≥1410	1.28	2.74
667	**********	≥930	≥2340	1.81	4.56
676	**********	≥1260	≥3610	2.46	7.02
683	**********	≥1570	≥5180	3.05	10.07
689	**********	≥1840	≥7020	3.57	13.64
695	**********	≥1990	≥9020	3.87	17.51
700	**********	≥2100	≥11120	4.07	21.58
704	**********	≥2250	≥13380	4.38	25.96
709	**********	≥2210	≥15590	4.30	30.26
712	**********	≥2150	≥17750	4.19	34.45
716	**********	≥2110	≥19860	4.10	38.55
719	**********	≥2100	≥21970	4.08	42.63
722	**********	≥2010	≥23990	3.91	46.55
725	**********	≥1890	≥25880	3.67	50.22
728	**********	≥1830	≥27720	3.56	53.78
731	**********	≥1740	≥29460	3.38	57.16
734	**********	≥1630	≥31090	3.17	60.33
736	**********	≥1560	≥32650	3.03	63.36
739	**********	≥1580	≥34240	3.08	66.44
742	**********	≥1410	≥35650	2.74	69.17
744	**********	≥1370	≥37020	2.67	71.84
746	**********	≥1310	≥38340	2.55	74.39
749	**********	≥1210	≥39550	2.35	76.74
751	**********	≥1210	≥40760	2.35	79.09
753	**********	≥1090	≥41850	2.12	81.21
756	**********	≥1030	≥42890	2.01	83.22
758	**********	≥970	≥43860	1.88	85.10
760	**********	≥860	≥44720	1.68	86.78
763	**********	≥790	≥45520	1.54	88.32
765	**********	≥790	≥46320	1.54	89.87
767	**********	≥730	≥47050	1.43	91.29
769	**********	≥690	≥47740	1.34	92.64
772	**********	≥590	≥48340	1.16	93.80
774	**********	≥560	≥48910	1.10	94.90
776	*********	≥450	≥49370	0.89	95.79
779	********	≥380	≥49750	0.74	96.53
781	*******	≥350	≥50100	0.69	97.22
783	******	≥300	≥50410	0.59	97.80
786	*****	≥240	≥50650	0.47	98.27
788	****	≥190	≥50840	0.37	98.65
791	***	≥170	≥51010	0.33	98.98
794	***	≥130	≥51150	0.26	99.24
797	**	≥100	≥51250	0.19	99.43
800	**	≥80	≥51330	0.16	99.60
803	*	≥60	≥51390	0.12	99.72
806	*	≥50	≥51450	0.11	99.83
809	*	≥20	≥51480	0.05	99.88
813		≥20	≥51500	0.04	99.93
817		≥10	≥51510	0.03	99.96
822		<10	≥51520	0.02	99.97
826		<10	≥51530	0.01	99.98
832		<10	≥51530	0.02	99.99
838		<10	≥51540	0.00	100.00
850		<10	≥51540	0.00	100.00

400 800 1200 1600 2000
Frequency

Table D.1.5 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 7 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 07

N	≥52330		
Mean	727.85	Median	726.00
Std deviation	31.85	Variance	1014.64
Skewness	0.1912	Kurtosis	-0.0278
Mode	703.00	Std Error Mean	0.1392
Range	200.00	Interquartile Range	41.00

Quantile	Estimate
100% Max	850
99%	805
95%	783
90%	770
75% Q3	748
50% Median	726
25% Q1	707
10%	690
5%	679
1%	651
0% Min	650

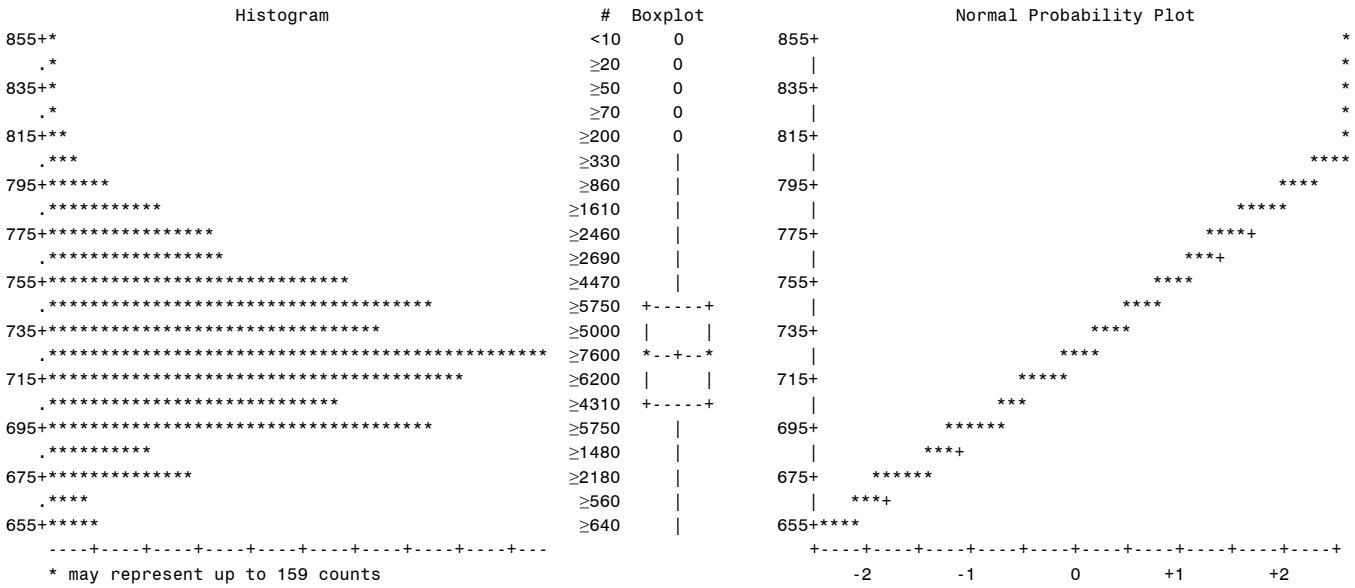


Table D.2.5 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 7 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 07

SCALE_SCORE	Freq	Cum. Freq	Percent	Cum. Percent
650	≥300	≥300	0.57	0.57
651	≥340	≥640	0.66	1.23
662	≥560	≥1210	1.08	2.31
671	≥960	≥2170	1.84	4.15
679	≥1220	≥3390	2.34	6.49
685	≥1480	≥4880	2.84	9.33
690	≥1740	≥6620	3.33	12.67
695	≥1980	≥8600	3.78	16.45
699	≥2020	≥10630	3.88	20.33
703	≥2160	≥12800	4.14	24.46
707	≥2140	≥14940	4.10	28.56
711	≥2090	≥17040	4.00	32.56
714	≥2040	≥19080	3.90	36.46
717	≥2070	≥21150	3.96	40.42
720	≥1980	≥23130	3.79	44.21
723	≥1950	≥25080	3.73	47.94
726	≥1840	≥26930	3.52	51.46
729	≥1830	≥28760	3.50	54.96
732	≥1780	≥30540	3.41	58.37
735	≥1630	≥32180	3.12	61.49
737	≥1580	≥33760	3.02	64.52
740	≥1540	≥35310	2.95	67.47
743	≥1440	≥36750	2.75	70.23
746	≥1420	≥38170	2.72	72.95
748	≥1340	≥39520	2.57	75.52
751	≥1260	≥40780	2.41	77.94
754	≥1160	≥41940	2.22	80.16
756	≥1040	≥42990	1.99	82.15
759	≥1000	≥43990	1.91	84.07
762	≥1040	≥45030	2.00	86.06
764	≥890	≥45930	1.70	87.77
767	≥750	≥46680	1.44	89.21
770	≥690	≥47380	1.33	90.54
772	≥640	≥48020	1.23	91.77
775	≥600	≥48620	1.15	92.92
778	≥520	≥49150	1.00	93.92
780	≥490	≥49640	0.94	94.85
783	≥400	≥50040	0.77	95.63
786	≥400	≥50440	0.77	96.39
788	≥320	≥50760	0.62	97.01
791	≥250	≥51020	0.49	97.50
794	≥240	≥51270	0.47	97.97
797	≥200	≥51470	0.40	98.36
799	≥150	≥51630	0.29	98.66
802	≥120	≥51760	0.25	98.91
805	≥110	≥51870	0.22	99.12
809	≥90	≥51960	0.17	99.30
812	≥80	≥52050	0.16	99.46
815	≥70	≥52120	0.13	99.59
819	≥40	≥52160	0.09	99.68
822	≥40	≥52210	0.08	99.76
826	≥30	≥52240	0.06	99.83
831	≥20	≥52260	0.05	99.88
835	≥30	≥52290	0.06	99.93
841	≥10	≥52310	0.03	99.96
847	≥10	≥52320	0.02	99.98
850	<10	≥52330	0.02	100.00

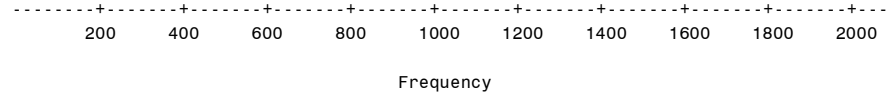


Table D.1.6 *Scale Score Descriptive Statistics and Plots for Spring 2021 Operational Science Grade 8 Test*

DESCRIPTIVE STATISTICS - SCALE SCORES
Science
ALL STUDENTS
GRADE 08

N	≥51850		
Mean	728.78	Median	729.00
Std deviation	31.51	Variance	993.11
Skewness	-0.0719	Kurtosis	-0.1789
Mode	708.00	Std Error Mean	0.1384
Range	200.00	Interquartile Range	42.00

Quantile	Estimate
100% Max	850
99%	799
95%	778
90%	768
75% Q3	750
50% Median	729
25% Q1	708
10%	687
5%	675
1%	650
0% Min	650

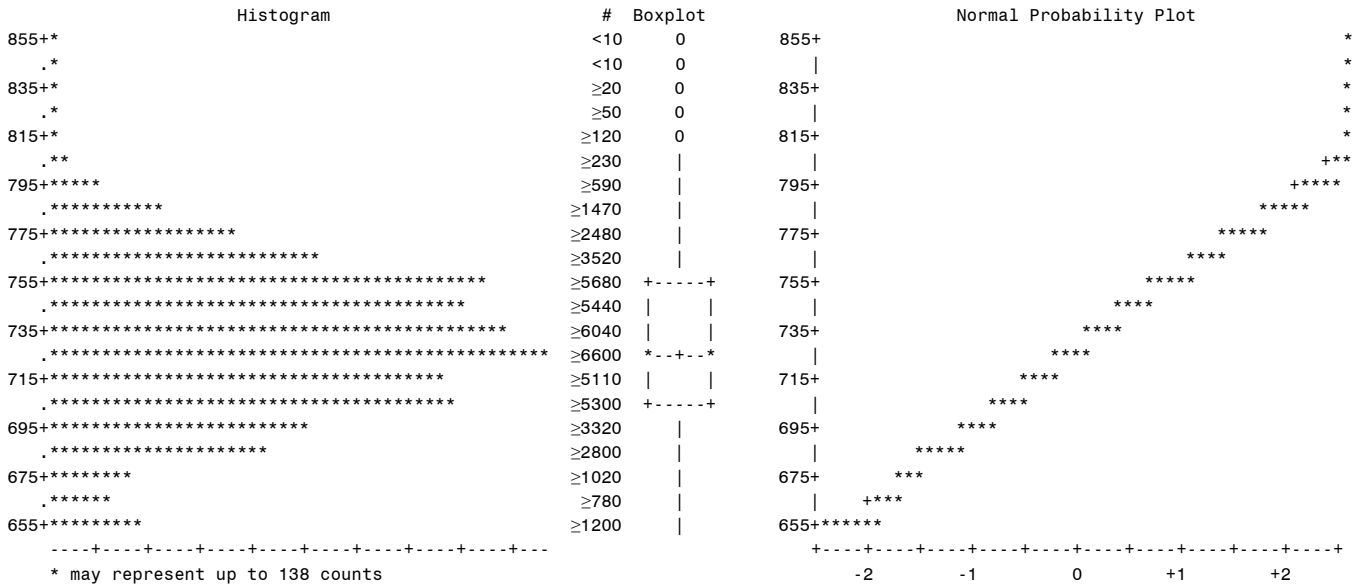


Table D.2.6 *Frequency Distribution of Scale Scores for Spring 2021 Operational Science Grade 8 Test*

FREQUENCY DISTRIBUTION - SCALE SCORES
Science
ALL STUDENTS
GRADE 08

SCALE_SCORE		Freq	Cum. Freq	Percent	Cum. Percent
650	************************	≥660	≥660	1.27	1.27
658	**********************	≥540	≥1200	1.05	2.32
668	******************************	≥780	≥1980	1.51	3.84
675	**************************************	≥1020	≥3010	1.98	5.82
682	**	≥1280	≥4300	2.47	8.29
687	**	≥1510	≥5810	2.93	11.22
692	**	≥1650	≥7470	3.19	14.41
697	**	≥1670	≥9140	3.23	17.64
701	**	≥1750	≥10900	3.38	21.03
705	**	≥1730	≥12630	3.34	24.37
708	**	≥1810	≥14440	3.49	27.86
712	**	≥1790	≥16240	3.47	31.33
715	**	≥1670	≥17910	3.22	34.55
718	**	≥1640	≥19560	3.18	37.73
721	**	≥1670	≥21230	3.23	40.96
723	**	≥1630	≥22870	3.16	44.11
726	**	≥1730	≥24600	3.34	47.45
729	**	≥1560	≥26160	3.01	50.46
731	**	≥1530	≥27700	2.97	53.43
733	**	≥1500	≥29200	2.90	56.33
736	**	≥1510	≥30720	2.93	59.25
738	**	≥1490	≥32210	2.87	62.13
741	**	≥1420	≥33630	2.75	64.88
743	**	≥1390	≥35030	2.69	67.57
745	**	≥1280	≥36310	2.48	70.04
747	**	≥1330	≥37650	2.58	72.62
750	**	≥1260	≥38920	2.44	75.06
752	**	≥1140	≥40060	2.21	77.27
754	**	≥1050	≥41120	2.04	79.31
756	**	≥1140	≥42260	2.20	81.51
759	**	≥1070	≥43330	2.07	83.58
761	**********************************	≥930	≥44270	1.81	85.38
763	**********************************	≥900	≥45170	1.75	87.13
766	**********************************	≥890	≥46070	1.73	88.86
768	******************************	≥780	≥46850	1.51	90.37
771	**************************	≥690	≥47550	1.35	91.72
773	************************	≥640	≥48200	1.25	92.96
776	************************	≥610	≥48810	1.18	94.15
778	**********************	≥520	≥49330	1.01	95.16
781	******************	≥450	≥49780	0.87	96.03
784	**********	≥380	≥50170	0.75	96.78
786	**********	≥330	≥50510	0.65	97.43
789	**********	≥290	≥50810	0.57	98.00
792	*********	≥220	≥51040	0.44	98.44
796	********	≥200	≥51240	0.39	98.82
799	*******	≥160	≥51400	0.32	99.15
803	*****	≥130	≥51540	0.26	99.40
806	****	≥90	≥51630	0.19	99.59
811	***	≥70	≥51710	0.15	99.74
815	**	≥40	≥51760	0.09	99.83
820	*	≥30	≥51790	0.06	99.90
825	*	≥10	≥51810	0.03	99.93
831	*	≥20	≥51830	0.04	99.97
838		<10	≥51840	0.01	99.98
846		<10	≥51840	0.01	99.99
850		<10	≥51850	0.01	100.00

200 400 600 800 1000 1200 1400 1600 1800
Frequency

Appendix E: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

Science

Contents
Table E.1 Reliability for Overall and Subgroups: Spring 2021 Operational Science Tests
Table E.2 Cronbach's Alpha Reliability: Spring 2021 Operational Science Tests
Table E.3.1 Classification Accuracy and Decision Consistency: Spring 2021 Operational Science Tests

- Because the spring 2021 tests were administered during the 2021 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table E.1

Reliability for Overall and Subgroups: Spring 2021 Operational Science Tests

Subgroup	3	4	5	6	7	8
All Students	0.859	0.866	0.857	0.851	0.869	0.873
Female	0.851	0.854	0.846	0.834	0.858	0.863
Male	0.867	0.876	0.867	0.864	0.878	0.881
African American	0.808	0.808	0.817	0.792	0.810	0.821
American Indian or Alaska Native	0.840	0.829	0.808	0.826	0.858	0.862
Asian	0.870	0.882	0.860	0.866	0.902	0.886
Hispanic/Latino	0.842	0.850	0.847	0.840	0.865	0.870
Multi-Racial	0.850	0.862	0.839	0.842	0.865	0.865
Native Hawaiian or Other Pacific Islander	0.814	0.887	N/A	0.821	0.907	0.861
White	0.854	0.857	0.840	0.841	0.865	0.860
Economically Disadvantaged: No	0.853	0.858	0.832	0.840	0.866	0.859
Economically Disadvantaged: Yes	0.832	0.841	0.838	0.825	0.842	0.852
EL: No	0.860	0.867	0.856	0.850	0.868	0.871
EL: Yes	0.739	0.749	0.788	0.731	0.775	0.781
Regular Education	0.858	0.864	0.850	0.845	0.865	0.868
Special Education	0.840	0.837	0.840	0.800	0.817	0.815
Section 504: No	0.860	0.867	0.857	0.851	0.869	0.873
Section 504: Yes	0.841	0.843	0.840	0.823	0.839	0.853

* N/A means no estimate is calculated since their *n* count is smaller than 30.

Table E.2

Cronbach's Alpha Reliability: Spring 2021 Operational Science Tests

Grade	Cronbach's Alpha
3	0.859
4	0.866
5	0.857
6	0.851
7	0.869
8	0.873

Table E.3

Classification Accuracy and Decision Consistency: Spring 2021 Operational Science Tests

Table E.3.1

Estimates of Accuracy and Consistency of Achievement-Level Classification

Grade	Accuracy	Consistency	PChance	Kappa
3	0.655	0.550	0.242	0.407
4	0.668	0.562	0.231	0.430
5	0.653	0.543	0.222	0.412
6	0.663	0.556	0.245	0.412
7	0.687	0.581	0.239	0.449
8	0.694	0.587	0.246	0.452

Table E.3.2

Accuracy of Classification at Each Achievement Level

Grade	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
3	0.810	0.701	0.595	0.517	0.618
4	0.837	0.605	0.622	0.643	0.658
5	0.822	0.637	0.552	0.636	0.672
6	0.814	0.608	0.614	0.651	*
7	0.803	0.635	0.615	0.740	0.755
8	0.809	0.713	0.629	0.685	0.665

* It was inestimable due to restricted sample size.

Table E.3.3

Accuracy of Dichotomous Categorizations (PAC Metric)

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.923	0.876	0.899	0.950
4	0.923	0.891	0.897	0.952
5	0.923	0.883	0.892	0.947
6	0.917	0.876	0.891	0.972
7	0.916	0.879	0.908	0.980
8	0.941	0.888	0.898	0.964

Table E.3.4

Consistency of Dichotomous Categorizations (PAC Metric)

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.888	0.829	0.856	0.937
4	0.891	0.848	0.855	0.935
5	0.889	0.837	0.848	0.925
6	0.881	0.829	0.846	0.961
7	0.879	0.832	0.870	0.971
8	0.914	0.844	0.856	0.953

Table E.3.5

Kappa of Dichotomous Categorizations (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	0.658	0.566	0.147
4	0.684	0.690	0.636	0.327
5	0.668	0.673	0.638	0.419
6	0.667	0.656	0.577	0.049
7	0.644	0.664	0.653	0.474
8	0.672	0.684	0.645	0.254

Table E.3.6

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

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.038	0.054	0.053	0.049
4	0.033	0.048	0.056	0.039
5	0.034	0.056	0.056	0.038
6	0.040	0.054	0.057	0.028
7	0.039	0.058	0.052	0.016
8	0.026	0.053	0.052	0.032

Table E.3.7

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

Grade	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.038	0.070	0.048	0.001
4	0.043	0.061	0.047	0.009
5	0.043	0.061	0.052	0.015
6	0.043	0.070	0.052	*
7	0.046	0.063	0.040	0.005
8	0.033	0.059	0.050	0.004

* It was inestimable due to restricted sample size.