LEAP 2025 U.S. History Technical Report: 2020–2021

Prepared by DRC, Pearson, and WestEd

EXECUTIVE SUMMARY

The Louisiana Educational Assessment Program 2025 (LEAP 2025) is composed of tests that are carefully constructed to fairly assess the achievement of Louisiana students. This technical report provides information on the operational test administrations, scoring activities, analyses, and results of the spring 2021 administration of the LEAP 2025 U.S. History test, which used intact forms based on previously administered operational forms see the 2018–2019 LEAP 2025 U.S. History Technical Report.

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

The chapters of this report outline general information about the administration and scoring activities of the LEAP 2025 assessments, CTT (Classical Test Theory) analysis results, and the interpretation of the scores on the tests. Additionally, because of conditions related to COVID-19, please use caution when making any inferences from the statistical results of the spring 2021 administration.

Table of Contents

EXECUTIVE SUMMARY ii

1. Introduction 5

2. Test Administration 6

- Training of School Systems 6
- Ancillary Materials 7
- Time 12
- Online Forms Administration 12
- Accessibility and Accommodations 12
- Testing Windows 14
- Test Security Procedures 14
- Data Forensic Analyses 15
- Alerts for Disturbing Content 16

3. Scoring Activities 17

- Constructed-Response and Extended-Response Scoring 19

4. Data Analysis 27

- Classical Item Statistics 27
- Differential Item Functioning 27
- Pre-Equating for Intact Forms 31
- Unidimensionality and Principal Component Analysis 31
- Scaling 32

5. Reliability and Validity 34

Internal Consistency Reliability Estimation 34

Student Classification Accuracy and Consistency 35

Validity 37

6. Statistical Summaries 39

References 41

Appendix A: Test Summary 44

Appendix B: Item Analysis Summary Report 47

Appendix C: Dimensionality 56

Appendix D: Scale Distribution and Statistical Report 59

Appendix E: Reliability and Classification Accuracy 62

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 Social Studies assessments in grades 3–8 and in U.S. History. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. Further, the LDOE and the BESE are committed to the development of rigorous assessments as one component of their comprehensive plan—Louisiana Believes—designed to ensure that every Louisiana student is on track to be successful in postsecondary education and the workforce.

The purpose of this technical report is to describe the processes for the spring 2021 administration of LEAP 2025 U.S. History. 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 2025 for High School (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. Reading Directions to Students
13.2. LEAP 2025 English I and English II
13.3. LEAP 2025 Algebra I and Geometry
13.4. LEAP 2025 Biology
13.5. 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. Spring 2021
3. LEAP 2025 High School Alerts
4. Pre-Administration Oath of Security and Confidentiality Statement
5. Post-Administration Oath of Security and Confidentiality Statement
6. General Information
6.1. DRC INSIGHT Portal (eDIRECT) and INSIGHT

7. LEAP 2025 High School
    7.1. Testing Requirements
8. Test Security
    8.1. Key Definitions
    8.2. Violations of Test Security
    8.3. Testing Guidelines
    8.4. Testing Conditions
    8.5. Testing Schedule
    8.6. Extended Time for Testing
    8.7. Extended Breaks
    8.8. Makeup Testing
9. LEAP 2025 High School and End-of-Course Testing Times
10. Roles and Responsibilities
    10.1. District Test Coordinator
    10.2. School Test Coordinator
    10.3. Chief Technology Officer
11. Managing Test Sessions and Tickets
    11.1. Student Transfers
    11.2. Locked Test Tickets
    11.3. Technical Issues
    11.4. Invalidating Test Tickets
12. Resources for Online Testing
    12.1. High School Test Administration Manual
    12.2. DRC INSIGHT Portal User Guide
    12.3. LEAP 2025 Accommodations and Accessibility Manual
    12.4. DRC INSIGHT Technology User Guide
    12.5. Student Tutorials
    12.6. Online Tools Training (OTT)
13. Post-Administration Rescoring Process for LEAP 2025 HS Assessments
14. Request for Rescoring
15. 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/504 Plan (IAP), or English Learner Plan (EL plan); allowing accommodations for students who do not have an IEP, IAP, or EL plan; or defining terms on the test.

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

The 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 they 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/IAP/EL plan.

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

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

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

In compliance with this standard, the TAM contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP 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 2020–2021 assessments for HS courses were administered to students within the state testing windows of December 1–18, 2020, or January 6–26, 2021 for fall administration, April 15–May 21, 2021 for spring administration, and June 21–25, 2021 for summer administration.

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 Fluctuation Analysis
- Item Exposure Monitoring
- 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 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.

Item Exposure Monitoring. The fall 2020 test administration included two testing windows; there was a testing window in December 2020 and a testing window in January 2021. Due to the same form being used in both windows, item performance was examined in the second window to ensure that item content had not been exposed. In addition to reviewing fit plots for good alignment of an item's performance across the windows, an item's moving *p*-value and point-biserial averages were produced daily.

During the January testing window, if an item's moving average *p*-value was larger than expected compared to the previous day's or the December average, the item was flagged. Similar methodology was also applied in the spring 2021 test window due to the reuse of the spring 2019 test forms.

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). 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

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, subtype, session, alignment, and prior use of item. WestEd reviews and confirms the contents of the DOTS file as part of test review rounds. The DOTS file is then provided to the LDOE for review and final approval. Once approved, the information contained in the DOTS is used in scoring the test and in reporting.

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

Scoring of Technology-Enhanced (TE) Items. All TE items are processed through DRC's autoscoring engine and scored according to the assigned scoring rules 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 rubrics describe 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 rubrics describe in detail the type of response that could receive credit for each score point.
- The information from each scoring rubric is entered into the scoring system within the item banking system so that the truth resides in one place along with the item image and other metadata. This scoring information designates specific

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

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

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

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

Constructed-Response and Extended-Response Scoring

Constructed-response items are scored by human raters trained by DRC. Extended-response items are scored by Project Essay Grade (PEG), an Artificial Intelligence (AI) scoring engine. Ten percent of the responses are scored twice to monitor and maintain inter-rater reliability. Scoring supervisors also conduct read-behinds and review all nonscores and alerts. Handscoring processing rules are detailed in the *LEAP 2025 Spring 2021 Handscoring/AI Documentation* document.

Selection of Scoring Evaluators. Standard 4.20 states the following:

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

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

Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2020–2021 LEAP 2025 HS 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. Whether training and scoring are conducted within a DRC facility or done remotely, security is essential to the handscoring process. When users log into DRC's secure, web-based scoring application, ScoreBoard, they are required to read and accept the security policy before they are allowed to access any project. For each project, scorers are also required to read and sign non-disclosure agreements, and during training emphasis is always given to what security means, the importance of maintaining security, and how this is accomplished.

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

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

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

Handscoring Training Process. Standard 6.9 specifies:

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

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

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

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

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

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

Monitoring the Scoring Process. Standard 6.8 states:

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

The following section explains the monitoring procedures that DRC uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored.

Detailed scoring rubrics, which specify the criteria for scoring, are available for all constructed- and extended-response items.

Reader Monitoring Procedures. Throughout the handscoring process, DRC project managers, scoring directors, and team leaders review the statistics that are generated daily. DRC uses one team leader for every 10 to 12 readers. If scoring concerns are apparent among individual scorers or if a scorer needs clarification on the scoring rules, team leaders address those issues on an individual basis. DRC supervisors typically monitor one out of five of the scorer's readings, making adjustments to that ratio as needed. If a supervisor disagrees with a reader's scores during monitoring, the supervisor provides retraining in the form of direct feedback to the reader, using rubric language and applicable training responses.

Validity Sets and Inter-Rater Reliability. In addition to the feedback that supervisors provide to readers during regular read-behinds and the continuous monitoring of inter-rater reliability and score point distributions, DRC also conducts validity scoring using LDOE-approved validity responses identified by DRC scoring supervisors during live scoring for newly operational items. Validity responses are inserted among the live student responses.

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

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

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

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

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

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

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

Inter-Rater Reliability. A minimum of 10% of the responses for constructed-and extended-response items are scored independently by a second reader. This is the case regardless of whether the first reader is a human rater or AI. The statistics for inter-rater reliability are calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the first and second scores is examined.

Tables 3.1–3.4 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the 2020–2021 forms.

Table 3.1

Inter-Rater Reliability for Operational Constructed-Response Items

Administration	Item	Inter-Rater Reliability*				
		2x	Total	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2020**	Item1	≥4,050	≥11,180	95	5	0
	Item 2	≥3,750	≥11,190	93	7	0
Spring 2021	Item1	≥13,480	≥43,560	93	7	0
	Item 2	≥11,940	≥43,240	88	12	0
Summer 2021	Item1	≥2,460	≥5,400	99	1	0
	Item 2	≥2,470	≥5,470	94	6	0

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

** Fall data includes both fall administration windows.

Table 3.2

Score Point Distributions for Operational Constructed-Response Items

Administration	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
Fall 2020***	Item 1	≥11,180	58	11	8	0	23
	Item 2	≥11,190	23	42	16	0	19
Spring 2021	Item 1	≥43,560	40	24	20	0	15
	Item 2	≥43,240	39	34	16	0	11
Summer 2021	Item 1	≥5,400	63	3	1	0	33
	Item 2	≥5,470	31	31	4	0	32

* 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.

*** Fall data includes both fall administration windows.

Table 3.3

Inter-Rater Reliability for Operational Extended-Response Items

Administration	Item	Inter-Rater Reliability*					
		2x	Total	Dimension	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2020**	Item 1	≥11,340	≥15,020	Content	96	4	0
				Claims	96	4	0
Spring 2021	Item 1	≥30,590	≥52,820	Content	93	7	0
				Claims	93	7	0
Summer 2021	Item 1	≥4,100	≥6,290	Content	95	5	0
				Claims	96	4	0

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

** Fall data includes both fall administration windows.

Table 3.4

Score Point Distributions for Operational Extended-Response Items

Admin.	Item	Total	Score Point Distribution*							
			Dimension	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	"3" Rating (%)	"4" Rating (%)	Blank (%)	Nonscore Codes (%)**
Fall 2020***	Item 1	≥15,020	Content	32	31	16	6	3	0	12
			Claims	43	25	13	5	2	0	12
Spring 2021	Item 1	≥52,820	Content	25	35	20	8	2	0	10
			Claims	32	29	19	8	2	0	10
Summer 2021	Item 1	≥6,290	Content	46	30	5	0	0	0	18
			Claims	54	23	4	0	0	0	18

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

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

*** Fall data includes both fall administration windows.

4. Data Analysis

Classical Item Statistics

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

Tables and figures that provide the information on classical item statistics for operational items for the spring 2021 test 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, 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 chi-square.

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

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

where N_{r1k} is the number of correct responses in the reference group at ability level *k*, N_{f0k} is the number of incorrect responses in the focal group at ability level *k*, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level *k*, and N_{r0k} is the number of incorrect responses in the reference group at ability level *k*. The *MH DIF* statistic is based on a 2×2×*M* (2 groups × 2 item scores × *M* strata) frequency table, in which students in the reference (male or white) and focal (female or black/Hispanic) groups are matched on their total raw scores.

The *ΔMH DIF* is then computed as

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

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

The *MH* chi-square statistic and the *ΔMH DIF* are 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 from 1.0 and is at least 1.5. Positive values are classified as "C+" and negative values as "C–."

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

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

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

The *MH DIF* statistic is based on a 2×(*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:

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

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

Table 4.2

DIF Categories for Polytomous Items

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

Three DIF analyses are conducted for the operational test items only: female/male, black/white, and Hispanic/white. That is, item score data are 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 are used to detect items on which both black/white and Hispanic/white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Table 4.3 provide the number of items flagged for DIF. Items flagged with A-

DIF show negligible DIF, items flagged with B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Very few operational test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the *MH* statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 *p*-value and *SMD* statistics. Because the spring 2021 test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 4.3
Summary of DIF Flags for Operational Items: Spring 2021 U.S. History

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	49	[1],[1]	[1],[1]
African American – White	52	[1],[0]	[0],[0]
Hispanic – White	51	[2],[0]	[0],[0]

Pre-Equating for Intact Forms

Because the spring 2021 test administration used an intact operational form from spring 2019, the pre-equating method was applied. That is, the existing spring 2019 scoring tables were used for score report and performance classifications.

Unidimensionality and Principal Component Analysis

Appendix C: Dimensionality provides information about principal component analysis of the LEAP 2025 U.S. History tests. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, in the case of scholastic achievement, a one-dimensional scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 2004; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to

be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983).

To check the unidimensionality of the spring 2021 test, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the Statistical Analysis System (SAS) program. The first and second principal component eigenvalues were compared *without rotation*. Table C.2.1 and Figure C.1.1 summarize the results of the first and second principal component eigenvalues of the assessments. A general guideline in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the tables and figures, the first component is substantially larger than the second eigenvalue for the spring 2021 test. Because the spring test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Scaling

Although the spring test used the preexisting scoring tables, general procedures for the scaling method are described here because 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 U.S. History 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 estimate is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the assumption that inter-item covariances constitute true-score variance and the fact that the average true score variance of items is greater than or equal to the average inter-item covariance. The formula for coefficient alpha is

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

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the ith item (or component), and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 U.S. History 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 Cronbach's alpha for the total test. Cronbach's alpha for the spring 2021 test was 0.94. Because the spring test was administered during the COVID-19 pandemic, however, statistical inferences should be cautiously drawn from these results. Additional reliabilities were calculated on various demographic subgroups using the population of students (see Appendix E: Reliability and Classification Accuracy). The subgroups are male/female, white/Black/Hispanic/Asian/American Indian or Alaska Native/Native Hawaiian or Other Pacific Islander/multi-racial, Economically Disadvantaged, English Learners, Education Classification, and Section 504.

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 are 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 index of the spring 2021 test was 0.751, the overall consistency index was 0.663 for the U.S. History test. Because the spring 2021 test was administered during the COVID-19 pandemic, however, great caution should be applied when any statistical inference is drawn.

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

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

where P is the probability of consistent classification and P_c is the probability of consistent classification by chance (Lee, Hanson, & Brennan, 2000). P is the sum of the diagonal elements, and P_c is the sum of the squared row totals. The PChance index was 0.236 for the spring 2021 U.S. History tests.

Kappa is a measure of "how much agreement exists beyond chance alone" (Fleiss, 1973), which means that it provides the proportion of consistent classifications between two forms after removing the proportion of consistent classifications expected by chance alone. The Kappa index was 0.559 for the spring 2021 U.S. History test.

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 for proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (AERA/APA/NCME, 2014). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process.

The spring 2021 U.S. History tests were designed and developed to provide fair and accurate scores that support appropriate, meaningful, and useful educational decisions. As the technical report progresses, it details the procedures and processes applied to the LEAP 2025 U.S. History 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 form used in the spring 2021 administration, please refer to the 2018–2019 LEAP 2025 U.S. History Technical Report. Because the spring 2021 test was administered during the COVID-19 pandemic, any validity evidence associated with the spring test should be carefully interpreted.

The knowledge, expertise, and professional judgment offered by Louisiana educators ultimately ensure that the content for the LEAP 2025 U.S. History 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 U.S. History test to derive valid inferences about Louisiana student performance.

Chapter 2 of the technical report describes the process, procedures, and policies that guide the administration of the LEAP 2025 assessments, including accommodations, test security, and detailed written procedures provided to test administrators and school personnel.

Chapter 3 describes scoring processes and activities for the LEAP 2025 U.S. History test. Although the spring 2021 test utilized a pre-equating method, Chapter 4 briefly describes classical data analysis, IRT, and scaling of the U.S. History tests, which derive scale scores from students' raw scores. In addition, Chapter 4 describes an analysis of DIF and includes gender and ethnicity DIF results. A summary of classical analysis and DIF results for the operational items is presented in Appendix B: Item Analysis Summary Report.

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

Chapter 6 reports the statistical summaries of the spring 2021 U.S. History test.

6. Statistical Summaries

For the spring 2021 U.S. History test, the lowest obtainable scale score (LOSS) is 650 and the highest obtainable scale score (HOSS) is 850. Test results are provided in Table 6.1. Scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state and are disaggregated by demographic groups. In addition to the descriptive statistics presented in Table 6.1, scale score frequency distributions are presented in Appendix D: Scale Distribution and Statistical Report. Finally, because the spring 2021 test was administered during the COVID-19 pandemic, great caution should be applied when any statistical inference is drawn.

Table 6.1

LEAP 2025 State Test Results: Spring 2021 Operational U.S. History

	Scale Score			% at Performance Level**				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥36,190	727.05	34.89	30	15	28	18	8
Gender								
Female	≥18,630	726.76	33.27	30	17	29	17	7
Male	≥17,550	727.36	36.52	31	14	27	19	9
Ethnicity								
African American	≥15,070	711.56	31.37	47	18	24	9	2
American Indian or Alaska Native	≥240	733.51	31.76	23	15	33	20	9
Asian	≥710	754.89	36.09	11	8	23	29	29
Hispanic/Latino	≥2,260	725.43	35.89	32	13	28	19	8
Multi-Racial	≥730	731.68	33.01	24	16	32	20	8
Native Hawaiian or Other Pacific Islander	≥30	734.53	39.98	24	12	21	29	15
White	≥17,130	739.44	31.92	17	13	32	26	13
Economically Disadvantaged*								
No	≥11,670	744.41	31.61	13	12	31	28	16
Yes	≥21,270	717.52	32.66	40	17	27	13	4
English Learner								
No	≥35,340	727.76	34.69	30	15	28	18	8
Yes	≥850	697.53	30.04	66	16	14	3	2
Education Classification								
Gifted or Talented	≥2,230	762.25	30.34	5	5	22	32	35
Regular	≥31,090	727.32	32.93	29	16	30	18	7
Special	≥2,850	696.51	31.54	68	13	13	5	1
Section 504								
No	≥33,030	728.15	34.74	29	15	28	19	8
Yes	≥3,150	715.49	34.39	44	16	24	11	5

* ≥3,250 students with no record of either No or Yes.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A: Test Summary

U.S. History

Contents
Table A.1.1 Item Type Summary: Spring 2021 Operational U.S. History
Table A.2.1 Raw Score Summary: Spring 2021 Operational U.S. History
Table A.3.1 Raw Score Summary by Reporting Category: Spring 2021 Operational U.S. History
Table A.4.1 Scale Score and Raw Score Summary: Spring 2021 Operational U.S. History

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

Table A.1.1

Item Type Summary: Spring 2021 Operational U.S. History

Administration	MC	MS	TE	CR	ER*
Spring 2021	40	3	7	2	1

* Classical 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: Spring 2021 Operational U.S. History

Admin.	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
Spring 2021	≥36,190	33.54	14.64	2	69	0.52719	0.48934	0.94	3.59

* Reliability is Cronbach's alpha.

Table A.3.1

Raw Score Summary by Reporting Category: Spring 2021 Operational U.S. History

Admin	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
Spring 2021	Standard 2	6.92	3.32	0	15	0.47942	0.46062	0.76	1.63
	Standard 3	4.93	2.29	0	9	0.54938	0.48075	0.70	1.25
	Standard 4	8.13	3.42	0	15	0.57023	0.46573	0.77	1.64
	Standards 5&6*	13.55	6.92	0	30	0.51938	0.52363	0.88	2.40

* Standards 5 and 6 were combined into one reporting category beginning with the 2018–2019 test administrations, resulting in a redistribution of the points for each reporting category.

Table A.4.1

Scale Score and Raw Score Summary: Spring 2021 Operational U.S. History

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥36,190	100.00	727.05	34.89	33.54	14.64
Female	≥18,630	51.50	726.76	33.27	33.27	14.11
Male	≥17,550	48.50	727.36	36.52	33.83	15.19
African American	≥15,070	41.65	711.56	31.37	26.92	12.53
American Indian or Alaska Native	≥240	0.67	733.51	31.76	36.11	13.89
Asian	≥710	1.96	754.89	36.09	45.26	14.62
Hispanic/Latino	≥2,260	6.26	725.43	35.89	33.11	14.88
Multi-Racial	≥730	2.03	731.68	33.01	35.35	14.11
Native Hawaiian or Other Pacific Islander	≥30	0.09	734.53	39.98	37.15	16.47
White	≥17,130	47.33	739.44	31.92	38.82	13.87
Economically Disadvantaged: No	≥11,670	32.25	744.41	31.61	40.96	13.70
Economically Disadvantaged: Yes	≥21,270	58.77	717.52	32.66	29.46	13.41
EL: No	≥35,340	97.64	727.76	34.69	33.83	14.60
EL: Yes	≥850	2.36	697.53	30.04	21.59	10.89
Gifted or Talented	≥2,230	6.18	762.25	30.34	48.53	12.38
Regular Education	≥31,090	85.92	727.32	32.93	33.57	14.03
Special Education	≥2850	7.89	696.51	31.54	21.50	11.51
Section 504: No	≥33,030	91.28	728.15	34.74	34.00	14.62
Section 504: Yes	≥3,150	8.72	715.49	34.39	28.72	13.95

Appendix B: Item Analysis Summary Report

Summary Statistics Reports

U.S. History

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

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

Table B.1.1

P-Value Summary by Item Type: Spring 2021 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.348	0.348	0.349	0.350	0.350
ER*	1	0.258	0.258	0.268	0.278	0.278
MC	40	0.321	0.523	0.582	0.674	0.789
MS	3	0.332	0.332	0.388	0.533	0.533
TEI	7	0.291	0.293	0.396	0.437	0.537

***** Classical 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: Spring 2021 Operational U.S. History

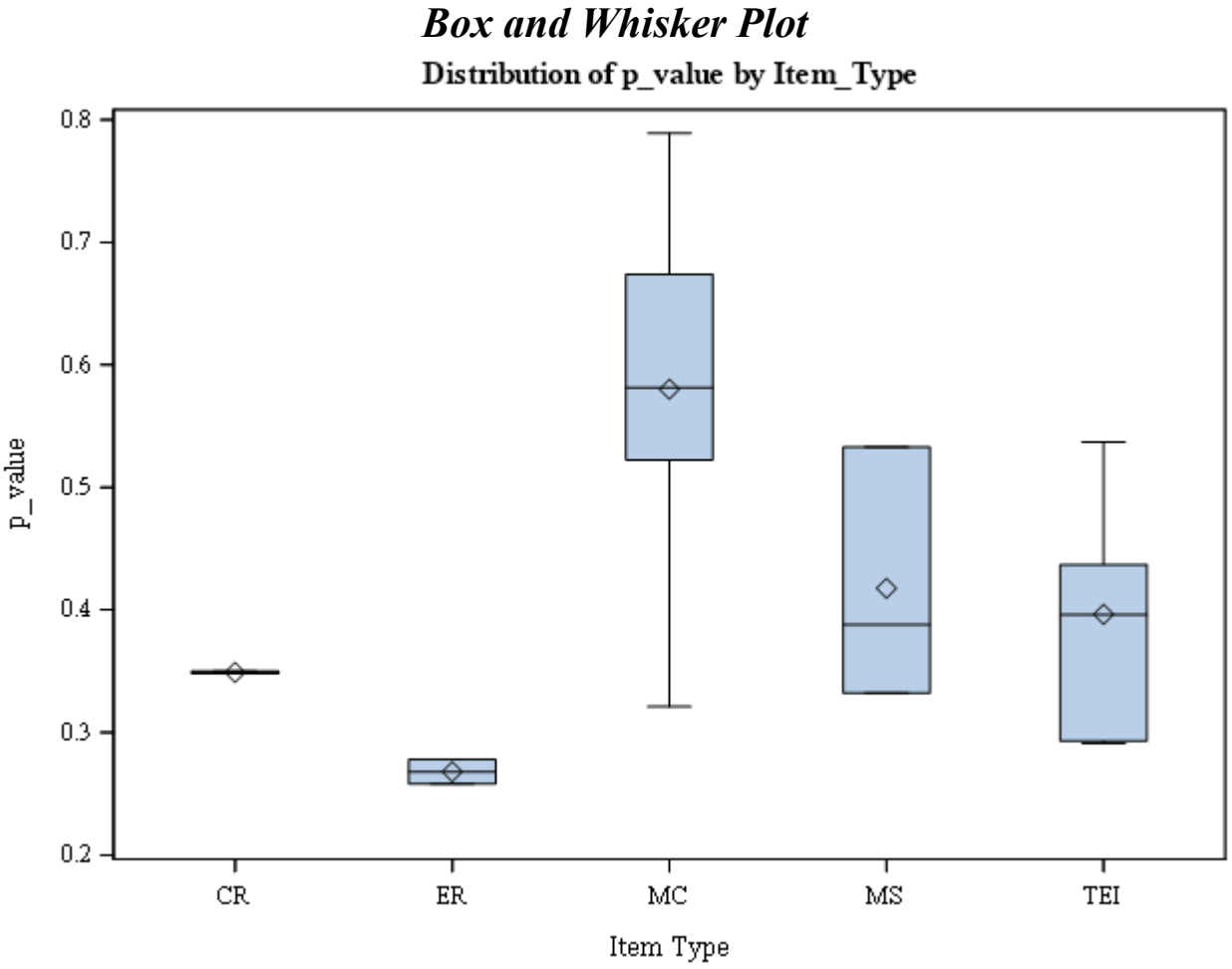


Table B.2.1

Item-Total Correlation Summary by Item Type: Spring 2021 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.615	0.615	0.677	0.739	0.739
ER*	1	0.776	0.776	0.779	0.783	0.783
MC	40	0.264	0.393	0.468	0.507	0.601
MS	3	0.486	0.486	0.516	0.559	0.559
TEI	7	0.525	0.528	0.547	0.610	0.635

* Classical 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: Spring 2021 Operational U.S. History

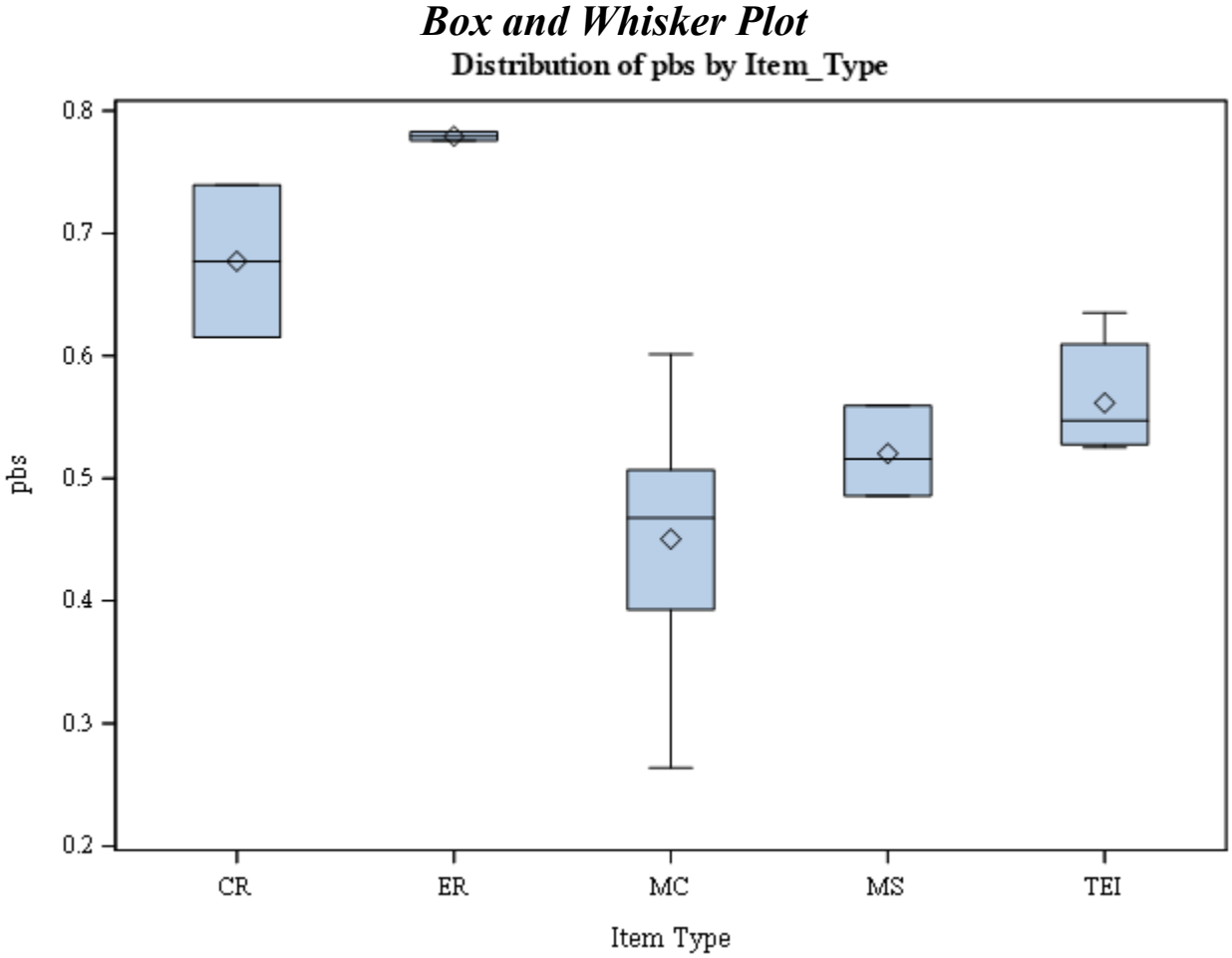


Table B.3.1

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2021 Operational U.S. History*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.582	0.582	0.647	0.713	0.713
ER**	1	0.745	0.745	0.749	0.754	0.754
MC	40	0.233	0.364	0.443	0.481	0.579
MS	3	0.460	0.460	0.490	0.536	0.536
TEI	7	0.483	0.495	0.518	0.576	0.606

***** 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 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 by Item Type: Spring 2021 Operational U.S. History

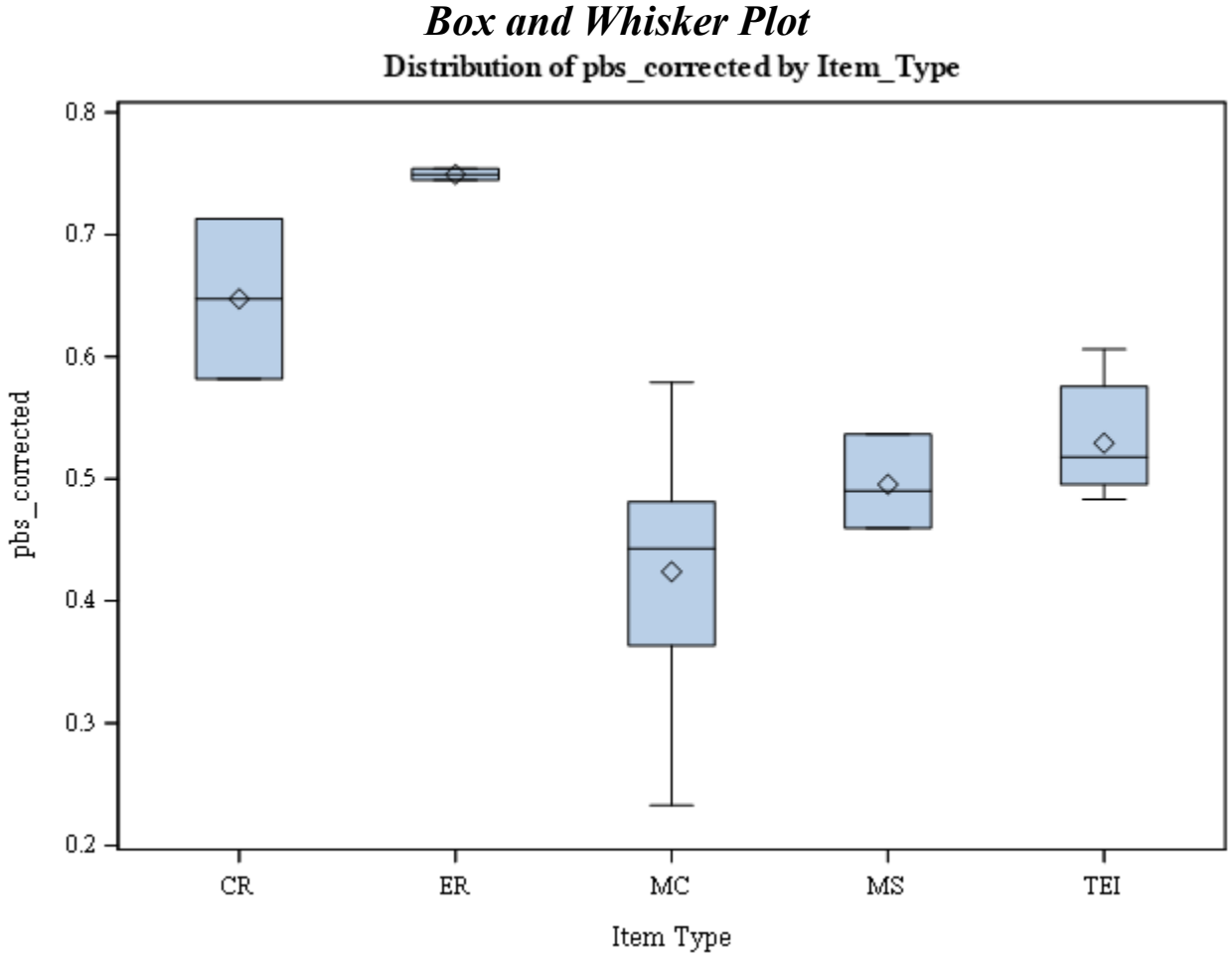


Table B.4.1

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2021 Operational U.S. History

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Standard 2	1	0.615	0.615	0.615	0.615	0.615
	Standards 5&6*	1	0.739	0.739	0.739	0.739	0.739
ER**	Standards 5&6*	1	0.776	0.776	0.779	0.783	0.783
MC	Standard 2	8	0.288	0.359	0.401	0.471	0.527
	Standard 3	5	0.354	0.447	0.490	0.496	0.519
	Standard 4	11	0.264	0.381	0.470	0.517	0.601
	Standards 5&6*	16	0.353	0.429	0.487	0.512	0.595
MS	Standard 2	1	0.559	0.559	0.559	0.559	0.559
	Standard 3	2	0.486	0.486	0.501	0.516	0.516
TE	Standard 2	2	0.528	0.528	0.537	0.547	0.547
	Standard 3	1	0.539	0.539	0.539	0.539	0.539
	Standard 4	2	0.548	0.548	0.579	0.610	0.610
	Standards 5&6*	2	0.525	0.525	0.580	0.635	0.635

* Standards 5 and 6 were combined into one reporting category beginning with the 2018–2019 test administrations, resulting in a redistribution of the points for each reporting category.

** Classical 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 Items by Item Type: Spring 2021 Operational U.S. History

Item Type	*N* OP Items	*N* Items Flagged for *P*-Value	*N* Items Flagged for Point-Biserial Correlation	*N* Items Flagged for DIF*	*N* Items Flagged for Omitting
CR	2	0	0	1	1
ER**	1	0	0	1	0
MC	40	0	0	4	0
MS	3	0	0	1	0
TE	7	0	0	0	0

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

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

Appendix C: Dimensionality

Dimensionality Reports

U.S. History

Contents
Table C.1.1 Intercorrelation Coefficients among Reporting Categories: Spring 2021 Operational U.S. History
Table C.2.1 First and Second Eigenvalues: Spring 2021 Operational U.S. History Figure C.1.1 Principal Component Analysis: Spring 2021 Operational U.S. History

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

Table C.1.1

Intercorrelation Coefficients among Reporting Categories: Spring 2021 Operational U.S. History

Reporting Category	Standard 2	Standard 3	Standard 4	Standards 5&6
Standard 2	1.00			
Standard 3	0.70	1.00		
Standard 4	0.76	0.73	1.00	
Standards 5&6*	0.80	0.76	0.81	1.00

* Standards 5 and 6 were combined into one reporting category beginning with the 2018–2019 test administrations, resulting in a redistribution of the points for each reporting category.

Table C.2.1

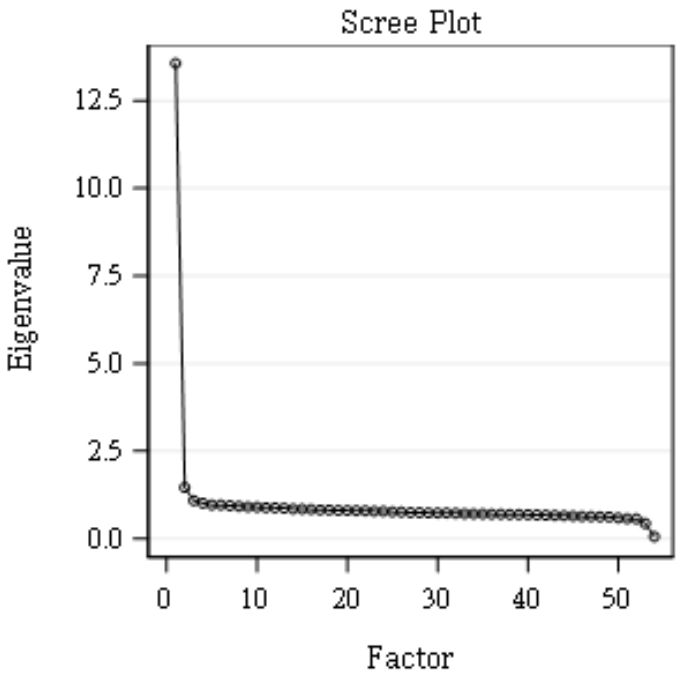
First and Second Eigenvalues: Spring 2021 Operational U.S. History*

First Eigenvalue	Second Eigenvalue
13.569	1.462

* The ratio of first and second eigenvalues is about 9.281.

Figure C.1.1

Principal Component Analysis Plot: Spring 2021 Operational U.S. History



Appendix D: Scale Distribution and Statistical Report

Contents
Table D.1.1 Scale Score Descriptive Statistics and Plots for Spring 2021 U.S. History
Table D.1.2 Frequency Distribution of Scale Scores for Spring 2021 U.S. History

- Because the spring 2021 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 Spring 2021 U.S. History*

DESCRIPTIVE STATISTICS - SCALE SCORES
U.S. HISTORY
ALL STUDENTS
Form ALL

N	≥36190		
Mean	727.05	Median	729.00
Std deviation	34.89	Variance	1217.18
Skewness	-0.0513	Kurtosis	-0.2220
Mode	650.00	Std Error Mean	0.1834
Range	200.00	Interquartile Range	48.00

Quantile	Estimate
100% Max	850
99%	807
95%	782
90%	771
75% Q3	751
50% Median	729
25% Q1	703
10%	683
5%	665
1%	650
0% Min	650

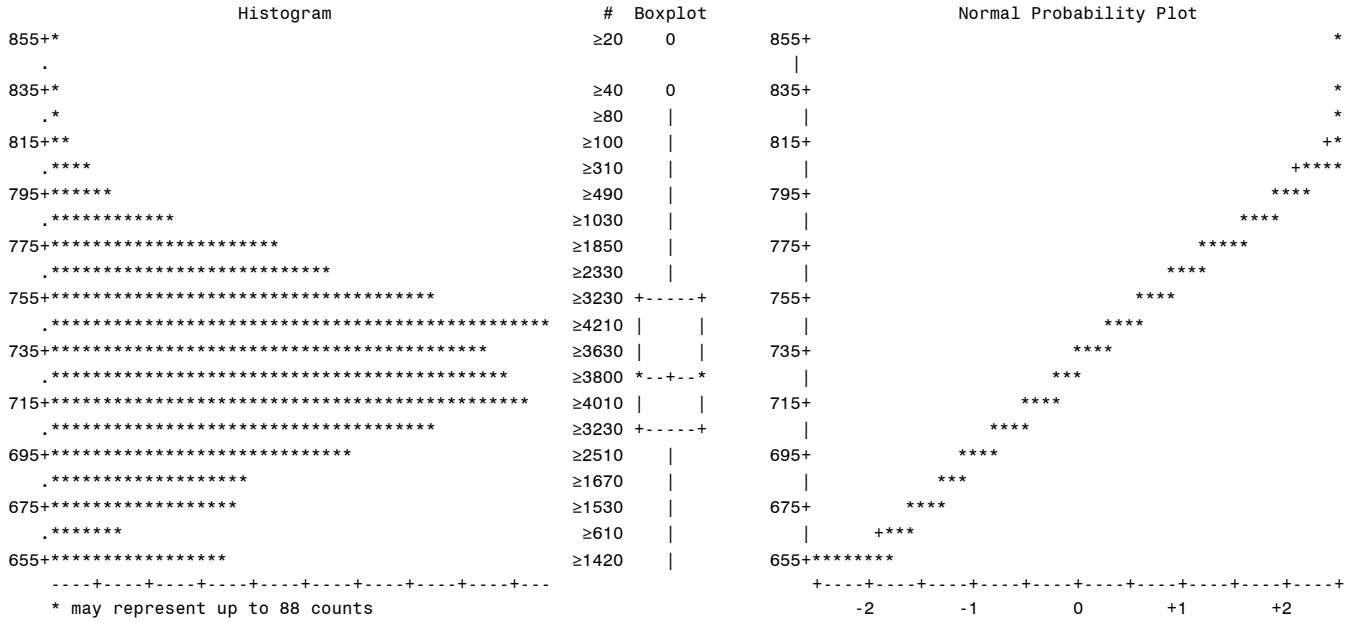


Table D.1.2 *Frequency Distribution of Scale Scores for Spring 2021 U.S. History*

FREQUENCY DISTRIBUTION - SCALE SCORES
U.S. HISTORY
ALL STUDENTS

SCALE_SCORE		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥900	≥900	2.50	2.50
655	*******	≥510	≥1420	1.42	3.93
665	********	≥610	≥2030	1.70	5.63
672	*********	≥720	≥2760	2.01	7.63
678	**********	≥800	≥3560	2.23	9.86
683	**********	≥850	≥4420	2.37	12.23
687	**********	≥810	≥5240	2.25	14.48
691	**********	≥800	≥6040	2.24	16.71
694	**********	≥880	≥6930	2.44	19.15
698	**********	≥820	≥7750	2.28	21.43
701	*********	≥770	≥8530	2.15	23.58
703	**********	≥810	≥9340	2.24	25.81
706	*********	≥770	≥10120	2.15	27.97
708	**********	≥870	≥10990	2.40	30.37
711	**********	≥800	≥11790	2.21	32.58
713	**********	≥820	≥12620	2.29	34.87
715	**********	≥790	≥13410	2.20	37.07
717	**********	≥790	≥14210	2.19	39.26
719	**********	≥790	≥15000	2.20	41.46
721	*********	≥760	≥15760	2.10	43.56
723	*********	≥750	≥16520	2.09	45.66
725	*********	≥740	≥17270	2.07	47.72
727	**********	≥780	≥18050	2.17	49.90
729	*********	≥750	≥18810	2.09	51.98
731	*********	≥770	≥19580	2.13	54.11
732	********	≥700	≥20280	1.94	56.05
734	*********	≥730	≥21020	2.03	58.08
736	********	≥700	≥21720	1.95	60.03
738	********	≥720	≥22450	2.00	62.03
740	********	≥720	≥23170	2.01	64.04
742	********	≥710	≥23890	1.97	66.01
743	********	≥670	≥24560	1.86	67.87
745	********	≥700	≥25260	1.94	69.81
747	********	≥690	≥25950	1.91	71.72
749	*********	≥710	≥26660	1.96	73.68
751	********	≥670	≥27340	1.87	75.55
753	********	≥660	≥28010	1.85	77.39
755	*******	≥620	≥28640	1.74	79.13
757	*******	≥620	≥29260	1.71	80.84
759	********	≥640	≥29900	1.77	82.62
761	*******	≥580	≥30480	1.62	84.24
763	********	≥650	≥31140	1.81	86.05
766	**********	≥540	≥31680	1.50	87.54
768	**********	≥550	≥32240	1.53	89.08
771	**********	≥480	≥32720	1.35	90.42
773	**********	≥500	≥33230	1.39	91.81
776	**********	≥450	≥33680	1.25	93.06
779	**********	≥410	≥34090	1.15	94.21
782	**********	≥360	≥34460	1.02	95.23
785	**********	≥340	≥34810	0.95	96.18
789	**********	≥320	≥35130	0.88	97.06
793	**********	≥240	≥35370	0.68	97.74
797	**********	≥250	≥35620	0.69	98.43
801	*********	≥180	≥35810	0.51	98.95
807	******	≥120	≥35940	0.35	99.30
813	*****	≥100	≥36040	0.29	99.59
821	****	≥80	≥36120	0.22	99.81
832	**	≥40	≥36160	0.11	99.93
850	*	≥20	≥36190	0.07	100.00

100 200 300 400 500 600 700 800 900

Frequency

Appendix E: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

U.S. History

Contents
Table E.1.1 Reliability for Overall and Subgroups: Spring 2021 Operational U.S. History
Table E.2.1 Cronbach's Alpha Reliability: Spring 2021 Operational U.S. History
Table E.3.1 Classification Accuracy and Decision Consistency: Spring 2021 Operational U.S. History

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

Table E.1.1

Reliability for Overall and Subgroups: Spring 2021 Operational U.S. History

Subgroup	Form
All Students	0.941
Female	0.936
Male	0.946
African American	0.922
American Indian or Alaska Native	0.934
Asian	0.943
Hispanic/Latino	0.943
Multi-Racial	0.936
Native Hawaiian or Other Pacific Islander	N/A
White	0.935
Economically Disadvantaged: No	0.934
Economically Disadvantaged: Yes	0.931
EL: No	0.941
EL: Yes	0.902
Gifted or Talented	0.927
Regular Education	0.936
Special Education	0.914
Section 504: No	0.941
Section 504: Yes	0.937

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

Table E.2.1

Cronbach's Alpha Reliability: Spring 2021 Operational U.S. History

Administration	Cronbach's Alpha
Spring 2021	0.941

Table E.3.1

Classification Accuracy and Decision Consistency: Spring 2021 Operational U.S. History

Table E.3.1.1

Estimates of Accuracy and Consistency of Achievement Level Classification

Accuracy	Consistency	PChance	Kappa
0.751	0.663	0.236	0.559

Table E.3.1.2

Accuracy of Classification at Each Achievement Level

Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
0.892	0.565	0.736	0.679	0.772

Table E.3.1.3

Accuracy of Dichotomous Categorizations (PAC Metric)

1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
0.937	0.921	0.930	0.958

Table E.3.1.4

Consistency of Dichotomous Categorizations (PAC Metric)

1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
0.911	0.890	0.902	0.941

Table E.3.1.5

Kappa of Dichotomous Categorizations (PAC Metric)

1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
0.793	0.779	0.753	0.602

Table E.3.1.6

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

1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
0.033	0.038	0.032	0.026

Table E.3.1.7

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

1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
0.029	0.041	0.037	0.017