LEAP 2025 U.S. History Technical Report: 2021–2022

Prepared by DRC, Pearson, and WestEd

EXECUTIVE SUMMARY

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

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

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

Table of Contents

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

The purpose of this technical report is to describe the process for the 2021–2022 operational (OP) administrations of the statewide summative social studies assessment for high school U.S. History. This report outlines the testing procedures, including forms construction, administration, scoring and analyses, and reporting of scores.

Summary of the 2018–2022 Activities

WestEd and Pearson, in partnership with the LDOE and Data Recognition Corporation (DRC), the administration vendor, developed a timeline to capture the major activities necessary to produce the Fall 2021 and Summer 2022 U.S. History operational forms and the spring 2022 operational forms with embedded field test (EFT) items. Table 1.1 summarizes the key activities along with the months during which the activities were completed.

Table 1.1

Key Activities from August 2018 to August 2022

Date	Activity
August 2018–May 2019	• The LDOE and WestEd planned item development and determined item sets and standalone items for revise and re-field test • WestEd and the LDOE worked to revise and develop sources and items
August–December 2019	• Data review of spring 2019 items • The LDOE, WestEd, and DRC prepared operational test form (Form F_S) for fall 2019 • The LDOE and WestEd constructed operational test form (Form G_S) for spring 2020, but the administration did not occur due to the COVID-19 pandemic • Fall test administration occurred
January–May 2020	• Source Review Committees convened • The LDOE staff conducted Source Review Committees • Spring 2020 administration of the LEAP 2025 assessments suspended due to COVID-19 pandemic
June 2020	• Item Content and Bias Review Committees convened
August–December 2020	• The LDOE, WestEd, and DRC repeated intact test form (Form C_S) for fall 2020 • The LDOE, WestEd, and DRC repeated intact test form (Form F_S) for spring 2021 • Fall test administration occurred • BESE authorized the review and revision of the Louisiana Student Standards for Social Studies (LSSSS)
January–May 2021	• Style guide updated • WestEd updated 2020–2021 Assessment Framework • Technical Advisory Committee convened • Spring 2021 test administered
March–August 2021	• The LDOE and WestEd repeated intact test forms (Form B_S, Form F_S) for fall 2021 (windows 1 and 2) • The LDOE and WestEd revisited and selected field test items and constructed field test forms for spring 2022 (Form G)

August 2021	• BESE approved the transition from a high school U.S. History assessment to a high school Civics assessment
November 2021	• Technical Advisory Committee convened
November 2021–January 2022	• Fall test administrations occurred
March–April 2022	• Technical Advisory Committee convened • BESE approved new LSSSS
April–May 2022	• Spring 2022 test administered, including field test items
August 2022	• Data review of spring 2022 items

2. Assessment Framework

The initial assessment framework developed at the start of the project included:

- proposed test designs;
- test blueprints;
- the range of standards and Grade-Level Expectations (GLEs) to be covered;
- reporting categories;
- percentages of assessment items and score points by reporting category;
- projected testing times; and
- the numbers of forms to be administered.

Before the 2021–2022 operational test forms were constructed, the Assessment Framework was updated to reflect any changes to the design and field test plan, as well as to clarify the criteria used to guide item and form selection.

3. Overview of the Development Process

This section describes the processes used to develop field test tasks, item sets, and standalone items to embed within the LEAP 2025 U.S. History assessment.

Item Development Plan

WestEd's proposed item development plans may include tasks, item sets, and standalone items. For the spring 2022 administration, field test items were selected from the items developed in 2019 and 2020. Table 3.1 shows the 2019–2020 revise and re-field test item development plan, and Table 3.2 shows the 2020–2021 item development plan for new and revise and re-field test items for U.S. History.

Table 3.1

Item Development Plan for Revise and Re-field Test Items, 2019–2020

		Total Sets	Total Items per Set	MC/MS	CR	TE	ER	Total Items
2020	Item Sets	4	13–15	49	4	8	–	61
	Standalone Items (MC/MS)	–	–	10	–	–	–	10
	TOTALS	4	–	59	4	8	–	71

Table 3.2

Item Development Plan for New and Revise and Re-field Test Items, 2020–2021

		Total Sets	Total Items per Set	MC/MS	CR	TE	ER	Total Items
2021	Item Sets	6	12–14	56	6	12	–	74
	Tasks	–	–	–	–	–	–	0
	Standalone Items (MC/MS)	–	–	10	–	–	–	10
	TOTALS	6	–	66	6	12	0	84

Key

MC: multiple choice MS: multiple select CR: constructed response
TE: technology enhanced ER: extended response

Proposal and Review of Topics and Sources

Determining Topics

The WestEd content lead reviewed the existing item bank, the LDOE instructional materials, and the U.S. History standards to help determine the content eligible for assessment and what was needed to support the development of the operational assessment. After studying these resources, the content lead made recommendations for which new and revise and re-field test item sets and standalone items should be developed.

When identifying possible topics, the WestEd content lead considers the following:

- Which topics have already been developed and which topics need development
- What content is eligible according to the companion document and scope and sequence document
- Whether proposed topics will support the required item types and number of items, including overage
- How GLEs will be combined to provide a meaningful assessment of content and concepts
- How a topic reflects the LDOE's goal of assessing larger ideas rather than discrete facts

Topics are chosen to represent the breadth of assessable U.S. History content while complementing the balance of topics in the existing pool. The process of choosing assessable GLEs for each topic is iterative and includes the identification of potential GLEs that could be assessed together. It also requires an understanding of the need to create an item pool with the broadest possible content coverage.

Tasks and Item Sets. Tasks and item sets contain multiple, related sources that provide the context from which students answer groups of questions. Sets allow students to delve deeply into a topic. To provide students with opportunities to make connections both

within and across time and place, item sets contain items aligned to different GLEs in a single reporting category, and tasks may include items aligned to GLEs across reporting categories.

Standalone Items. Standalone items assess content that may or may not be connected to a source. A goal in standalone item development is to have a source for 80% of the standalone items to best support students in answering questions. All standalone items are selected-response (SR) items (multiple choice, multiple select). Standalone items are included in the test design to provide greater coverage of the assessable content and GLEs and to provide flexibility in meeting the blueprint and test characteristic curve targets across test administration. Content leads select topics for standalone items based on content and GLEs that may not be sufficiently covered across the sets, with the goal of providing maximum flexibility during test construction. Consequently, the standalone items are typically developed last.

GLE Coverage

By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 35 assessable GLEs associated with Standards 2–6. It also aligned as a secondary alignment at least 1 item to GLEs 1.2, 1.4, and 1.5 that are associated with Standard 1. Although Standard 1 is not part of the reporting category structure, it does contain important content and skills needed to successfully answer items assessed under Standards 2–6. Because of this, many items have a secondary alignment to Standard 1 GLEs, with at least 1 item aligned to GLEs 1.2, 1.4, and 1.5. Having already developed items aligned to all of the assessable GLEs associated with Standards 2–6, WestEd targeted item development to key content during the item development cycles for 2019 and beyond.

Obtaining LDOE Approval for Topics

For tasks and item sets, WestEd submits lists of proposed topics to the LDOE for review prior to new item development. These lists describe the topics and possible related sources so that the LDOE can review and approve them simultaneously. The proposed

topic lists also include the GLEs that might be assessed by the tasks and item sets. Once the LDOE approves the topics to be developed for the development cycle, source searching and development of tasks and item sets begin.

For standalone items, there is no separate approval phase for the topics or sources. However, WestEd and the LDOE have a process to identify the appropriate alignment of the standalone items.

For revised and re-field tested item sets and standalone items, WestEd submits lists of previously developed and field tested item sets and standalone items to the LDOE for review. Working with WestEd and reviewing the field test data, the LDOE determines which sets should be revised, including their sources, and re-field tested.

Identifying Sources

The LEAP 2025 U.S. History assessment focuses on the use of authentic historical and contemporary documents, including letters, speeches, photographs, paintings, reports, and other primary source documents. The assessment also includes secondary source documents, such as authentic newspaper articles and book excerpts. These documents are supplemented by timelines, maps, tables, charts, and graphic organizers created by WestEd's Design Team.

Experienced internal editors locate appropriate sources for tasks, item sets, and standalone items. Before the source searchers begin, WestEd trains them on the search process, on the LDOE's objectives, and on best practices, including bias and sensitivity training. For an outline of the training, see the LEAP 2025 U.S. History Source Search Training Agenda in Appendix A.

All are selected by WestEd staff and reviewed for evaluation for alignment and appropriateness for the approved topics. Based on this evaluation, the WestEd Content lead selects the final sources to propose to the LDOE.

Public Domain versus Permissioned Work. WestEd endeavors to maintain a ratio of 80% royalty-free sources from the public domain or created internally, to a maximum of 20% permissioned work. The actual percentages vary from year to year, depending on the needs of the content in development. Before administration of the assessment, WestEd's permissions coordinator obtains permissions from the rights holders for five years of use of any work that was not in the public domain or created internally.

Evaluating the Readability of Sources. WestEd performs both a Lexile analysis and an ATOS analysis on each passage in the tasks and item sets to obtain a quantitative measure of the readability of the texts. The Lexile Analyzer, developed by MetaMetrics, analyzes the semantic and syntactic features of a text and assigns it a Lexile measure. MetaMetrics also provides grade-level ranges corresponding to Lexile ranges. It should be noted that the grade-level ranges include overlap across grade levels. The ATOS readability tool, developed by Renaissance, also analyzes the reading level of passages. It focuses on elements of text complexity, such as average sentence length, average word length, and word difficulty. Using the Lexile and ATOS measurements provides important statistical information to determine if the passages are grade-level appropriate. Besides the Lexile and ATOS measurements, the *Children's Writer's Word Book* (Mogilner, 2006) and the *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies* (Steck-Vaughn, 1989) are used as additional measures of grade-level appropriateness. WestEd and the LDOE also draw on the professional experience of Louisiana educators during content reviews to verify that sources are accessible to students and make changes based on their feedback.

Most of the sources chosen as part of the development cycles for 2019 and beyond were found to be below or at grade level; however, some of the authentic historical documents were evaluated as above grade level. In those cases, additional support was added, such as footnotes for words that were above grade level and for words or phrases that were thought to be sources of potential confusion for students.

Obtaining LDOE Approval for Tasks, Item Sets, and Sources

As sources for tasks and item sets are reviewed and approved for submission to the LDOE, WestEd content leads finalize set overviews, which outline the content of the sets, identify the GLEs and sources associated with each item, and provide rough drafts of the item stems. WestEd then submits the set overviews and sources to the LDOE for another round of approval before beginning item writing.

For standalone items, WestEd submits the items along with their corresponding sources.

Item Writing and Review Process

WestEd employs item writers and editors for U.S. History. Some of the WestEd writers have been part of item development since the first development cycle in 2016–2017. WestEd secures the required approval from the LDOE for each writer during their first development cycle. Writers and editors receive training from WestEd that outlines lessons learned from previous development cycles, the LDOE's expectations, and best practices for item development, including bias and sensitivity. For an outline of the information covered at the training, see Appendix A for the LEAP 2025 U.S. History Item Editor Training Agenda.

After the training, item writers and editors are provided with approved set overviews or documentation, which identify the set topics, list the GLEs to be addressed, specify the number and type of items to be written, and offer specific guidance about how the content for each item within a set should be assessed. The use of set overviews allows WestEd to control the quality of the tasks and item sets.

Once written, items go through two rounds of content editing, one round of proofreading, and a final round of review before being submitted to the LDOE for their first round of review. The LDOE has two rounds of review prior to content and bias review committee meetings. WestEd revises items based on feedback provided by the LDOE assessment staff.

Item Development Platform. Items are developed in Assessment Banking and Building solutions for Interoperable assessment (ABBI), Pearson's proprietary item development platform. In addition to the items and sources, the platform captures item metadata and allows viewers to preview items using Pearson's format viewer (TestNav 8). In this view, items appear together with their associated sources. The ability to examine the items and sources together is critical in the item review and in the evaluation of the content and cognitive demands on students.

Style Guidelines. The *LEAP Social Studies and Science Content Style Guide* is updated immediately following test construction to reflect final formatting decisions made by the LDOE. Throughout the development and review process, when questions of style arise that are unanswered by existing documentation, WestEd consults the LDOE, and approved changes are added to the Style Guide.

LDOE Content Review. As writing and editing for batches of tasks, item sets, and standalone items are completed, the batches are sent to the LDOE for content lead review. Feedback from the LDOE review is implemented before educator committees convene for content and bias review.

Content and Bias Review Committees. After the completion of item development and the initial rounds of the LDOE review, virtual content and bias review meetings are held. The LDOE recruits educators from different parts of Louisiana, who represent all Louisiana students, to serve on the committees. The meetings are led jointly by facilitators from the LDOE and WestEd. Table 3.3 provides information about the representation of educators who participated in the content and bias reviews in June 2020.

Table 3.3

Representation of Educators Participating in June 2020 Content and Bias Reviews

Grade Level	Number of Committee Participants	Classroom Teacher	Special Education Teacher	Instructional Lead or Supervisor	Visually Impaired Teacher	EL Teacher/ Supervisor
USH	10	5	1	2	1	1

*One of the participants was also a Native American tribal representative.

Training and Security for Virtual Content and Bias Review. The virtual format of content and bias review allows participants to access the item development platform and vote on sources and items individually before coming together in an online meeting format to discuss the items and sources as a group. Prior to accessing the platform, WestEd provides training to explain the content and bias review process and to review the security protocols associated with the virtual pre-review and review. To orient educators to the process, WestEd describes the criteria for evaluating items for content and bias considerations, explains how to use ABBI for item review, and shows educators how to individually review the items and record their recommendation to accept, accept with edits, or reject an item.

Committee members are provided with a pre-review day during which they access the items using ABBI and vote on the items. Comments are compiled and shared with the LDOE and WestEd facilitators prior to the joint virtual committee review. When the committee convenes as a group, the committee members revisit and discuss items and sources. A WestEd recorder takes detailed notes about discussions and records the final committee recommendations. These notes are compiled for reconciliation with the LDOE and post-review implementation. Access to the items is tightly controlled by WestEd, with password access shutting off immediately following the close of each pre-review and review session. At the close of each session, committee members are instructed to clear their internet browser history. In addition, all participants complete a nondisclosure agreement prior to accessing any items.

Results of Content and Bias Review. The results of the reviewers' individual recommendations are captured in ABBI. Table 3.4 provides the results based on the

participants' individual votes following their initial review of the sources and items. Table 3.5 shows the results of the group votes after discussing and reaching a consensus on the disposition of the sources and items.

Table 3.4

Vote Totals Based on Individual Votes Following Initial Review of Sources and Items

Grade	Number of Sources/ Items	Accept	Accept with Edits	No Vote	Reject	Grand Total
USH	84	744	65	4	21	834

* Votes cast as "Accept with Reconciliation" were counted as "Accept with Edits" since this vote was not used during this round of review.

Table 3.5

Vote Totals for Items Based on Group Consensus for Sources and Items

Grade	Number of Sources/Items	Accept	Accept with Edits	No Vote	Reject
USH	84	59	25	0	0

Post-Committee Finalization. At the conclusion of the content and bias reviews, WestEd content leads consult the LDOE to reconcile any unresolved committee feedback. Following the implementation of the committee's feedback, the LDOE and WestEd content leads meet virtually for final item reconciliation. WestEd provides records of all the implemented changes to the LDOE prior to the virtual reconciliation meetings. During the reconciliation meetings, the leads review the items to ensure that they were correctly edited. Once content considerations are resolved, all items and sources go through a final formal fact-checking round and two additional rounds of proofreading. Any changes resulting from these reviews are submitted to the LDOE for approval.

Data Review Process and Results

During data review of EFT items, content experts and psychometric support staff review field tested items with accompanying data to make judgments about the appropriateness

of items for use on operational test forms. Statistically flagged items are not rejected on the sole basis of statistics; only items with identifiable flaws are rejected.

The data review meetings begin with presentation of the general guidelines for reviewing data. The presentation includes a review of item statistics (difficulty, discrimination, DIF, score distributions), appropriate interpretations and inferences, what would be considered reasonable values, and how the values might differ across different item types.

Facilitators from WestEd and Pearson lead the data review. Statistical information for each item is evaluated to determine whether the item functions as intended. Each item's suitability for future operational tests is then evaluated in the context of field test statistics. Judgments to accept, accept with edits (or "revise/field test"), or reject are then recorded. If the decision is to edit or to reject an item, additional information is captured to document the reason for the decision. Table 3.6 summarizes the decisions by item type for data-reviewed items field tested in spring 2022.

Table 3.6
FT Item Decisions by Item Type, 2022 Data Review

Item Type		Number of Items			
	Field Tested	Accepted	Accept with Edits	Reject	% of Total
MC	102	97	-	5	66.67
MS	15	15	0	0	9.80
TE	24	22	0	2	15.68
CR	12	11	0	1	7.84
ER	–	–	–	–	–
Total	153	145	0	8	100.00

Note: % of Total means percent of total # of items.

4. Construction of Test Forms

Initial Construction

The purpose of the forms construction activities is to create operational (OP) forms and to embed field test items for potential use in future OP assessments. This section describes the process used to create operational and field test forms.

2021–2022 Operational Forms

WestEd and the LDOE content staff worked together to complete item selection for one new form (Form G). This form was originally selected for the spring 2020 OP administration. It was not used because of the suspension of that administration due to the COVID-19 pandemic. The decision was made to use the form for the spring 2022 OP administration.

To construct the form, content specialists drew from a pool that included data of review-approved items from previous embedded field tests and operational administrations. WestEd submitted the form to Pearson psychometricians for consideration before formal submission to the LDOE. The OP form was designed to adhere to the blueprint for U.S. History and exhibit the broadest possible balance of content and breadth of GLE and content coverage. The task was selected first, followed by item sets with CRs, other item sets, and standalone items. Test form developers worked to avoid cueing and clanging between items. Cueing occurs when the content in one item provides clues to the answer of another item. Clanging refers to the overlap or similarity of content. Because the content was purposely distributed across sessions, cueing and clanging were intended to have been avoided; however, developers also conducted a separate review of the forms to check for inadvertent cueing and clanging. During item selection, test maps were created to capture details of the forms, including each item's unique identification number (UIN), test session, item sequence, item descriptions, and associated item metadata. Table 4.1 provides the test composition for the U.S. History spring 2022 OP form.

Table 4.1

U.S. History Test Composition for Spring 2022 OP Form

Sets and Standalone Items	Total Sets	Total Items per Set	Total Points per Set	SR	CR	TE	ER	Total Items	Total Points
6-Item Set with TE and CR	2	6	8	8	2	2	0	12	16
5-Item Set	5	5	6	20	0	5	0	25	30
Standalone Items	0	0	0	11	0	0	0	11	11
Task	1	5	12	4	0	0	1	5	12
Total	8			43	2	7	1	53	69

Table 4.2 provides the number of total points and points by item type for each standard and reporting category as well as the standards and reporting categories assessed by the task for the tests administered in 2021–2022. The table also shows the number of points excluding the task and the CRs, which are not included in the reporting category percentages for the blueprint because the standards addressed by the task and the CRs may vary by form.

Table 4.2

U.S. History Operational Test Composition for the 2021–2022 Forms shown by Order of Administration (Fall 2021 Window 1, Fall 2021 Window 2, Spring 2022 OP)

Standard	Task Alignment	SR	CR	TE	ER	Total Points
2. Western Expansion to Progressivism	√/×/×	10/9/9	2/2/2	4/4/4	0/0/0	16/15/15
3. Isolationism through the Great War	×/×/×	6/7/9	0/0/0	2/2/2	0/0/0	8/9/11
4. Becoming a World Power through World War II	√/√/×	9/11/12	0/0/2	6/4/4	0/0/8	15/15/26
5. & 6. Cold War Era and the Modern Era	√/√/√	14/16/13	2/2/0	6/4/4	8/8/0	30/30/17
Total Points Excluding Task and CRs		37/39/39	0/0/0	16/14/14	0/0/0	53/53/53
Total Points		41/43/43	4/4/4	16/14/14	8/8/8	69/69/69

Spring 2022 Field Test Forms

Twelve item sets were field tested in spring 2022. Sets were placed on multiple field test forms, with different combinations of items on each form to ensure field testing of the maximum number of items in each set. Twenty-one standalone items were embedded across twenty-four field test forms. Each form included one item set with 4 SR, 1 TE, and 1 CR and three standalone items. Standalone items were repeated on field test forms as necessary to fill all available positions.

Revision and Review

Psychometric Approval of Operational Forms

Prior to submitting the forms to the LDOE staff for review, Pearson psychometricians and WestEd content specialists participate in an iterative process of reviewing and revising the forms. The psychometric review consists of comparisons of the expected representation and the actual representation of reporting categories (Standards 2–6) and item types—selected response (SR), constructed response (CR), technology enhanced (TE), and extended response (ER)—on the operational forms. The answer keys for multiple-choice (MC) items also are examined, to determine whether any forms have significantly non-uniform distributions of correct responses (A, B, C, and D). Spreadsheets are used to generate frequency tables of reporting categories, item types, and MC answer keys for each form. They are also used to compare the operational forms from previous years. Deviations from the blueprint are identified and addressed. Test characteristic curves (TCC) based on item response theoretic models are applied to data, and conditional standard errors of measurement are computed for each iteration during the test construction process to evaluate how well a proposed test form matches psychometric targets. Psychometric approval from Pearson is provided for all the forms prior to submission to the LDOE for their review.

Table 4.3

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

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

Note: Detailed information can be found in the 2018–2019 Framework and Test Construction Document. It should be noted that these values are psychometric recommendations. Actual item decision occurs by content staff based on these recommendation criteria.

LDOE Review

Following the psychometric reviews, the test maps and constructed sets are delivered to the LDOE for approval. Forms are reviewed by both the LDOE content and psychometric staff. Based on the LDOE review, sets or items are replaced and resequenced as requested. After these changes, the overall balance of answer choices and key runs is re-evaluated, and final adjustments are made to achieve the appropriate balance. Finalized test maps are used to create PDF versions of forms, or constructed sets, which are reviewed by WestEd's proofreaders before the items are transferred from ABBI to DRC.

Online and Paper Versions

All forms are delivered online. One form is designated by the LDOE as the accommodated version to be used with students who have accommodations. The accommodated version is available in print form to students who require paper testing. The accommodated version is also rendered in braille. To support students with low or no vision, additional text (alternate text) is provided to describe the graphic components of the assessment. Content specialists evaluate the graphics and draft the alternate text.

5. Test Administration

This chapter describes processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) *Standards for Educational and Psychological Testing* (hereafter the *Standards*), "The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions" (111). This chapter examines how test administration procedures implemented for the Louisiana 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.

Table of Contents for *LEAP 2025 High School Test Administration Manual* (TAM)

a. Notes and Reminders
b. Pre-Administration Oath and Security Confidentiality Statement
c. Post-Administration Oath and Security Confidentiality Statement
d. Overview
e. Test Security
    i. Secure Test Materials
    ii. Testing Irregularities and Security Breaches
    iii. Testing Environment
    iv. Violations of Test Security
    v. Voiding Student Tests
f. Test Administrator Responsibilities
    i. Software Tools and Features for Test Administrators
g. Test Administration Checklists
    i. Before Testing
    ii. During Testing
    iii. After Testing (Daily)
    iv. After Testing (Last Day)
h. Test Materials
    i. Receipt of Test Materials
i. Testing Guidelines
    i. Testing Eligibility

    ii. Testing Schedule
    iii. LEAP 2025 Testing Time
    iv. Extended Time for Testing
    v. Makeup Test Procedures
    vi. Testing Conditions
    vii. Accessibility Features
j. Special Populations and Accommodations
    i. IDEA Special Education Students
    ii. Students with One or More Disabilities According to Section 504
    iii. Gifted and Talented Special Education Students
    iv. Test Accommodations for Special Education and Section 504 Students
    v. Special Considerations for Students Who Are Deaf or Hearing Impaired
    vi. English Learners (ELs)
k. Directions for Administering the LEAP 2025 Tests
l. LEAP 2025 Testing Times
m. General Instructions for LEAP 2025
    i. Reading Directions to Students
    ii. LEAP 2025 English I and English II
    iii. LEAP 2025 Algebra I and Geometry
    iv. LEAP 2025 Biology
    v. LEAP 2025 U.S. History
n. Post-Test Procedures
    i. Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    ii. Returning Test Materials to the School Test Coordinator
o. 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.

Table of Contents for *Test Coordinators Manual* (TCM)

5. Student Tutorials
6. Online Tools Training (OTT)

12. Post-Administration Rescoring Process for LEAP 2025 HS Assessments
13. Request for Rescoring
14. Void Notification

The 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.

Standard 4.15. The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented. (90)

The 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 TAM describe the following: general rules of online testing; assessment duration, timing, and sequencing information; and the materials required for testing.

Standard 6.1. Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (114)

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the LEAP 2025 tests were administered according to the prescribed TAM. 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 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, 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 2021–2022 assessments for HS courses were administered to students within the state testing windows of November 30–December 17, 2021, or January 5–24, 2022, for fall administration, April 14–May 13, 2022, for the spring administration, and June 20–24, 2022, for the 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.

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.

6. Scoring Activities

Directory of Test Specifications (DOTS) Process. DRC creates a DOTS file, based on the approved test selection. The DOTS is a document containing information about each item on a test form, such as item identifier, item sequence, answer key, score points, 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 the LDOE and WestEd content staff for key verification. The staff reviews the list of flagged MC and MS items to confirm that the answer keys are accurate. The scoring of MC and MS items is also evaluated at data review.

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

- A scoring rubric is created for each TE item. The 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, the 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 2022 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 2021–2022 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 the 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 6.1–6.6 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the 2021–2022 forms.

Table 6.1

Inter-Rater Reliability for Operational Constructed-Response Items

Administration	Item	Inter-Rater Reliability*				
		2x	Total	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2021 Window 1	Item 1	≥4,410	≥10,960	97	3	0
	Item 2	≥4,380	≥11,080	93	7	0
Fall 2021 Window 2	Item 1	≥1,660	≥3,840	96	4	0
	Item 2	≥1,420	≥3,810	95	5	0
Spring 2022	Item 1	≥12,670	≥42,790	87	13	0
	Item 2	≥14,130	≥43,490	90	10	0
Spring 2022 Senior	Item 1	≥3,940	≥1,390	93	7	0
	Item 2	≥3,850	≥1,340	96	3	0
Summer 2022	Item 1	≥3,060	≥1,820	99	1	0
	Item 2	≥2,940	≥1,490	97	3	0

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

Table 6.2

Score Point Distributions for Operational Constructed-Response Items

Administration	Item	Score Point Distribution*					
		Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
Fall 2021 Window 1	Item 1	≥10,960	61	8	4	0	27
	Item 2	≥11,080	27	39	8	0	25
Fall 2021 Window 2	Item 1	≥3,840	37	29	13	0	31
	Item 2	≥3,810	37	29	10	0	23
Spring 2022	Item 1	≥42,790	36	33	17	0	13
	Item 2	≥43,490	38	33	13	0	17
Spring 2022 Senior	Item 1	≥3,940	28	37	7	1	28
	Item 2	≥3,850	60	7	3	1	30
Summer 2022	Item 1	≥3,060	40	8	1	0	50
	Item 2	≥2,940	35	21	1	0	41

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

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

Table 6.3

Inter-Rater Reliability for Operational Extended-Response Items

Administration	Item	Inter-Rater Reliability*					
		2x	Total	Dimension	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Fall 2021 Window 1	Item 1	≥9,050	≥13,430	Content	96	4	0
				Claims	96	4	0
Fall 2021 Window 2	Item 1	≥3,080	≥4,660	Content	96	4	0
				Claims	97	3	0
Spring 2022	Item 1	≥52,590	≥63,400	Content	97	3	0
				Claims	96	4	0
Spring 2022 Senior	Item 1	≥4,160	≥5,320	Content	98	2	0
				Claims	98	2	0
Summer 2022	Item1	≥3,400	≥2,390	Content	96	4	0
				Claims	98	2	0

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

Table 6.4

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 2021 Window 1	Item 1	≥13,430	Content	31	33	15	5	2	0	13
			Claims	38	27	14	6	2	0	13
Fall 2021 Window 2	Item 1	≥4,660	Content	31	31	14	7	1	0	15
			Claims	38	25	14	6	1	0	15
Spring 2022	Item 1	≥63,400	Content	21	39	20	9	3	0	9
			Claims	21	39	19	9	3	0	9
Spring 2022 Senior	Item 1	≥5,320	Content	41	27	6	1	1	1	22
			Claims	53	18	4	1	0	1	22
Summer 2022	Item 1	≥3,400	Content	47	19	2	0	0	1	30
			Claims	54	12	2	0	0	1	30

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

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

Table 6.5

Constructed-Response Inter-Rater Reliability for Field Test Items

Administration	Item	Inter-Rater Reliability				
		Total	2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
Spring 2022 FT	1	≥1,690	≥390	82	18	0
	2	≥1,750	≥500	93	7	0
	3	≥1,730	≥460	81	19	0
	4	≥1,770	≥550	94	6	0
	5	≥1,700	≥400	77	22	1
	6	≥1,740	≥480	84	16	0
	7	≥1,750	≥500	90	9	1
	8	≥1,710	≥420	89	11	0
	9	≥1,720	≥450	79	21	0
	10	≥1,770	≥550	88	12	0
	11	≥1,700	≥400	73	25	2

Note: Total Exact + Adjacent + Non-adjacent does not always add up to 100% due to rounding.

Table 6.6

Constructed-Response Score Point Distributions for Field Test Items

Admin	Item	Total	"0" Rating (%)	"1" Rating (%)	"2" Rating (%)	Blank (%)	Nonscore Codes (%)**
Spring 2022 FT	1	≥1,690	43	40	10	0	6
	2	≥1,750	54	26	8	0	11
	3	≥1,730	31	47	12	0	10
	4	≥1,770	56	20	6	1	16
	5	≥1,700	26	49	19	0	5
	6	≥1,740	60	21	5	0	14
	7	≥1,750	50	29	9	0	13
	8	≥1,710	56	26	7	0	10
	9	≥1,720	33	42	14	0	9
	10	≥1,770	32	34	20	0	15
	11	≥1,700	50	26	17	0	7

7. Data Analysis

Classical Item Statistics

This section describes the classical item analysis for data obtained from the operational LEAP 2025 HS U.S. History. 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 additional information on classical item statistics for the spring 2022 test can be found in Appendix C: Item Analysis Summary Report. Tables C.1–C.4 and C.6 show the summaries of classical item statistics. As a measure of item difficulty, p (or "the p-value") indicates the average proportion of total points earned on an item. For example, if p = 0.50 on an MC item, then half of the examinees earned a score of 1. If p = 0.50 on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). A measure of point-biserial correlation indicates the correlation between an item score and the total test score. Items with higher item-total correlations provide better information about how well items discriminate between lower- and higher-performing students. It should be also noted that a corrected point-biserial correlation indicates the correlation between an item score and the total test score, where the item score is not included in the total score. The results can be found in Tables C.1-C4. By the way, the statistical analysis results for field test (FT) items are stored in Pearson's Assessment Banking and Building solutions for Interoperable assessment (ABBI) system.

Differential Item Functioning

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

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

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

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

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

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

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

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

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

The *ΔMH DIF* is then computed as

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

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

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

Table 7.1
DIF Categories for Dichotomous Items

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

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

average of D_m using the weights $w_m = N_{fm}$ (the number of students in the focal group with raw score m), which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

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

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

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

Table 7.2
DIF Categories for Polytomous Items

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

Three DIF analyses were conducted for the operational test items only: female/male, black/white, and Hispanic/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly,

given their performance on the full assessment. The same methods were used to detect items on which both black/white and Hispanic/white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Table 7.3 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. Neither analysis flagged any of the operational test items for C-DIF.

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

Table 7.3
Summary of DIF Flags: Spring 2022 U.S. History Operational Items

Comparison Groups	A	[B+],[B-]	[C+],[C-]
Female – Male	50	[1],[2]	[0],[0]
African American – White	53	[0],[0]	[0],[0]
Hispanic – White	51	[1],[1]	[0],[0]

Measurement Models

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

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

In that model, $p_i(\theta_j)$ is the probability that student *j* would earn a score of 1 on item *i*, b_i is the difficulty parameter for item *i*, a_i is the slope (or discrimination) parameter for item *i*, c_i is the pseudo-chance (or guessing) parameter for item *i*, and *D* is the constant 1.7.

Since the U.S. History test also included polytomous items scored higher than 1 point, the generalized partial credit model (GPCM) (Muraki, 1992) was used to estimate the parameters of these items:

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

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

Calibration and Linking

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

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

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

Thus, the primary purpose of form equating is to establish score equivalency between two (or more) forms. Equivalency is established by first building the forms to be equated according to content specifications. Then the form scores are placed on the same scale (by equating), such that students performing on two scaled assessments at the same level

of underlying achievement should receive the same scale score on both forms, although they may not receive the same number-correct score (or raw score). The LDOE and Pearson strive to maintain equivalent samples or use near-census samples over the years, minimizing the potential differences caused by the different samples.

It should be noted that the spring 2018 test is the first operational administration for U.S. History, and in the spring of 2021, the forms used were intact and were originally administered in 2019.

Table 7.4 provides scale scores at selected percentiles that can be used to compare the distributional characteristics of the spring 2022 test form to previous administrations. Although these scale scores are rounded values, there were differences in the scale score values for a given percentile across the forms. These variations could arise for several reasons: (1) differences in the proficiency (i.e., achievement) of the students in the samples or growth in student achievement across years; (2) unevenness in the respective distributions that combine with the number-correct-to-scale-score scoring method, leaving “gaps” in the scale; or (3) other sources of equating error. In general, however, the test characteristic function equating techniques will “level” the equated forms through the raw-to-scale-score adjustment.

Table 7.4

Comparisons of Scale Scores at Selected Percentiles: U.S. History Operational Forms

Percentile	2018 Spring Form B	2018 Spring Form C	2018 Spring Form D	2018 Spring Form E	2019 Spring Form F	2021 Spring Form F	2022 Spring Form G
99	805	805	805	800	801	807	803
95	781	781	781	781	782	782	782
90	770	770	770	770	771	771	772
85	763	763	763	763	763	763	763
80	756	758	756	758	759	757	757
75	752	754	752	752	753	751	751
70	746	748	748	748	749	747	748
65	742	744	744	744	745	742	742
60	738	740	740	740	740	736	738
55	733	735	735	735	736	732	734
50	729	731	731	731	732	729	730
45	725	729	727	727	729	723	726
40	721	725	723	723	725	719	721
35	716	721	719	719	719	715	717
30	714	716	714	714	715	708	712
25	707	712	709	709	711	703	706
20	701	707	704	704	706	698	701
15	696	699	699	699	698	691	694
10	685	692	689	689	691	683	686
5	671	676	676	676	672	665	670
1	650	650	650	650	650	650	650

Operational Item Parameters

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

Item Fit

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

It should be noted that Yen's Q_1 statistic (Yen, 1981) was calculated to evaluate item fit for both operational and field test items by comparing observed and expected item performance. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. For dichotomous items, Q_1 is computed as

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

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

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

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

The generalization of Q_1 for items with multiple response categories is

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

where

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

Both Q_1 and generalized Q_1 results are transformed to ZQ_1 and are compared to a criterion $ZQ_{1,crit}$ to determine whether fit is acceptable. The conversion formulas are

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}}$$

and

$$ZQ_{1,crit} = \frac{N}{1500} * 4,$$

where *df* is the degrees of freedom (the number of intervals minus the number of independent item parameters). Items are categorized as exhibiting either fit or misfit.

A summary of IRT item parameter statistics and item fit for operational items is displayed in Appendix D: Dimensionality.

Dimensionality and Local Item Independence

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

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

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

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

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

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

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

With n items, there are $n(n - 1)/2$ Q_3 statistics. If an assessment consists of 48 items, for example, there are 1,128 Q_3 values. The Q_3 values should all be small. Summaries of the distributions of Q_3 are provided in Appendix D: Dimensionality. Specifically, Q_3 data are summarized by minimum, 5th percentile, median, 95th percentile, and maximum values

for LEAP 2025 U.S. History. To add perspective to the meaning of Q_3 distributions, the average zero-order correlation (simple intercorrelation) among item responses is also shown. If the achievement construct accounts for the relationships between items, Q_3 values should be much smaller than the zero-order correlations. The Q_3 summary tables in the dimensionality reports in Appendix D show for the 2022 U.S. History test that at least 90% (between the 5th and 95th percentiles) of the items are expectedly small. These data, coupled with the Q_1 data, indicate that the unidimensional IRT model provides a reasonable solution to capture the essence of student science achievement defined by the selected set of items for each grade level.

Scaling

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

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

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

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

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

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

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

which are rearranged as

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

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

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

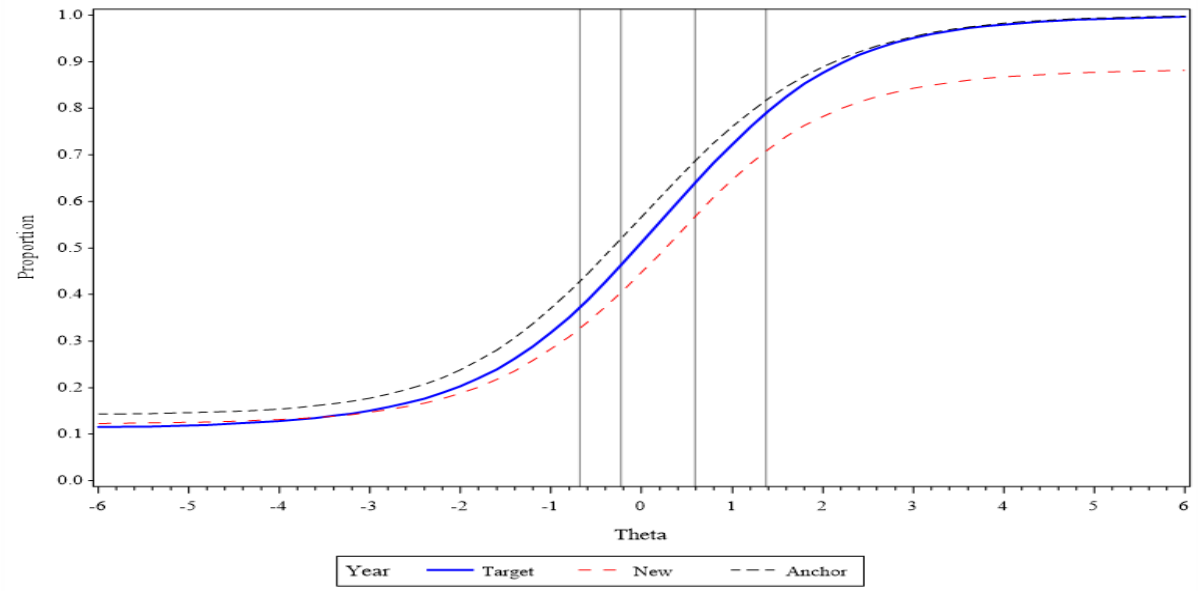
The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same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

Test Characteristic Curve

Additional evidence of comparability can be found by reviewing the test characteristic curves (TCCs) across administrations of the LEAP 2025 assessments, as can be seen in the following figure. As seen from Plot 7.1, the TCCs between two years were similar across ability ranges. By the way, Plot 9.1 also indicates that the SEMs between two years are similar across ability ranges, especially in the middle ability ranges; each theta cut matches the scale score of each performance-level cut (i.e., 711, 725, 750, and 774).

Plot 7.1

Test Characteristic Curve: Spring 2022 Operational U.S. History



Note: The scale is on theta. Each theta cut matches the scale score of each performance cut: 711, 725, 750, and 774; Target = 2018 Operational Form; New = 2022 Operational Form; Anchor = Anchor Form.

Test Information Curve, Score Distribution, and IRT Difficulty Distribution

In this section, students' U.S. History score distribution, IRT item difficulty (i.e., b-parameter) distribution, and item information curve are presented. Compared to the base year (i.e., 2018 U.S. History test), the 2022 U.S. History test provides more test information around the middle range of theta than other rages, as can be observed from Table 7.5 and Plot 7.2.

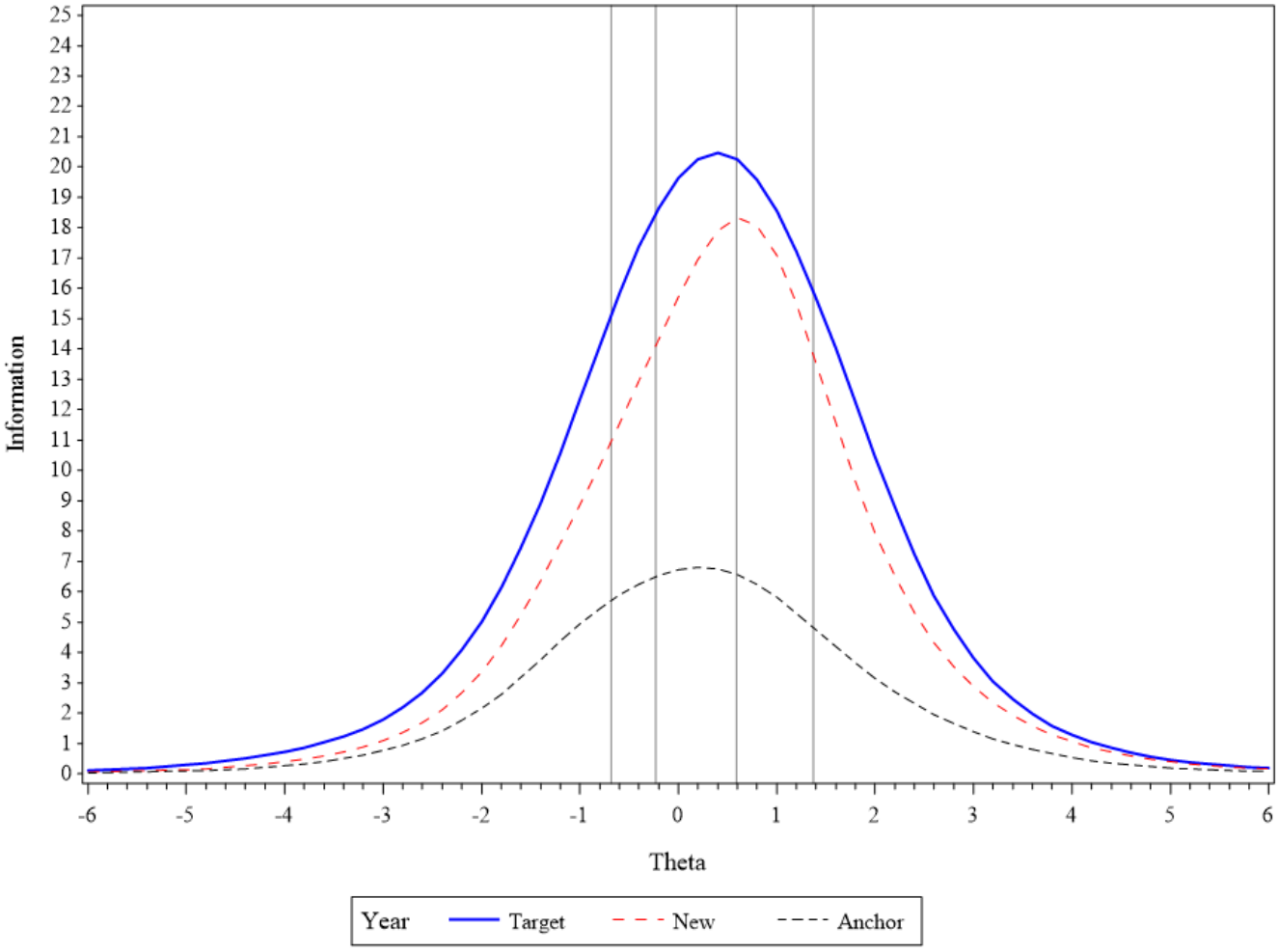
Table 7.5

Students' Score and IRT B-Parameter Distribution: Spring 2022 Operational U.S. History

Percent of Students' Theta	Theta Range	Number of Items of IRT-B
1.10	theta < -3.5	0
0.79	-3.5 <= theta < -3.0	0
1.05	-3.0 <= theta < -2.5	0
3.02	-2.5 <= theta < -2.0	0
3.95	-2.0 <= theta < -1.5	0
11.85	-1.5 <= theta < -1.0	3
12.51	-1.0 <= theta < -0.5	7
17.29	-0.5 <= theta < 0.0	8
18.06	0.0 <= theta < 0.5	11
14.23	0.5 <= theta < 1.0	20
9.50	1.0 <= theta < 1.5	4
4.29	1.5 <= theta < 2.0	1
1.44	2.0 <= theta < 2.5	0
0.75	2.5 <= theta < 3.0	0
0.12	3.0 <= theta < 3.5	0
0.05	3.5 <= theta	0
-6.00	**Minimum**	-1.37
6.00	**Maximum**	1.88
-0.14	**Mean**	0.26
1.20	**SD**	0.68
≥35,950	**Total**	54

Plot 7.2

Test Information Curve: Spring 2022 Operational U.S. History



Note: The scale is on theta. Each theta cut matches the scale score of each performance cut: 711, 725, 750, and 774; Target = 2018 Operational Form; New = 2022 Operational Form; Anchor = Anchor Form.

Field Test Data Review

The process used to complete the field test item equating is an anchored item equating process. In this process, the item parameters from the operational items from the 2018 administration were fixed as constant (i.e., to calculate Stocking-Lord equating constant) and the item parameters for the field test items were freely calibrated, placing the item parameters for the field test items on the same scale as the operational items.

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

8. Test Results and Score Reports

This chapter provides information on the results of the spring LEAP U.S. History test. The scale score results and achievement level information are also presented here. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative description of student achievement. The levels are Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory. The results in Table 8.1 are presented as evidence of the reliability and validity of the scores from the LEAP 2025 U.S. History assessment.

Demographic Characteristics of Students

The operational U.S. History assessment was administered to all eligible students in the appropriate grade level during the first administration window in spring 2022. Spring 2022 operational score results were reviewed based on the following student characteristics:

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

Test Results

For the spring 2022 U.S. History test, the lowest obtainable scale score (LOSS) on the tests is 650 and the highest obtainable scale score (HOSS) is 850. Scale score means and

standard deviations as well as the percentages of students in each performance level are reported for the state and disaggregated into various demographic groups. In addition to the descriptive statistics presented in Table 8.1, scale score frequency distributions are presented in Appendix E: Scale Distribution and Statistical Report. Finally, because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table 8.1

LEAP 2025 State Test Results: Spring 2022 Operational U.S. History

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥35,950	728.58	33.83	29	15	27	19	9
Gender								
Female	≥18,200	728.94	31.89	27	17	29	19	8
Male	≥17,750	728.21	35.71	31	14	25	20	10
Ethnicity								
African American	≥14,510	714.69	30.88	44	18	24	11	3
American Indian or Alaska Native	≥220	728.29	30.22	27	15	32	19	7
Asian	≥700	756.51	37.07	11	6	20	29	33
Hispanic/Latino	≥2,560	724.14	35.48	34	13	26	19	7
Two or More Races	≥840	733.65	31.75	23	17	28	23	10
Native Hawaiian or Other Pacific Islander	≥20	736.09	36.09	17	13	30	30	9
White	≥17,070	739.65	31.10	17	13	31	26	13
Economically Disadvantaged*								
No	≥14,000	743.29	31.47	14	12	29	28	16
Yes	≥20,780	719.33	31.78	39	17	26	14	4

Table 8.1 (continued)

	Scale Score			% at Performance Level				
	N	Mean	SD	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
English Learner								
No	≥34,970	729.45	33.55	28	15	28	20	9
Yes	≥980	697.70	29.09	68	12	15	4	NR
Education Classification								
Gifted/Talented	≥2,310	763.67	30.44	4	5	19	32	39
Regular	≥30,740	728.74	31.65	27	16	29	20	7
Special	≥2,890	698.79	31.02	68	13	12	5	2
Section 504								
No	≥32,690	729.54	33.65	28	15	28	20	9
Yes	≥3,260	719.02	34.14	42	15	23	13	6
Migrant								
No	≥35,900	728.61	33.83	29	15	27	19	9
Yes	≥50	713.96	34.98	47	17	19	9	8
Homeless Status								
No	≥35,420	728.74	33.85	29	15	27	20	9
Yes	≥530	718.10	30.90	40	20	25	11	4
Military Affiliation								
No	≥35,460	728.39	33.79	29	15	27	19	9
Yes	≥490	742.73	33.72	15	12	28	26	18
Foster Care Status								
No	≥35,890	728.62	33.83	29	15	27	19	9
Yes	≥60	710.23	33.01	56	10	20	13	2

* Economic status was not available for all students.

Effect Size

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

$$d = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}}$$

where is $\bar{x}_a$ is the mean score of group A, $\bar{x}_b$ is the mean score of group B, s_a^2 is the variance of group A, s_b^2 is the variance of group B, n_a is the number of students in group A, and n_b is the number of students in group B.

Cohen's *d*, then, expresses the difference in group means in terms of the standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the *d* statistic: *d* = 0.20 is a small ES, *d* = 0.50 is a medium ES, and *d* = 0.80 is a large ES. Based on Cohen's (1988) guidelines, certain trends are observable in Table B.6 in Appendix B. Although no big difference in U.S. History test scores was seen between females and males, mean raw scores and ESes show that Asian and White students tend to outperform other ethnicity groups. There were clear performance differences among regular education, gifted/talented education, and special education students in Education Classification and Non-English Learner and English Learner in EL status. Performance differences were also observed from Economically Disadvantaged status, Homeless status, Foster Care status, and Military Affiliation status.

Uses of Test Scores

To understand whether a test score is being used properly, one must understand the purpose of the test. The intended uses of the LEAP 2025 test scores include the following:

- evaluating students' overall proficiency of the Louisiana Student Standards
- identifying students' strengths and weaknesses
- evaluating programs at the school, school system, and/or state level

- informing stakeholders, including students, teachers, school administrators, school system administrators, LDOE staff members, parents, and the public, of the status of students' progress toward meeting college and career readiness standards.

This technical report refers to the uses of the test-level scores (i.e., scale scores and achievement levels), and reporting category-level scores and achievement level classifications.

Score Reports

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

School Roster Report. A School Roster Report, which provides summary information about student performance on the LEAP 2025 high school U.S. History assessment, is available to school systems and schools through eDIRECT. Total test scores and achievement level indicators are shown for the test of interest. Category and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by category and subcategory are summarized. More details can be found in the *LEAP 2025 High School Interpretive Guide (iGUIDE) 2021–2022*.

Individual Student-Level Report. The ISR is another type of report available through the eDIRECT system. ISRs may be downloaded and printed by schools to be sent home to parents. At the top of the page, overall student performance is reported by scale score and achievement level. In the middle of the page, category and subcategory performance

indicators are reported. When a student does not receive a scale score, their achievement level will be left blank. ISRs for students whose scores were invalidated will display a blank scale score for a given course.

LEAP 2025 High School Interpretive Guide (iGUIDE) 2021–2022. The *LEAP 2025 High School Interpretive Guide (iGUIDE) 2021–2022* was written to help school administrators, teachers, and parents in the Louisiana school system, and the general public, understand the LEAP 2025 U.S. History test. The *LEAP 2025 High School Interpretive Guide (iGUIDE) 2021–2022* was developed collaboratively by the LDOE and DRC staff. The LDOE staff had opportunities to review the guide, provide feedback, and give final approval. The elements of the table of contents are provided below:

- Introduction to the Interpretive Guide
    - Overview
        - Purpose of the Interpretive Guide
    - Test Design
    - Scoring
        - Item Types and Scoring
    - Interpreting Scores and Achievement Levels
        - Scale Score
        - Achievement Level Definitions
        - Student Rating by Reporting Category and Subcategory
- Student-Level Reports
    - Sample Student Report: Explanation of Results and Terms
    - Sample Student Report A
    - Sample Student Report B
    - Parent Guide to the LEAP 2025 High School Student Reports
- School Roster Report
    - Sample School Roster Report: Explanation of Results and Terms
    - Sample School Roster Report

Achievement Level Policy Definitions and Cut Scores

Achievement level policy definitions for the LEAP 2025 U.S. History assessment are shown in Table 8.2. The titles and descriptions of the achievement levels were defined to be part of a cohesive assessment system, and the achievement levels indicate a student's ability to demonstrate proficiency on the LSSS defined for a specific course. The standard-setting section of the LEAP 2025 U.S. History 2017–2018 technical report contains comprehensive information.

Table 8.2
Achievement Level Policy Definitions for LEAP 2025

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

It should be noted that the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information address multiple best practices of the testing industry. Table 8.3 shows the cut of each performance level, and the CSEM for each performance level can be found at Table 9.1 in Chapter 9, Reliability. The standard-setting

section of the LEAP 2025 U.S. History 2017–2018 technical report contains comprehensive information.

Table 8.3

Performance Level Cuts at the Approaching Basic, Basic, Mastery, and Advanced: Operational 2022 LEAP U.S. History

Approaching Basic	Basic	Mastery	Advanced
Cut Score	Cut Score	Cut Score	Cut Score
711	725	750	774

9. Reliability

Internal Consistency Reliability Estimation

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

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

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the *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 analysis results are discussed in Chapter 7, Data Analysis. The reliability and classification accuracy reports in Appendix F: Reliability and Classification Accuracy provide coefficient alpha and IRT model-based or "marginal reliability" (Thissen, Chen, & Bock, 2003) for the total test.

Both coefficient alpha and marginal alpha values were 0.93 for the 2022 U.S. History test. Marginal reliability is described as "an average reliability over levels of θ or theta" (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 "posterior standard deviations" (SEMs) in the IRTPRO output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31

θs. Marginal reliability is the average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals. The formula for marginal reliability is

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

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

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

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

Classical Standard Error of Measurement

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

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

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

It is important to note that the SEM provides an estimate of the average test score error for all students regardless of their individual proficiency levels. It is generally accepted that the SEM varies across the range of student proficiencies (Peterson, Kolen, & Hoover, 1989). For this reason, it is useful to report test-level SEM, and SEM for 2022 U.S. History was 3.63, as seen in Table B.4 in Appendix B.

Conditional Standard Error of Measurement

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

Table 9.1

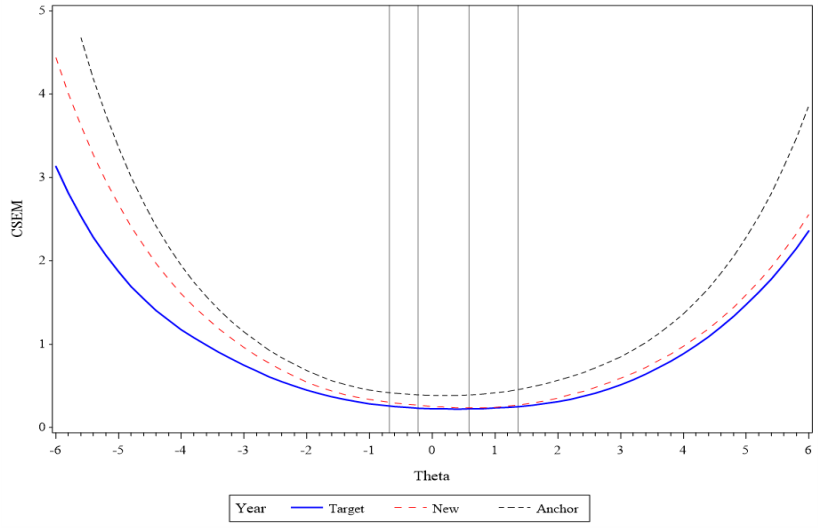
Conditional Standard Errors of Performance Level Cuts: Spring 2022 Operational U.S. History

Approaching Basic		Basic		Mastery		Advanced	
Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
711	8	725	7	750	6	774	7

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

Plot 9.1

CSEM Curves: Spring 2022 Operational U.S. History



Note: Although the CSEM values in the plot are placed on the theta scale, when these CSEM values are converted using the scaling constants, they translate to 8, 7, 6, and 7 as shown in Table 9.1.; Target = 2018 test; New = 2022 OP form; Anchor = anchor items.

Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their scale scores. It is important to know the reliability of student scores in any examination; assessing the reliability of the classification decisions based on these scores is of even greater importance. Classification decision reliability is estimated by the probabilities of correct and consistent classification of students. Procedures from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) were used 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)" (Livingston and & Lewis, 1995, p. 189). A true score is also referred to as a hypothetical mean of scores from all possible forms of the test if they could be somehow obtained (Young & Yoon, 1998).

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

Accuracy and Consistency Indices. Three types of accuracy and consistency indices were generated: *overall*, *conditional-on-level*, and *cut point*, provided in Appendix F: Reliability and Classification Accuracy. The *overall accuracy* of performance-level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels. It is a proportion (or percentage) of correct classification across all the levels. While the overall accuracy index was 0.729, the overall consistency index was 0.637 for the LEAP 2025 U.S. History.

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 0.229 for the 2022 U.S. History test.

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 0.529 for the 2022 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 the level.

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

10. Validity

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

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

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

Chapter 7 describes classical data analysis and item response theoretic calibration, scaling, and equating methods, as well as processes and procedures to clean data to ensure replicable, iterative calibrations and scaling of the 2022 U.S. History test to derive scale scores from students’ raw scores. Some references to introductory and advanced

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

Evidence for Construct-Related Validity

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

Internal Structure of Reporting Categories

The 2022 U.S. History test contains three reporting categories: *Investigate, Evaluate, and Reason Scientifically.* Table D.3 in Appendix D shows that moderate correlations were observed among the reporting categories; since we used distinct items for each reporting category, a moderate correlation was anticipated.

Content-Related Evidence

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

Dimensionality and Principal Component Analysis

Appendix D: Dimensionality provides information about principal component analysis of the 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 2022 assessment, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the Statistical Analysis System (SAS) program. The first and second principal component eigenvalues were compared without rotation. Table D.4 and Plot D.1 summarize the results of the first and second principal component eigenvalues of the assessments. A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the table and figure, the first component is substantially larger than the second eigenvalue for the spring 2022 test. Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Evidence Based on Relations to Other Variables

Evidence based on *relations to other variables* is a typical utility of criterion-related validity evidence to measure concurrent or predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as

multitrait-multimethod studies (Campbell & Fiske, 1959). Thus, external variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores on other tests) to evaluate the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades).

A significant number of students who took the LEAP U.S. History test also took the LEAP Biology test. For the total student group, in general, moderate correlation was observed between the U.S. History and Biology exams. In general, however, the English Learner group reported a slightly lower correlation coefficient than other groups. A separate report, *External Validity Study: SPR 2022*, that was submitted to the LDOE has more specific information.

Item Development and Field Test Analysis

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

Field test items were embedded and administered in one of 10 test forms. Once these items were scored, the LDOE and WestEd/Pearson conducted additional item analysis and content review. Any field test items that exhibited statistical results that suggested potential problems were carefully reviewed by both the LDOE and WestEd/Pearson content specialists. A determination was then made as to whether an item should be accepted, rejected, and revised/re-field tested. Information on statistical analyses for field test items can be obtained in Chapter 7, Data Analysis.

Additional, corroborating evidence consistent with the validity, reliability, and consistency of the LEAP 2025 U.S. History assessment has been documented in the LEAP U.S. History framework, test development plans, and the 2018 U.S. History standard-setting technical report. Finally, Table 10.1 summarizes the sources of validity evidence and indicates where the evidence can be found in the technical report.

Table 10.1

Evidence of Validity and the Corresponding Technical Report Chapter

Source of Validity	Related Information	Related Chapter/Source
Evidence Based on Test Content	Item Development Process	Chapter 3 LEAP 2025 High School U.S. History Assessment Frameworks
	Test Blueprint and Item Alignment to Curriculum and Standards	Chapters 2 & 3 Appendix A LEAP 2025 High School U.S. History Assessment Frameworks
	Item Bias, Sensitivity, and Content Appropriateness	Chapter 3
	Accommodations	Chapter 4
Evidence Based on Response Processes	Field Test Analysis Data Review	Chapters 3, 7, & 9 LEAP 2025 High School U.S. History Assessment Frameworks
	Classical Item analysis IRT Analysis	Chapter 7
Evidence Based on Internal Structure	Differential Item Functioning	Chapter 7
	Reliability and Standard Errors of Measurement	Chapter 9
	Correlation among Reporting Categories	Chapter 10
	Dimensionality Analysis	Chapter 10
Evidence Based on Relations to Other Variables	Correlation Analysis between LEAP U.S. History and Biology Tests	Chapter 10
Evidence Based on the Consequences of Testing	Scale Score and Performance Level Information	Chapter 8
	Test Interpretive Guide	Chapter 8

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mogilner, A., & Mogilner, T. (2006). *Children's Writer's Word Book*, Cincinnati, OH: Writer's Digest Books

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, Stanford (1989). EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies, Orlando, FL: Steck-Vaughn Company.

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

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing:* A *primer* (pp. 161–186). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). MULTILOG (version 7) [Computer software]. In Mathilda du Toit (Ed.), *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. Chicago: Scientific Software International.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5(2),* 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8(2),* 125–145.

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 inEducation*, 10(4), 321–344.

Appendix A: Training Agendas

LEAP 2025 Social Studies Source Search Training Agenda

I. **Introductions**

II. **Source Set Overviews**

a. Task and Item Set Topics

i. Themes of the task or item set that will need to be developed and supported by sources and items

ii. Reporting Categories

iii. Potential Assessable GLEs

1. Sources should support these GLEs

iv. Potential Types of Sources

1. The overview contains recommended stimuli that will support the task or item set
2. Searchers can propose other sources that support the task or item set

v. Source Internet Source Links

1. The overview contains specific websites that can be used to find sources or specific sources

b. Bias and Sensitivity

i. Bias: Avoid sources that cannot be aligned to GLEs. The focus on content aligned to the GLEs reduces the potential for bias that can occur by including content that is not aligned to instruction. This could give an advantage to one student group over other student groups.

ii. Sensitivity: Avoid topics in sources that may upset or offend students in items (e.g., references to graphic violence, nudity, alcohol, drugs, recent natural disasters, caricature representation of ethnic groups).

iii. Universal Design and Visual Impairment

III. **Receiving Source Search Assignments**

IV. **Submitting Sources for Assignments**

a. Text-Based Sources

i. Readability Measurements

1. Lexile

a. Lexile bands

2. ATOS

ii. Originals and marked-up copies of texts

iii. Text Complexity

iv. Range of Textual Evidence

v. Levels of Inference

b. Graphic-Based Sources

i. PDFs with source of graphic and location

ii. Word document with caption

iii. Gifs and JPEGs

V. **Completing Webforms**

VI. **Using Box**

VII. **Additional Resources**

LEAP 2025 U.S. History Item Writer and Editor Training Agenda

I. **Louisiana Student Standards and GLEs**

a. High School
    i. Reporting Categories and Standards
    ii. Grade-Level Expectations (GLEs)

II. **Item Types and Overviews**

a. Selected-Response Items (Multiple Choice, Multiple Select)
b. Constructed-Response Items (item sets only)
c. Technology-Enhanced Items (item sets only)
d. Extended-Response Items (tasks only)
e. Item Sets
    i. Sources (each set will have multiple sources)
    ii. Item Set Overviews
        1. Item stems provided for each item
        2. Metadata associated with each item
        3. Answer options and the nature of distractors
f. Task
    i. Sources (each task will have multiple sources)
    ii. Task Overviews
        1. Item stems provided for each item
        2. Metadata associated with each item
        3. Answer options and the nature of distractors
g. Standalone Items
    i. Purpose
    ii. Sources

III. **Writing and Editing Rubrics and Scoring Guides**

a. Constructed-Response Item Scoring Rubrics
b. Constructed-Response Item Scoring Information
c. Extended-Response Scoring Rubrics
    i. Content
    ii. Claims
d. Extended-Response Scoring Information

IV. **Item Metadata**

a. Range of Textual Evidence
b. Levels of Inference
c. Depth of Knowledge: Items should be DOK 2 or DOK 3

V. **Examples of Items**

VI. **Item Writing Reminders**

a. Grade Appropriate Language: Make sure the vocabulary of the items does not exceed the grade level of the students (exception: content-specific vocabulary that is part of the state standards).
b. Plausible and Logical Distracters: Distracters should address misconceptions that the students have about the topic.
c. Cueing and Clanging of answer options:
    i. Items should avoid using key terms from the sources or in the stem that direct students to specific answer options.
    ii. Items in tasks should avoid cueing each other, either in the stems or in the answer options.
d. Outliers in answer options. Answer options should not stand out because they appear different from the other answer options.
    i. Capitalized words, use of numerals
    ii. Grammatical differences in answer options
e. Bias and Sensitivity
    i. Bias: Avoid information in items that may give an advantage to one group over another group in answering the item (e.g., information that is not part of the curriculum, standards).

ii. Sensitivity: Avoid topics that may upset or offend students in items (e.g., references to graphic violence, nudity, alcohol, drugs, recent natural disasters, group stereotypes).

VII. ABBI Item Development Platform

a. Functionality of the ABBI platform
b. Creating items in ABBI
c. Attaching scoring information in ABBI
d. Checking scoring of Technology-Enhanced items

VIII. Receiving item assignments via Smartsheet

IX. Graphic Arts Requests (editing only)

a. Using the Smartsheet form
b. Attaching marked-up graphics in ABBI
c. Confirming graphic edits have been made

X. Alerting the coordinator that you have completed the item-writing or item-editing assignment and are ready for another assignment

XI. Constructed-Response Item Sample Prompt, Rubric, and Scoring Notes:

Scoring for SOXXXXXXXXXXXXXX

Stem: Based on the sources and your knowledge of social studies, describe two different ways that World War II affected Louisiana.

Scoring Information	
Score Points	**Description**
2	Student's response correctly describes two different ways that World War II affected Louisiana.
1	Student's response correctly describes one way that World War II affected Louisiana.
0	Student's response does not correctly describe one way that World War II affected Louisiana.

Scoring Notes:

- People in Louisiana migrated from rural to urban areas because many jobs in war industries were in the cities.
- The number of employees increased in Louisiana businesses that produced goods for the war.
- Louisiana helped train and mobilize U.S. forces.
- Individuals from Louisiana served in the war.

Accept other reasonable answers.

XII. Selected-Response (Multiple-Choice, Multiple-Select Items)

a. Reference sources in stems where appropriate. Use the language Sources 1 and 2 rather than Source 1 and Source 2. When referring to all of the sources, say “all of the sources.” Refer to the source in the stem, where it is most appropriate.
b. Make sure MS items are in the correct format:

    Which natural resources inspired Americans to migrate westward?
    Select the **two** correct answers.

c. Make sure the item scores correctly.

XIII. Editorial Process

a. Move the items to Content Editor 2 or to Proofing 1, depending on the editorial status of the item or the direction of the coordinator.

LEAP 2025 U.S. History and Grades 3–8 Data Review Training Agenda

I. What is a Data Review?

a. Statistical Definition: Classical Test Theory

1. P-value
2. Point-Biserial
3. Option/Distribution Analysis
4. Differential Item Function (DIF)
5. Flagging Value

Statistics	**Flagging Value**
P-value	≤ 0.25 or > 0.90
Omit Percentage	> 4%
Point-biserial Correlation	< 0.20
Distractor Percentage (MC only)	> 40%
Distractor Point-biserial Correlation (MC only)	> 0.00
DIF	B, C

b. Statistical Definition: Item Response Theory (IRT)

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

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

II. Judgment Task in ABBI

a. Accept
b. Accept with Edits
c. Reject

Appendix B: Test Summary

U.S. History

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

- Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table B.1

Percentage of Points by Reporting Category (includes Task Items): Spring 2022 Operational U.S. History

Reporting Category	Form G
Standard 2	21.7%
Standard 3	15.9%
Standard 4	37.7%
Standard 5&6	24.6%

Table B.2

Standard Coverage: Spring 2022 Operational U.S. History

Reporting Categories		No. of Items					% of Test
		TEI	MS	MC	ER	CR	
		N	N	N	N	N	
Standard 2	US.2.3			1			1.89
	US.2.4	1		2		1	7.55
	US.2.5			1			1.89
	US.2.6	1		2			5.66
	US.2.7			1			1.89
	US.2.8			2			3.77
	Sub-Total	**2**		**9**		**1**	**22.64**
Standard 3	US.3.1	1	2	2			9.43
	US.3.2			2			3.77
	US.3.3			1			1.89
	US.3.5			1			1.89
	US.3.6			1			1.89
	Sub-Total	**1**	**2**	**7**			**18.87**
Standard 4	US.4.2			1			1.89
	US.4.3	1		3			7.55
	US.4.4			1			1.89
	US.4.5			1			1.89
	US.4.6			2	1		5.66
	US.4.7		1	2		1	7.55
	US.4.8	1					1.89
	US.4.10			1			1.89
	Sub-Total	**2**	**1**	**11**	**1**	**1**	**30.19**
Standard 5&6	US.5.1			1			1.89
	US.5.2			1			1.89
	US.5.3			1			1.89
	US.5.4			1			1.89
	US.5.5	1		4			9.43
	US.6.4	1		4			9.43
	US.6.5			1			1.89
	Sub-Total	**2**		**13**			**28.30**
Total		**7**	**3**	**40**	**1**	**2**	**100.00**

Table B.3

Item Type Summary: Spring 2022 Operational U.S. History

Admin.	MC	MS	TEI	CR	ER*
Spring 2022	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 B.4

Raw Score Summary: Spring 2022 Operational U.S. History

Admin.	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability*	SEM
Spring 2022	≥35,950	33.03	13.73	0	69	0.51	0.46	0.93	3.63

* Reliability is Cronbach's alpha.

Table B.5

Raw Score Summary by Reporting Category: Spring 2022 Operational U.S. History

Admin	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
Spring 2022	Standard 2	7.43	3.56	0	15	0.52	0.49	0.79	1.63
	Standard 3	5.57	2.63	0	11	0.50	0.45	0.70	1.44
	Standard 4	11.39	5.43	0	26	0.51	0.49	0.84	2.17
	Standard 5&6	8.64	3.52	0	17	0.52	0.40	0.73	1.83

Table B.6

Scale Score and Raw Score Summary: Spring 2022 Operational U.S. History

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD	Effect Size
Total	≥35,950	100	728.58	33.83	33.03	13.73	-
Female	≥18,200	50.63	728.94	31.89	33.01	13.11	**0.00**
Male	≥17,750	49.37	728.21	35.71	33.05	14.34	-
African American	≥14,510	40.38	714.69	30.88	27.3	11.83	**0.81**
American Indian or Alaska Native	≥220	0.62	728.29	30.22	32.74	12.51	**0.36**
Asian	≥700	1.97	756.51	37.07	44.68	14.79	**-0.54**
Hispanic/Latino	≥2,560	7.13	724.14	35.48	31.47	14.02	**0.46**
Multi-Racial	≥840	2.35	733.65	31.75	34.99	13.32	**0.19**
Native Hawaiian or Other PacificIslander	≥20	0.06	736.09	36.09	35.74	13.9	**0.14**
White	≥17,070	47.48	739.65	31.1	37.55	13.19	-
Economically Disadvantaged: No	≥14,000	38.94	743.29	31.47	39.11	13.35	**-0.77**
Economically Disadvantaged: Yes	≥20,780	57.8	719.33	31.78	29.18	12.5	-
EL: No	≥34,970	97.26	729.45	33.55	33.36	13.68	**-0.89**
EL: Yes	≥980	2.74	697.7	29.09	21.23	9.78	-
Gifted or Talented	≥2,310	6.45	763.67	30.44	47.72	12.44	**-2.25**
Regular Education	≥30,740	85.51	728.74	31.65	32.98	13.03	**-0.87**
Special Education	≥2,890	8.04	698.79	31.02	21.77	10.8	-
Section 504: No	≥32,690	90.93	729.54	33.65	33.41	13.71	**-0.30**
Section 504: Yes	≥3,260	9.07	719.02	34.14	29.24	13.42	-
Migrant: No	≥35,900	99.85	728.61	33.83	33.04	13.73	**-0.41**
Migrant: Yes	≥50	0.15	713.96	34.98	27.43	13.44	-
Homeless: No	≥35,420	98.51	728.74	33.85	33.1	13.74	**-0.33**
Homeless: Yes	≥530	1.49	718.1	30.9	28.52	12.23	-
Military Affiliation: No	≥35,460	98.63	728.39	33.79	32.94	13.71	**0.44**
Military Affiliation: Yes	≥490	1.37	742.73	33.72	39.02	14.01	-
Foster Care: No	≥35,890	99.83	728.62	33.83	33.04	13.73	**-0.51**
Foster Care: Yes	≥60	0.17	710.23	33.01	25.97	12.62	-

Appendix C: Item Analysis Summary Report

Contents
Table C.1 P-Value Summary: Spring 2022 Operational U.S. History Table C.1.1 P-Value Summary by Item Type: Spring 2022 Operational U.S. History Plot C.1 P-Value Summary by Item Type: Spring 2022 Operational U.S. History
Table C.2. Item-Total Correlation Summary: Spring 2022 Operational U.S. History Table C.2.1 Item-Total Correlation Summary by Item Type: Spring 2022 Operational U.S. History Plot C.2 Item-Total Correlation Summary by Item Type: Spring 2022 Operational U.S. History
Table C.3. Corrected Point-Biserial Correlation Summary: Spring 2022 Operational U.S. History Table C.3.1 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational U.S. History Plot C.3 Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational U.S. History
Table C.4 Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2022 Operational U.S. History
Table C.5.1 IRT-A Parameter Summary by Reporting Category: Spring 2022 Operational U.S. History Table C.5.2 IRT-B Parameter Summary by Reporting Category: Spring 2022 Operational U.S. History Table C.5.3 IRT Parameter Summary by Item Type: Spring 2022 Operational U.S. History
Plot C.5.1 IRT Parameter Summary: Spring 2022 Operational U.S. History: A-Parameter Plot C.5.2 IRT Parameter Summary: Spring 2022 Operational U.S. History: B-Parameter Plot C.5.3 IRT Parameter Summary: Spring 2022 Operational U.S. History: C-Parameter
Table C.6 Statistically Flagged Items by Item Type: Spring 2022 Operational U.S. History

- Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table C.1

P-Value Summary: Spring 2022 Operational U.S. History

Form	No. of Items	0 ≤ p < 0.2	0.2 ≤ p < 0.4	0.4 ≤ p < 0.6	0.6 ≤ p < 0.8	0.8 ≤ p ≤ 1.0
D	54	0	7	33	12	2

Table C.1.1

P-Value Summary by Item Type: Spring 2022 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.316	0.316	0.338	0.359	0.359
ER*	1	0.286	0.286	0.287	0.287	0.287
MC	40	0.268	0.457	0.534	0.619	0.835
MS	3	0.400	0.400	0.492	0.503	0.503
TEI	7	0.277	0.397	0.428	0.479	0.535

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

Plot C.1
P-Value Summary by Item Type: Spring 2022 Operational U.S. History

Box and Whisker Plot

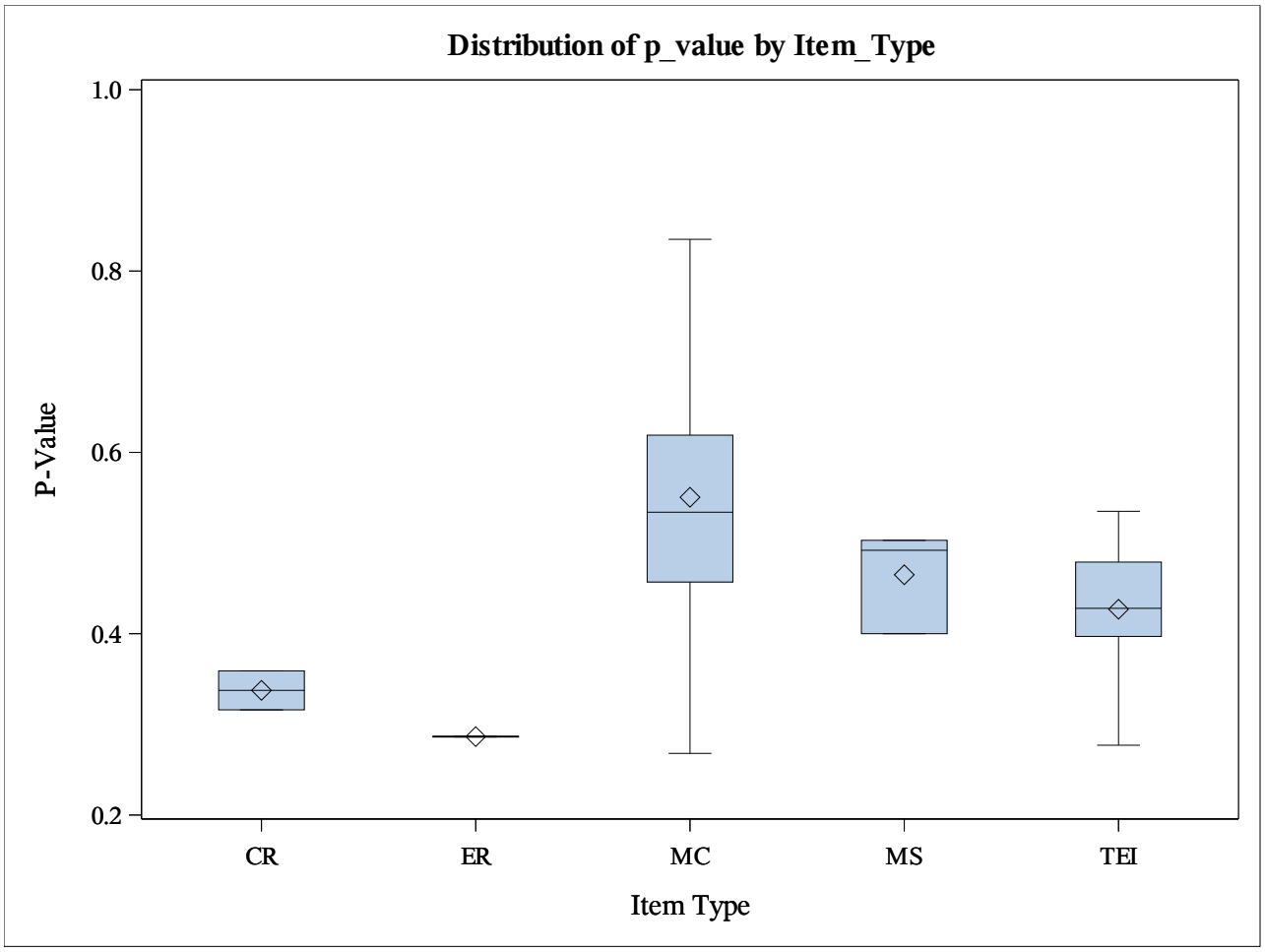


Table C.2

Item-Total Correlation Summary: Spring 2022 Operational U.S. History

No. ofItems	r < 0	0.0 ≤ r < 0.2	0.2 ≤ r < 0.3	0.3 ≤ r < 0.4	0.4 ≤ r < 0.5	r ≥ 0.5
54	0	1	2	14	24	13

Table C.2.1

Item-Total Correlation Summary by Item Type: Spring 2022 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.636	0.636	0.674	0.712	0.712
ER*	1	0.777	0.777	0.778	0.778	0.778
MC	40	0.193	0.359	0.430	0.464	0.558
MS	3	0.475	0.475	0.498	0.520	0.520
TEI	7	0.478	0.479	0.532	0.593	0.601

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

Plot C.2
Item-Total Correlation Summary by Item Type: Spring 2022 Operational U.S. History

Box and Whisker Plot

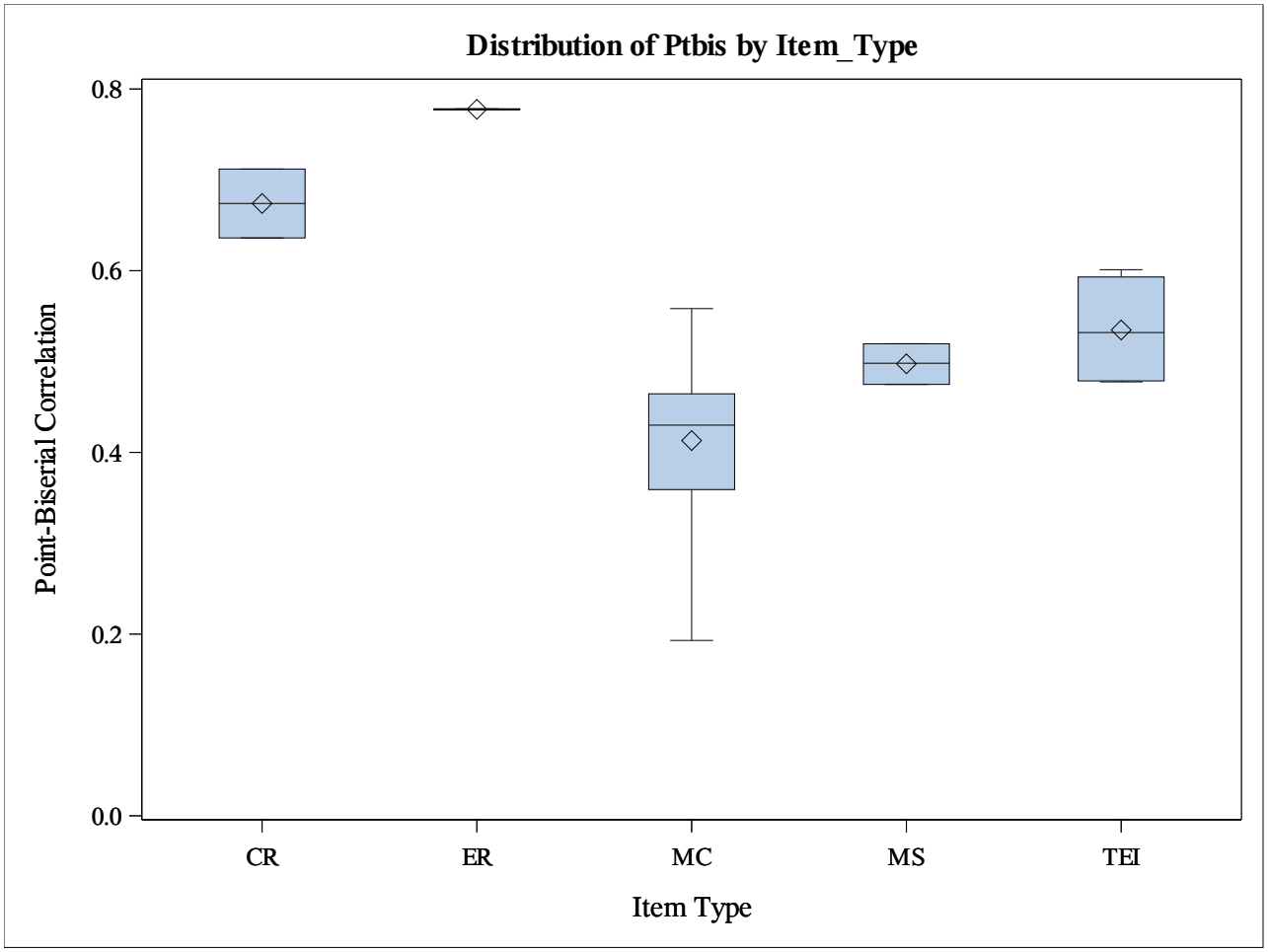


Table C.3

Corrected Point-Biserial Correlation Summary: Spring 2022 Operational U.S. History*

No. ofItems	r < 0	0.0 ≤ r < 0.2	0.2 ≤ r < 0.3	0.3 ≤ r < 0.4	0.4 ≤ r < 0.5	r ≥ 0.5
54	0	1	5	13	26	9

Table C.3.1

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational U.S. History*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.603	0.603	0.643	0.683	0.683
ER**	1	0.744	0.744	0.745	0.745	0.745
MC	40	0.158	0.328	0.401	0.436	0.534
MS	3	0.447	0.447	0.470	0.492	0.492
TEI	7	0.432	0.443	0.499	0.558	0.570

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

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

Plot C.3

Corrected Point-Biserial Correlation Summary by Item Type: Spring 2022 Operational U.S. History

Box and Whisker Plot

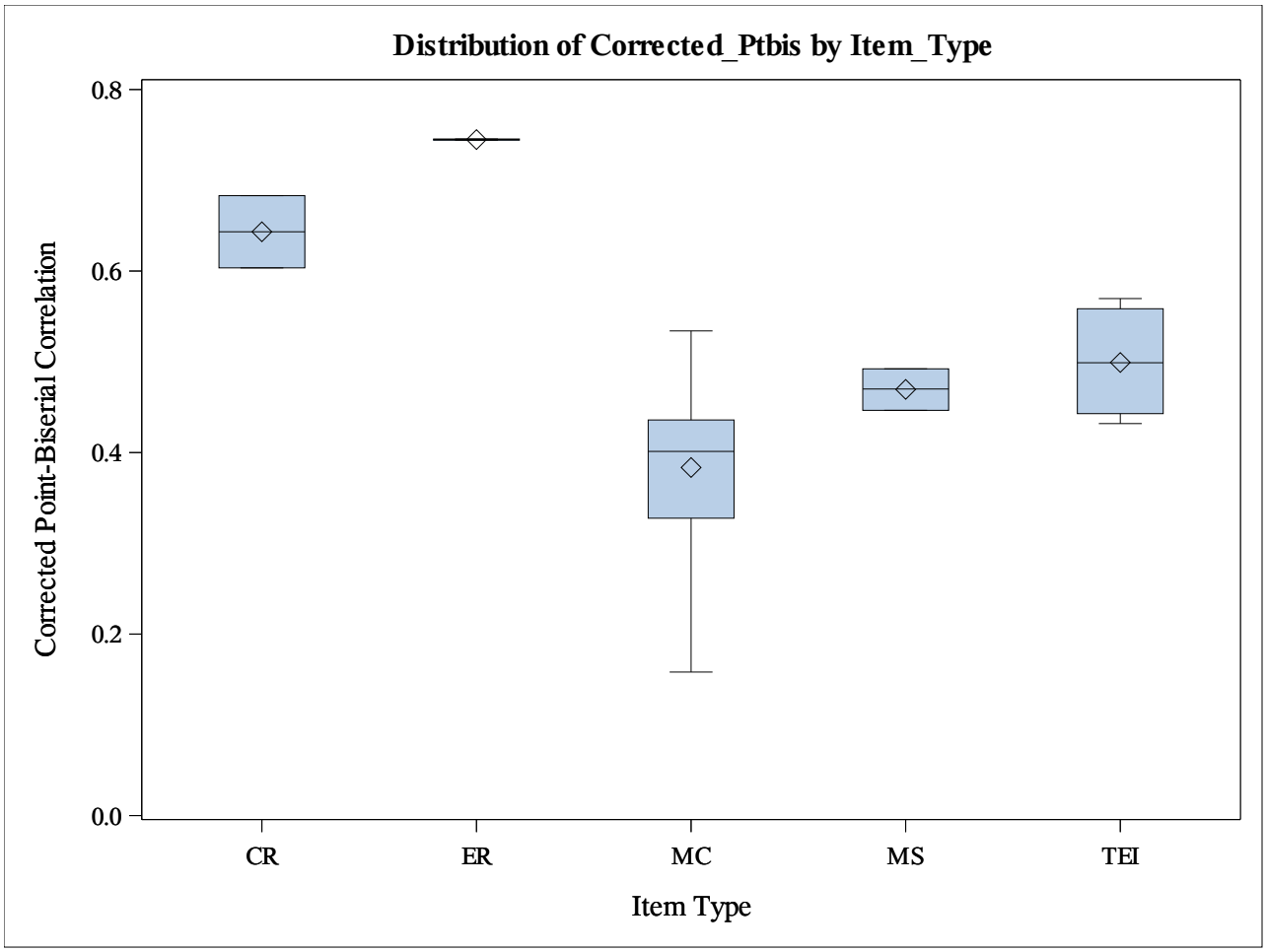


Table C.4

Item-Total Correlation Summary by Reporting Category and Item Type: Spring 2022 Operational U.S. History

***Item* Type**	**Reporting Category**	**No. of Items**	**Minimum**	**25th Percentile**	**Median**	**75th Percentile**	**Maximum**
CR	Standard 2	1	0.712	0.712	0.712	0.712	0.712
	Standard 4	1	0.636	0.636	0.636	0.636	0.636
ER	Standard 4	1	0.777	0.777	0.778	0.778	0.778
MC	Standard 2	9	0.379	0.405	0.447	0.497	0.558
	Standard 3	7	0.344	0.354	0.391	0.469	0.547
	Standard 4	11	0.283	0.359	0.439	0.452	0.495
	Standard 5&6	13	0.193	0.321	0.399	0.450	0.475
MS	Standard 3	2	0.475	0.475	0.497	0.520	0.520
	Standard 4	1	0.498	0.498	0.498	0.498	0.498
TEI	Standard 2	2	0.516	0.516	0.555	0.593	0.593
	Standard 3	1	0.532	0.532	0.532	0.532	0.532
	Standard 4	2	0.478	0.478	0.511	0.544	0.544
	Standard 5&6	2	0.479	0.479	0.540	0.601	0.601

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

Table C.5.1

IRT-A Parameter Summary by Reporting Category: Spring 2022 Operational U.S. History

IRT-a Range	Standard 2	Standard 3	Standard 4	Standard 5&6	Number of Items
a < 0.0	0	0	0	0	0
0.0 <= a < 0.2	0	0	0	0	0
0.2 <= a < 0.4	0	0	1	1	2
0.4 <= a < 0.6	1	0	1	4	6
0.6 <= a < 0.8	4	2	5	3	14
0.8 <= a < 1.0	1	7	3	4	15
1.0 <= a < 1.2	2	0	4	3	9
1.2 <= a < 1.4	3	1	3	0	7
1.4 <= a < 1.6	1	0	0	0	1
1.6 <= a < 1.8	0	0	0	0	0
1.8 <= a < 2.0	0	0	0	0	0
2.0 <= a	0	0	0	0	0
Minimum	0.58	0.67	0.39	0.36	0.36
Maximum	1.48	1.23	1.29	1.16	1.48
Mean	0.98	0.90	0.89	0.77	0.88
SD	0.32	0.15	0.28	0.25	0.27
Number of Items	12	10	17	15	54

Table C.5.2

IRT-B Parameter Summary by Reporting Category: Spring 2022 Operational U.S. History

IRT-b Range	Standard 2	Standard 3	Standard 4	Standard 5&6	Number of Items
b < -3.5	0	0	0	0	0
-3.5 <= b < -3.0	0	0	0	0	0
-3.0 <= b < -2.5	0	0	0	0	0
-2.5 <= b < -2.0	0	0	0	0	0
-2.0 <= b < -1.5	0	0	0	0	0
-1.5 <= b < -1.0	0	0	3	0	3
-1.0 <= b < -0.5	4	0	2	1	7
-0.5 <= b < 0.0	0	3	1	4	8
0.0 <= b < 0.5	4	1	2	4	11
0.5 <= b < 1.0	3	6	6	5	20
1.0 <= b < 1.5	1	0	3	0	4
1.5 <= b < 2.0	0	0	0	1	1
2.0 <= b < 2.5	0	0	0	0	0
2.5 <= b < 3.0	0	0	0	0	0
3.0 <= b < 3.5	0	0	0	0	0
3.5 <= b	0	0	0	0	0
Minimum	-0.91	-0.34	-1.37	-0.54	-1.37
Maximum	1.23	0.86	1.33	1.88	1.88
Mean	0.12	0.38	0.18	0.37	0.26
SD	0.72	0.44	0.84	0.60	0.68
Number of Items	12	10	17	15	54

Table C.5.3

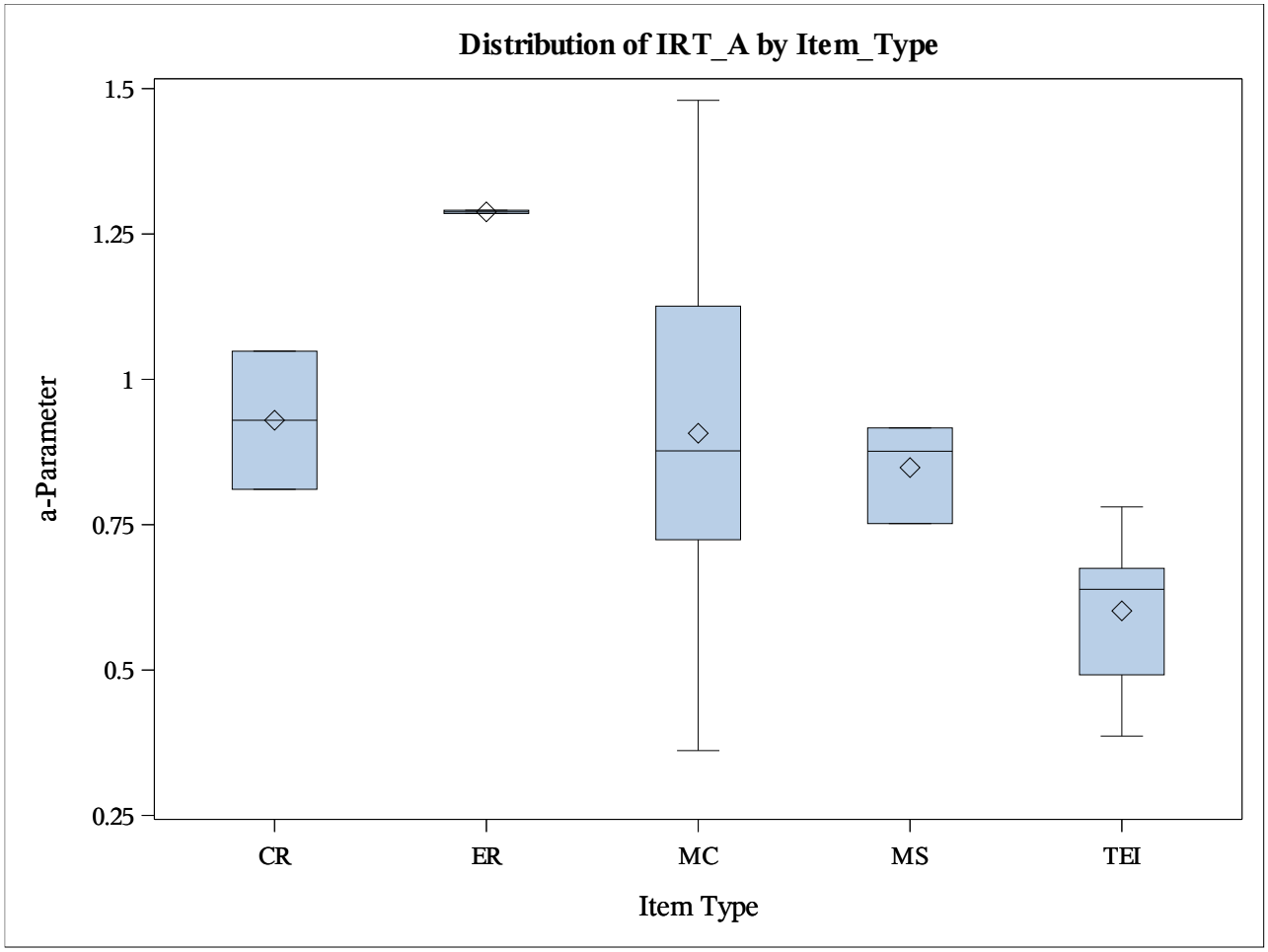
IRT Parameter Summary by Item Type: Spring 2022 Operational U.S. History

Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	A	2	0.811	0.811	0.93	1.048	1.048
	B	2	0.474	0.474	0.515	0.556	0.556
ER*	A	1	1.285	1.285	1.288	1.291	1.291
	B	1	0.85	0.85	0.85	0.851	0.851
MC	A	40	0.361	0.724	0.877	1.126	1.48
	B	40	-1.37	-0.326	0.41	0.74	1.878
	C	40	0.019	0.132	0.207	0.259	0.374
MS	A	3	0.752	0.752	0.876	0.916	0.916
	B	3	0.032	0.032	0.095	0.611	0.611
	C	3	0.04	0.04	0.097	0.109	0.109
TEI	A	7	0.386	0.492	0.639	0.675	0.781
	B	7	-0.337	0.012	0.365	0.623	1.088

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

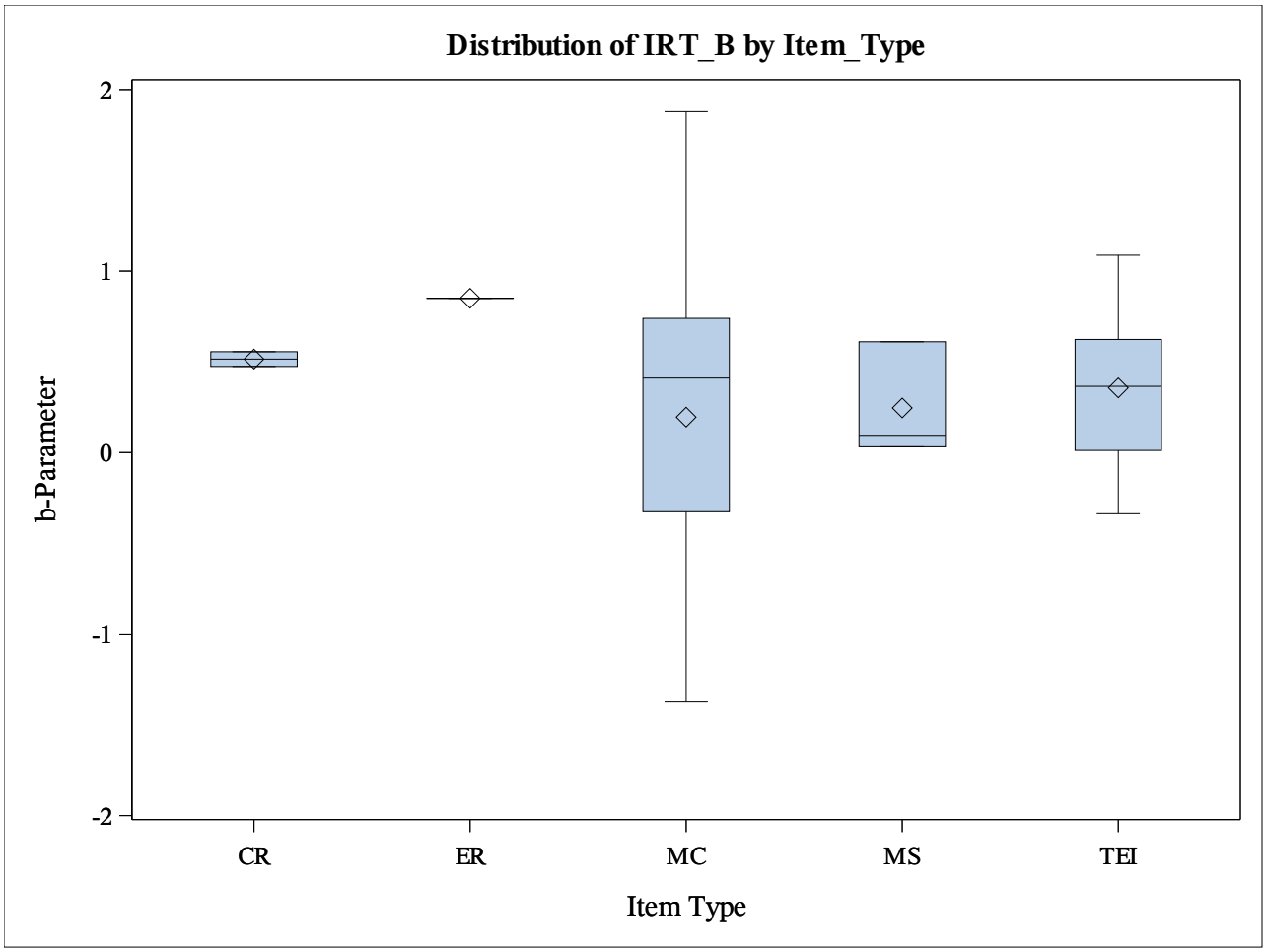
Plot C.5.1

IRT Item Parameter Summary for Spring 2022 Operational U.S. History: A-Parameter



Plot C.5.2

IRT Item Parameter Summary for Spring 2022 Operational U.S. History: B-Parameter



Plot C.5.3

IRT Item Parameter Summary for Spring 2022 Operational U.S. History: C-Parameter

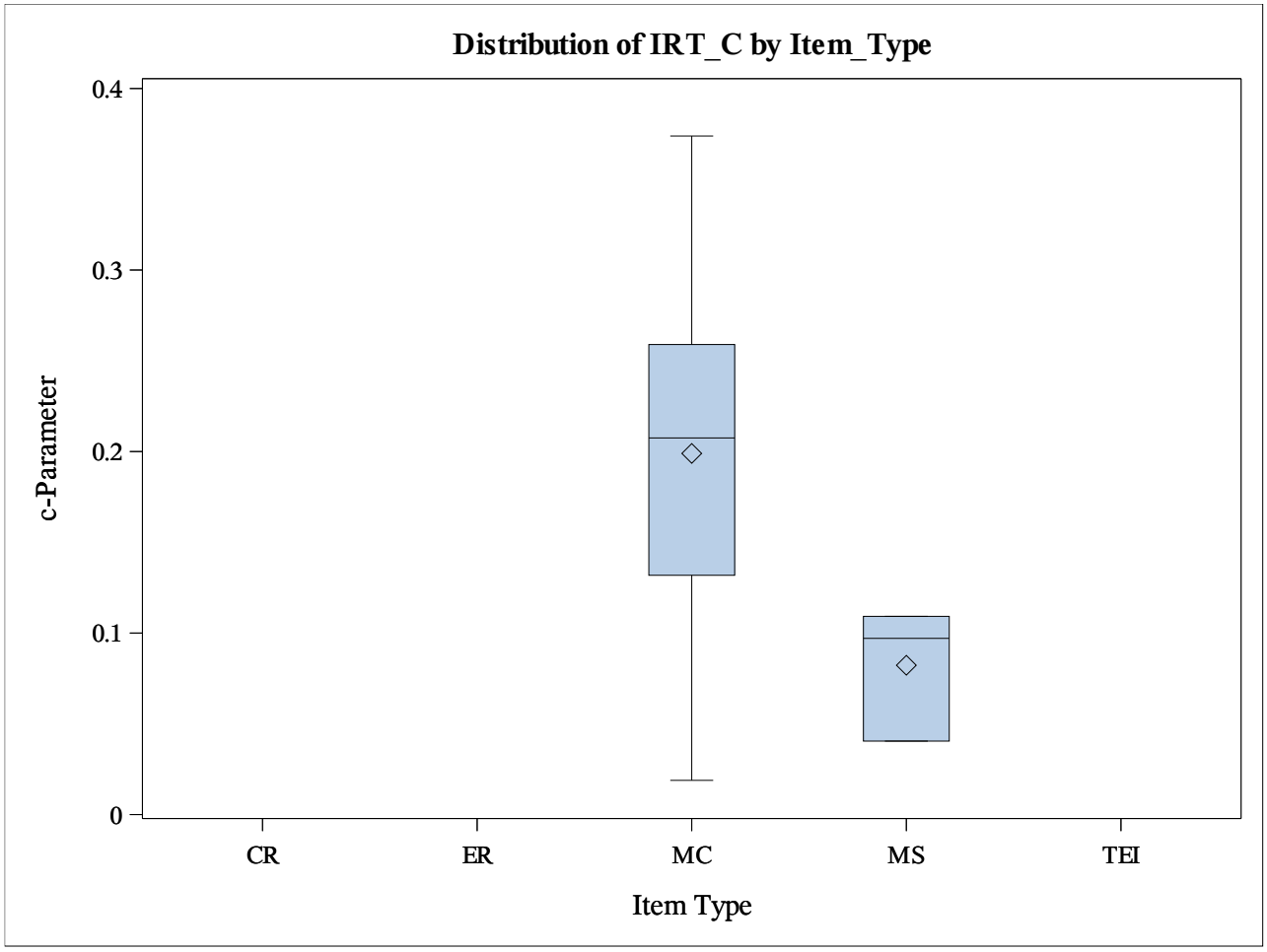


Table C.6

Statistically Flagged Operational Items: Spring 2022 Operational U.S. History

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

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

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

Appendix D: Dimensionality

Dimensionality Reports
U.S. History

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

- Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table D.1

Zq1 Statistics and Summary Data: Spring 2022 Operational U.S. History

Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
CR	24.99	24.99	28.98	32.98	32.98	0
ER	28.15	28.15	28.39	28.63	28.63	0
MC	0.12	2.63	4.47	7.82	67.92	1
MS	3.14	3.14	5.01	5.33	5.33	0
TEI	10.49	19.44	46.48	89.90	144.33	4

Table D.2

Q3 Statistics and Summary Data: Spring 2022 Operational U.S. History

Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
0.195	-0.259	-0.087	-0.016	0.061	0.962

Table D.3

Reporting Category Intercorrelation Coefficients: Spring 2022 Operational U.S. History

Reporting Category	Standard 2	Standard 3	Standard 4	Standard 5&6
Standard 2	1.00			
Standard 3	0.74	1.00		
Standard 4	0.81	0.75	1.00	
Standard 5&6	0.74	0.71	0.76	1.00

Table D.4

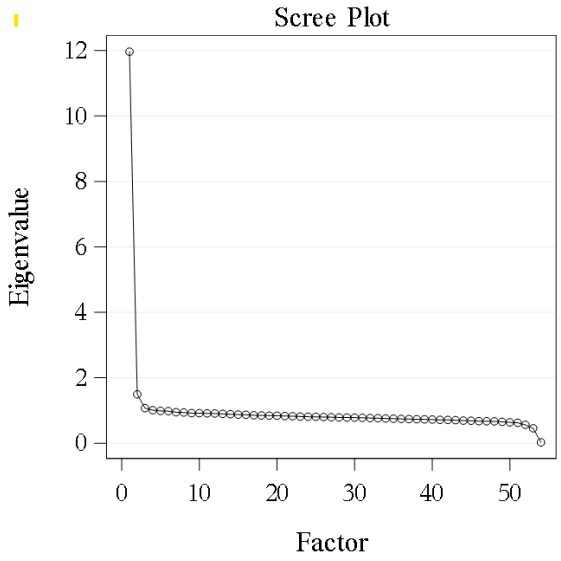
First and Second Eigenvalue: Spring 2022 Operational U.S. History*

Administration	First Eigenvalue	Second Eigenvalue	Ratio*
Spring 2022	11.964	1.494	8.008

* The ratio of first and second eigenvalues.

Plot D.1

Principal Component Analysis Plot: Spring 2022 Operational U.S. History



Appendix E: Scale Distribution and Statistical Report

U.S. History

Contents
Table E.1 Scale Score Descriptive Statistics and Plots: Spring 2022 Operational U.S. History
Table E.2 Frequency Distribution of Scale Scores: Spring 2022 Operational U.S. History

- Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table E.1

Scale Score Descriptive Statistics and Plots: Spring 2022 Operational U.S. History

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

N	≥35950		
Mean	728.58	Median	730.00
Std deviation	33.83	Variance	1144.64
Skewness	-0.0751	Kurtosis	-0.1367
Mode	724.00	Std Error Mean	0.1784
Range	200.00	Interquartile Range	45.00

Quantile	Estimate
100% Max	850
99%	803
95%	782
90%	772
75% Q3	751
50% Median	730
25% Q1	706
10%	686
5%	670
1%	650
0% Min	650

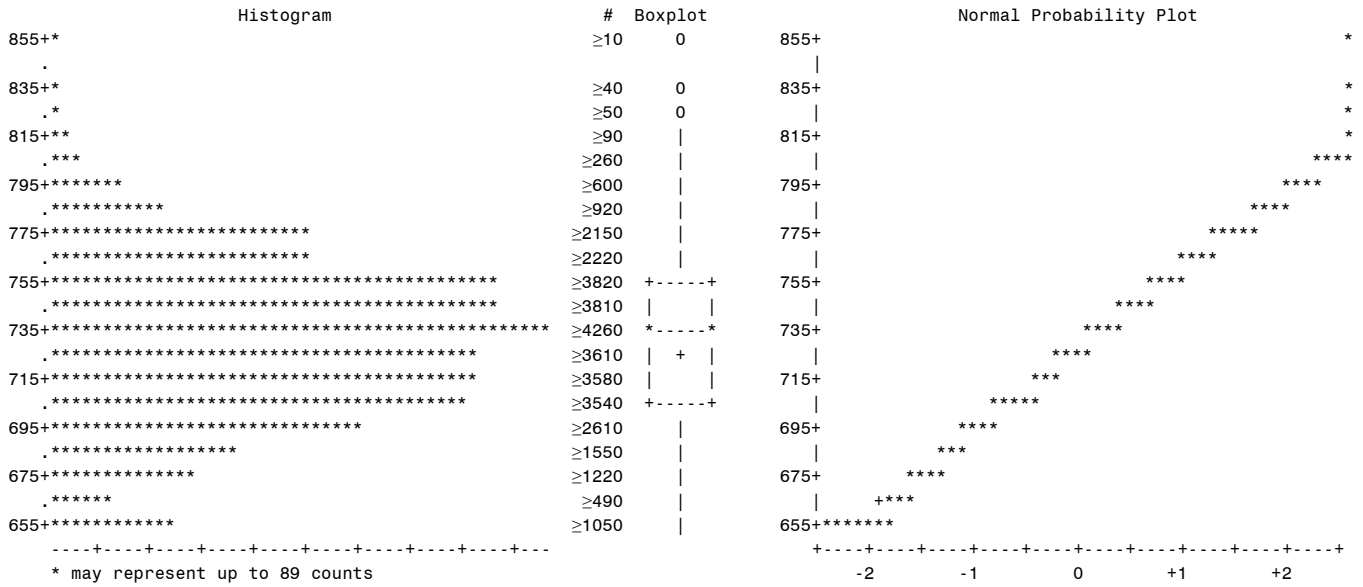
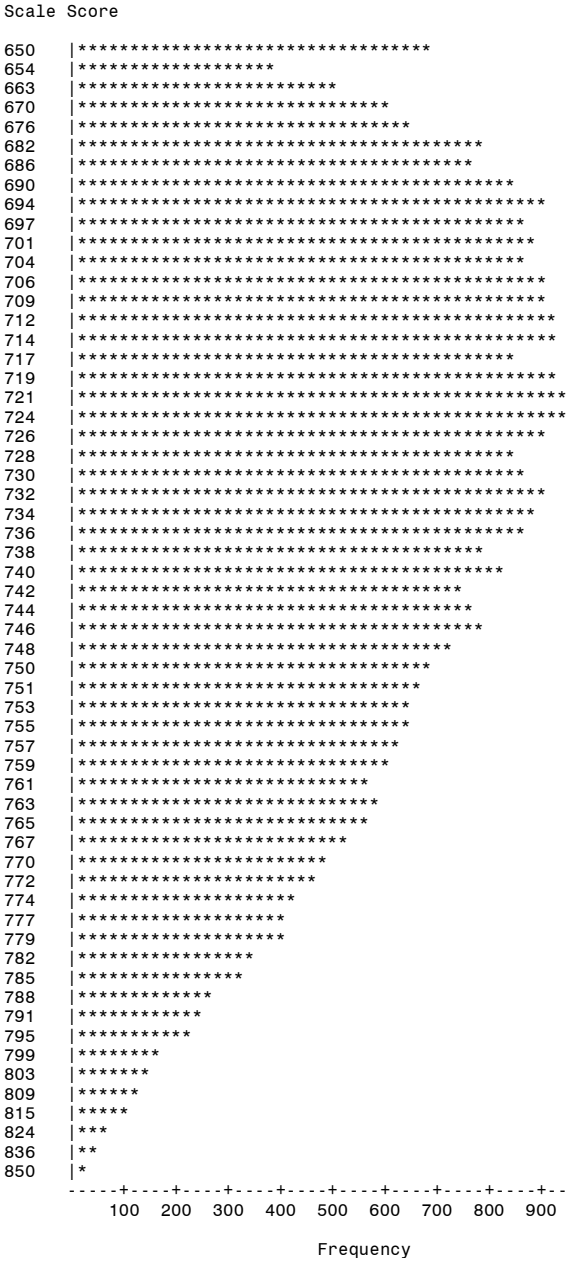


Table E. 2

Frequency Distribution of Scale Scores: Spring 2022 Operational U.S. History



Scale Score	Freq	Cum. Freq	Percent	Cum. Percent
650	≥670	≥670	1.89	1.89
654	≥370	≥1050	1.05	2.93
663	≥490	≥1540	1.37	4.31
670	≥590	≥2140	1.65	5.96
676	≥630	≥2770	1.76	7.72
682	≥780	≥3560	2.19	9.91
686	≥760	≥4320	2.12	12.04
690	≥840	≥5170	2.36	14.40
694	≥890	≥6070	2.49	16.89
697	≥860	≥6930	2.41	19.30
701	≥880	≥7820	2.47	21.77
704	≥850	≥8680	2.39	24.16
706	≥890	≥9580	2.49	26.65
709	≥900	≥10480	2.51	29.16
712	≥920	≥11400	2.57	31.73
714	≥910	≥12320	2.55	34.28
717	≥830	≥13160	2.33	36.60
719	≥910	≥14070	2.54	39.14
721	≥930	≥15000	2.59	41.72
724	≥930	≥15930	2.61	44.33
726	≥900	≥16830	2.50	46.83
728	≥840	≥17680	2.34	49.18
730	≥850	≥18530	2.38	51.56
732	≥890	≥19430	2.50	54.06
734	≥870	≥20310	2.43	56.49
736	≥860	≥21170	2.40	58.90
738	≥770	≥21940	2.14	61.04
740	≥810	≥22760	2.28	63.32
742	≥730	≥23490	2.03	65.35
744	≥760	≥24250	2.12	67.47
746	≥770	≥25030	2.16	69.63
748	≥720	≥25760	2.03	71.65
750	≥680	≥26440	1.90	73.56
751	≥650	≥27100	1.82	75.37
753	≥630	≥27730	1.76	77.13
755	≥640	≥28370	1.80	78.93
757	≥610	≥28990	1.71	80.64
759	≥590	≥29580	1.64	82.28
761	≥560	≥30150	1.58	83.85
763	≥570	≥30720	1.61	85.47
765	≥550	≥31280	1.55	87.01
767	≥520	≥31800	1.45	88.46
770	≥480	≥32290	1.35	89.81
772	≥450	≥32740	1.28	91.08
774	≥420	≥33170	1.19	92.27
777	≥390	≥33560	1.08	93.36
779	≥390	≥33950	1.08	94.44
782	≥330	≥34280	0.92	95.37
785	≥320	≥34610	0.91	96.27
788	≥260	≥34870	0.73	97.01
791	≥230	≥35100	0.64	97.65
795	≥210	≥35320	0.61	98.26
799	≥150	≥35480	0.43	98.69
803	≥140	≥35620	0.39	99.08
809	≥120	≥35740	0.33	99.42
815	≥90	≥35830	0.26	99.67
824	≥50	≥35890	0.16	99.83
836	≥40	≥35930	0.12	99.95
850	≥10	≥35950	0.05	100.00

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports U.S. History

Contents
Table F.1. Reliability and SEM for Overall and Subgroups: Spring 2022 Operational U.S. History
Table F.2. Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational U.S. History
Table F.3. Classification Accuracy and Decision Consistency: Spring 2022 Operational U.S. History

- Because the spring 2022 test was administered under conditions related to COVID-19, great caution should be applied when any statistical inference is drawn.

Table F.1

Reliability and SEM for Overall and Subgroups: Spring 2022 Operational U.S. History

Subg	Reliability	SEM
All Students	0.931	3.607
Female	0.923	3.638
Male	0.938	3.571
African American	0.909	3.569
American Indian or Alaska Native	0.918	3.582
Asian	0.943	3.531
Hispanic/Latino	0.934	3.602
Multi-Racial	0.927	3.599
Native Hawaiian or Other Pacific Islander	0.933	3.598
White	0.925	3.612
Economically Disadvantaged: No	0.927	3.607
Economically Disadvantaged: Yes	0.918	3.579
English Learner: No	0.930	3.619
English Learner: Yes	0.876	3.444
Gifted or Talented	0.924	3.429
Regular Education	0.923	3.616
Special Education	0.900	3.415
Section 504: No	0.930	3.627
Section 504: Yes	0.929	3.576
Migrant: No	0.931	3.607
Migrant: Yes	0.930	3.556
Homeless: No	0.931	3.609
Homeless: Yes	0.914	3.587
Military Affiliation: No	0.930	3.627
Military Affiliation: Yes	0.935	3.572
Foster Care: No	0.931	3.607
Foster Care: Yes	0.923	3.502

Table F.2

Cronbach's Alpha and Marginal Reliability: Spring 2022 Operational U.S. History

Administration	Cronbach's Alpha	Marginal Reliability
Spring 2022	0.931	0.930

Table F.3

Classification Accuracy and Decision Consistency: Spring 2022 Operational U.S. History

Accuracy Matrix: Spring 2022 Operational U.S. History

Adm.	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
Spring 2022	1	0.26	0.03	0.00	0.00	0.00	0.29
	2	0.03	0.08	0.04	0.00	0.00	0.15
	3	0.00	0.04	0.19	0.04	0.00	0.27
	4	0.00	0.00	0.04	0.14	0.03	0.21
	5	0.00	0.00	0.00	0.02	0.06	0.08
	Total	0.29	0.15	0.27	0.19	0.09	1.00

Consistency Matrix: Spring 2022 Operational U.S. History

Adm.	Level	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)	Total
Spring 2022	1	0.25	0.04	0.01	0.00	0.00	0.30
	2	0.03	0.06	0.05	0.00	0.00	0.14
	3	0.01	0.05	0.16	0.05	0.00	0.26
	4	0.00	0.00	0.06	0.11	0.03	0.20
	5	0.00	0.00	0.00	0.03	0.06	0.09
	Total	0.29	0.15	0.27	0.19	0.09	1.00

Table F.3.1

Estimates of Accuracy and Consistency of Achievement Level Classification

Adm.	Accuracy	Consistency	PChance	Kappa
Spring 2022	0.729	0.637	0.229	0.529

Table F.3.2

Accuracy of Classification at Each Achievement Level

Adm.	Unsatisfactory(1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
Spring 2022	0.884	0.537	0.703	0.668	0.759

Table F.3.3

Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Adm.	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
Spring 2022	0.934	0.915	0.921	0.953

Table F.3.4

Consistency of Dichotomous Categorizations by Form (PAC Metric)

Adm.	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
Spring 2022	0.906	0.882	0.889	0.933

Table F.3.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Adm.	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
Spring 2022	0.777	0.762	0.732	0.595

Table F.3.6

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

Adm.	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
Spring 2022	0.034	0.041	0.038	0.028

Table F.3.7

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

Adm.	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
Spring 2022	0.032	0.044	0.041	0.020

Appendix G: Guidelines for Accommodated Print and Braille

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

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

- Highlight-text items or item parts in the online environment require a student to click on the selected text, which highlights the selected word, phrase, or sentence. In the AP and braille forms, the text is presented in the same format and the student is asked to circle the answer. Where only certain words or phrases are selectable in the online system, those options are underlined in the AP and braille forms to indicate which words and/or phrases the student should select from.
- Drop-down menu items in the online environment have answer options in a drop-down menu format, oftentimes as part of a complete sentence. The AP and braille forms display the item with a blank line in place of the drop-down menu in the sentence, with all the answer options for the drop-down menu presented vertically below the sentence. The directions are then modified to ask the student to circle the word/phrase that belongs in the blank.
- Short-answer items in the online environment require a student to type the answer in a box. In the AP and braille forms, a box is provided for the student to write the response.
- Keypad input items in the online environment require a student to enter a numeric response including all rational and irrational numbers as well as expressions and equations. In the AP and braille forms, a box is provided for the student to write the response.
- Graphing items, including coordinate planes, number lines, line plots, and bar graphs, in the online environment require a student to complete a graph by plotting points, adding Xs to create a line plot, or raising/lowering bars to create a bar graph or histogram. In the AP and braille forms, the student is provided with the same coordinate plane, number line, line plot, or bar graph as in the online item, including titles, axis labels, and keys, and is asked to complete the graph.

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

AP forms are thoroughly reviewed by the LDOE and DRC content experts, and braille forms are reviewed by an outside third-party braille expert. Students respond to their

accommodated print and braille test using the same online test as used by the general population, either through use of a scribe or by themselves if able. This ensures a valid and reliable assessment for students who are unable to participate in the online assessment.

Appendix H: Ongoing Quality Control

A system for monitoring, maintaining, and increasing the quality of its assessment system, including precise and technically sound criteria for the analyses of all of the assessments in its assessment system, is crucial and critical for keeping a high quality of assessments. The places where information about monitoring, maintaining, and improving quality is incorporated are included in the following table.

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