LEAP 2025 Biology
Technical Addendum: 2019–2020

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 fall 2019 and summer 2020 administrations of the LEAP 2025 Biology test, which used intact forms based on previously administered operational forms. For information on the development and forms construction processes for these forms, see the 2019 LEAP 2025 Biology Technical Report.

While this technical addendum 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) and in the new edition, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

The chapters of this addendum 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. Due to the COVID-19 pandemic, the spring 2020 administration did not occur. Additionally, because of conditions related to COVID-19, please use caution when making any inferences from the statistical results of the summer 2020 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 Biology. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. Further, the LDOE and the BESE are committed to the development of rigorous assessments as one component of their comprehensive plan—Louisiana Believes—designed to ensure that every Louisiana student is on track to be successful in postsecondary education and the workforce.

The purpose of this technical addendum is to describe the processes for the fall 2019 and the summer 2020 administrations of LEAP 2025 Biology. 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 Educational Assessment Program for High School 2025 (LEAP 2025 HS) strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations.

Training of School Systems

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

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

Ancillary Materials

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

For each test administration, Data Recognition Corporation (DRC) produces an administration manual, the *LEAP 2025 High School Test Administration Manual* (TAM). The TAM provides detailed instructions for administering the LEAP 2025 HS assessments. The manual includes information on test security, test administrator responsibilities, test preparation, administration of online tests, and post-test procedures.

Test Administrators Manual Table of Contents

1. Notes and Reminders
2. Pre-administration Oath and Security Confidentiality Statement
3. Post-administration Oath and Security Confidentiality Statement
4. Overview
5. Test Security
   - 5.1. Secure Test Materials
   - 5.2. Testing Irregularities and Security Breaches
   - 5.3. Testing Environment
   - 5.4. Violations of Test Security
   - 5.5. Voiding Student Tests
6. Test Administrator Responsibilities
   - 6.1. Software Tools and Features for Test Administrators
7. Test Administration Checklists
   - 7.1. Before Testing
   - 7.2. During Testing
   - 7.3. After Testing (Daily)
   - 7.4. After Testing (Last Day)
8. Test Materials
   - 8.1. Receipt of Test Materials
9. Testing Guidelines
   - 9.1. Testing Eligibility
   - 9.2. Testing Schedule

    9.3. LEAP 2025 Testing Time
    9.4. Extended Time for Testing
    9.5. Makeup Test Procedures
    9.6. Testing Conditions
    9.7. Accessibility Features
10. Special Populations and Accommodations
    10.1. IDEA Special Education Students
    10.2. Students with One or More Disabilities According to Section 504
    10.3. Gifted and Talented Special Education Students
    10.4. Test Accommodations for Special Education and Section 504 Students
    10.5. Special Considerations for Students Who Are Deaf or Hearing Impaired
    10.6. English Learners (ELs)
11. Directions for Administering the LEAP 2025 Tests
12. LEAP 2025 Testing Times
13. General Instructions for LEAP 2025
    13.1. LEAP 2025 English I and English II
    13.2. LEAP 2025 Algebra I and Geometry
    13.3. LEAP 2025 Biology
    13.4. LEAP 2025 U.S. History
14. Post-Test Procedures
    14.1. Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    14.2. Returning Test Materials to the School Test Coordinator
15. Index

DRC also produces a Test Coordinator Manual (TCM). The TCM provides detailed instructions for district and school test coordinators' responsibilities for distributing, collecting, and returning test materials.

Test Coordinators Manual Table of Contents

1. Key Dates
2. LEAP 2025 High School Alerts
3. Pre-Administration Oath of Security and Confidentiality Statement
4. Post-Administration Oath of Security and Confidentiality Statement
5. General Information
    5.1. DRC INSIGHT Portal (eDIRECT) and INSIGHT

6. LEAP 2025 High School
   6.1. Testing Requirements
7. Test Security
   7.1. Key Definitions
   7.2. Violations of Test Security
   7.3. Testing Guidelines
   7.4. Testing Conditions
   7.5. Testing Schedule
   7.6. Extended Time for Testing
   7.7. Extended Breaks
   7.8. Makeup Testing
8. LEAP 2025 High School and End-of-Course Testing Times
9. Roles and Responsibilities
   9.1. District Test Coordinator
   9.2. School Test Coordinator
   9.3. Chief Technology Officer
10. Managing Test Tickets
    10.1. Student Transfers
    10.2. Locked Test Tickets
    10.3. Technical Issues
    10.4. Invalidating Test Tickets
11. Resources for Online Testing
    11.1. High School Test Administration Manual
    11.2. DRC INSIGHT Portal User Guide
    11.3. LEAP 2025 Accommodations and Accessibility Manual
    11.4. DRC INSIGHT Technology User Guide
    11.5. Student Tutorials
    11.6. Online Tools Training (OTT)
12. Post-Administration Rescoring Process for LEAP 2025 HS Assessments
13. Request for Rescoring
14. Void Notification

LDOE assessment staff review, provide feedback, and give final approval for the manuals. The manuals are inclusive of LEAP 2025 HS assessments in English Language Arts (ELA), Mathematics, Social Studies, and Science.

The *Standards* contain multiple references relevant to test administration. Information in the TAM 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 TAM provides 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 online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

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

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

Department staff release annual test security reports that describe a wide range of improper activities that may occur during testing, including the following: copying and reviewing test questions with students; cueing students during testing, verbally or with written materials on the classroom walls; cueing students nonverbally, such as by tapping or nodding the head; allowing students to correct or complete answers after tests have been submitted; splitting sessions into two parts; ignoring the standardized directions for the assessment; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the Individualized Education Program (IEP), Individual Accommodation Plan (IAP), or EL Checklist; allowing accommodations for students who do not have an IEP, IAP, or EL Checklist; or defining terms on the test.

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

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

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

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

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

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

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

Time

Each session of each content area test is timed to provide sufficient time for students to attempt all items. The manuals provide examiners with timing guidelines for the assessments.

Online Forms Administration

The online forms are administered via DRC's INSIGHT online assessment system. School system and school personnel set up test sessions via DRC's online testing portal, DRC INSIGHT Portal (eDIRECT), and print test tickets. Students enter 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.

Accessibility and Accommodations

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

- Access for All features are available to all students taking an assessment.

- Accessibility Features are available to students when deemed appropriate by a team of educators.
- Accommodations must appear in a student’s IEP/504/EL plan.

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

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

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

In compliance with this standard, the TAM contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP 2025 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 Accessibility and Accommodations Manual*.

Testing Windows

The 2019–2020 assessments were administered to students within the state testing windows of December 2–18, 2019, and July 13–24, 2020.

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 HS assessments. Test security procedures are discussed throughout the TCM and TAM.

Test coordinators and administrators are instructed to keep all test materials in locked storage, except during actual test administration, and access to secure materials must be restricted to authorized individuals 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 HS assessments and must account for all test materials and supervise the test administrations at all times.

Data Forensic Analyses

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

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

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

Response Change Analysis

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

Score Fluctuation Analysis

It is anticipated that performance on the LEAP 2025 HS assessments will improve over time for reasons such as changes in the curriculum and improvement in instruction. However, large and unexpected score changes may be a sign of testing impropriety. The LDOE applies an approach where the state's level of change in performance from one year to the next is compared to a schools' and test administrators' change in student performance during the same time frame. Schools and test administrators are identified when the level of change is statistically unexpected.

Web Monitoring

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

Plagiarism Detection

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

Alerts for Disturbing Content

Scorers for the LEAP 2025 HS assessments also have the ability to apply an alert flag to student responses that may indicate disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). All alerted responses are automatically routed to the scoring director who reviews and forwards appropriate responses to senior project staff for review. If it is concluded that a response warrants an alert, project management will contact the LDOE to take the necessary action. At no point during this process do scorers or staff have access to demographic information for any students participating in the assessment.

3. Scoring Activities

Answer Key Verification

After a targeted number of tests are administered, DRC conducts an answer key verification. The purpose of this verification is to verify that the correct answers are being properly applied during the scoring process.

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

Selected-Response Item Keycheck. TRIAN, a standardized Pearson program that calculates MC item statistics, is used to verify that MC field-test items are keyed correctly (i.e., that the true correct response is applied during scoring). Items are flagged if their 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 items, with the reasons for flagging, are provided to LDOE and WestEd content staff for key verification. Scoring of MS items is evaluated at data review.

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

- A scoring rubric is created for each TE item. The rubric describes the one and only correct answer for dichotomously scored items (i.e., items scored as either right or wrong). If partial credit is possible, the rubric describes in detail the type of response that could receive credit for each score point.
- The information from the 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 test forms, student responses to 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 Directory of Test Specifications (DOTS), which includes scoring information for all item types.

The adjudication process focuses on detecting possible errors in scoring TE and MS items. DRC provides a report listing the frequency distributions of TE item responses and MS items. Members of the LDOE and WestEd content staff examine the TE and MS response distributions and the auto-frequency reports to evaluate whether the items are scored appropriately. In the event that scoring issues are identified, WestEd content staff and the LDOE committee 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. The constructed- and extended-response items are scored by human raters trained by DRC. Project Essay Grade (PEG), an Artificial Intelligence (AI) scoring engine, also scores the extended-response items. 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.

Selection of Scoring Evaluators. Standard 4.20 states the following:

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

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

The Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2019–2020 LEAP 2025 HS Biology 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, and 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. Each DRC scoring center is a secure facility. All employees are issued photo identification badges and are required to wear them in plain view at all times. Access to scoring centers is limited to badge-wearing staff and to visitors accompanied by authorized staff. All readers are made aware that no scoring materials may leave the scoring center and must sign legally binding confidentiality agreements before work begins. DRC retains these agreements for the duration of the contract. To prevent the unauthorized duplication of secure materials, cell phone and camera use within the scoring rooms is strictly forbidden. Readers only have access to the student responses they are qualified to score. Each scorer is assigned a unique username and password to access the DRC imaging system and must qualify before viewing any live student responses. DRC maintains full control of who may access the system and which item each scorer may score. No demographic data is available to scorers at any time.

Handscoring Training Process. Standard 6.9 specifies:

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

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

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

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

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

Table 3.1

Biology Training Set Composition

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

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

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

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

Table 3.2

Biology Qualifying Standards

Course and Item Type	Qualifying Standard	
Biology 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.
Biology 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 Biology ERs, the appropriate qualifying standard should be achieved on each part of the item. For example, if an item has Part A with a top score of 6 and Part B with a top score of 3, a scorer would need to achieve 60% perfect agreement on Part A and 70% perfect agreement on Part B on one or more of the qualifying sets. A scorer may qualify on one part in the first qualifying set and the other part in the second qualifying set.

Monitoring the Scoring Process. Standard 6.8 states:

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

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

Reader Monitoring Procedures. Throughout the handscoring process, DRC project managers, scoring directors, and team leaders reviewed the statistics that were generated daily. DRC used one team leader for every 10 to 12 readers. If scoring concerns were apparent among individual scorers, team leaders dealt with those issues on an individual basis. If a scorer appeared to need clarification of the scoring rules, DRC supervisors typically monitored one out of five of the scorer's readings, adjusting to 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 live 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 score point apart, divided by the number of responses that were scored twice

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

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

Subject	Score Point Range	Perfect Agreement	Perfect Agreement + Adjacent
Biology CR	0–2	80%	95%
Biology (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 regarding 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 Biology were scored independently by a second reader. The statistics for the inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the first and second scores was examined.

Tables 3.4–3.7 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the 2019–2020 forms.

Table 3.4
Inter-Rater Reliability for Operational Constructed-Response Items

Admin.	Item	Inter-Rater Reliability*				
		2x	Total	Exact Agreement (%)	Adjacent Agreement (%)	Non-Adjacent (%)
Fall 2019	Item 1	≥3,730	≥13,940	98	2	0
	Item 2	≥5,340	≥14,660	95	4	0
	Item 3	≥4,750	≥14,380	95	5	0
Summer 2020	Item 1	≥420	≥1,390	100	0	0
	Item 2	≥680	≥1,510	100	0	0
	Item 3	≥620	≥1,470	97	2	0

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

Table 3.5

Score Point Distributions for Operational Constructed-Response Items

Administration	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
Fall 2019	Item1	≥13,940	82	5	3	0	9
	Item2	≥14,660	59	13	7	0	20
	Item3	≥14,380	61	21	3	0	14
Summer 2020	Item1	≥1,390	86	1	0	1	12
	Item2	≥1,510	67	5	0	0	28
	Item3	≥1,470	59	17	2	0	22

* 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

Admin.	Item	Inter-Rater Reliability*					
		2X	Total	Part	Exact Agreement (%)	Adjacent Agreement (%)	Non-Adjacent (%)
Fall 2019	Item 1	≥4,490	≥14,300	Part A (0–3)	92	7	0
				Part B (0–6)	89	9	2
Summer 2020	Item 1	≥630	≥1,490	Part A (0–3)	98	2	0
				Part B (0–6)	98	2	0

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

Table 3.7

Score Point Distributions for Operational-Extended Response Items

Admin.	Item	Total	Score Point Distribution*									
			Part	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	"5" (%)	"6" (%)	Blank (%)	Nonscore Codes (%)**
Fall 2019	Item 1	≥14,300	Part A (0–3)	16	49	16	6	N/A	N/A	N/A	0	13
			Part B (0–6)	40	21	14	5	4	1	1	0	13
Summer 2020	Item 1	≥1,490	Part A (0–3)	18	54	8	0	N/A	N/A	N/A	0	20
			Part B (0–6)	47	23	9	1	0	0	0	0	20

* 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 HS Biology. The classical analysis includes statistical analysis based on the following types of items: multiple-choice/multiple-selected items, rule-based machine-scored items such as technology-enhanced items, and hand-scored 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 information on classical item statistics for operational items for fall 2019 and summer 2020 tests can be found in Appendix B: Item Analysis Summary Report. Tables B.1.1–B.5.2 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. 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) 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.

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

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

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

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

The *p*-value associated with the Mantel χ^2 statistic and the *SMD* (on an effect-size metric) are used to determine DIF classifications. Table 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 |*SMD/SD*| ≤ 0.17
B (slight to moderate)	Mantel χ^2 *p*-value < 0.05 and 0.17<|*SMD/SD*| < 0.25
C (moderate to large)	Mantel χ^2 *p*-value < 0.05 and |*SMD/SD*| ≥ 0.25

Three DIF analyses were conducted for the operational test items only: female/male, black/white, and Hispanic/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The same methods were used to detect items on which both black/white and Hispanic/white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Tables 4.3.1 and 4.3.2 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 and 4.3.2 summarize the operational-test DIF statistics for the operational items appearing on the fall 2019 and summer 2020 test forms, respectively. Because the summer 2020 test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 4.3.1
Summary of DIF Flags for Fall 2019 Biology Operational Items

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	40	[1],[0]	[0],[0]
African American – White	41	[0],[0]	[0],[0]
Hispanic – White	41	[0],[0]	[0],[0]

Table 4.3.2
Summary of DIF Flags for Summer 2020 Biology Operational Items

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	38	[2],[1]	[0],[0]
African American – White	40	[0],[1]	[0],[0]
Hispanic – White	39	[0],[1]	[1],[0]

Pre-Equating for Intact Forms

In general, the LEAP 2025 Biology 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 fall 2019 and summer 2020 administrations, intact forms based on previously administered operational forms were given; therefore, the pre-equating method was applied and existing scoring tables were used for score classifications.

Unidimensionality and Principal Component Analysis

Appendix C: Dimensionality provides information about principal component analysis of the Biology 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 fall 2019 and summer 2020 assessments, 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 the second principal component eigenvalues were compared *without rotation*. Tables C.2.1 and C.2.2 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 fall 2019 test, not for the summer 2020 test. Because the summer test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Scaling

Although both the fall 2019 and the summer 2020 tests used the preexisting scoring tables, general procedures for 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 Biology 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 *i*th item or component, and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 Biology 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. While the coefficient alpha value for the fall 2019 assessment was 0.88, that of the summer 2020 assessment was 0.59. Because the summer test was administered during the COVID-19 pandemic, any statistical inferences should be cautiously drawn from these results. 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, and English Learners.

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 of the fall 2019 test is 0.714, that of the summer 2020 test is 0.682. Because the summer 2020 test was administered during the COVID-19 pandemic, 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 are 0.278 and 0.468 for the fall 2019 and the summer 2020 Biology tests, respectively.

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 are 0.476 to 0.229 for the fall 2019 and the summer 2020 Biology tests, respectively.

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 fall 2019 and the summer 2020 Biology 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 LEAP 2025 Biology test 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 fall 2019 and summer 2020, please refer to the 2019 LEAP 2025 Biology Technical Report.
Because the summer 2020 test was administered during the COVID-19 pandemic, any validity evidence associated with the summer 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 Biology test 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 Biology test 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 Biology assessment.

Although the fall 2019 and the summer 2020 tests are based on a pre-equating method, Chapter 4 briefly describes classical data analysis, IRT, and scaling of the Biology 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 and marginal alpha as measures of internal consistency and also describes analysis procedures for classification consistency and classification accuracy.

Chapter 6 reports the statistical summaries of the fall 2019 and the summer 2020 Biology tests.

6. Statistical Summaries

For the fall 2019 and the summer 2020 Biology 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 6.1.1 and 6.1.2. Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters. Scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state and disaggregated into various demographic groups. In addition to the descriptive statistics presented in the tables 6.1.1-6.1.2, scale score frequency distributions are presented in Appendix D: Scale Distribution and Statistical Report. Finally, because the summer 2020 test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 6.1.1

LEAP 2025 State Test Results: Fall 2019 Operational Biology

	Scale Score			% at Performance Level				
	N	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥12,700	717.79	27.53	38	27	20	11	4
Gender								
Female	≥6,070	720.16	26.73	34	28	22	12	4
Male	≥6,630	715.62	28.06	42	25	19	10	3
Ethnicity								
African American	≥7,010	709.23	22.53	49	30	16	5	1
American Indian or Alaska Native	≥80	730.55	25.72	21	23	26	26	4
Asian	≥230	739.15	32.62	20	15	23	26	18
Hispanic/Latino	≥1,190	712.71	26.73	47	24	18	9	2
Multi-Racial	≥190	726.47	29.32	22	32	23	16	7
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR	NR	NR
White	≥3,970	732.52	28.39	19	22	28	21	8
Economically Disadvantaged								
No	≥3,080	734.03	29.39	20	21	27	23	10
Yes	≥9,620	712.59	24.74	44	29	18	7	2
LEP Status (English Learner)								
No	≥11,810	719.09	27.65	36	27	21	12	4
Yes	≥890	700.59	18.69	65	26	8	NR	NR

Table 6.1.2

LEAP 2025 State Test Results: Summer 2020 Operational Biology

	Scale Score			% at Performance Level				
	N	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥1,240	702.90	17.34	58	34	7	NR	NR
Gender								
Female	≥530	703.88	16.57	55	37	7	NR	NR
Male	≥710	702.16	17.87	60	32	8	NR	NR
Ethnicity								
African American	≥900	701.91	16.83	60	34	6	NR	NR
American Indian or Alaska Native	≥10	707.80	29.27	50	40	NR	NR	10
Asian	<10	NR	NR	NR	NR	NR	NR	NR
Hispanic/Latino	≥92	698.47	18.79	65	29	5	NR	NR
Multi-Racial	≥10	713.64	20.44	36	36	27	NR	NR
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR	NR	NR
White	≥220	707.80	16.99	49	37	14	NR	NR
Economically Disadvantaged								
No	≥220	705.50	18.98	54	35	11	NR	NR
Yes	≥1,020	702.33	16.92	59	34	7	NR	NR
LEP Status (English Learner)								
No	≥1,150	703.38	17.18	58	34	8	NR	NR
Yes	≥80	696.64	18.21	67	29	3	NR	NR

References

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

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

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.

Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, 63(4), 602–614.

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

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

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

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

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

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

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

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

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.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 452-461.

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.

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

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.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E.P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies: A revised core vocabulary*. Austin, TX: Steck-Vaughn.

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

Biology

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer 2020 test was administered during the 2020 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table A.1.1

Item Type Summary: Fall 2019 and Summer 2020 Operational Biology

Administration	MC	MS	TE*	CR	ER**	TPD	TPI
Fall 2019	13	6	8	3	1	8	2
Summer 2020	13	6	8	3	1	8	2

* One of the TE items is a multiple-part, selected-response (MPSR) item.

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

Table A.2.1

Raw Score Summary: Fall 2019 and Summer 2020 Operational Biology

Admin.	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
Fall 2019	≥12,700	19.94	10.55	0	62	0.30	0.41	0.88	3.65
Summer 2020	≥1,240	14.03	5.13	3	45	0.22	0.23	0.59	3.28

* Reliability is Cronbach’s alpha.

Table A.3.1

Raw Score Summary by Reporting Category: Fall 2019 and Summer 2020 Operational Biology

Admin	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
Fall 2019	Investigate	2.34	1.55	0	8	0.33	0.30	0.24	1.35
	Evaluate	5.95	3.40	0	18	0.32	0.45	0.70	1.86
	Reason Scientifically	9.69	5.76	0	32	0.29	0.41	0.80	2.58
Summer 2020	Investigate	1.90	1.28	0	7	0.27	0.19	-0.01	1.29
	Evaluate	4.10	2.02	0	15	0.22	0.23	0.26	1.74
	Reason Scientifically	6.73	3.18	0	24	0.21	0.24	0.49	2.27

Table A.4.1

Scale Score and Raw Score Summary: Fall 2019 Operational Biology

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥12,700	100.00	717.79	27.53	19.94	10.55
Female	≥6,070	47.82	720.16	26.73	20.73	10.45
Male	≥6,630	52.18	715.62	28.06	19.22	10.59
African American	≥7,010	55.23	709.23	22.53	16.51	7.88
American Indian or Alaska Native	≥80	0.66	730.55	25.72	24.70	10.79
Asian	≥230	1.84	739.15	32.62	28.94	13.31
Hispanic/Latino	≥1,190	9.40	712.71	26.73	18.08	9.89
Multi-Racial	≥190	1.55	726.47	29.32	23.40	11.61
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥3,970	31.25	732.52	28.39	25.77	11.67
Economically Disadvantaged	≥9,620	75.73	712.59	24.74	17.84	9.06
English Language Learners	≥890	7.01	700.59	18.69	13.47	5.62

Table A.4.2

Scale Score and Raw Core Summary: Summer 2020 Operational Biology

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥1,240	100.00	702.90	17.34	14.03	5.13
Female	≥530	42.80	703.88	16.57	14.26	5.03
Male	≥710	57.20	702.16	17.87	13.86	5.20
African American	≥900	72.57	701.91	16.83	13.69	4.78
American Indian or Alaska Native	≥10	0.80	707.80	29.27	16.50	10.95
Asian	<10	NR	NR	NR	NR	NR
Hispanic/Latino	≥90	7.40	698.47	18.79	12.85	5.16
Multi-Racial	≥10	0.88	713.64	20.44	17.82	7.04
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥220	17.70	707.80	16.99	15.56	5.61
Economically Disadvantaged	≥1,020	82.14	702.33	16.92	13.82	4.92
English Language Learners	≥80	7.16	696.64	18.21	12.29	4.50

Appendix B: Item Analysis Summary Report

Summary Statistics Reports

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer 2020 test was administered during the 2020 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table B.1.1

P-Value Summary by Item Type: Fall 2019 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.056	0.056	0.132	0.142	0.142
ER*	1	0.168	0.168	0.238	0.307	0.307
MC	13	0.188	0.299	0.372	0.521	0.602
MS	6	0.066	0.076	0.118	0.237	0.365
TE	8	0.079	0.199	0.313	0.394	0.572
TPD	8	0.199	0.243	0.279	0.442	0.572
TPI	2	0.161	0.161	0.353	0.544	0.544

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

Table B.1.2

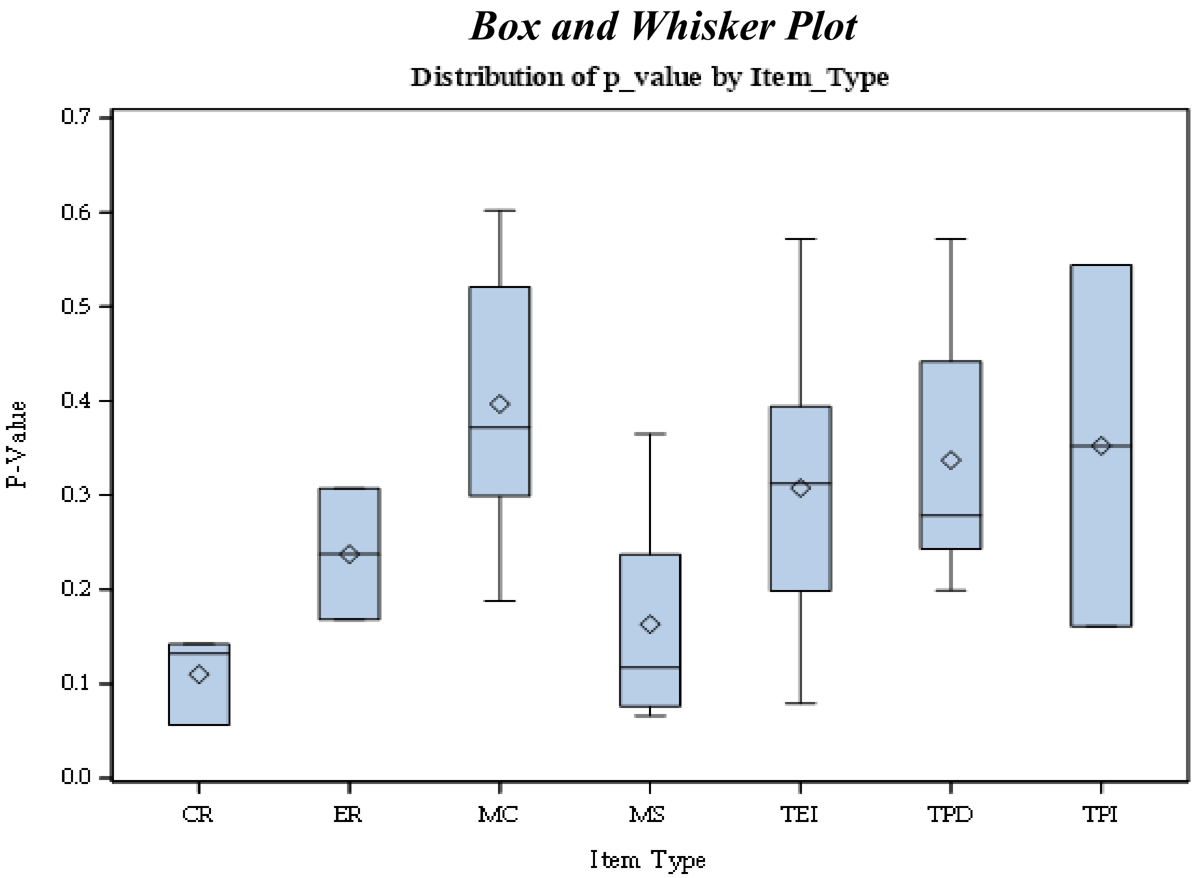
P-Value Summary by Item Type: Summer 2020 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.005	0.005	0.027	0.102	0.102
ER*	1	0.079	0.079	0.163	0.246	0.246
MC	13	0.151	0.204	0.292	0.381	0.500
MS	6	0.030	0.035	0.070	0.123	0.304
TE	8	0.058	0.114	0.210	0.301	0.424
TPD	8	0.159	0.169	0.216	0.301	0.410
TPI	2	0.111	0.111	0.259	0.407	0.407

* 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 Item Type: Fall 2019 Operational Biology



Plot B.1.2

P-Value Summary by Item Type: Summer 2020 Operational Biology

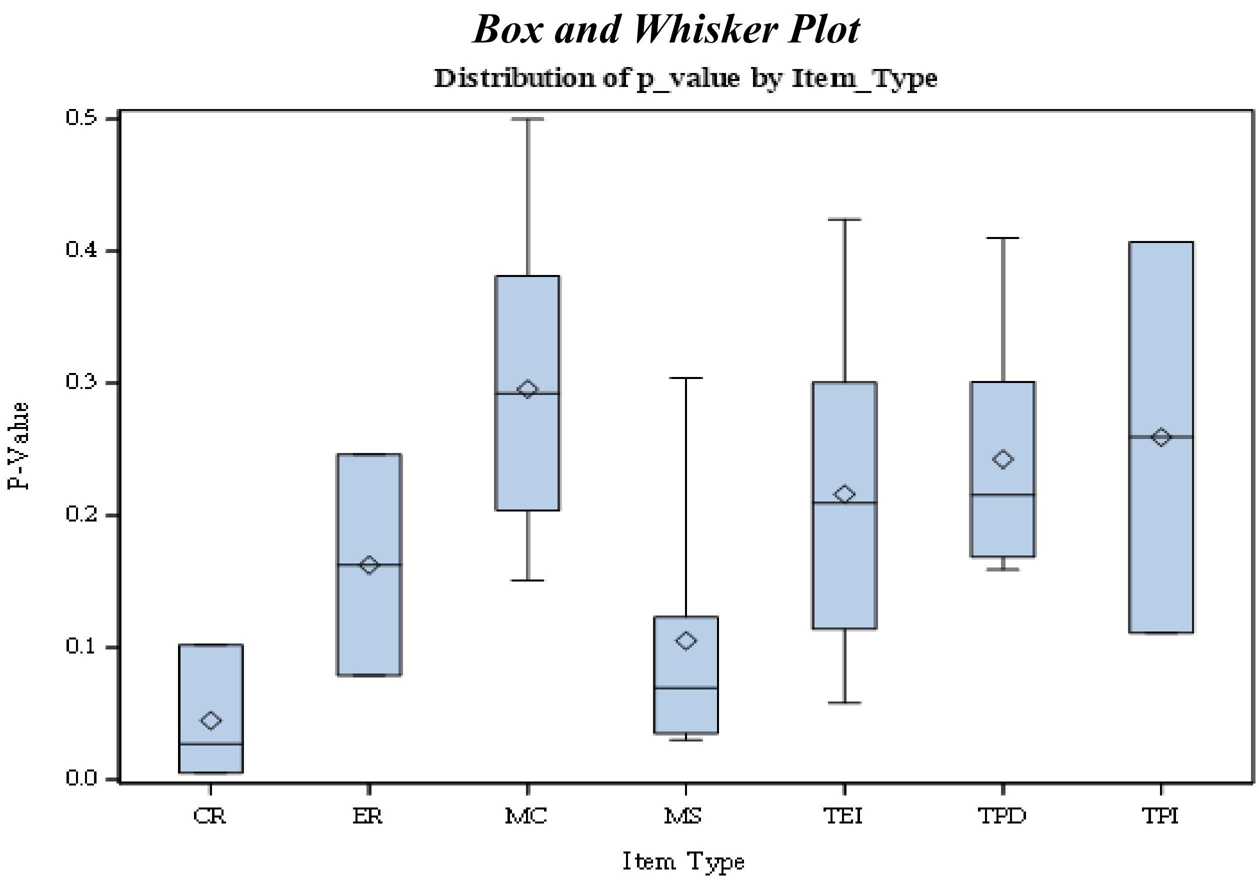


Table B.2.1

Item-Total Correlation Summary by Item Type: Fall 2019 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.427	0.427	0.529	0.661	0.661
ER*	1	0.713	0.713	0.718	0.723	0.723
MC	13	0.099	0.229	0.392	0.452	0.505
MS	6	0.139	0.255	0.332	0.391	0.471
TE	8	0.249	0.374	0.433	0.513	0.578
TPD	8	0.231	0.296	0.416	0.548	0.612
TPI	2	0.363	0.363	0.474	0.585	0.585

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

Table B.2.2

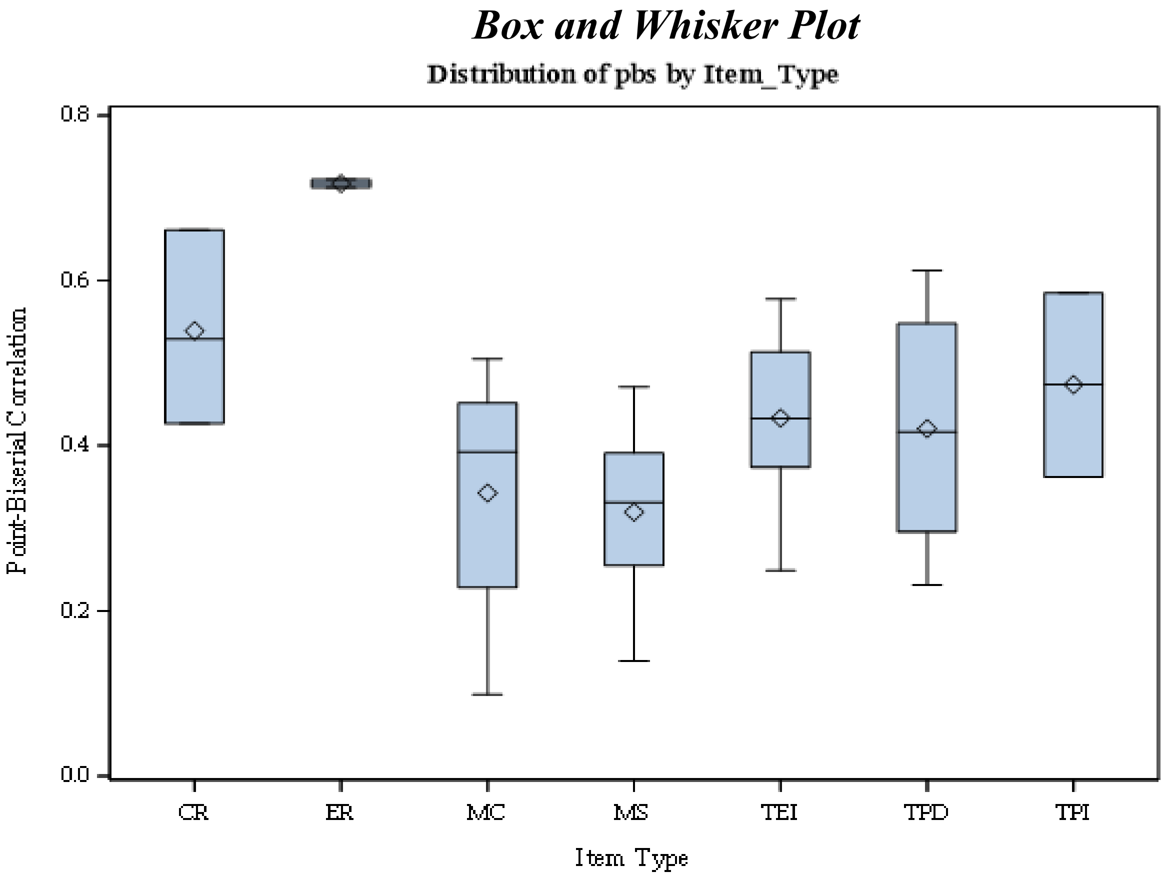
Item-Total Correlation Summary by Item Type: Summer 2020 Operational Biology

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.187	0.187	0.266	0.269	0.269
ER*	1	0.487	0.487	0.493	0.498	0.498
MC	13	0.015	0.136	0.212	0.263	0.345
MS	6	0.026	0.062	0.136	0.226	0.232
TE	8	0.093	0.156	0.251	0.349	0.403
TPD	8	0.113	0.153	0.234	0.275	0.513
TPI	2	0.163	0.163	0.255	0.348	0.348

* 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 Item Type: Fall 2019 Operational Biology



Plot B.2.2

Item-Total Correlation Summary by Item Type: Summer 2020 Operational Biology

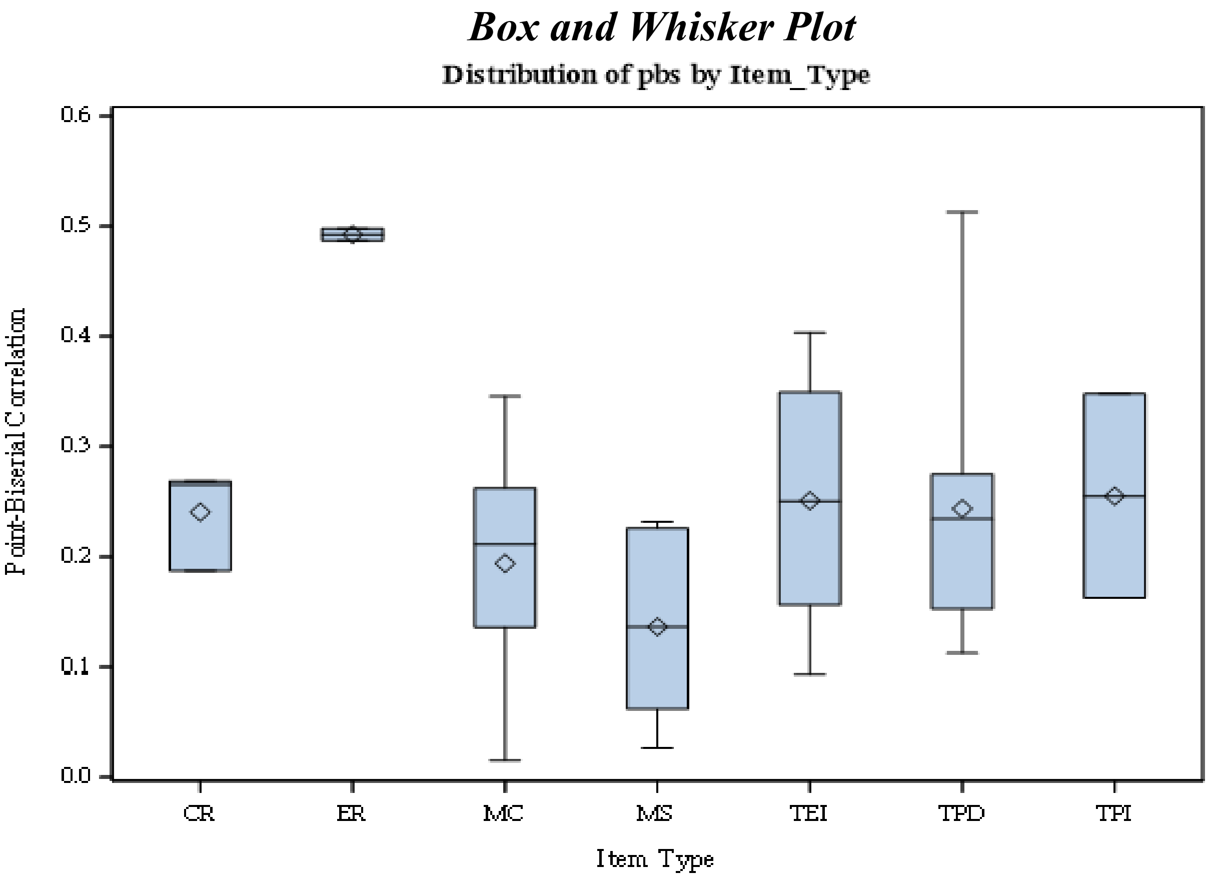


Table B.3.1

Corrected Point-Biserial Correlation Summary by Item Type: Fall 2019 Operational Biology*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.387	0.387	0.502	0.628	0.628
ER**	1	0.656	0.656	0.663	0.670	0.670
MC	13	0.056	0.185	0.351	0.413	0.471
MS	6	0.116	0.212	0.307	0.365	0.439
TE	8	0.225	0.325	0.386	0.463	0.527
TPD	8	0.163	0.239	0.348	0.495	0.561
TPI	2	0.318	0.318	0.426	0.534	0.534

* 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.

Table B.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Summer 2020 Operational Biology*

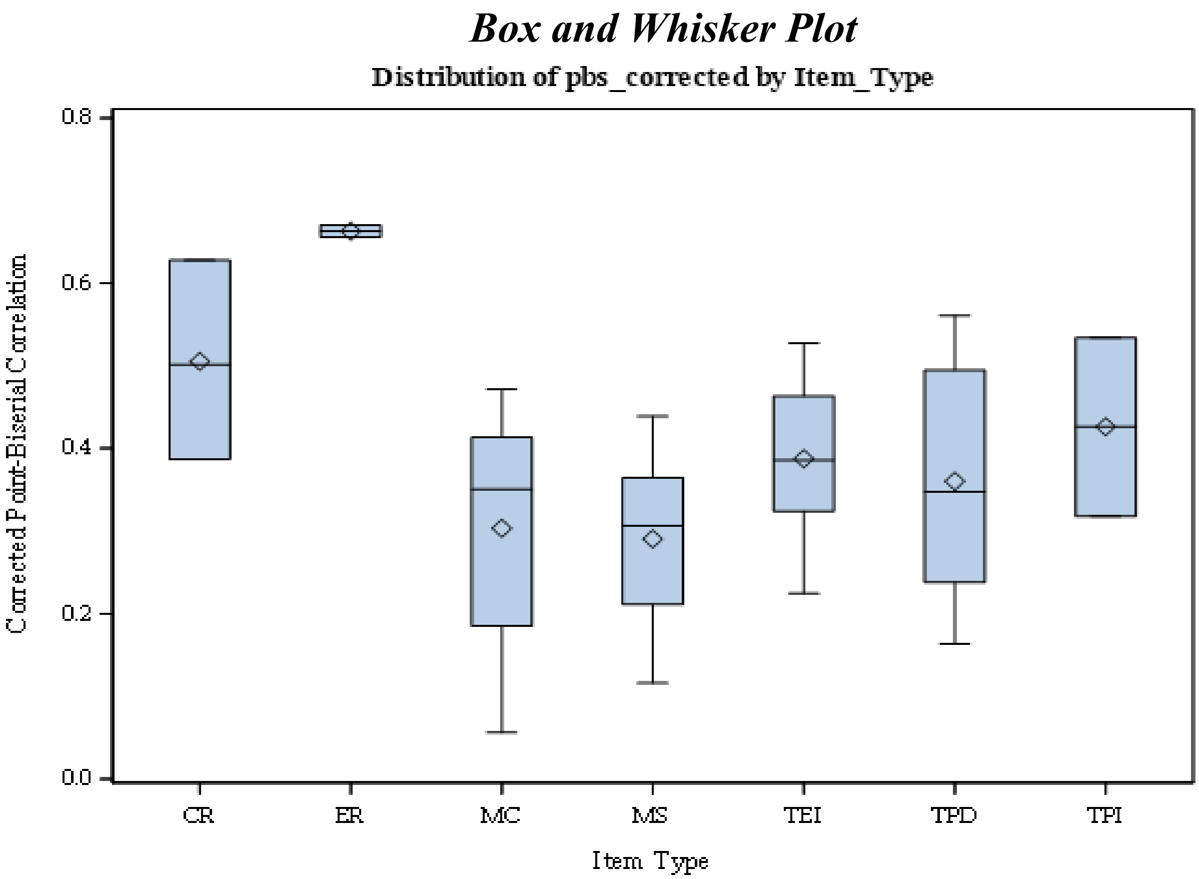
Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	3	0.167	0.167	0.183	0.225	0.225
ER**	1	0.363	0.363	0.380	0.398	0.398
MC	13	-0.071	0.065	0.141	0.178	0.257
MS	6	-0.018	0.008	0.102	0.145	0.164
TE	8	0.044	0.083	0.140	0.241	0.280
TPD	8	-0.018	0.039	0.120	0.136	0.373
TPI	2	0.078	0.078	0.150	0.221	0.221

* 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 Item Type: Fall 2019 Operational Biology



Plot B.3.2

Corrected Point-Biserial Correlation Summary by Item Type: Summer 2020 Operational Biology

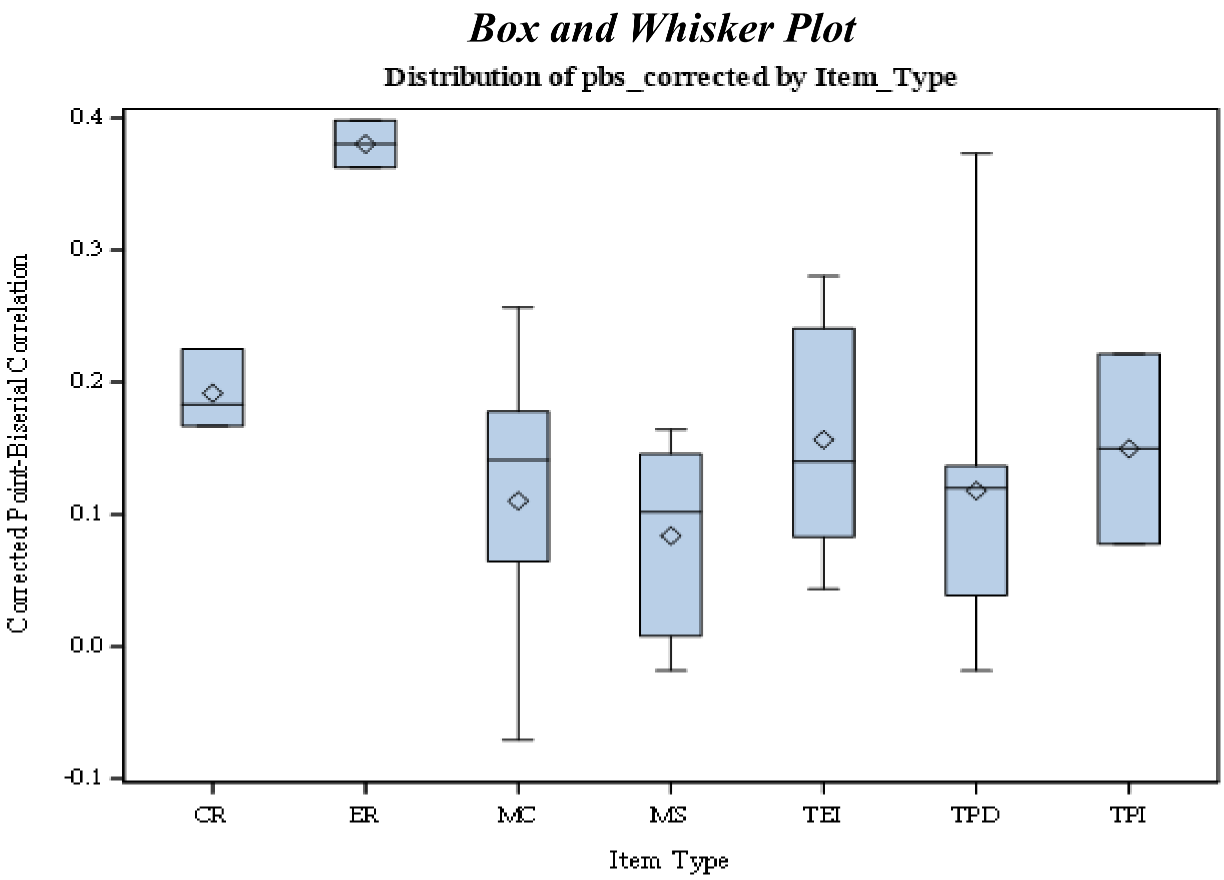


Table B.4.1

Item-Total Correlation Summary by Reporting Category and Item Type: Fall 2019 Operational Biology

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Evaluate	1	0.427	0.427	0.427	0.427	0.427
	Reason Scientifically	1	0.529	0.529	0.529	0.529	0.529
ER*	Reason Scientifically	1	0.713	0.713	0.718	0.723	0.723
MC	Evaluate	3	0.271	0.271	0.459	0.505	0.505
	Investigate	1	0.415	0.415	0.415	0.415	0.415
	Reason Scientifically	7	0.099	0.216	0.378	0.452	0.469
MS	Evaluate	1	0.367	0.367	0.367	0.367	0.367
	Investigate	1	0.255	0.255	0.255	0.255	0.255
	Reason Scientifically	4	0.139	0.218	0.344	0.431	0.471
TEI	Evaluate	3	0.396	0.396	0.534	0.578	0.578
	Reason Scientifically	3	0.249	0.249	0.353	0.492	0.492
TPD	Evaluate	3	0.342	0.342	0.484	0.612	0.612
	Investigate	3	0.231	0.231	0.250	0.351	0.351
	Reason Scientifically	2	0.482	0.482	0.547	0.611	0.611
TPI	Reason Scientifically	2	0.363	0.363	0.474	0.585	0.585
	Evaluate	1	0.427	0.427	0.427	0.427	0.427

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

Table B.4.2

Item-Total Correlation Summary by Reporting Category and Item Type: Summer 2020 Operational Biology

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Evaluate	1	0.266	0.266	0.266	0.266	0.266
	Reason Scientifically	1	0.187	0.187	0.187	0.187	0.187
ER*	Reason Scientifically	1	0.487	0.487	0.493	0.498	0.498
MC	Evaluate	3	0.084	0.084	0.209	0.250	0.250
	Investigate	1	0.212	0.212	0.212	0.212	0.212
	Reason Scientifically	7	0.015	0.063	0.249	0.276	0.345
MS	Evaluate	1	0.092	0.092	0.092	0.092	0.092
	Investigate	1	0.232	0.232	0.232	0.232	0.232
	Reason Scientifically	4	0.026	0.044	0.121	0.203	0.226
TEI	Evaluate	3	0.224	0.224	0.304	0.403	0.403
	Reason Scientifically	3	0.093	0.093	0.122	0.394	0.394
TPD	Evaluate	3	0.175	0.175	0.213	0.276	0.276
	Investigate	3	0.113	0.113	0.131	0.256	0.256
	Reason Scientifically	2	0.275	0.275	0.394	0.513	0.513
TPI	Reason Scientifically	2	0.163	0.163	0.255	0.348	0.348
	Evaluate	1	0.266	0.266	0.266	0.266	0.266

* 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: Fall 2019 Operational Biology

Item Type	*N* OP Items	*N* Items Flagged for *P*-Value	*N* Items Flagged for Mean	*N* Items Flagged for Point-Biserial Correlation	*N* Items Flagged for DIF*	*N* Items Flagged for Omitting
CR	3	3	0	0	1	0
ER**	1	1	0	0	0	0
MC	13	2	0	2	0	0
MS	6	5	0	1	0	0
TEI	8	2	0	0	0	0
TPD	8	2	0	0	0	0
TPI	2	1	0	0	0	0

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

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

Table B.5.2

Statistically Flagged Operational Items: Summer 2020 Operational Biology

Item Type	*N* OP Items	*N* Items Flagged for *P*-Value	*N* Items Flagged for Mean	*N* Items Flagged for Point-Biserial Correlation	*N* Items Flagged for DIF*	*N* Items Flagged for Omitting
CR	3	3	0	1	2	0
ER**	1	2	0	0	1	0
MC	13	5	0	5	0	0
MS	6	5	0	4	0	0
TEI	8	4	0	3	2	0
TPD	8	5	0	2	1	0
TPI	2	1	0	1	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

Biology

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer 2020 test was administered during the 2020 COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table C.1.1

Reporting Category Intercorrelation Coefficients for Fall 2019 Operational Biology

Reporting Category	**Investigate**	**Evaluate**	**Reason Scientifically**
Investigate	1.00		
Evaluate	0.42	1.00	
Reason Scientifically	0.46	0.76	1.00

Table C.1.2

Reporting Category Intercorrelation Coefficients for Summer 2020 Operational Biology

Reporting Category	**Investigate**	**Evaluate**	**Reason Scientifically**
Investigate	1.00		
Evaluate	0.07	1.00	
Reason Scientifically	0.21	0.35	1.00

Table C.2.1

First and Second Eigenvalue: Fall 2019 Operational Biology*

Form	First Eigenvalue	Second Eigenvalue
B	8.078	1.324

* The ratio of first and second eigenvalues is about 6.101.

Plot C.1.1

Principal Component Analysis Plot: Fall 2019 Operational Biology

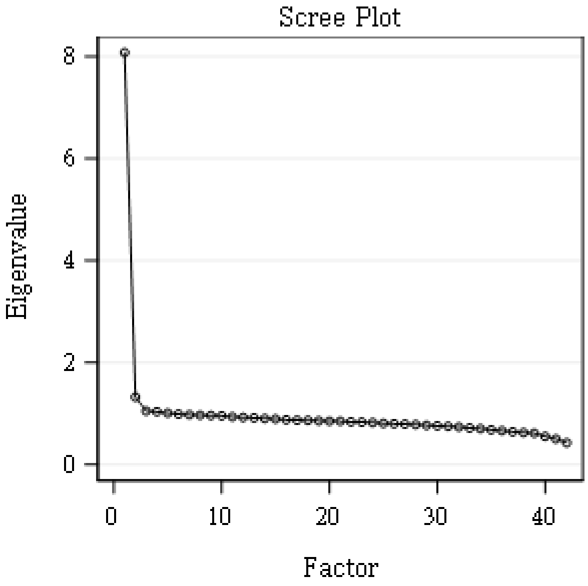


Table C.2.2

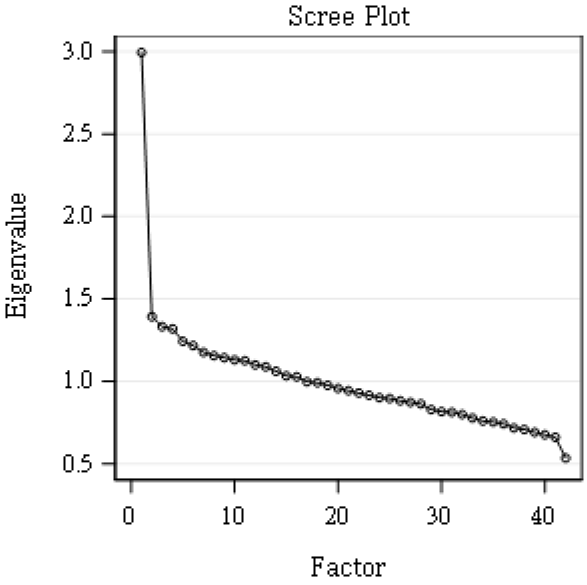
First and Second Eigenvalue: Summer 2020 Operational Biology*

Form	First Eigenvalue	Second Eigenvalue
B	2.994	1.394

* The ratio of first and second eigenvalues is about 2.147.

Plot C.1.2

Principal Component Analysis Plot: Summer 2020 Operational Biology



Appendix D: Scale Distribution and Statistical Report

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer 2020 test was administered during the 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 Fall 2019 Biology

```
                                        DESCRIPTIVE STATISTICS - SCALE SCORES
                                                   BIOLOGY
                                                   ALL STUDENTS
                                                     Form B


                         N                      ≥12700
                         Mean                   717.79      Median                714.00
                         Std deviation           27.53      Variance              757.75
                         Skewness               0.3569      Kurtosis             -0.0170
                         Mode                   706.00      Std Error Mean        0.2442
                         Range                  192.00      Interquartile Range    36.00

                                             Quantile       Estimate

                                             100% Max          842
                                             99%               784
                                             95%               769
                                             90%               757
                                             75% Q3            735
                                             50% Median        714
                                             25% Q1            699
                                             10%               687
                                             5%                675
                                             1%                654
                                             0% Min            650
```

```
                      Histogram                           #  Boxplot                        Normal Probability Plot
845+*                                                   <10     0         845+                                                 *
   .                                                                         |
   .*                                                   ≥10     0            |                                                 *
   .*                                                   ≥10     0            |                                                 *
   .*                                                   ≥10     0            |                                                 *
   .**                                                  ≥40     0            |                                                 *
   .*****                                              ≥180     |            |                                             *****
   .********                                           ≥300     |            |                                         *****+
   .*************                                      ≥530     |            |                                      ****+
   .******************                                 ≥750     |            |                                   ****+
   .***********************                            ≥960     |            |                                ****
   .***********************                            ≥970  +-----+         |                             ****
   .*********************************                 ≥1380  |     |         |                         +****
   .************************************************  ≥2050  *--+--*         |                      *****
   .**********************************************    ≥1950  |     |         |                   *****
   .**********************************************    ≥1950  +-----+         |             *******
   .**********************                             ≥920     |            |         *****+
   .*******                                            ≥270     |            |      ****++
   .*****                                              ≥180     |            |   ****+
655+*****                                              ≥190     |         655+****+
    ----+----+----+----+----+----+----+----+----+---                         +----+----+----+----+----+----+----+----+----+----+
    * may represent up to 43 counts                                              -2        -1         0        +1       +2
```

Table D.1.2 Frequency Distribution of Scale Scores for Fall 2019 Biology

```
                                            FREQUENCY DISTRIBUTION - SCALE SCORES
                                                       BIOLOGY
                                                       ALL STUDENTS
                                                        Form B
SCALE_SCORE                                                                            Cum.            Cum.
                                                                               Freq    Freq  Percent   Percent
650    |*********                                                               ≥80     ≥80     0.68      0.68
654    |**********                                                             ≥100    ≥190     0.81      1.50
666    |******************                                                     ≥180    ≥370     1.43      2.93
675    |****************************                                           ≥270    ≥640     2.16      5.09
681    |****************************************                               ≥390   ≥1040     3.14      8.23
687    |****************************************************                   ≥520   ≥1560     4.10     12.33
691    |***************************************************************        ≥630   ≥2190     4.97     17.30
695    |******************************************************************     ≥660   ≥2850     5.19     22.50
699    |*******************************************************************    ≥660   ≥3520     5.24     27.74
702    |******************************************************************     ≥650   ≥4180     5.17     32.91
706    |*******************************************************************    ≥670   ≥4850     5.28     38.19
709    |***************************************************************        ≥620   ≥5470     4.92     43.11
711    |*************************************************************          ≥600   ≥6080     4.79     47.90
714    |***************************************************                    ≥500   ≥6590     4.00     51.90
717    |**********************************************                         ≥460   ≥7050     3.65     55.55
719    |************************************************                       ≥470   ≥7530     3.74     59.29
722    |*************************************                                  ≥360   ≥7900     2.90     62.20
724    |************************************                                   ≥360   ≥8260     2.87     65.06
726    |***********************************                                    ≥350   ≥8610     2.75     67.82
729    |******************************                                         ≥290   ≥8910     2.34     70.15
731    |******************************                                         ≥300   ≥9210     2.39     72.55
733    |**********************                                                 ≥220   ≥9440     1.76     74.31
735    |**********************                                                 ≥220   ≥9660     1.75     76.06
738    |***********************                                                ≥220   ≥9880     1.78     77.84
740    |***********************                                                ≥220  ≥10110     1.77     79.61
742    |*********************                                                  ≥200  ≥10320     1.63     81.24
744    |*******************                                                    ≥180  ≥10500     1.46     82.70
746    |****************                                                       ≥160  ≥10660     1.26     83.96
748    |*******************                                                    ≥190  ≥10850     1.50     85.46
751    |***************                                                        ≥140  ≥11000     1.14     86.60
753    |*******************                                                    ≥180  ≥11190     1.47     88.08
755    |***************                                                        ≥150  ≥11340     1.21     89.29
757    |*************                                                          ≥130  ≥11470     1.04     90.33
759    |**************                                                         ≥140  ≥11610     1.10     91.43
761    |**************                                                         ≥130  ≥11750     1.06     92.49
763    |***********                                                            ≥100  ≥11850     0.83     93.32
765    |**********                                                              ≥90  ≥11950     0.75     94.07
767    |**********                                                             ≥100  ≥12050     0.80     94.87
769    |**********                                                             ≥100  ≥12150     0.79     95.66
771    |**********                                                             ≥100  ≥12250     0.79     96.45
773    |********                                                                ≥70  ≥12330     0.62     97.07
775    |*******                                                                 ≥60  ≥12400     0.53     97.60
778    |******                                                                  ≥50  ≥12450     0.46     98.06
780    |******                                                                  ≥50  ≥12510     0.46     98.52
782    |****                                                                    ≥30  ≥12550     0.30     98.82
784    |***                                                                     ≥20  ≥12580     0.22     99.04
787    |***                                                                     ≥30  ≥12610     0.27     99.31
789    |***                                                                     ≥20  ≥12640     0.20     99.51
792    |**                                                                      ≥20  ≥12660     0.17     99.69
795    |*                                                                       <10  ≥12670     0.06     99.75
798    |*                                                                       ≥10  ≥12680     0.11     99.86
802    |*                                                                       <10  ≥12690     0.05     99.91
805    |                                                                        <10  ≥12690     0.03     99.94
809    |                                                                        <10  ≥12700     0.02     99.96
814    |                                                                        <10  ≥12700     0.02     99.98
825    |                                                                        <10  ≥12700     0.01     99.99
842    |                                                                        <10  ≥12700     0.01    100.00
       -----+----+----+----+----+----+----+----+----+----+----+----+----+--
           50  100  150  200  250  300  350  400  450  500  550  600  650
```

Table D.2.1 Scale Score Descriptive Statistics and Plots for Summer 2020 Biology

```
                                          DESCRIPTIVE STATISTICS - SCALE SCORES
                                                     BIOLOGY
                                                     Form B


                           N                        ≥1240
                           Mean                    702.90     Median                  706.00
                           Std deviation            17.34     Variance                300.65
                           Skewness               -0.3677     Kurtosis                0.9561
                           Mode                    691.00     Std Error Mean          0.4918
                           Range                   125.00     Interquartile Range      23.00


                                               Quantile       Estimate

                                               100% Max          775
                                               99%               744
                                               95%               729
                                               90%               724
                                               75% Q3            714
                                               50% Median        706
                                               25% Q1            691
                                               10%               681
                                               5%                675
                                               1%                654
                                               0% Min            650


                    Histogram                     #        Boxplot                             Normal Probability Plot
777.5+*                                          <10           0            777.5+                                                  *
     .                                                                           |
     .                                                                           |
     .*                                          <10           0                 |                                                  *
     .                                                                           |
752.5+*                                          <10           0            752.5+                                                  *
     .**                                         <10           |                 |                                                 **
     .**                                         <10           |                 |                                               ***
     .***                                        ≥10           |                 |                                            +**
     .*****                                      ≥20           |                 |                                         +***
727.5+********                                   ≥30           |            727.5+                                      +****
     .*************                              ≥60           |                 |                                   +****
     .************************                  ≥110           |                 |                                +****
     .*******************************           ≥150        +-----+              |                             *****
     .***************************************   ≥190        *-----*              |                         *****
702.5+*********************                     ≥100        |  +  |         702.5+                       ***+
     .***************************************   ≥190        |     |              |                   *****
     .**********************                    ≥100        +-----+              |                ****+
     .***************                            ≥70           |                 |             ****+
     .**********                                 ≥40           |                 |           ***+
677.5+********                                   ≥30           |            677.5+        ****
     .                                                         |                 |       ++
     .******                                     ≥20           |                 |    +****
     .                                                                          | +++
     .                                                                          |+
652.5+******                                     ≥20           0            652.5+******
      ----+----+----+----+----+----+----+----                                    +----+----+----+----+----+----+----+----+----+----+
      * may represent up to 5 counts                                                 -2        -1         0        +1        +2
```

Table D.2.2 Frequency Distribution of Scale Scores for Summer 2020 Biology

```
                                                  FREQUENCY DISTRIBUTION - SCALE SCORES
                                                             BIOLOGY
                                                             ALL STUDENTS
                                                              Form B


SCALE_SCORE                                                                           Cum.     Cum.
                                                                     Freq   Freq    Percent  Percent
650   |******                                                         ≥10    ≥10      0.88     0.88
654   |*********                                                      ≥10    ≥20      1.45     2.33
666   |**************                                                 ≥20    ≥50      2.25     4.59
675   |*******************                                            ≥30    ≥90      3.06     7.64
681   |************************                                       ≥40   ≥140      3.86    11.50
687   |************************************                           ≥70   ≥210      5.79    17.30
691   |******************************************************        ≥100   ≥320      8.61    25.91
695   |**************************************************            ≥100   ≥420      8.05    33.95
699   |************************************************                ≥90   ≥510      7.64    41.59
702   |****************************************************          ≥100   ≥620      8.37    49.96
706   |****************************************************          ≥100   ≥720      8.29    58.25
709   |**********************************************                  ≥90   ≥810      7.32    65.57
711   |*******************************************                     ≥80   ≥900      6.92    72.49
714   |*********************************                               ≥60   ≥960      5.23    77.72
717   |************************************                            ≥70  ≥1030      5.71    83.43
719   |************************                                        ≥40  ≥1080      3.78    87.21
722   |*****************                                               ≥30  ≥1110      2.65    89.86
724   |***************                                                 ≥30  ≥1140      2.41    92.28
726   |***********                                                     ≥20  ≥1160      1.77    94.05
729   |*********                                                       ≥10  ≥1180      1.37    95.41
731   |******                                                          ≥10  ≥1190      0.88    96.30
733   |*******                                                         ≥10  ≥1210      1.13    97.43
735   |****                                                            <10  ≥1210      0.56    97.99
738   |***                                                             <10  ≥1220      0.48    98.47
740   |***                                                             <10  ≥1220      0.40    98.87
742   |*                                                               <10  ≥1230      0.08    98.95
744   |*                                                               <10  ≥1230      0.16    99.12
746   |**                                                              <10  ≥1230      0.32    99.44
748   |**                                                              <10  ≥1230      0.24    99.68
751   |*                                                               <10  ≥1240      0.08    99.76
753   |*                                                               <10  ≥1240      0.08    99.84
761   |*                                                               <10  ≥1240      0.08    99.92
775   |*                                                               <10  ≥1240      0.08   100.00
-----+----+----+----+----+----+----+----+----+----+----
10   20   30   40   50   60   70   80   90  100

Frequency
```

Appendix E: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

Biology

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table E.1

Reliability for Overall and Subgroups: Fall 2019 Operational Biology

Subgroup	Form B
All Students	0.882
Female	0.878
Male	0.886
African American	0.810
American Indian or Alaska Native	0.875
Asian	0.914
Hispanic/Latino	0.873
Multi-Racial	0.898
Native Hawaiian or Other Pacific Islander	N/A*
White	0.893
Economically Disadvantaged	0.849
English Learners	0.679

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

Table E.2

Reliability for Overall and Subgroups: Summer 2020 Operational Biology

Subgroup	Form B
All Students	0.594
Female	0.576
Male	0.607
African American	0.541
American Indian or Alaska Native	N/A*
Asian	N/A*
Hispanic/Latino	0.619
Multi-Racial	N/A*
Native Hawaiian or Other Pacific Islander	N/A*
White	0.634
Economically Disadvantaged	0.561
English Learners	0.501

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

Table E.2.1

Cronbach Alpha Reliability: Fall 2019 and Summer 2020 Operational Biology

Administration	Cronbach's Alpha
Fall 2019	0.882
Summer 2020	0.594

Table E.3.1

Classification Accuracy and Decision Consistency: Fall 2019 Operational Biology

Table E.3.1.1

Estimates of Accuracy and Consistency of Achievement-Level Classification for Form

Form	Accuracy	Consistency	PChance	Kappa
B	0.714	0.622	0.278	0.476

Table E.3.1.2

Accuracy of Classification at Each Achievement Level for Form

Form	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
B	0.849	0.624	0.601	0.676	0.775

Table E.3.1.3

Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.887	0.899	0.943	0.979

Table E.3.1.4

Consistency of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.844	0.858	0.922	0.97

Table E.3.1.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.673	0.695	0.665	0.567

Table E.3.1.6

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

Form	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.059	0.046	0.039	0.014

Table E.3.1.7

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

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.054	0.055	0.017	.006

Table E.3.2
Classification Accuracy and Decision Consistency: Summer 2020 Operational Biology

Table E.3.2.1
Estimates of Accuracy and Consistency of Achievement-Level Classification for Form

Form	Accuracy	Consistency	PChance	Kappa
B	0.682	0.590	0.468	0.229

Table E.3.2.2
Accuracy of Classification at Each Achievement Level for Form

Form	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
B	0.799	0.536	0*	0*	0*

* Zero value reflects there were very few students in *Basic*, *Mastery*, and *Advanced* performance levels.

Table E.3.2.3
Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.750	0.923	0.997	0*

* Zero value reflects there were very few students in *Basic*, *Mastery*, and *Advanced* performance levels.

Table E.3.2.4
Consistency of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.670	0.868	0.997	0*

* Zero value reflects there were very few students in *Basic*, *Mastery*, and *Advanced* performance levels.

Table E.3.2.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.663	0.672	0.630	0.377

Table E.3.2.6

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

Form	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.025	0.047	0.058	0.045

Table E.3.2.7

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

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
B	0.033	0.058	0.056	0.014

Appendix F: Comparison of 2018, 2019, and 2020 Assessments

Biology

Contents

- Keep in mind that the fall administration includes both retesters and initial testers, while the summer administration includes primarily retesters.
- Because the summer 2020 test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table F.1

Subgroup N Count of Selected Assessments: Biology

Subgroup	Fall 2018	Fall 2019	Summer 2019	Summer 2020
	N	*N*	*N*	*N*
Total	≥6,940	≥12,700	≥250	≥1,240
Gender				
Female	≥3,410	≥6,070	≥100	≥530
Male	≥3,530	≥6,630	≥150	≥710
Ethnicity				
African American	≥3,310	≥7,010	≥170	≥900
American Indian or Alaska Native	≥50	≥80	<10	≥10
Asian	≥160	≥230	<10	<10
Hispanic/Latino	≥620	≥1,190	≥10	≥90
Multi-Racial	≥130	≥190	<10	≥10
Native Hawaiian or Other Pacific Islander	<10	<10	<10	<10
White	≥2,650	≥3,970	≥50	≥220
Economically Disadvantaged				
No	≥2,240	≥3,080	≥60	≥220
Yes	≥4,690	≥9,620	≥180	≥1,020
LEP Status (English Learner)				
No	≥6,650	≥11,810	≥230	≥1,150
Yes	≥280	≥890	≥20	≥80