LEAP 2025 U.S. History Technical Report: 2018–2019

Prepared by DRC, Pearson, and WestEd

FOREWORD

Improving student achievement is a primary goal of any educational assessment program such as the Louisiana Educational Assessment Program 2025 (LEAP 2025). 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 technical information herein is 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, 2009) and in the new edition, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Table of Contents

1. Introduction

The Louisiana Department of Education (LDOE) has a long and distinguished history in the development and administration of assessments that support its state accountability system and are aligned to its state content standards. Per state law, the LDOE is to administer statewide summative Social Studies assessments in grades 3–8 and in U.S. History. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. Further, the LDOE and the BESE are committed to the development of rigorous assessments as one component of their comprehensive plan—Louisiana Believes—designed to ensure that every Louisiana student is on track to be successful in postsecondary education and the workforce.

The purpose of this Technical Report is to describe the process for the operational administration 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 2018–2019 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 2018 and summer 2019 U.S. History operational forms, and the spring 2019 operational forms with embedded field test (EFT) Items. Table 1.1 summarizes the key activities during which the activities were completed.

Table 1.1

Key Activities from November 2017 to August 2019

Date	Activity
November 2017	• Started item development planning for spring 2019 field test
December 2017–February 2018	• Item development plans approved • Style guide updated • LDOE staff conducted source review committees • WestEd began item writing and development • WestEd updated 2018–2019 Assessment Framework document
March 2018	• 2018–2019 Assessment Framework document proposed
February–May 2018	• LDOE staff reviewed proposed content development
June–July 2018	• Item Content/Bias Review Committee convened • Reconciliation meeting held between LDOE and WestEd staff • 2018–2018 Assessment Framework document approved • Test construction activities began • Data review for spring 2018 field test and operational items • Standard setting committee convened
August–October 2018	• Fall 2018 content delivered to administration vendor • Biannual planning meeting held • LDOE staff reviewed proposed spring 2019 operational and field test selections
October–November 2018	• Initial batch of spring 2019 content delivered to administration vendor, including online forms and accommodated print
November 2018	• Technical Advisory Committee Meeting convened
November–December 2018	• Fall 2018 tests administered
December 2018	• Remaining batches of content delivered to administration vendor
January 2019	• Biannual planning meeting held
March 2019	• Technical Advisory Committee Meeting convened
April–May 2019	• Spring 2019 tests administered, including field test items
August 2019	• Data reviewed to verify accuracy of spring 2019 field test 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 2018–2019 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. Tables 3.1 and 3.2 show the item development plan for U.S History in 2018–2019.

Table 3.1
Item Development Plan for New Field Test Items, 2018–2019

		Total Sets	Total Items per Set	MC/MS	CR	TE	ER	Total Items
2019	Tasks	5	12	50	–	–	10	60
	Standalone Items (MC/MS)	–	–	10	–	–	–	10
	TOTALS	5	–	60	0	0	10	70

Table 3.2
Item Development Plan for Revise and Re-field Test Items, 2018–2019

		Total Sets	Total Items per Set	MC/MS	CR	TE	ER	Total Items
2019	Item sets	5	6–13	39	5	10	–	54
	Tasks	1	11	9	–	–	2	11
	Standalone Items (MC/MS)	–	–	–	–	–	–	0
	TOTALS	6	–	48	5	10	2	65

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, 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 tasks 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 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 stimuli 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 stimulus. A goal in standalone item development is to have a stimulus 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.

Obtaining LDOE Approval for Topics

For tasks and item sets, WestEd submits lists of proposed topics to the LDOE for review prior to item development. These lists describe the topics, and possible related stimuli 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, stimulus searching and development of tasks and item sets begin.

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

Identifying Stimuli

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.

Both experienced internal and external editors locate appropriate stimuli for tasks, item sets, and standalone items. Before the stimuli 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 Stimulus Search Training Agenda (2018–2019) in Appendix A.

All stimuli are submitted to WestEd 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 stimuli from the public domain or created internally to a maximum of 20% permissioned work. The actual percentage of permissioned work for the 2018–2019 development cycle was 19% permissioned work and 81% in the public domain or created internally by WestEd. 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 Stimuli. 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 educators during content reviews to verify that sources are accessible to students and make changes based on their feedback.

Most of the stimuli chosen as part of the 2018–2019 development cycle 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 such as footnotes was added 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 Stimuli

As stimuli 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 stimuli associated with each item, and provide rough drafts of the item stems. WestEd then submits the set overviews and stimuli to the LDOE for another round of approval before beginning item writing.

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

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 the editors receive training from WestEd that outlines lessons learned from previous development cycles, LDOE expectations, and best practices for item development, including bias and sensitivity. For an outline of the information covered at the 2018–2019 training see Appendix A for the LEAP 2025 U.S. History Item Editor Training Agenda (2018–2019).

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 stimuli, 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 stimuli. The ability to examine the items and stimuli 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 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 2018.

Table 3.3
Representation of Educators Participating in June 2018 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	3	1	0

*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 stimuli and items individually before coming together in an online meeting format to discuss the items and stimuli 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 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 stimuli. 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 stimuli and items. Table 3.5 shows the results of the group votes after discussing and reaching consensus on the disposition of the stimuli and items.

Table 3.4

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

Grade	Number of Stimuli/Items	Accept	Accept with Edits	No Vote	Reject	Grand Total
USH	107	941	90*	5	24	1060

*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 Stimuli and Items

Grade	Number of Stimuli/Items	Accept	Accept with Edits	No Vote	Reject
USH	107	59	48	0	0

Post-Committee Finalization. At the conclusion of the content and bias reviews, WestEd content leads consult with the LDOE to reconcile any unresolved committee feedback. Following implementation of the committee's feedback, the LDOE and WestEd content leads meet virtually for final item reconciliation. WestEd provides records of all 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 stimuli 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.

4. Construction of Test Forms

Initial Construction

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

2018–2019 Operational Forms

In 2018–2019, reorganized versions of the 2017–2018 test forms were given for all administrations except the spring 2019 operational (OP) administration during which a new form was given. The 2017–2018 forms were altered prior to the 2018–2019 administrations to better reflect updates to the design. Changes to these forms included movement of some item sets from one session to another within the test forms and addition of one placeholder item per form. Placeholder items were introduced to match the adjusted sequence and to make the testing experience comparable for students across administrations. Table 4.1 provides the test composition for these forms.

Table 4.1

U.S. History Test Composition 2018–2019 Forms, excluding Spring 2019 OP

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
6-Item Set with TE	1	6	8	4	0	2	0	6	8
5-Item Set	3	5	6	12	0	3	0	15	18
4-Item Set	1	4	5	3	0	1	0	4	5
Standalone Items	0	0	0	10	0	0	0	10	10
Task	1	5	12	4	0	0	1	5	12
Total	**8**			**41**	**2**	**8**	**1**	**52**	**69**

WestEd and the LDOE content staff worked together to complete item selection for one new form for the spring 2019 OP administration. Content specialists drew from a pool that included data 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 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 content in one item provides clues to the answer of another item. Clanging refers to overlap or similarity of content. Because 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 4.2 provides the test composition for the U.S. History spring 2019 OP form.

Table 4.2

U.S. History Test Composition for Spring 2019 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
6-Item Set with TE	1	6	7	5	0	1	0	6	7
5-Item Set	3	5	6	12	0	3	0	15	18
4-Item Set	1	4	5	3	0	1	0	4	5
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 4.3 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 2018–2019. The table also shows the number of points excluding the task and the CRs, which are not included in the reporting category percentages in the blueprint because the standards addressed by the task and the CRs may vary by form.

Table 4.3

U.S. History Operational Test Composition for the 2018–2019 Forms shown by Order of Administration (fall 2018 OP, fall 2018 AE, spring 2019 OP)

Standard	Task Alignment	SR	CR	TE	ER	Total Points
2. Western Expansion to Progressivism	√/×/×	12/8/9	2/0/2	2/4/4	0/0/0	16/12/15
3. Isolationism through the Great War	×/×/×	6/6/7	0/0/0	2/2/2	0/0/0	8/8/9
4. Becoming a World Power through World War II	√/√/×	9/11/11	0/2/0	6/4/4	0/0/0	15/17/15
5. & 6. Cold War Era and the Modern Era	√/√/√	14/16/16	2/2/2	6/6/4	8/8/8	30/32/30
Total Points Excluding Task and CRs		37/37/39	0/0/0	16/16/14	0/0/0	53/53/53
Total Points		**41/41/43**	**4/4/4**	**16/16/14**	**8/8/8**	**69/69/69**

Spring 2019 Field Test Forms

Six tasks were field tested in spring 2019. Sets were placed on multiple field test forms, with a different combination of items on each form, to ensure field testing of the maximum number of items in each set. Two versions of each task and ten standalone items were embedded across twelve field test forms. Each form included one 5-item task with 4 SR and 1 ER and four 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 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 to 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 forms prior to submission to the LDOE for their review.

LDOE Review

Following the psychometric reviews, the test maps and constructed sets are delivered to the LDOE for approval. Forms are reviewed by both 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 for High School 2025 (LEAP 2025 HS) strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations.

Training of School Systems

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

Test Administrators Manual Table of Contents

1. Notes and Reminders
2. Pre-administration Oath and Security Confidentiality Statement
3. Post-administration Oath and Security Confidentiality Statement
4. Overview
5. Test Security
   - 5.1. Secure Test Materials
   - 5.2. Testing Irregularities and Security Breaches
   - 5.3. Testing Environment
   - 5.4. Violations of Test Security
   - 5.5. Voiding Student Tests
6. Test Administrator Responsibilities
   - 6.1. Software Tools and Features for Test Administrators
7. Test Administration Checklists
   - 7.1. Before Testing
   - 7.2. During Testing
   - 7.3. After Testing (Daily)
   - 7.4. After Testing (Last Day)
8. Test Materials
   - 8.1. Receipt of Test Materials
9. Testing Guidelines
   - 9.1. Testing Eligibility
   - 9.2. Testing Schedule
   - 9.3. LEAP 2025 Testing Time

9.4. EOC Testing Time

9.5. Extended Time for Testing

9.6. Makeup Test Procedures

9.7. Testing Conditions

9.8. Accessibility Features

10. Special Populations and Accommodations

10.1. IDEA Special Education Students

10.2. Students with One or More Disabilities According to Section 504

10.3. Gifted and Talented Special Education Students

10.4. Test Accommodations for Special Education and Section 504 Students

10.5. Special Considerations for Students who Are Deaf or Hearing Impaired

10.6. English Learners (ELs)

11. Directions for Administering the LEAP 2025 Tests

12. LEAP 2025 Testing Times

13. General Information for LEAP 2025

13.1. LEAP 2025 English I and English II

13.2. LEAP 2025 Algebra I and Geometry

13.3. LEAP 2025 Biology

13.4. LEAP 2025 U.S. History

14. Directions for Administering End-of-Course Tests

15. End-of-Course Suggested Testing Times

16. General Instructions for EOC

16.1. End-of-Course English III

16.2. End-of-Course Biology

17. Post-Test Procedures

17.1. Test Administrator Post-Administration Oath of Security and Confidentiality Statement

17.2. Returning Test Materials to the School Test Coordinator

18. Index

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

Test Coordinators Manual Table of Contents

1. Key Dates
2. Resources Available in eDIRECT
3. LEAP 2025 and EOC High School Alerts
4. Pre-Administration Oath of Security and Confidentiality Statement
5. Post-Administration Oath of Security and Confidentiality Statement
6. General Information
    - 6.1. eDIRECT and INSIGHT
7. LEAP 2025/EOC High School
    - 7.1. Testing Requirements
8. Test Security
    - 8.1. Key Definitions
    - 8.2. Violations of Test Security
    - 8.3. Testing Guidelines
    - 8.4. Testing Conditions
    - 8.5. Testing Schedule
    - 8.6. Extended Time for Testing
    - 8.7. Extended Breaks
    - 8.8. Makeup Testing
    - 8.9. LEAP 2025 High School and End-of-Course Testing Times
9. Roles and Responsibilities
    - 9.1. District Test Coordinator
    - 9.2. School Test Coordinator
    - 9.3. Chief Technology Officer
10. Managing Test Tickets
    - 10.1. Student Transfers
    - 10.2. Locked Test Tickets
    - 10.3. Technical Issues
    - 10.4. Invalidating Test Tickets
11. Resources for Online Testing
    - 11.1. High School Test Administration Manual
    - 11.2. eDIRECT User Guide
    - 11.3. LEAP 2025 Accommodations and Accessibility User Guide

    11.4. INSIGHT Technology User Guide
    11.5. Student Tutorials
    11.6. Online Tools Training (OTT)
12. Post-administration Rescoring Process for LEAP 2025/EOC Tests
13. Request for Rescoring
14. Void Notification

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

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

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

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

Furthermore, the standardized procedures addressed in the TAM need to be followed, as the *Standards* state in Standard 6.1: “Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user” (114). To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it was essential that the LEAP 2025 tests were administered according to the prescribed test administration manual. It should be noted that adhering to the test schedule is also a critical component. The test coordinator manuals 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 about testing concerns observed during monitoring visits. These reports describe a wide range of improper activities that may occur during testing, including copying and reviewing test questions with students or using a calculator on parts of the test where it is not allowed.

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

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

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

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

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

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

Throughout the manuals, test coordinators and examiners are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted. Detailed information about test security procedures is presented under “Test Security” in the test administration 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, eDIRECT, and print test tickets. Students enter their ticket information to access the test in INSIGHT. In addition, students have access to Online Tools Training before the testing window, which allows them to practice using tools and features within INSIGHT. Tutorials with online video clips that demonstrate features of the system are also available to students before testing.

Accessibility and Accommodations

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

- Access for All features are available to all students taking an assessment.
- Accessibility Features are available to students when deemed appropriate by a team of educators.
- Accommodations must appear in a student’s IEP/504/EL plan.

Accommodations may be used with students who qualify under the Individuals with Disabilities Education Act (IDEA) and have an IEP or Section 504 of the Americans with

Disabilities Act and have a Section 504 plan, or who are identified as English Learners (ELs).

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

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

In compliance with this standard, the TAM contains the list of Universal Tools, Designated Supports, and Accommodations permissible for the LEAP 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 2018–2019 assessments were administered to students within the state testing windows of November 28 to December 14, 2018; April 15 to May 17, 2019; and June 17–21, 2019.

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

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

Response Change Analysis

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

Score Fluctuation Analysis

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

Web Monitoring

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

Plagiarism Detection

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

6. Scoring Activities

Answer Key Verification

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

Directory of Test Specifications (DOTS) Process. DRC creates a DOTS file, based on the approved test selection. The DOTS is a document containing information about each item on a test form, such as item identifier, item sequence, answer key, score points, 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. Scoring of SR items is evaluated with TRIAN, a standardized Pearson program that calculates MC item statistics, to verify that MC items are keyed correctly (i.e., that the true correct response is applied during scoring). Items are flagged if item statistics fall outside expected ranges. For example, items are flagged if few students select the correct response (*p*-value less than 0.15), if the item does not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) select a certain incorrect response. Lists of flagged MC and MS items, with the reasons for flagging, are provided to LDOE and WestEd content staff for key verification. WestEd staff review the list of flagged MC and MS items to confirm that the answer keys are accurate. Scoring of MS items is also evaluated at data review.

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

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

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

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

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

Constructed-Response and Extended-Response Scoring

The constructed- and extended-response items were scored by human raters trained by DRC. Human scorers provided second reads to 10% of these responses as well as handscoring supervisory reviews.

Selection of Scoring Evaluators. Standard 4.20 states the following:

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

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

The Recruitment and Interview Process. DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2018–2019 LEAP 2025 high school Biology test responses had at least a four-year college degree.

DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, or recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to demonstrate their proficiency in writing by responding to a DRC writing topic and their proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and also demonstrate his or her ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

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

Handscoring Training Process. Standard 6.9 specifies:

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

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

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

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

Tables 6.1–6.4 provide the inter-rater reliability and score-point distributions for the constructed-response and extended-response items administered in the 2018–2019 forms.

Table 6.1

Operational Constructed-Response Inter-Rater Reliability

Administration	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
Fall 2018 OP	USH_Item1	≥1,720	80	20	0
	USH_Item2	≥1,690	88	12	0
Spring 2019 OP	USH_Item1	≥11,660	88	12	0
	USH_Item2	≥13,490	91	9	0
Spring 2019 SR	USH_Item1	≥590	94	6	0
	USH_Item2	≥670	91	9	0
Summer 2019 OP	USH_Item1	≥1,560	99	1	0
	USH_Item2	≥1,750	96	4	0

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

Table 6.2

Operational Constructed-Response Score Point Distributions

Administration	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
Fall 2018 OP	USH_Item1	≥9,370	43	39	18	0
	USH_Item2	≥9,200	75	15	10	0
Spring 2019 OP	USH_Item1	≥44,400	31	39	21	0
	USH_Item2	≥44,970	35	29	22	0
Spring 2019 SR	USH_Item1	≥3,060	54	7	5	0
	USH_Item2	≥3,180	60	18	3	0
Summer 2019 OP	USH_Item1	≥3,400	59	3	1	0
	USH_Item2	≥3,520	33	23	4	0

Table 6.3

Operational Extended-Response Inter-Rater Reliability

Administration	Item		Inter-Rater Reliability			
		2x	Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
Fall 2018 OP	USH_Item1	≥8,050	Content	91	9	0
			Claims	92	8	0
Spring 2019 OP	USH_Item1	≥29,250	Content	93	7	0
			Claims	93	7	0
Spring 2019 SR	USH_Item1	≥3,510	Content	97	2	0
			Claims	98	2	0
Summer 2019 OP	USH_Item1	≥2,740	Content	95	5	0
			Claims	97	3	0

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

Table 6.4

Operational Extended-Response Score Point Distributions

Administration	Item	Total	Score Point Distribution						
			Dimension	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent "3" Rating	Percent "4" Rating	Percent Blank
Fall 2018 OP	Item 1	≥12,510	Content	39	32	19	5	2	0
			Claims	45	27	18	5	1	0
Spring 2019 OP	Item 1	≥53,430	Content	22	35	22	10	2	0
			Claims	29	29	21	9	2	0
Spring 2019 SR	Item 1	≥4,620	Content	39	24	6	1	1	0
			Claims	50	16	4	1	0	0
Summer 2019 OP	Item 1	≥4,030	Content	44	26	2	0	0	0
			Claims	52	19	2	0	0	0

Table 6.5

Field Test Extended-Response Inter-Rater Reliability

Administration	Item		Inter-Rater Reliability			
		2x	Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
Spring 2019 FT	1	≥5,000	Content	77	22	1
			Claims	78	21	1
	2	≥5,000	Content	71	28	1
			Claims	71	27	1
	3	≥5,000	Content	76	23	1
			Claims	74	25	1
	4	≥5,000	Content	77	21	2
			Claims	73	24	3
	5	≥5,000	Content	76	22	2
			Claims	76	22	2
	6	≥5,000	Content	76	22	2
			Claims	76	22	2
	7	≥5,000	Content	69	29	2
			Claims	69	29	2
	8	≥5,000	Content	70	28	1
			Claims	71	28	1
	9	≥5,000	Content	67	31	2
			Claims	68	30	2
	10	≥5,000	Content	70	28	2
			Claims	69	29	2
	11	≥5,000	Content	72	26	1
			Claims	71	27	1
	12	≥5,000	Content	74	25	1
			Claims	73	26	1

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

Table 6.6

Field Test Extended-Response Score Point Distributions

Administration	Item	Total	Score Point Distribution						
			Dimension	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent "3" Rating	Percent "4" Rating	Percent Blank
Spring 2019 FT	1	≥5,000	Content	35	31	11	5	2	0
			Claims	35	31	11	5	2	0
	2	≥5,000	Content	25	43	19	7	2	0
			Claims	26	41	19	7	2	0
	3	≥5,000	Content	12	54	22	6	3	0
			Claims	18	49	21	6	3	0
	4	≥5,000	Content	20	52	18	5	3	0
			Claims	36	37	17	5	3	0
	5	≥5,000	Content	33	35	18	7	2	0
			Claims	35	34	18	6	2	0
	6	≥5,000	Content	41	30	16	5	2	0
			Claims	43	29	15	5	2	0
	7	≥5,000	Content	35	38	19	4	1	0
			Claims	39	34	18	4	1	0
	8	≥5,000	Content	32	37	19	6	2	0
			Claims	33	35	19	6	2	0
	9	≥5,000	Content	22	40	24	8	3	0
			Claims	18	43	23	8	3	0
	10	≥5,000	Content	25	40	22	8	2	0
			Claims	24	42	22	7	2	0
	11	≥5,000	Content	33	38	17	5	1	0
			Claims	30	41	17	5	1	0
	12	≥5,000	Content	40	36	15	3	1	0
			Claims	40	37	14	3	1	0

7. Data Analysis

Classical Item Statistics

Appendix C: Item Analysis Summary Report includes tables and figures that provide the information on classical item statistics for operational items. Tables C.1–C.5 show summaries of classical item statistics. As a measure of item difficulty, *p* (or "the *p*-value") indicates the average proportion of total points earned on an item. For example, if $p = 0.50$ on an MC item, then half of the examinees earned a score of 1. If $p = 0.50$ on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). The corrected point-biserial correlation is a measure of item discrimination. Items with higher item-total correlations provide better information about how well items discriminate between lower- and higher-performing students. It should be noted that statistical results of FT items are stored in Pearson ABBI.

Differential Item Functioning

Differential item functioning (DIF) analyses are intended to statistically signal potential item bias. DIF is defined as a difference between similar ability groups' (e.g., males or females that attain the same total test score) probability of getting an item correct. Because test scores can reflect many sources of variation, the test developers' task is to create assessments that measure the intended knowledge and skills without introducing construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores may reflect those extraneous elements in addition to what the test is purported to measure. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). Different cultural and socioeconomic experiences are among some factors that can confound test scores intended to reflect the measured construct.

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

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

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

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

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

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

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

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

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

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

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

Three DIF analyses were conducted for operational test items: 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 black or 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 B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Very few operational test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the *MH* statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 *p*-value and *SMD* statistics. Table 7.3 summarizes the operational-test DIF statistics for the operational items on the 2019 spring test forms.

All items exhibiting statistical DIF were reviewed by the LDOE and WestEd content staff. Per the LDOE's standard practice, if multiple items exhibiting statistical DIF must be used on a test, the items to be used are purposefully reviewed and selected to ensure that the DIF flags do not consistently favor or disfavor the same comparison group. At the 2019 data review, no items were found to exhibit bias, and no items were rejected from the prospective item pool strictly on the basis of DIF analysis results and content reviews.

Table 7.3
Summary of DIF Flags for Operational Items for U.S. History

Comparison Groups	A	[B],[B-]	[C],[C-]
Female – Male	48	[1],[1]	[2],[1]
African American – White	38	[6],[5]	[2],[2]
Hispanic – White	50	[2],[1]	[0],[0]

The results of classical test theoretic data analyses—item *p*-values, item discrimination indices, and *MH DIF* indices—and analyses based on item theoretic methods are reviewed by committees of Louisiana educators for potential bias. It should be noted that for data review on field test item analysis results, particularly, any statistically flagged items evaluated for and determined to present potential bias are rejected from inclusion in the item pool.

Item Calibration and Scaling

The LEAP 2025 U.S. History assessment is a standards-based assessment that has been constructed to align rigorously to the Louisiana Student Standards for Social Studies, 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 what students need to master in order to achieve proficiency. Constructing tests that are aligned 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 for the LEAP 2025 U.S. History test were independently replicated by Pearson staff as an added quality-control check.

Scaling is the process whereby we associate student performance with some ordered value, typically a number. 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, 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. Typically, a scaled metric is used, on which test forms from different years are equated.

Measurement Models

IRTPRO, a software application for item calibration and test scoring, was used to estimate item response theory (IRT) parameters from LEAP 2025 data. Multiple-Choice (MC) and Multiple-Select (MS) items were both scored dichotomously (0/1), so the 3-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.

This operational test also included three types of polytomous items: TEs scored 0–2, CR items scored 0–2, and ER items scored on two 0–4 traits. Data from polytomous items were used to estimate parameters for the generalized partial credit model (GPCM) (Muraki, 1992):

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

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 M_i is the number of score categories of item *i* with possible item scores as consecutive integers from 0 to M_i – 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.).

Operational Item Parameters

The distributions of item parameters are summarized in Table C.6. Figures in Appendix C provide graphical displays of the distributions of IRT parameter estimates for each grade. The IRT *a*-parameter, or the discrimination parameter, represents the relationship between the probability of a correct response and increasing ability. The IRT *b*-parameter, or the location parameter, represents the difficulty of the item on the latent trait scale. The IRT *c*-parameter, or the pseudo-guessing parameter, represents an item’s lower asymptote. TE, CR, and ER items have no *c* parameters because they are polytomous items and are therefore modeled using the GPCM. A desired range of item parameters can be found in the framework used for test construction. It should be noted that statistical results of FT items are stored in 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.

Yen's Q$_1$ statistic (Yen, 1981) was calculated to evaluate item fit for 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

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

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

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

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 $ZQ_{1,\mathrm{crit}}$ 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). As reported in Appendix D: Dimensionality, the number of operational items flagged by the Q_1 statistic is 0 for grades 7 and 8 and 1 to 3 for other grades, which is quite negligible.

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 strong, single 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 suggests a relatively straightforward 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 can be thought of as what is left in performance after accounting for underlying achievement. If performance on an item is driven by a single achievement construct, then not only will the residual be small (as tested by the Q_1 statistic), but the correlation between residuals of the pair of items also will be small. These correlations are analogous to partial correlations, which can be interpreted as the relationship between two variables (items) after the effects of a third variable (underlying achievement) are held constant or "accounted for." 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

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

A correlation between d_i and d_j values is a 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 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.

Unidimensionality and Principal Component Analysis

It should be noted that Appendix D provides information about principal component analysis of grades 3–9 science. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, in the case of scholastic achievement, 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 2019 LEAP assessments, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using the SAS program. The first and the second principal component eigenvalues were compared *without rotation*. Table D.4 and Figure D.4 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 figures, the first component is substantially larger than the second eigenvalue across the assessments: the first eigenvalue was at least 8 times as big as the second eigenvalue. In addition, the figure indicates that the second component sharply drops from the first and gets flat. As a result, we could conclude that the unidimensionality assumption of 2019 assessment was met.

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 are subsequently interpolated. 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.

8. Reporting for U.S. History

Score reports are the primary means of communicating test scores to appropriate school system personnel (e.g., testing coordinators or superintendents), teachers, and parents.

Standard 6.10 of the *Standards* states:

> When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (119)

Standard 5.1 is related to Standard 6.10. It states:

> Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (102)

Interpretations of test scores from each administration are disseminated in two ways: the individual score report and the *LEAP Interpretive Guide*.

In addition to providing interpretations of test results, the LDOE and DRC must ensure that those interpretations are understandable for the target audience. Standard 7.0 states:

> Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (125)

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 2018–2019 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.

In this section, descriptions of the School Roster Report and the ISR are provided.

In compliance with AERA, APA, & NCME (2014) Standard 12.18, the LEAP 2025 score reports provide clear information about the results of individual students and of specific groups of students. Standard 12.18 states:

> In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (200)

School Roster Report

A School Roster Report, which provides summary information about student performance on the LEAP 2025 high school ELA and mathematics assessments, 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 Interpretive Guide.

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.

A data file referred to as the Louisiana Department of Education Student File (LDESTD) was provided to the LDOE by DRC. It contains one record for every student tested; each record contains demographic information, responses for multiple-choice (MC) items, scores for items that are not MC items, raw scores, content and process standard raw scores, scale scores, and performance-level data for each content area.

The LEAP Interpretive Guide was written to help Louisiana school system and school administrators, teachers, parents, and the general public understand the LEAP 2025 ELA and mathematics tests. The *LEAP Interpretive Guide* was developed collaboratively by DRC and LDOE staff. LDOE staff had opportunities to review the guide, provide feedback, and give final approval.

The LEAP Interpretive Guide has three sections. The first section presents an introduction and an overview of key terms and test-related concepts. The second section discusses assessment terms and types of scores that are presented on the ISRs. Sample ISRs are included in the guide. The third section discusses information that is presented on the School Roster Report and an example of the report.

9. 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 9 summarizes the decisions by item type for data-reviewed items field tested in spring 2019.

Table 9
FT Item Decisions by Item Type, 2019 Data Review

Item Type		Number of Items			
	Field Tested	Accepted	Accept with Edits	Reject	% of Total
MC	57	52	4	1	81.43
MS	1	–	–	1	1.43
TE	–	–	–	–	–
CR	–	–	–	–	–
ER	12	8	3	1	17.14
Total	70	60	8	2	100.00

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

10. Reliability and Validity

Internal Consistency Reliability Estimation

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

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

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the ith item (or component), and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 U.S. History tests is evidenced through a dimensionality analysis. Dimensionality analyses results are discussed in “Chapter 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. Coefficient alpha value was 0.93, and the marginal alpha value was 0.93. Marginal reliability is described as “an average reliability over levels of θ or theta” (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 “posterior standard deviations” (SEMs) in the IRTPRO output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31 θs. Marginal reliability is the average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals. The formula for marginal reliability is

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

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

Additional reliabilities were calculated on various demographic subgroups[1] using the population of students (see Appendix F: Reliability and Classification Accuracy). Included with coefficient alpha in the tables is 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.

Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their scale scores. It is important to know the reliability of student scores in any examination, but assessing the reliability of the classification decisions based on these scores is of even greater importance. Classification decision reliability is estimated by the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis

[1] The subgroups are male/female, white/Black/Hispanic/Asian/American Indian or Alaska Native/Native Hawaiian or Other Pacific Islander/multi-racial, and English Learners.

(1995) and Lee, Hanson, and Brennan (2000) to derive accuracy and consistency classification measures.

Accuracy of Classification. According to Livingston and Lewis (1995, p. 180), the classification accuracy is “the extent to which the actual classifications of the test takers agree with those that would be made on the basis of their true scores, if their true scores could somehow be known.” Accuracy estimates are calculated from cross-tabulations between “classifications based on an observable variable (scores on a test) and classifications based on an unobservable variable (the test takers’ true scores).” True score is also referred to as a hypothetical mean of scores from all possible forms of the test if they could be somehow obtained (Young & Yoon, 1998).

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

Accuracy and Consistency Indices. Three types of accuracy and consistency indices were generated: *overall*, *conditional-on-level*, and *cut point*, provided in Appendix F: Reliability and Classification Accuracy. The *overall accuracy* of performance-level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels. It is a proportion (or percentage) of correct classification across all the levels. The overall accuracy index is 0.739 for the LEAP 2025 U.S. History Assessment.

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 is 0.255 for the LEAP 2025 U.S. History Assessment.

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 is 0.542 across forms.

Consistency conditional-on-level is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all the students classified into that level (marginal entry).

Accuracy conditional-on-level is analogously computed. The only difference is that in the consistency table both row and column marginal sums are the same, whereas in the accuracy table, the sum that is based on true status is used as a total for computing accuracy conditional on level.

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

Validity

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (AERA/APA/NCME, 2009; 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 2018–2019 LEAP 2025 U.S. History test was designed and developed to provide fair and accurate scores that support appropriate, meaningful, and useful educational decisions. Validity evidence may be found in the following portions: Chapter 2

(Assessment Frameworks), Chapter 3 (Overview of the Test Development Process), Chapter 4 (Construction of Test Forms), Chapter 5 (Test Administration), Chapter 6 (Scoring Activities), Chapter 7 (Data Analysis), Chapter 8 (Reporting for U.S. History), Chapter 9 (Data Review Process and Results), Chapter 10 (Reliability and Validity), and Chapter 11 (Statistical Summaries). As the technical report has evolved, chapter by chapter, it reflects phases of the testing cycle. Each part of the technical report details the procedures and processes applied in the creation of LEAP 2025 and their results.

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 3 and 4 of the technical report address test-form development. Chapter 3 presents 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.

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 2018–2019 LEAP 2025 U.S. History test to derive scale scores from students' raw scores. Some references to introductory and advanced discussions of IRT are provided. In addition, Chapter 7 describes an analysis of DIF and includes gender and ethnicity DIF results. A summary of DIF results for the operational items is presented in Appendix C.

Chapter 8 of the technical report summarizes the test results, score distributions, and achievement level information.

Chapter 9 describes the data review process and results.

Chapter 10 addresses Cronbach's alpha and marginal alpha as measures of internal consistency and also describes analysis procedures for classification consistency and classification accuracy.

Chapter 11 reports the statistical summaries of the LEAP 2025 U.S. History assessment for 2018–2019.

Additional, corroborating evidence consistent with the validity, reliability, and consistency of the LEAP 2025 U.S. History assessment has previously been documented in the earlier LEAP U.S. History and standard setting technical reports.

11. Statistical Summaries

For all LEAP 2025 assessments including U.S. History, the lowest obtainable scale score (LOSS) on the social studies tests is 650 and the highest obtainable scale score (HOSS) is 850. Test results are provided in Table 11.1. Scale score means and standard deviations as well as the percentages of students in each performance level are reported for the state and are disaggregated by demographic groups. In addition to the descriptive statistics presented in Table 11.1, scale score frequency distributions are presented in Appendix E.

Measurement implies order and magnitude along a single dimension (Andrich, 1989). Consequently, in the case of scholastic achievement, 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, 1988; 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, the relative sizes of the eigenvalues associated with a principal component analysis of the item set will be examined using the SAS program. The first and the second principal component eigenvalues will be then compared *without rotation*. The current years' unidimensionality results can be found in Appendix D. We will continue to conduct a principal component analysis.

Table 11.1

LEAP 2025 State Test Results: 2019 Spring Operational U.S. History

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥34,550	731.17	32.29	24	16	32	20	8
Gender								
Female	≥17,750	730.57	30.78	24	17	33	19	7
Male	≥16,800	731.81	33.81	25	14	30	21	9
Ethnicity								
Hispanic/Latino	≥1,920	728.85	34.65	28	15	29	19	9
American Indian or Alaska Native	≥210	737.51	27.97	17	14	36	24	9
Asian	≥610	752.40	36.46	12	6	25	30	27
Black	≥14,850	718.17	29.93	37	20	29	12	2
Native Hawaiian or Other Pacific Islander	≥30	753.06	34.65	12	12	21	21	35
White	≥16,330	742.12	29.27	13	12	35	28	12
Multi-Racial	≥570	737.35	32.32	19	13	32	25	11
Economically Disadvantaged								
No	≥14,600	742.59	30.93	13	12	32	28	14
Yes	≥19,950	722.82	30.68	32	18	31	15	4
LEP Status								
Fully English Proficient	≥33,780	731.85	32.04	23	16	32	21	8
English Learner	≥770	701.30	28.89	62	18	15	5	0

References

AERA/APA/NCME. (1999/2014). *Standards for educational and psychological testing.* Washington, DC: Author.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mantel, N. (1963). Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association,* 58, 690–700.

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

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

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

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

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

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, 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, 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 in Education*, 10(4), 321–344.

Appendix A: Training Agendas

LEAP 2025 Social Studies Stimulus Search Training Agenda
Item Development Cycle for 2018–2019

I. **Introductions**

II. **Stimulus 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 stimuli and items

      ii. Reporting Categories

      iii. Potential Assessable GLEs

         1. Stimuli should support these GLEs

      iv. Potential Types of Stimuli

         1. The overview contains recommended stimuli that will support the task or item set

         2. Searchers can propose other stimuli that support the task or item set

      v. Stimulus Internet Source Links

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

   b. Bias and Sensitivity

      i. Bias: Avoid stimuli 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 stimuli 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 stimulus search assignments**

IV. **Submitting stimuli for assignments**

   a. Text-based stimuli

      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 stimuli

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
Item Development Cycle for 2018–2019

I. **Louisiana Student Standards and GLEs**
    a. High School
        i. Reporting Categories and Standards
        ii. Grade-Level Expectation (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. Stimuli

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 stimuli 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.

Appendix B: Test Summary

U.S. History

Contents

Table B.1

Test Blueprint Distribution by Reporting Category for Spring 2019 Operational U.S. History: Percentage of Points by Reporting Category (includes Task Items)

Reporting Category	Form F
Standard 1	22.6%
Standard 2	15.1%
Standard 3	24.5%
Standard 4	37.7%

Table B.2

GLE Coverage by Item Type: Spring 2019 Operational U.S. History

Reporting Categories		No. of Items					% of Test
		TEI	MS	MC	ER	CR	
		N	N	N	N	N	
Standard 1	US.2.1			1			1.89
	US.2.4	1		1			3.77
	US.2.5			1			1.89
	US.2.6	1		2		1	7.55
	US.2.8		1	3			7.55
	Sub-Total	2	1	8		1	22.64
Standard 2	US.3.1	1	2	2			9.43
	US.3.2			1			1.89
	US.3.5			1			1.89
	US.3.6			1			1.89
	Sub-Total	1	2	5			15.09
Standard 3	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	1					1.89
	US.4.8			1			1.89
	US.4.9			3			5.66
	US.4.10			1			1.89
	Sub-Total	2		11			24.53
Standard 4	US.5.1	1		4		1	11.32
	US.5.2			1			1.89
	US.5.3			1			1.89
	US.5.4			2			3.77
	US.5.5			1			1.89
	US.6.2			1			1.89
	US.6.3	1		3			7.55
	US.6.4			2	1		5.66
	US.6.5			1			1.89
	Sub-Total	2		16	1	1	37.74
Total		7	3	40	1	2	100.00

Table B.3

Summary of Spring 2019 EFT Item Development Field Tested Items by Item Type

Item Type	Item Count	Percent
ER	12	17%
MC	57	81%
MS	1	1%

Table B.4

Item Type Summary: Spring 2019 Operational U.S. History

Form	MC	MS	TE	CR	ER
F	40	3	7	2	1

Table B.5

Raw Score Summary: Spring 2019 Operational U.S. History

Form	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
F	≥34,550	35	14	1	69	0.50	0.47	0.93	3.68

Note: Reliability is coefficient alpha.

Table B.6

Raw Score Summary by Reporting Category: Spring 2019 Operational U.S. History

Test Form	Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
F	Standard 1	7.35	3.22	0	15	0.44	0.46	0.74	1.64
	Standard 2	5.12	2.20	0	9	0.55	0.43	0.68	1.24
	Standard 3	8.56	3.22	0	15	0.53	0.48	0.74	1.64
	Standard 4	14.17	6.68	0	30	0.50	0.50	0.88	2.31

Table B.7

Scale Score and Raw Score Summary: Spring 2019 Operational U.S. History

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥34,550	100.00	731.17	32.29	35	14
Female	≥17,750	51.38	730.57	30.78	35	13
Male	≥16,800	48.62	731.81	33.81	36	14
African American	≥14,850	42.99	718.17	29.93	29	12
American Indian or Alaska Native	≥210	0.63	737.51	27.97	38	13
Asian	≥610	1.77	752.40	36.46	44	15
Hispanic/Latino	≥1,920	5.57	728.85	34.65	34	15
Multi-Racial	≥570	1.67	737.35	32.32	38	14
Native Hawaiian or Other Pacific Islander	≥30	0.10	753.06	34.65	44	15
White	≥16,330	47.28	742.12	29.27	40	13
Economically Disadvantaged	≥19,950	57.74	722.82	30.68	32	13
English Language Learners	≥770	2.23	701.30	28.89	23	11

Note: These tables report the number of students, scaled-score means, and standard deviations for subgroups.

Appendix C: Item Analysis Summary Report

Summary Statistics Reports

U.S. History

Table C.1

P-Value Summary by Item Type: Spring 2019 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.275	0.275	0.290	0.305	0.305
ER	1	0.224	0.224	0.234	0.244	0.244
MC	40	0.334	0.479	0.576	0.644	0.744
MS	3	0.338	0.338	0.389	0.536	0.536
TE	7	0.257	0.290	0.391	0.446	0.540

Plot C.1

P-Value by Item Type: Spring 2019 Operational U.S. History

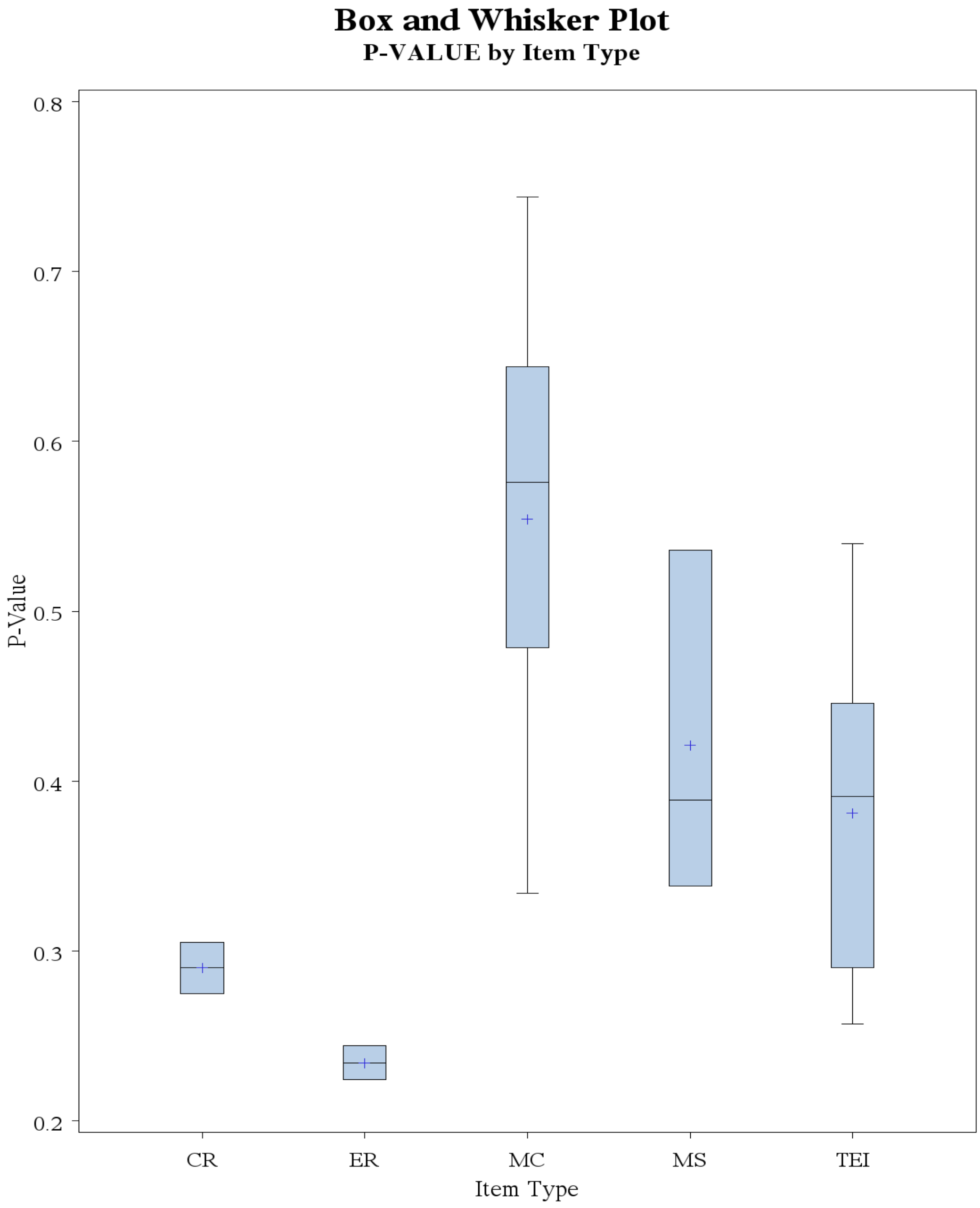


Table C.2

Item-Total Correlation by Item Type: Spring 2019 Operational U.S. History

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.552	0.552	0.629	0.706	0.706
ER	1	0.762	0.762	0.767	0.772	0.772
MC	40	0.247	0.395	0.437	0.506	0.620
MS	3	0.411	0.411	0.433	0.462	0.462
TE	7	0.448	0.458	0.551	0.572	0.620

Plot C.2

Item-Total Correlation by Item Type: Spring 2019 Operational U.S. History

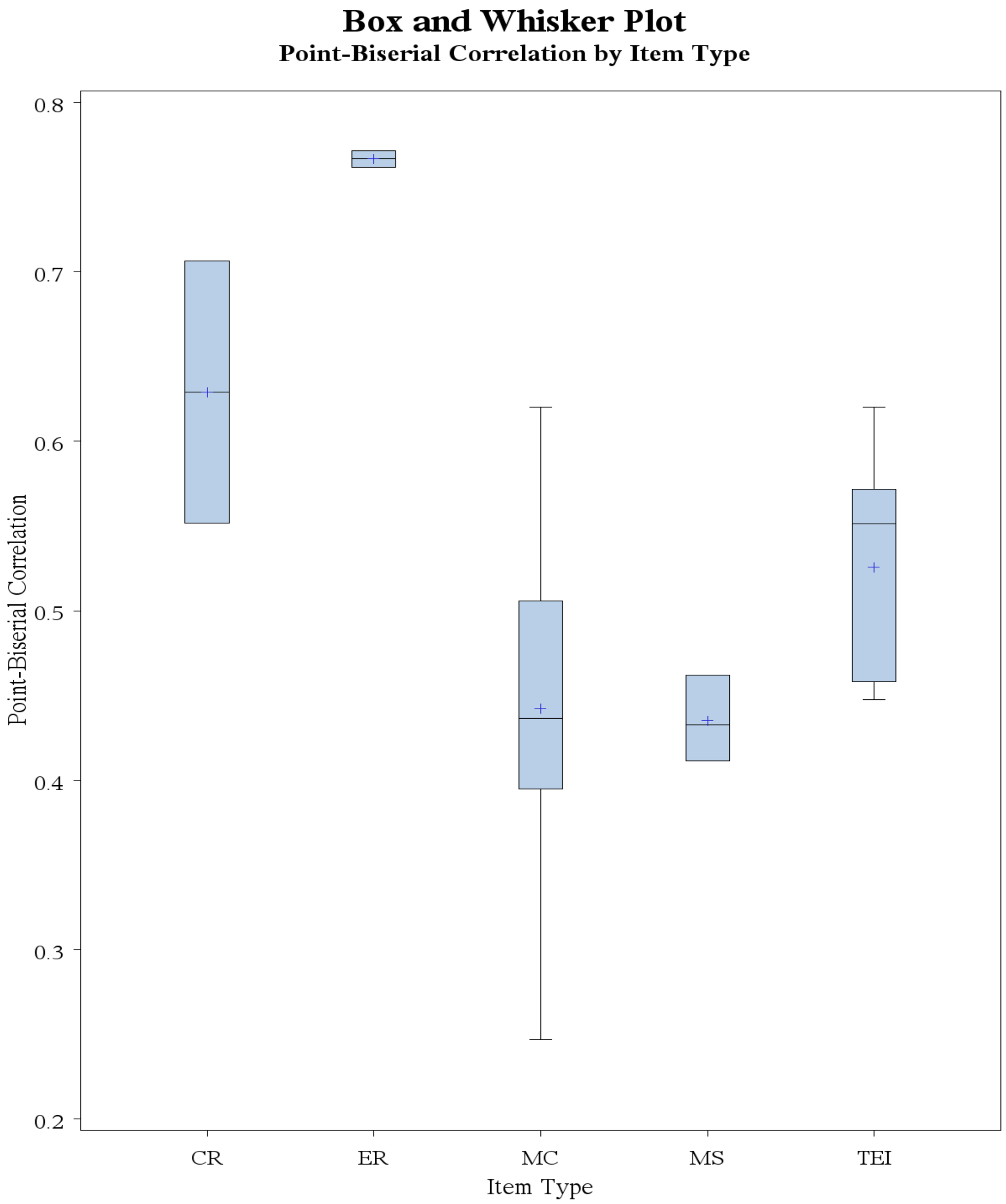


Table C.3

Corrected Point-Biserial Correlation: Spring 2019 Operational U.S. History*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.515	0.515	0.596	0.677	0.677
ER	1	0.729	0.729	0.735	0.741	0.741
MC	40	0.213	0.366	0.409	0.480	0.598
MS	3	0.382	0.382	0.403	0.435	0.435
TE	7	0.414	0.421	0.517	0.540	0.587

Note: *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.

Plot C.3

Corrected Point-Biserial Correlation by Item Type: Spring 2019 Operational U.S. History*

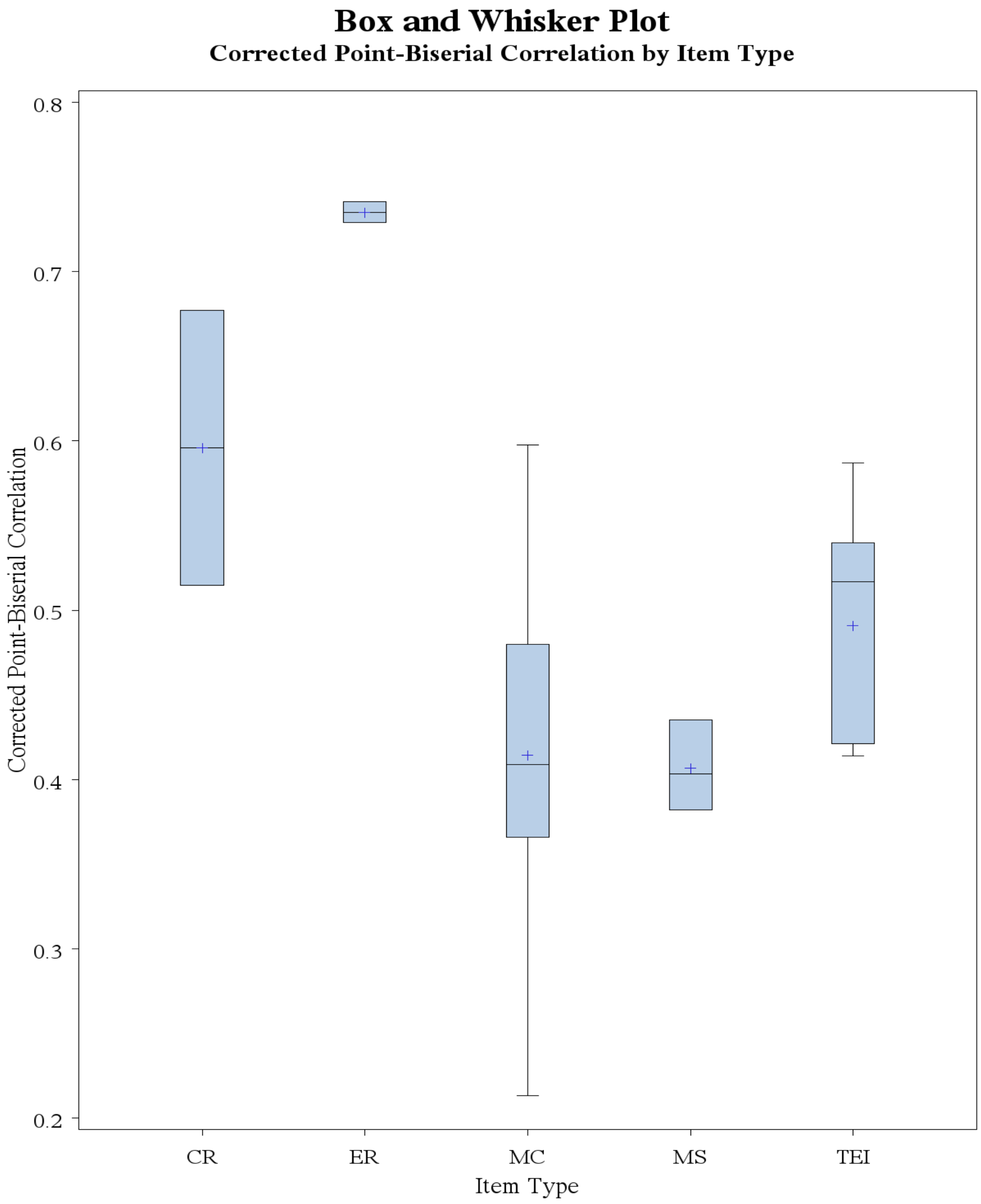


Table C.4

Item-Total Correlation by Reporting Category and Item Type: Spring 2019 Operational U.S. History

Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Standard 1	1	0.552	0.552	0.552	0.552	0.552
	Standard 4	1	0.706	0.706	0.706	0.706	0.706
ER	Standard 4	1	0.762	0.762	0.767	0.772	0.772
MC	Standard 1	8	0.247	0.345	0.440	0.519	0.574
	Standard 2	5	0.320	0.428	0.430	0.435	0.501
	Standard 3	11	0.275	0.379	0.469	0.527	0.620
	Standard 4	16	0.333	0.405	0.435	0.485	0.578
MS	Standard 1	1	0.462	0.462	0.462	0.462	0.462
	Standard 2	2	0.411	0.411	0.422	0.433	0.433
TE	Standard 1	2	0.458	0.458	0.515	0.572	0.572
	Standard 2	1	0.448	0.448	0.448	0.448	0.448
	Standard 3	2	0.563	0.563	0.591	0.620	0.620
	Standard 4	2	0.467	0.467	0.509	0.551	0.551

Table C.5

Statistically Flagged Items by Item Type: Spring 2019 Operational U.S. History

Item Type	***N* OP Items**	***N* Items Flagged for *P*-Value**	***N* Items Flagged for Mean**	***N* Items Flagged for Point-Biserial Correlation**	***N* Items Flagged for DIF**	***N* Items Flagged for Omitting**
CR	2	0	0	0	1	0
ER	1	1	1	0	1	0
MC	40	0	0	0	13	0
MS	3	0	0	0	0	0
TE	7	0	0	0	3	0

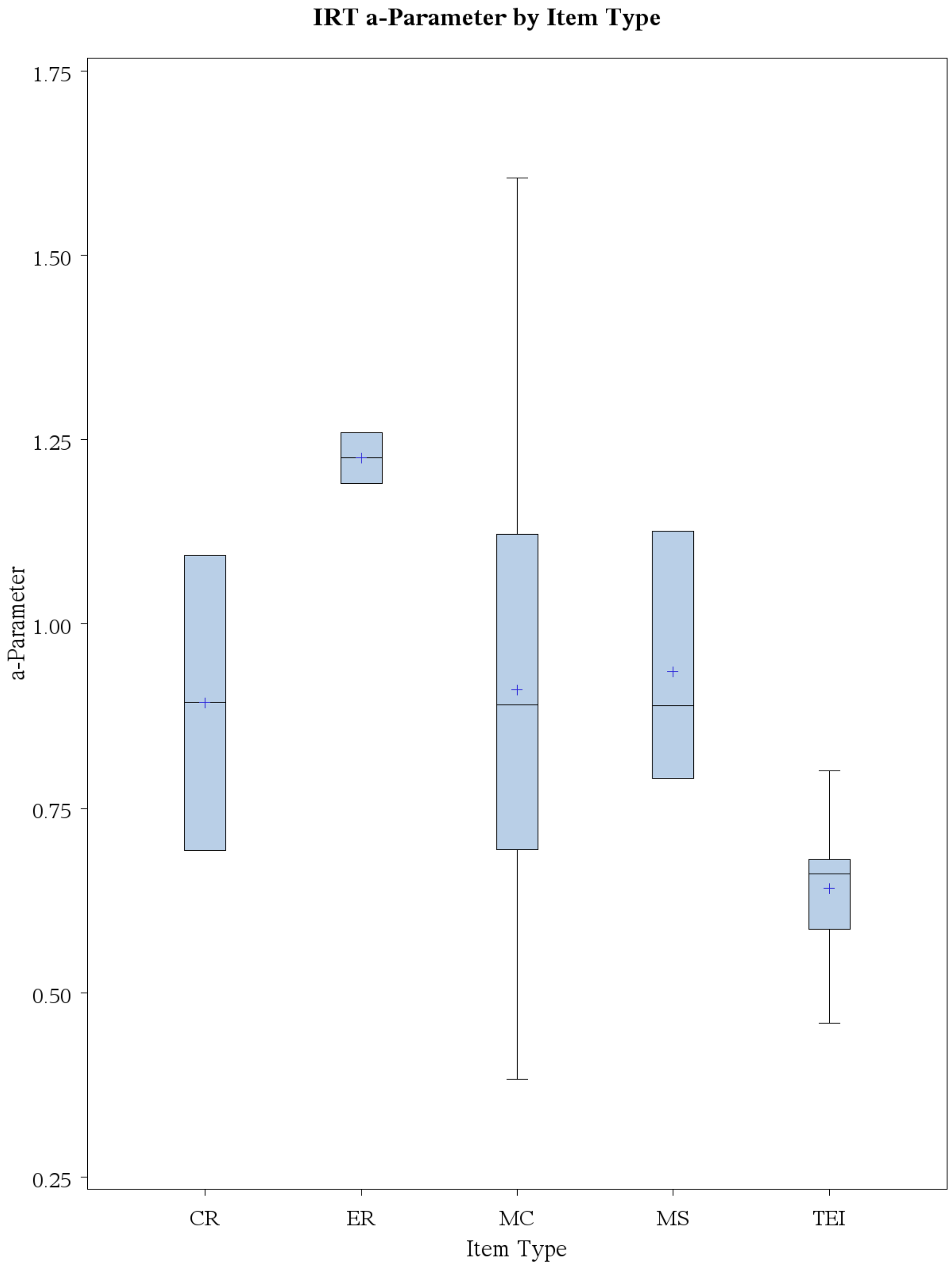
Table C.6

IRT Parameters by Item Type: Spring 2019 Operational U.S. History

Item Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	*a*	2	0.693	0.693	0.893	1.093	1.093
	b	2	0.456	0.456	0.466	0.477	0.477
ER	*a*	1	1.190	1.190	1.225	1.260	1.260
	b	1	1.019	1.019	1.046	1.073	1.073
MC	*a*	40	0.382	0.694	0.890	1.122	1.605
	b	40	-1.312	-0.495	0.012	0.492	1.815
	c	40	0.022	0.156	0.205	0.250	0.361
MS	*a*	3	0.791	0.791	0.889	1.126	1.126
	b	3	0.155	0.155	0.668	0.741	0.741
	c	3	0.066	0.066	0.077	0.146	0.146
TE	*a*	7	0.458	0.586	0.661	0.680	0.801
	b	7	-0.302	0.376	0.425	0.829	0.978

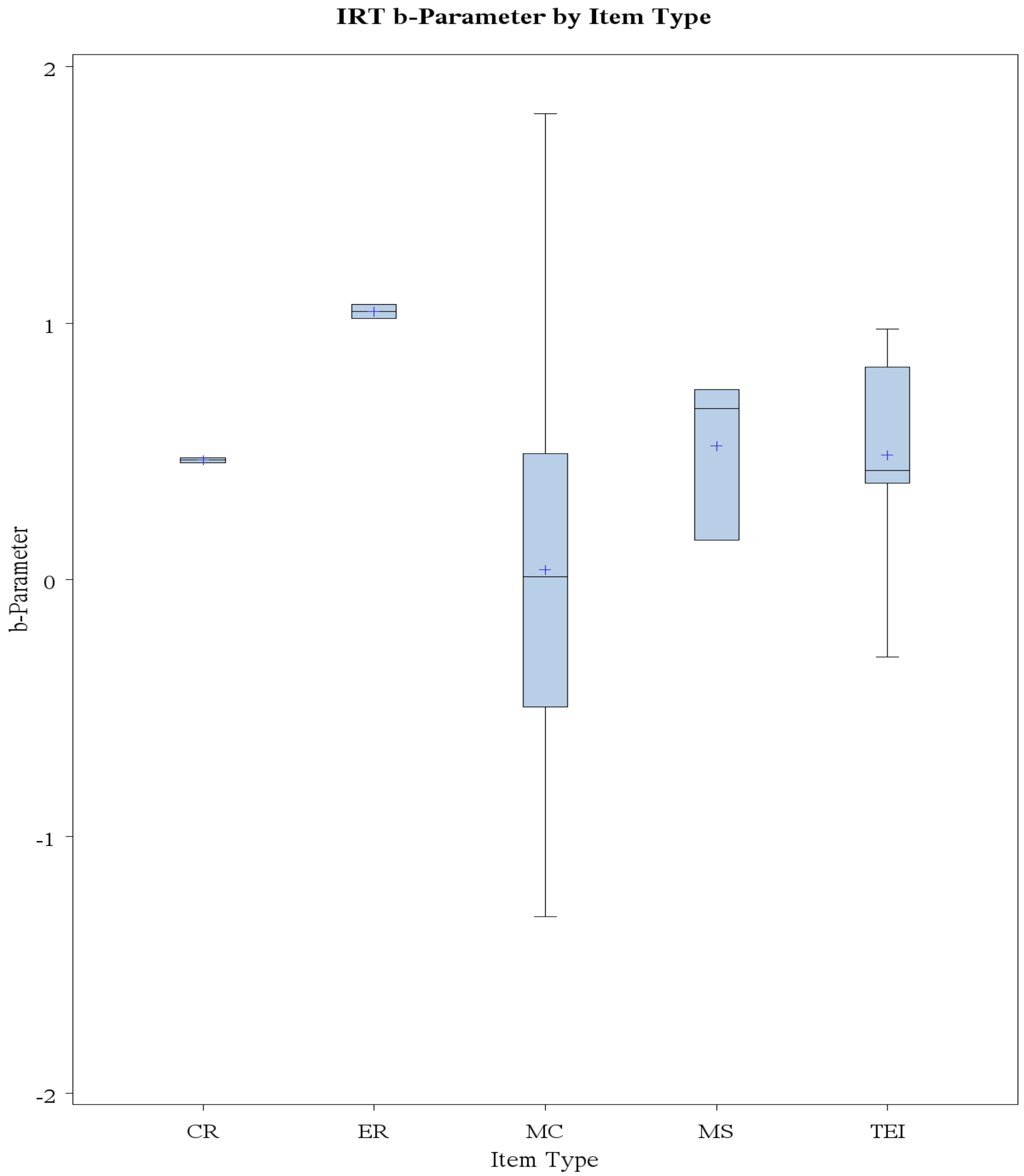
Plot C.4

IRT a-Parameter: Spring 2019 Operational U.S. History



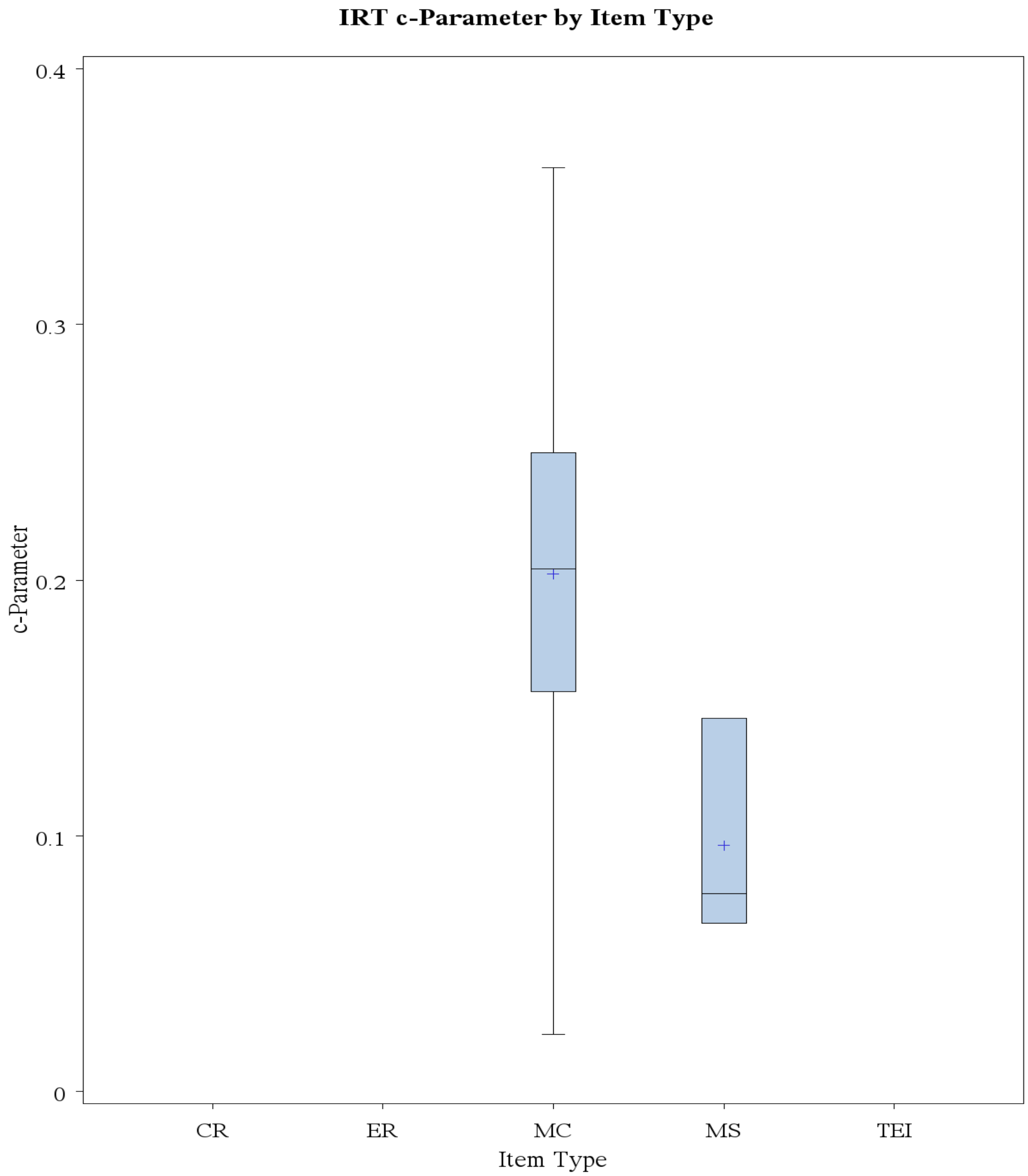
Plot C.5

IRT b-Parameter: Spring 2019 Operational U.S. History



Plot C.6

IRT c-Parameter: Spring 2019 Operational U.S. History



Note: Only dichotomous items (scored 0 or 1) have c parameters.

Appendix D: Dimensionality

Dimensionality Reports

U.S. History

Table D.1

Zq1 Statistics by Item Type: Spring 2019 Operational U.S. History

Form	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
F	CR	31	31	39	47	47	0
	ER	19	19	24	28	28	0
	MC	0	3	4	9	38	0
	MS	2	2	6	10	10	0
	TE	4	9	2	44	47	0

Table D.2

Q3 Statistics and Summary Data: Spring 2019 Operational U.S. History

Form	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
F	0.206	-0.103	-0.038	-0.017	0.025	0.908

Table D.3

Intercorrelation Coefficients among Reporting Categories: Spring 2019 Operational U.S. History

Reporting Category	Standard 1	Standard 2	Standard 3	Standard 4
Standard 1	1.00			
Standard 2	0.67	1.00		
Standard 3	0.73	0.70	1.00	
Standard 4	0.77	0.74	0.79	1.00

Table D.4

First and Second Eigenvalues: Spring 2019 Operational U.S. History

Grade	Form	First Eigenvalue	Second Eigenvalue
History	F	12.487	1.411

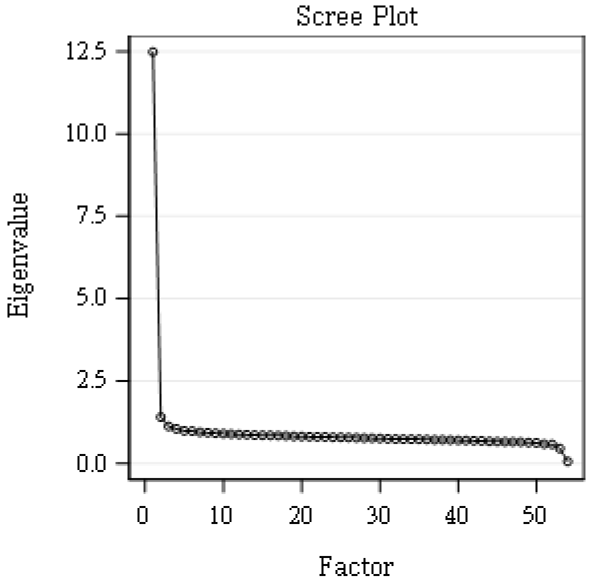


Figure D.4

Principal Component Analysis Plot: Spring 2019 Operational U.S. History

Appendix E: Scale Distribution and Statistical Report

Table E.1 *Scale Score Descriptive Statistics and Plots*

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

                         N                      ≥34550
                         Mean                   731.17      Median                732.00
                         Std deviation           32.29      Variance             1042.88
                         Skewness              -0.1604      Kurtosis              0.0239
                         Mode                   742.00      Std Error Mean        0.1737
                         Range                  200.00      Interquartile Range    42.00

                                             Quantile       Estimate

                                             100% Max          850
                                             99%               801
                                             95%               782
                                             90%               771
                                             75% Q3            753
                                             50% Median        732
                                             25% Q1            711
                                             10%               691
                                             5%                672
                                             1%                650
                                             0% Min            650

                      Histogram                          #  Boxplot                        Normal Probability Plot
855+*                                                   ≥10     0         855+                                                  *
   .                                                                        |
835+*                                                   ≥20     0         835+                                                  *
   .*                                                   ≥50     0            |                                                  *
815+*                                                   ≥80     |         815+                                                  *
   .***                                                ≥270     |            |                                               ****
795+*****                                              ≥460     |         795+                                            ****
   .***********                                       ≥1030     |            |                                         ****
775+********************                              ≥1910     |         775+                                     *****
   .**************************                        ≥2430     |            |                                  ****
755+*************************************             ≥3540  +-----+      755+                               ****
   .************************************************  ≥4620  |     |         |                           *****
735+*****************************************         ≥3920  *--+--*      735+                        ****
   .******************************************        ≥4040  |     |         |                     ****
715+****************************************          ≥3790  +-----+      715+                  ****
   .******************************                    ≥2860     |            |               ****
695+*********************                             ≥2030     |         695+            ****
   .************                                      ≥1140     |            |          ***
675+***********                                       ≥1000     |         675+       ****
   .****                                               ≥370     |            |   +++**
655+**********                                         ≥900     |         655+*******
    ----+----+----+----+----+----+----+----+----+---                         +----+----+----+----+----+----+----+----+----+----+
    * may represent up to 97 counts                                              -2        -1         0        +1        +2
```

Table E.2 *Scale Score Descriptive Statistics and Plots*

```
                                          FREQUENCY DISTRIBUTION - SCALE SCORES
                                                     U.S. HISTORY
                                                     ALL STUDENTS
                                                        Form F
SCALE_SCORE                                                                                                                Cum.     Cum.
                                                                                                     Freq    Freq   Percent  Percent
650   |**********************************************************                              ≥570    ≥570    1.68     1.68
655   |*********************************                                                       ≥320    ≥900    0.94     2.62
665   |**************************************                                                  ≥370   ≥1270    1.09     3.70
672   |************************************************                                        ≥470   ≥1750    1.38     5.08
678   |****************************************************                                    ≥520   ≥2270    1.51     6.59
683   |*********************************************************                               ≥560   ≥2840    1.64     8.23
687   |**********************************************************                              ≥570   ≥3420    1.67     9.90
691   |******************************************************************                      ≥660   ≥4080    1.91    11.81
694   |*******************************************************************                     ≥670   ≥4750    1.95    13.76
698   |**********************************************************************                  ≥690   ≥5450    2.02    15.78
701   |*********************************************************************                   ≥690   ≥6140    2.00    17.77
703   |*********************************************************************                   ≥680   ≥6830    1.99    19.76
706   |***************************************************************************             ≥740   ≥7570    2.16    21.92
708   |**************************************************************************              ≥740   ≥8310    2.14    24.06
711   |*************************************************************************               ≥720   ≥9040    2.10    26.16
713   |*************************************************************************               ≥730   ≥9770    2.12    28.28
715   |***************************************************************************             ≥750  ≥10520    2.18    30.46
717   |*********************************************************************************       ≥800  ≥11330    2.34    32.79
719   |******************************************************************************          ≥770  ≥12110    2.25    35.05
721   |************************************************************************************    ≥830  ≥12940    2.42    37.47
723   |********************************************************************************        ≥800  ≥13750    2.32    39.79
725   |*******************************************************************************         ≥790  ≥14540    2.30    42.09
727   |*********************************************************************************       ≥800  ≥15350    2.34    44.42
729   |********************************************************************************        ≥800  ≥16150    2.32    46.74
731   |********************************************************************************        ≥800  ≥16950    2.32    49.06
732   |*************************************************************************               ≥730  ≥17680    2.11    51.18
734   |*********************************************************************************       ≥800  ≥18490    2.34    53.52
736   |**********************************************************************************      ≥810  ≥19310    2.36    55.88
738   |*****************************************************************************           ≥770  ≥20080    2.23    58.10
740   |*******************************************************************************         ≥790  ≥20870    2.29    60.40
742   |************************************************************************************** ≥860  ≥21730    2.50    62.89
743   |***********************************************************************                 ≥710  ≥22440    2.06    64.95
745   |**************************************************************************              ≥740  ≥23180    2.15    67.10
747   |*****************************************************************************           ≥770  ≥23950    2.23    69.33
749   |***************************************************************************             ≥740  ≥24700    2.16    71.49
751   |************************************************************************                ≥720  ≥25420    2.09    73.58
753   |**************************************************************************              ≥740  ≥26170    2.15    75.73
755   |*************************************************************************               ≥720  ≥26890    2.10    77.83
757   |***********************************************************************                 ≥700  ≥27600    2.05    79.87
759   |*****************************************************************                       ≥640  ≥28250    1.88    81.75
761   |**************************************************************                          ≥620  ≥28870    1.79    83.54
763   |******************************************************************                      ≥650  ≥29520    1.90    85.44
766   |***********************************************************                             ≥590  ≥30110    1.72    87.15
768   |*********************************************************                               ≥560  ≥30680    1.64    88.79
771   |****************************************************                                    ≥510  ≥31200    1.50    90.29
773   |****************************************************                                    ≥510  ≥31720    1.50    91.79
776   |**********************************************                                          ≥450  ≥32170    1.33    93.11
779   |******************************************                                              ≥420  ≥32590    1.22    94.33
782   |****************************************                                                ≥390  ≥32990    1.15    95.48
785   |***********************************                                                     ≥340  ≥33340    1.01    96.49
789   |*****************************                                                           ≥280  ≥33630    0.83    97.32
793   |***************************                                                             ≥260  ≥33900    0.78    98.10
797   |********************                                                                    ≥200  ≥34100    0.58    98.68
801   |****************                                                                        ≥160  ≥34260    0.46    99.14
807   |***********                                                                             ≥110  ≥34370    0.32    99.46
813   |********                                                                                 ≥80  ≥34450    0.24    99.70
821   |*****                                                                                    ≥50  ≥34510    0.16    99.86
832   |***                                                                                      ≥20  ≥34540    0.08    99.95
850   |**                                                                                       ≥10  ≥34550    0.05   100.00
      -----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+-
          50  100  150  200  250  300  350  400  450  500  550  600  650  700  750  800  850

      Frequency
```

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

U.S. History

Table F.1

Reliability for Overall and Subgroups: 2019 Spring Operational U.S. History

Subgroup	Form F
All Students	0.932
Female	0.927
Male	0.938
African American	0.917
American Indian or Alaska Native	0.916
Asian	0.939
Hispanic/Latino	0.937
Multi-Racial	0.931
Native Hawaiian or Other Pacific Islander	0.935
White	0.923
Ethnicity Unknown	0.916
English Learners	0.895

Table F.2

Cronbach's Alpha and Marginal Reliability: 2019 Spring Operational U.S. History

Form	Cronbach's Alpha	Marginal Reliability
F	0.93	0.93

Table F.3

Classification Accuracy and Decision Consistency: 2019 Spring Operational U.S. History

Table F.3.1

Estimates of Accuracy and Consistency of Achievement Level Classification

Form	Accuracy	Consistency	PChance	Kappa
F	0.739	0.659	0.255	0.542

Table F.3.2

Accuracy of Classification at Each Achievement Level for Each Form

Form	Unsatisfactory (1)	Approaching Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
F	0.888	0.597	0.757	0.608	0.702

Table F.3.3

Accuracy of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
F	0.936	0.916	0.928	0.955

Table F.3.4

Consistency of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
F	0.908	0.884	0.898	0.95

Table F.3.5

Kappa of Dichotomous Categorizations by Form (PAC Metric)

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
F	0.789	0.766	0.709	0.125

Table F.3.6

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

Form	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
F	0.034	0.038	0.03	0.044

Table F.3.7

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

Form	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
F	0.03	0.046	0.041	.001