LEAP 2025 U.S. History Technical Report: 2017–2018 Operational and Field Test

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

FOREWORD ii

1. Introduction 1

Summary of the 2017–2018 Activities 2

2. Framework of Assessments 3

3. Overview of the Test Development Process 4

Item Development Plan 4

Proposal and Review of Topics and Sources 5

Determining Topics and Choosing Grade Level Expectations (GLEs) 5

GLE Coverage 5

Obtaining LDOE Approval for Revision of Item Sets 6

Identifying Stimuli 6

Obtaining LDOE Approval for Item Sets and Stimuli 7

Item Writing and Review Process 7

4. Construction of Test Forms 11

Initial Construction 11

Operational Forms 11

Field Test Versions 15

Revision and Review 17

Psychometric Approval of Operational Forms 17

LDOE Review 17

Streamlining of Stimuli 18
Revisiting Design 18

5. Test Administration 19

Training of School Systems 19
Ancillary Materials 20
Time 23
Online Forms Administration 23
Accessibility and Accommodations 23
Testing Windows 24
Test Security Procedures 25

6. Scoring Activities 26

Answer Key Verification 26
Constructed-Response and Extended-Response Scoring 28

7. Data Analysis 32

Classical Item Statistics 32
Differential Item Functioning 35
Item Calibration 39
Measurement Models 39
Field Test Item Parameters 40
Item Fit 44

8. Reporting for U.S. History 46

U.S. History Standard Setting 47

Standard Setting Executive Report **48**

LEAP 2025 U.S. History Standard Setting Process and Results 48

General Method 50

Results for LEAP 2025 U.S. History 52

9. Data Review Process and Results 53

10. Reliability and Validity 55

Internal Consistency Reliability Estimation 55

Student Classification Accuracy and Consistency 56

Accuracy of Classification 57

Consistency of Classification 57

Accuracy and Consistency Indices 57

Validity 58

11. Statistical Summaries 61

References 62

Appendix A: Training Agendas 66

Appendix B: Test Summary 74

Appendix C: Item Analysis Summary Report 84

Appendix D: Dimensionality 96

Appendix E: Scale Distribution and Statistical Report 101

Appendix F: Reliability and Classification Accuracy 111

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, standard setting, and reporting of scores.

Summary of the 2017–2018 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 2017 and summer 2018 U.S. History operational forms, and the spring 2018 operational forms with embedded field test. Table 1.1 summarizes those key activities along with the months during which the activities were completed.

Table 1.1
Key Activities from December 2016 to August 2018

Date	Activity
December 2016	• Started item development planning for spring 2018 EFT
January–March 2017	• Item development plans approved • Content development specifications and style guide updated • WestEd began item writing and development
March 2017	• 2017–2018 Framework and Test Construction Document proposed
March–June 2017	• LDOE staff reviewed proposed content
July 2017	• Item Content/Bias Review Committee convened • 2017–2018 Framework and Test Construction Document approved; test construction activities began
August–October 2017	• Data for spring 2017 results reviewed • LDOE staff reviewed proposed 2017–2018 test selections • Fall 2017 OP and AE materials delivered to administration vendor • Reconciliation meeting held between LDOE and WestEd staff • Planning Meeting held
October–November 2017	• LDOE staff reviewed proposed spring 2018 EFT test selections • Online content delivered to administration vendor
November 2017	• Technical Advisory Committee Meeting convened • Began planning for standard setting
November–December 2017	• Fall 2017 test administered
December 2017	• Remaining spring 2018 materials delivered to administration vendor
January 2018	• Planning Meeting held
March 2018	• Technical Advisory Committee Meeting convened • Teacher Survey
April 2018	• Pre-policy Meeting convened • Spring 2018 test administered, including EFT
July 2018	• Standard setting initiated
August 2018	• Data for spring 2018 results reviewed

2. Framework of Assessments

The assessment framework includes information on both the spring 2017 embedded field test and the first operational year. It includes:

- proposed test designs;
- test blueprints;
- the range of standards 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.

Based on the results of the Data Review for the spring 2017 field test, the 2017–2018 operational test forms were constructed. The Assessment Framework was revised to include tables reflecting the actual structure of the 2017–2018 test forms as constructed, as well as the statistical item criteria used to guide item and form selection. In addition, the field test development plan for the embedded field test items was revised to reflect plans for the current year.

3. Overview of the Test Development Process

This section describes the processes used to develop field test item sets and standalone items to embed within the spring 2018 LEAP 2025 U.S. History assessment.

Item Development Plan

WestEd proposed to revise eight item sets from the existing item bank and develop 10 standalone items, for a total of 120 items. The table below provides the overview of the development plans for item sets and standalone items. Revised item sets included a combination of revised items and new items. All of the standalone items were new items.

Table 3.1

Item Development Plan for Embedded Field Test

		Total Item Sets	Total Items per Set	SR	CR	TE	ER	Total Items
2018 EFT	6–item w/ 2CR	–	–	–	–	–	–	–
	6–item w/ 2TE	–	–	–	–	–	–	–
	6–item w/ 1CR +2TE	8	~13	86	8	16	0	**110**
	Standalone Items (MC/MS)	0	–	10	0	0	0	**10**
	TOTALS ACROSS ALL ITEM SETS	**8**	–	**96**	**8**	**16**	**0**	**120**

Proposal and Review of Topics and Sources

Determining Topics and Choosing Grade Level Expectations (GLEs)

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 item sets should be revised and re-field tested. The content lead also generated a list of topics for standalone items.

When identifying a topic, the WestEd content lead considered:

- what content is eligible for each GLE per the U.S. History standards;
- which GLEs were not covered in great numbers in the item pool;
- which GLEs could appear together in sets to provide meaningful assessment of content and concepts; and
- how a topic could tie into the LDOE's goal of assessing larger themes rather than discrete facts.

Topics were chosen to represent the breadth of assessable U.S. History content. Choosing which GLEs to assess was central to determining standalone items. The process of choosing GLEs was iterative and included the identification of potential GLEs that could be assessed together within a single topic, as well as an understanding of the need to address as many GLEs as possible in the field test.

GLE Coverage

By the end of the second year of development in U.S. History, WestEd had developed at least 1 item aligned to each of the 35 assessable GLEs that are 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, alignment to Standard 1 GLEs demonstrates that the items address a range of historical thinking skills.

Obtaining LDOE Approval for Revision of Item Sets

For item sets, WestEd submitted lists of proposed item sets from the existing item bank to the LDOE for review to determine which ones would be eligible for revision. These lists described topics, the number of items and stimuli associated with each item set, the reporting category for each item set, and the development requirements for revising each item set so that it would be compliant with current LDOE expectations. Once the LDOE approved the proposed item sets to be revised and expanded for the development cycle, stimulus searching and development of item sets began.

Identifying Stimuli

The stimuli for the LEAP 2025 U.S. History assessment included authentic historical and contemporary documents, including letters, speeches, photographs, paintings, reports, and other primary source documents, such as authentic newspaper articles. The assessment also included secondary sources, such as book excerpts. These materials are supplemented by timelines, maps, tables, charts, and graphic organizers created by WestEd's design team. WestEd used both experienced internal and external U.S. History assessment editors as searchers to locate appropriate stimuli for item sets and standalone items. Before the editors began the process, the WestEd content lead trained them on the stimulus searching process, on the LDOE's objectives, and on best practices for accessibility, as well as on bias and sensitivity issues. For an outline of the training, see Appendix A for the LEAP 2025 U.S. History Stimulus Search Training Agenda (2017–2018).

Public Domain versus Permissioned Work. WestEd endeavored to maintain a ratio of 80% of the stimuli as royalty free, drawing from sources in the public domain or created internally, to 20% permissioned work. The actual percentage of permissioned work in the field test was 5% permissioned work and 95% in the public domain or created internally by WestEd. Before administration of the assessment, WestEd's permissions coordinator obtained permission from the rights holders for five years of use of any work that required permission.

Evaluating the Reading Level of Stimuli. WestEd performed a Lexile analysis on each passage of the revised 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. In addition to the Lexile measure, the *Children's Writer's Word Book* (Mogilner, 2006) and the *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies* (Steck-Vaughn, 1989) were used as additional measures of grade-level appropriateness. WestEd and the LDOE also drew on the professional experience of educators, during Content and Bias Committee review, to verify that sources would be accessible to students, and made changes based on their feedback. Most of the stimuli chosen for the assessments 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. The appropriateness of the stimuli for both content and readability was an explicit part of the content review process with Louisiana teachers.

Obtaining LDOE Approval for Item Sets and Stimuli

The WestEd content lead worked with LDOE staff to confirm that any new stimuli were appropriate for use as the sets were revised.

For standalone items, WestEd submitted the items along with their corresponding stimuli; there was no separate stimulus approval phase for the topics or stimuli for these items.

Item Writing and Review Process

WestEd's best practice is to employ external writers and editors for U.S. History. WestEd's general process is to get LDOE approval for new writers on a project. The writers and the editors received training from WestEd that outlined lessons learned from the previous cycle of development, reviewed LDOE expectations, and presented best practices for item development, including bias and sensitivity. The LDOE did not participate in the writer and editor training. For an outline of the information covered, see Appendix A for the LEAP 2025 U.S. History Item Writer and Editor Training Agenda (2017–2018).

After the training, item writers were provided with approved item overviews, which identified the topics, listed the GLEs to be addressed, and offered specific guidance to the

item writer about how the content for each item should be addressed. The use of item overviews allowed WestEd to control the quality of the standalone items and item sets during the item development cycle. Once written, items went 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 had two rounds of review prior to Content and Bias Review Committee meetings. WestEd revised items based on the feedback provided by the LDOE assessment staff.

After the training, editors were provided with documentation for the item sets to be revised and expanded. They were instructed to revise the item sets to match the current LDOE style as well as to develop new items to meet current LDOE expectations and requirements for field testing two versions of each item set. Each item went through two rounds of editing, one round of proofreading, and one final round of review before being submitted to the LDOE for their first round of review. The LDOE had two rounds of review prior to Content and Bias Review Committee meetings. WestEd revised items based on the feedback provided by the LDOE assessment staff.

Item Development Platform. Items were 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 captured item metadata and allowed viewers to preview items using Pearson's format viewer (TestNav 8). In this view, items appeared together with their associated stimuli in the set. The ability to examine the items and stimuli as a set was critical in the item review and in the evaluation of each set's content and cognitive demands on students.

Style Guidelines. The *LEAP 2025 U.S. History Content Style Guide* was updated immediately following spring 2017 test construction to reflect final formatting decisions made by the LDOE. Throughout the development phase, when questions of style arose that were unanswered by existing documentation, WestEd consulted the LDOE, and approved changes were added to the Style Guide throughout the development and review process.

LDOE Content Review. As writing and editing for batches of item sets and standalone items were completed, the batches were sent to the LDOE for content lead review. Feedback from the LDOE review was implemented before content and bias review.

Content and Bias Review. After the completion of item development and the initial rounds of LDOE review, WestEd coordinated virtual content and bias review meetings. The meetings were led jointly by facilitators from the LDOE and WestEd. Participants included current classroom teachers, content specialists, and school administrators. The LDOE recruited the participating educators, who represented schools across the state. Table 3.2 provides information about the representation of educators who participated in the content and bias reviews.

Table 3.2

Representation of Educators Participating in 2017–2018 Content and Bias Reviews

Grade Level	Teacher	School System Coordinator	VI/HI Teacher
USH	5	1	1

Training and Security for Online Review. The virtual format of content and bias review allowed 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 provided 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 described the criteria for evaluating items for content and bias considerations, explained how to use ABBI for item review, and showed educators how to individually review the items and record their recommendation to accept, accept with edits, or reject an item.

Committee members were provided a pre-review day during which they accessed the items using the ABBI tool and voted on the items. Comments were compiled and shared with the LDOE and WestEd facilitators prior to the joint virtual committee review. When the committee convened as a group, the committee members revisited and discussed all items and stimuli presented. A WestEd recorder took detailed notes about discussions and recorded the final committee recommendations. These notes were compiled for reconciliation with the LDOE and post-review implementation. Access to the items was 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 were instructed to clear their internet browser history. In addition, all participants completed a nondisclosure agreement prior to accessing any items.

Results of Content Review. The results of the reviewers' individual recommendations were captured in ABBI. Table 3.3 provides the results based on the participants' individual votes following their initial review of the stimuli and items. Table 3.4 shows the results of the group votes after discussing and reaching consensus on the disposition of the stimuli and items.

Table 3.3

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	160	827	85*	3	1	916

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

Table 3.4

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	160	115	45	0	0

Post-Review Finalization. At the conclusion of the content and bias reviews, WestEd content leads consulted with the LDOE to reconcile any unresolved committee feedback. Following implementation of the committee's feedback, the LDOE and WestEd content leads met virtually for final item reconciliation. WestEd provided records of all implemented changes to the LDOE prior to the virtual reconciliation meetings. During the reconciliation meetings, the leads reviewed the items to ensure that they were finalized for inclusion in the embedded field test. Once all content considerations were resolved, all items and stimuli went through a final formal fact-checking round and two additional rounds of proofreading. Any changes resulting from these reviews were submitted to the LDOE for approval.

4. Construction of Test Forms

Initial Construction

The purpose of the fall 2017, spring 2018, and summer 2018 forms construction activities was to create operational forms using items from the spring 2017 embedded field test and to embed field test items in the spring 2018 form for potential use in future operational assessments. This section describes the process used to create operational and field test forms.

Operational Forms

Data review-approved items from the spring 2017 embedded field test were available for use on the fall 2017 and spring 2018 operational assessments. (See the *2016–2017 LEAP 2025 U.S. History Technical Report: 2016–2017 Field Test* for results from the data review and reconciliation of the spring 2017 field test items.)

WestEd completed item selection for one operational (OP) form and one administrative error (AE) form for the fall 2017 administration. The designation of these forms was reversed for the spring 2018 administration so that the fall operational assessment became the spring 2018 administrative error form and the fall administrative error form became the basis for creating the spring operational form. Four operational forms were created for the spring administration. The difference between the four forms was the task that was selected. Otherwise, the item sets and standalone items were the same across the four forms. WestEd worked with the LDOE content staff to select items for the forms following the data review meeting in August and submitted these forms to Pearson psychometricians for consideration before formal submission to the LDOE for approval. The operational and administrative error forms were designed to adhere to the blueprint for U.S. History and exhibit the broadest possible balance of content and breadth of GLE coverage. Based on these considerations, the WestEd content lead selected the task first and followed with a combination of item sets and standalone items that would ensure that the relative distribution of score points by reporting category would meet the blueprint for the operational assessment and administrative error forms for U.S. History while avoiding similar content and topics across the balance of items and item types.

Placeholder items were included on the fall operational and administrative error forms to match the location and item types of the field test items that would appear on the spring 2018 forms. The spring 2018 administrative error form included placeholder items. Table 4.1 provides the original operational test design for U.S. History as the forms were initially constructed for fall 2017 and spring 2018.

Table 4.1

Original U.S. History Operational Test Composition for 2017–2018

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	2	6	7	10	0	2	0	12	14
6-Item Set with CR	2	6	8	8	2	2	0	12	16
5-Item Set	4	5	6	16	0	4	0	20	24
Standalone Items	0	0	0	10	0	0	0	10	10
Task Set	1	5	12	4	0	0	1	5	12
Total	**9**			**48**	**2**	**8**	**1**	**59**	**76**

However, as a result of discussions between the LDOE, WestEd, and Pearson, a decision was reached to adjust the length of the U.S. History test. Based on the testing time data from the spring 2017 embedded field test and conversations with stakeholders, the LDOE and WestEd reduced the length of the operational test by one item set and decreased the number of field-test items by one item set and three standalone items. As a result, the number of total points on the test was reduced from 76 to 69 and the composition of the test design was modified. Points were reduced from each reporting category. The number of points in the reporting category Western Expansion to Progressivism was reduced from 13 to 12; the number of points in Isolationism through the Great War was reduced from 9 to 8 points; the number of points in Becoming a Great Power through World War II was reduced from 15 to 14 points; the number of points in the reporting category The Cold War Era was reduced from 15 to 12; and the number of points in the reporting category The Modern Age was reduced from 8 to 7 points. Table 4.2 provides the operational test design for U.S. History as the forms were constructed for the 2017–2018 administrations.

Table 4.2

U.S. History Operational Test Composition for 2017–2018

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 2 TEs	1	6	8	4	0	2	0	6	8
6-Item Set with CR	2	6	8	8	2	2	0	12	16
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**

Tables 4.3 and 4.4 provide the number of points per standard for the operational tests. Points for the tasks are presented separately from the other item types. The points for CRs are included in the standard rows, although they are not included in the reporting category.

Table 4.3

U.S. History Operational Test Point Compositions by Reporting Category for Fall 2017

Standard	Task Standard	SR	CR	TE	ER	Total Points
S2. Western Expansion to Progressivism		8	0	4	0	12
S3. Isolationism through the Great War		6	0	2	0	8
S4. Becoming a World Power through World War II		10	2	4	0	16
S5. The Cold War Era		8	0	4	0	12
S6. The Modern Age		5	2	2	0	9
Total Points Excluding Task		37	4	16	0	57
Task	S4. Becoming a World Power through World War II	1	0	0	0	1
	S5. The Cold War Era	3	0	0	8	11
Total		**41**	**4**	**16**	**8**	**69**

Table 4.4

U.S. History Operational Test Point Composition by Reporting Category for Spring 2018

Standard	Task Standard	SR	CR	TE	ER	Total Points
S2. Western Expansion to Progressivism		10	2	2	0	14
S3. Isolationism through the Great War		6	0	2	0	8
S4. Becoming a World Power through World War II		8	0	6	0	14
S5. The Cold War Era		8	2	4	0	14
S6. The Modern Age		5	0	2	0	7
Total Points Excluding Task		37	4	16	0	57
Task Forms 1–7	S2. Western Expansion to Progressivism	2	0	0	0	2
	S4. Becoming a World Power through World War II	1	0	0	0	1
	S5. The Cold War Era	1	0	0	8	9
Task Forms 8–14	S4. Becoming a World Power through World War II	1	0	0	0	1
	S5. The Cold War Era	3	0	0	8	11
Task Forms 15–20	S5. The Cold War Era	2	0	0	0	2
	S6. The Modern Age	2	0	0	8	10
Task Forms 21–26	S3. Isolationism through the Great War	2	0	0	0	2
	S4. Becoming a World Power through World War II	1	0	0	8	9
	S5. The Cold War Era	1	0	0	0	1
Total		**41**	**4**	**16**	**8**	**69**

Field Test Versions

Twenty-six embedded field test forms were administered in spring 2018 for U.S. History. This number is greater than the number of item sets that were revised from the existing item bank for field-testing. Because standards for the U.S. History assessment were to be

set in summer 2018 following the spring 2018 administration, there was a need to field test as many items as possible so that those items would appear on the same scale for future operational assessments. All of the items selected for the fall 2017 operational administration were embedded in field test positions. Additional item sets that had been field tested in spring 2017 were also embedded and re-field tested. One or two versions of each item set were field tested as needed.

- 7 item sets from the fall 2017 operational form were embedded in field test positions
- 6 revised item sets from the existing item bank were field tested
- 5 item sets previously field tested in spring 2017 were re-field tested
- 1 item set developed but not field tested in spring 2017 was field tested
- 10 standalone items from the fall 2017 administration were field tested
- 10 standalone items developed for spring 2018 were field tested
- 41 standalone items previously field tested were re-field tested

The following embedded field-test design was developed for the spring 2018 administration:

- One 6-item set with 2 TE or 1 TE & 1 CR
- Three standalone items

Because fewer standalone items were developed than positions were available across the twenty-six field test forms, standalone items were repeated as necessary across the forms.

In addition to content balance, the WestEd content lead was careful 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 purposefully distributed across the forms, cueing and clanging were intended to have been avoided; however, developers also conducted a separate review of the forms to check for inadvertent cueing or clanging.

Following the final item placement by the WestEd content lead, test maps containing each item's unique identification number (UIN) were created. The test maps captured details about each proposed form, including test session, item sequence, unique item number,

and associated item metadata. Item descriptions were also included for each item, to aid in the review of the selection and placement of individual items.

Revision and Review

Psychometric Approval of Operational Forms

Prior to submitting the forms to LDOE staff for review, Pearson psychometricians and WestEd content specialists participated in an iterative process of reviewing and revising the forms. The psychometric review consisted 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 were examined, to determine whether any forms had significantly non-uniform distributions of correct responses (A, B, C, and D). Spreadsheets were used to generate frequency tables of reporting categories, item types, and MC answer keys for each form and across forms. Deviations from the blueprint were identified and addressed. Test characteristic curves (TCC) based on item response theoretic models were applied to data, and conditional standard errors of measurement were computed for each iteration during the test construction process to evaluate how well a proposed test form matched psychometric targets. Psychometric approval from Pearson was 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 were delivered to the LDOE for approval. Forms were reviewed by both LDOE content and psychometric staff. Based on the LDOE review, sets or items were replaced and the sequence of answer choices (for field test items) and the sequence of items within sets were revised as requested. Following these changes, the overall balance of answer choices and key runs was re-evaluated, and final adjustments were made to achieve the appropriate balance. Finalized test maps were used to create PDF versions of paper forms, which were reviewed by WestEd's proofreaders before the items were transferred from ABBI to DRC.

Streamlining of Stimuli

Following the spring 2017 field test administration, the LDOE evaluated the student testing times. The Department expressed concern that the length of stimuli in item sets, particularly text-based documents, was a contributing factor to students taking longer than anticipated to complete the tests. As a result, the LDOE requested that WestEd revisit the stimuli of existing field test item sets and reduce the length of the stimuli to the extent possible. Focus was also to be given to reducing the reading level of stimuli where possible to further ensure that the sources would not impede student performance on the assessments.

Revisiting Design

As part of the effort to reduce the testing time on the assessment, WestEd worked with the LDOE to reduce the length of the test for the 2017–2018 administration. The total number of operational points was reduced from 76 to 69, and the number of operational item sets was decreased from 8 to 7. The number of embedded field-tested item sets was reduced from 2 to 1, and the number of embedded standalone field test items was reduced from 6 to 3 on each form.

5. Test Administration

This section 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 mandates, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE provides train-the-trainer opportunities for school system test coordinators, who in turn convey test administration training to schools within their school system. The LDOE conducts quality-assurance visits during testing to ensure school system adherence to the standardized administration of the tests.

The school system 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 the spring 2018 test administration, DRC produced an administration manual, the *LEAP 2025/EOC Test Administration Manual* (TAM), which serves for the LEAP 2025 administrations.

DRC also produced a test coordinator manual. LDOE assessment staff review and provide feedback and final approval for the test administration and test coordinator manuals. The manuals are inclusive of all LEAP 2025 HS assessments in ELA, mathematics, social studies, and science. The manual provides detailed instructions for school system and school test coordinators' responsibilities to distribute and collect test materials and to return test materials to DRC when appropriate.

The TAM provides detailed instructions for administering the LEAP 2025 HS assessments. The manual includes instructions for test security, test administrator responsibilities, test preparation, administration of online tests, and post-test procedures.

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

The manual's directions for test administration address Standard 4.15 from the *Standards*, 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 before-, during-, and after-testing activities 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 administration manuals included instructions for scheduling the test within the state testing window. The test administration manual also contained the schedule for timing each test session.

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

Department staff 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 classroom environment

testing 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 was timed to provide sufficient time for students to attempt all items. The test administration manuals provided examiners with timing guidelines for the assessments.

Online Forms Administration

The online forms were 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 printed test tickets. Students entered their ticket information to access the test in INSIGHT. In addition, students had access to Online Tools Training, which allowed them to practice using tools and features within INSIGHT. Students were required to experience the Online Tools Training (OTT) before the computer-based test administration. The OTT allows students to observe and practice features of the Online Assessment Software prior to an actual test administration. Students were also required to view the Student Tutorials, which present visual and verbal descriptions of the properties and features of the DRC INSIGHT Online Assessment Software.

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

Operational and field test items were administered during the online testing, which was available from Monday, April 23, through Friday, May 18, 2018.

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 trained in test security (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.

The LDOE routinely conducts comprehensive data forensics with the administration vendor. Incidents that warrant further investigation with prospective voided test results include plagiarism, excessive wrong-to-right response changes, and patterns of unusual school-level gains.

6. Scoring Activities

Answer Key Verification

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

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

Multiple-Choice Item Keycheck. Scoring of SR items is evaluated with TRIAN, a standardized Pearson program that calculates MC item statistics, to verify that MC items were keyed correctly (i.e., that the true correct response was 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 WestEd content staff for key verification. Scoring of MS items was evaluated at data review.

Scoring of TEs and Adjudication. All TE and MS items were processed through DRC's autoscoring engine and scored according to the assigned scoring rules as established during content creation by WestEd in conjunction with the LDOE. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring any TE items. DRC established an adjudication process for technology-enhanced items to verify that correct answers were identified. DRC's technology-enhanced scoring process included the following procedures:

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

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

TE items and other eligible items identified in the test map were automatically scored as tests were processed. TE items were scored according to scoring rules in the Directory of Test Specifications (DOTS), which includes scoring information for all item types.

The adjudication process focuses on detecting possible errors in scoring TE and MS items. DRC provides a report listing the frequency distributions of TE item responses and MS items. Members of the LDOE and WestEd content staff examine the TE and MS response distributions and the auto-frequency reports to evaluate whether the items were scored appropriately. In the event that scoring issues are identified, WestEd content staff and the LDOE review 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

A full description of the methods used to score constructed and extended responses can be found in the *LEAP Processing Rules—Scoring (LEAP Spring 2018)* document. The document table of contents is listed below.

- Schedule, Locations, and Staffing
  - Training and Scoring Schedule
  - Scorer Degree Requirements
- Training
  - Social Studies Training Materials
- Qualifying
- Reader Monitoring Procedures
  - Team Leader Read-Behinds
  - Validity Responses
  - Recalibration Sets
  - Inter-Rater Reliability
  - Handscoring Quality Control Reports
    - Scoring Summary Report Sample—8-Point, Two-Trait Extended-Response Item
    - Scoring Summary Report Sample—2-Point Constructed-Response Item
  - Expected Agreement Rates (Inter-Rater Reliability and Validity)
    - Spring 2018 Operational Extended-Response Items—2017 Field Test Handscoring Data
    - Spring 2018 Operational Constructed-Response Items—2017 Field Test Handscoring Data
  - Reader Feedback Logs
- Handscoring Rules
  - AI Scoring of Grade 5–8 ER Items
  - Scoring of CR Items and Grade 3 & 4 ER Items
- Handling Unusual Responses
  - Nonscore Blanks

        - Grade 3–8 CR and ER Nonscore Code
    - Alerts
- Artificial Intelligence Scoring
    - Model Building
    - 2018 Model Enhancement Process
    - Evaluation Metric
    - Scoring Responses with the AI Engine
    - Quality Control of the AI Engine (MI)
    - Scoring (DRC)
    - Identifying Responses for Human Review
        - Alert Detection System
        - Identification of Non-Alert Responses Requiring Human Review
        - Identifying Copied Text and Plagiarism with the AI Engine
- Appendix

It should be noted that Tables 6.1 through 6.4 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the 2017–2018 forms.

Table 6.1

Constructed-Response Inter-Rater Reliability

Administration	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
Spring 2018	USH_Item1	≥7,240	78	22	0
	USH_Item2	≥7,600	80	19	0

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

Table 6.2

Constructed-Response Score Point Distributions Spring 2018

Item	Score Point Distribution				
	Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
USH_Item1	≥39,700	22	55	23	0
USH_Item2	≥39,700	55	24	21	0

Table 6.3

Extended-Response Inter-Rater Reliability Spring 2018

Item		Inter-Rater Reliability			
	2x	Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
USH_Item1	≥5,440	Content	94	6	0
		Claim	93	7	0
USH_Item2	≥7,110	Content	95	5	0
		Claim	94	6	0
USH_Item3	≥9,990	Content	96	4	0
		Claim	95	5	0
USH_Item4	≥10,140	Content	93	7	0
		Claim	93	7	0

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

Table 6.4

Extended-Response Score Point Distributions

Item	Total	Score Point Distribution						
		Dimension	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent "3" Rating	Percent "4" Rating	Percent Blank
USH_Item1	≥10,520	Content	16	32	31	15	3	0
		Claim	21	28	30	15	3	0
USH_Item2	≥11,340	Content	31	32	21	11	5	0
		Claim	29	35	21	10	4	0
USH_Item3	≥14,060	Content	21	34	26	11	6	0
		Claim	31	33	21	10	3	0
USH_Item4	≥16,720	Content	37	30	26	11	4	0
		Claim	32	27	25	10	4	0

7. Data Analysis

Classical Item Statistics

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 is a measure of item discrimination. Items with higher item-total correlations provide better information about overall student ability (i.e., they discriminate between lower- and higher-ability students). Table 7.1 summarizes several key classical item statistics from 2018 field-tested items. The majority of items field tested were MC, complemented by other item types that comprised 25% of the field test item pool. The summary table and the box plots that follow illustrate the range of item difficulties by item type, and good item-total discriminating power for all item types overall.

Table 7.1
Summary of Classical Statistics for Field Test Items for U.S. History

Item Type	*N* Items	*p*-value Mean	*p*-value SD	Item-Total Correlation Mean	Item-Total Correlation SD	Percent with B-level DIF	Percent with C-level DIF
MC	162	0.57	0.13	0.38	0.09	15%	1%
MS	11	0.36	0.15	0.39	0.13	18%	0%
CR	10	0.21	0.10	0.42	0.12	30%	10%
TE	28	0.42	0.10	0.46	0.11	21%	4%

Figure 7.1 *Box Plot of Item p-Values*

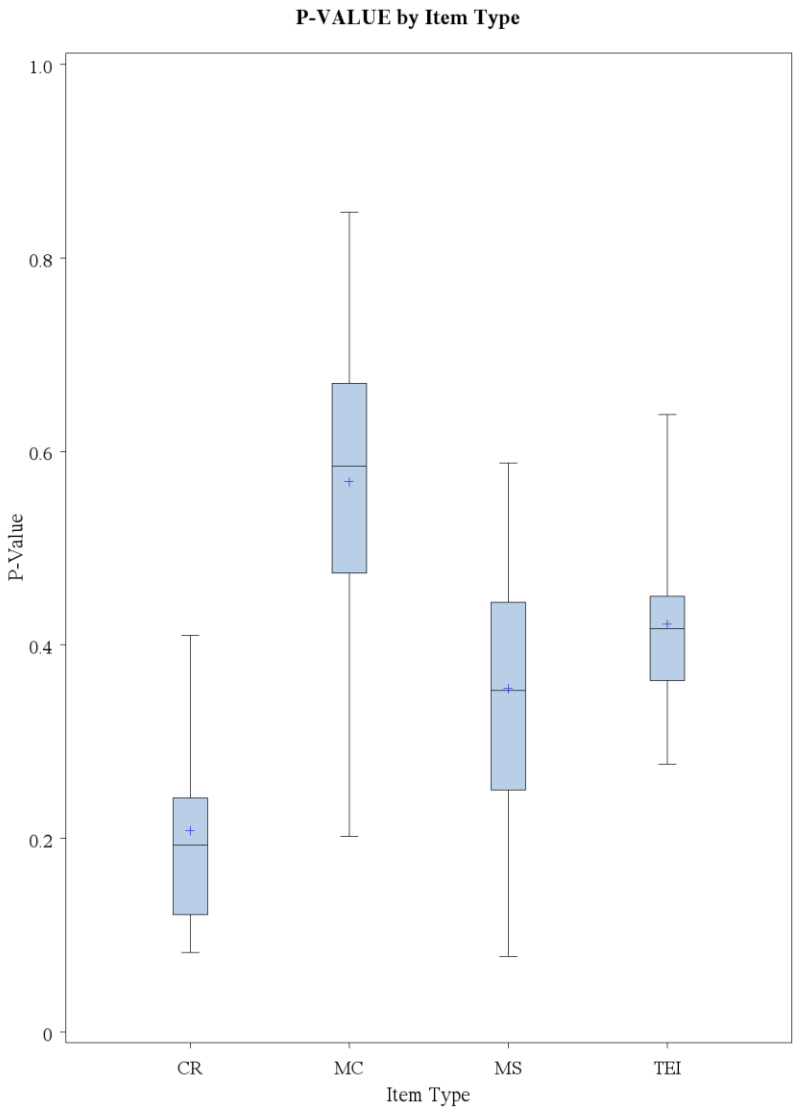
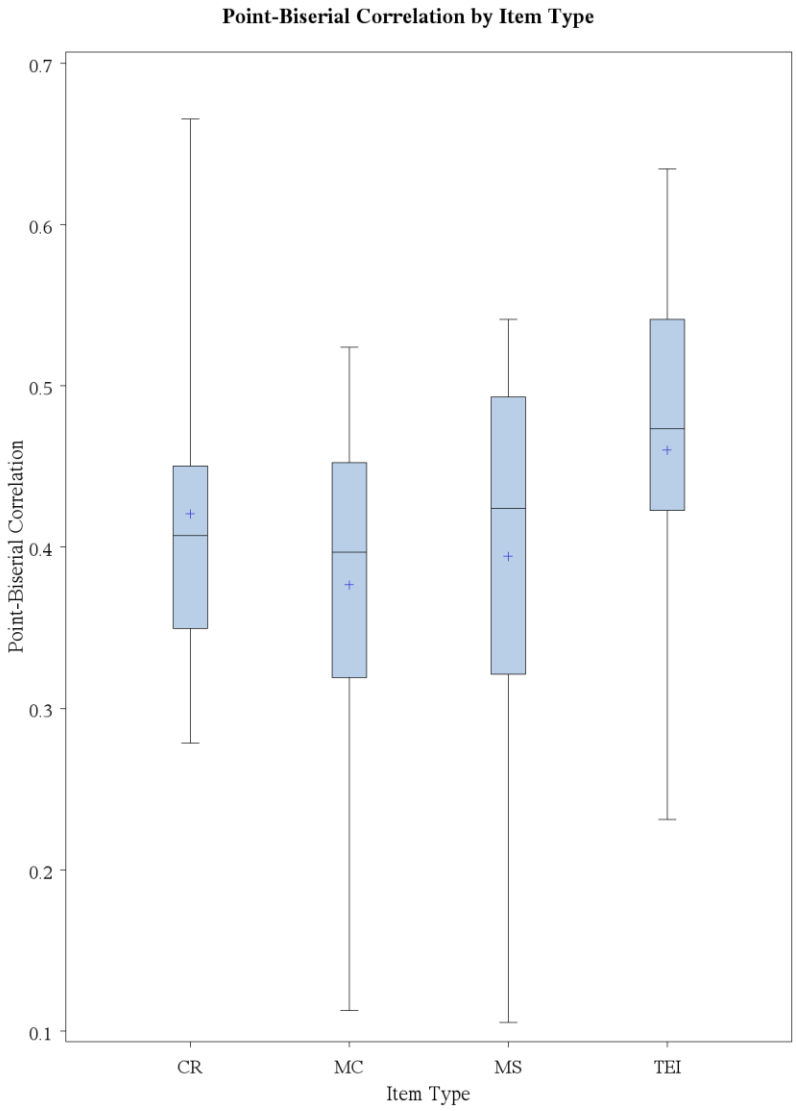


Figure 7.2 *Box Plot of Item-Total Correlations/Point Biserial (PBIS)*



Differential Item Functioning

Differential item functioning (DIF) analyses are designed to detect statistical evidence of potential item bias. Because test scores can have many sources of variation, the test developers' task is to create assessments that measure the intended abilities and skills without introducing extraneous elements or construct-irrelevant variance. When tests measure something other than what they are intended to measure, test scores will reflect these unintended skills and knowledge, as well as what is purportedly assessed by the test. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975; Zumbo, 1999). One of the factors that may render test scores as biased is differing cultural and socioeconomic experiences.

Analysis of DIF is a statistical method to detect potential bias of an item. DIF is defined as a difference between groups (e.g., male and female) in the probability of getting an item correct. These analyses are conditioned on the ability that the assessment is intended to measure.

The DIF methodology for dichotomous items used the Mantel–Haenszel (*MH*) DIF statistic (Holland & Thayer, 1988; Mantel & Haenszel, 1959). The *MH* method is frequently used and is efficient in terms of statistical power (Clauser & Mazor, 1998). The *MH* chi-square statistic is computed as

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

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 (Δ*MH*) was computed. The Educational Testing Service (ETS) first developed the Δ*MH DIF* statistic. To compute the Δ*MH* DIF, the *MH* alpha (the odds ratio) is first computed:

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

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 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), whereas 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 the field test items that exhibit strong, weak, or no DIF (Zieky, 1993). Table 7.2 defines the DIF categories for dichotomous items.

Table 7.2
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 $2\times(T+1)\times M$ (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 7.3 defines the DIF categories for polytomous items.

Table 7.3

DIF Categories for Polytomous Items

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

Two DIF analyses were conducted for field test items: female/male and black/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 7.4 provide the percentages 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 field 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 7.4

Summary of DIF Flags for Field Test Items for U.S. History

Comparison Groups	A	B,[B-]	C,[C-]
Female – Male	191	9,[9]	0,[2]
African American – White	185	2,[23]	0,[1]

All items exhibiting 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 2018 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.

Item Calibration

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 how much of the content standards students need to master in order to achieve proficiency. Constructing tests to content standards enables the tests to assess the same constructs from one year to the next.

Item Response Theory (IRT) models were used in the item calibration for the LEAP 2025 U.S. History test. The LEAP 2025 U.S. History test was calibrated independent of the EOC U.S. History test. 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):

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$, $p_{im}(\theta_j)$ is the probability of an examinee with θ_j getting score *m* on item *i*, and 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.).

Field Test Item Parameters

The distributions of item parameters are summarized in Table 7.5. Figures 7.3–7.5 provide Box Plot displays of the distributions of IRT parameter estimates by item type. 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. 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.

Table 7.5
Summary of IRT Statistics for Field Test Items for U.S. History

Item Type	N Items	*b* Mean	*b* SD	*a* Mean	*a* SD	*c* Mean*	*c* SD*	% Fit
MC	162	0.24	0.76	0.90	0.30	0.20	0.06	95%
MS	11	0.90	0.84	0.86	0.22	0.07	0.03	91%
CR	10	0.68	0.54	0.87	0.17			80%
TE	28	0.57	0.74	0.60	0.19			50%

Note. *Only dichotomous items (i.e., scored 0 or 1) have *c* parameters.

Figure 7.3 *Box Plot of IRT a-Parameters*

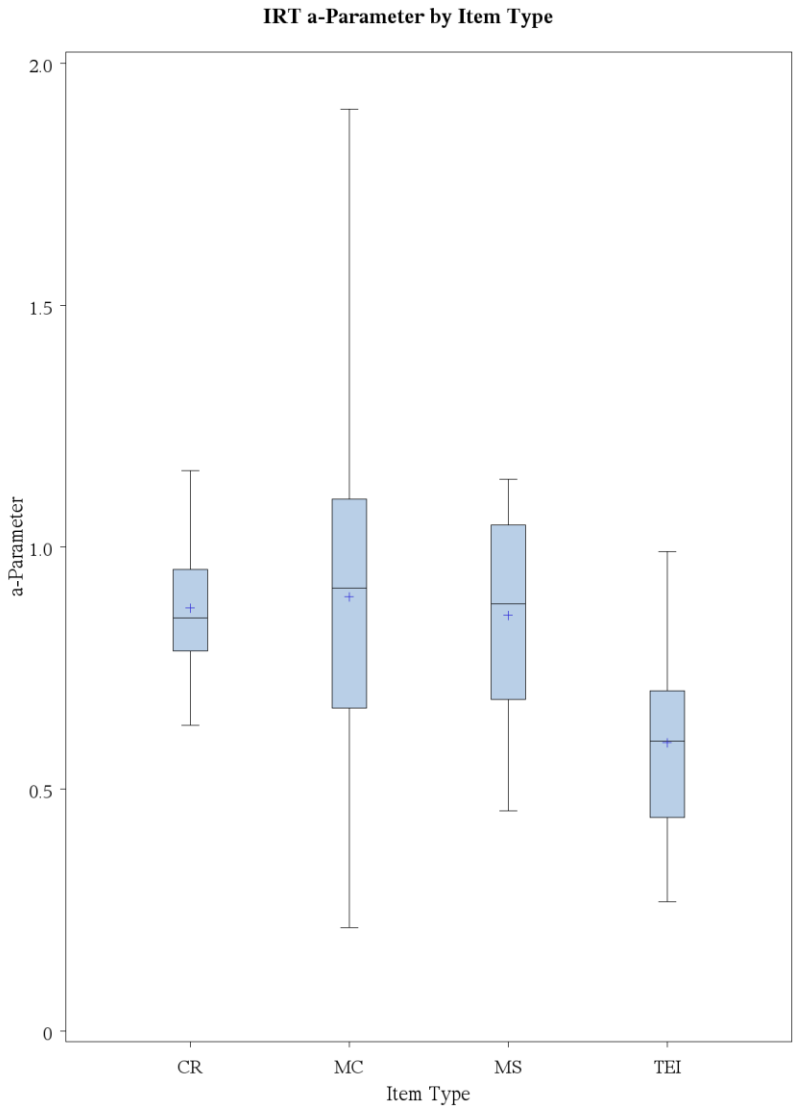


Figure 7.4 *Box Plot of IRT b-Parameters*

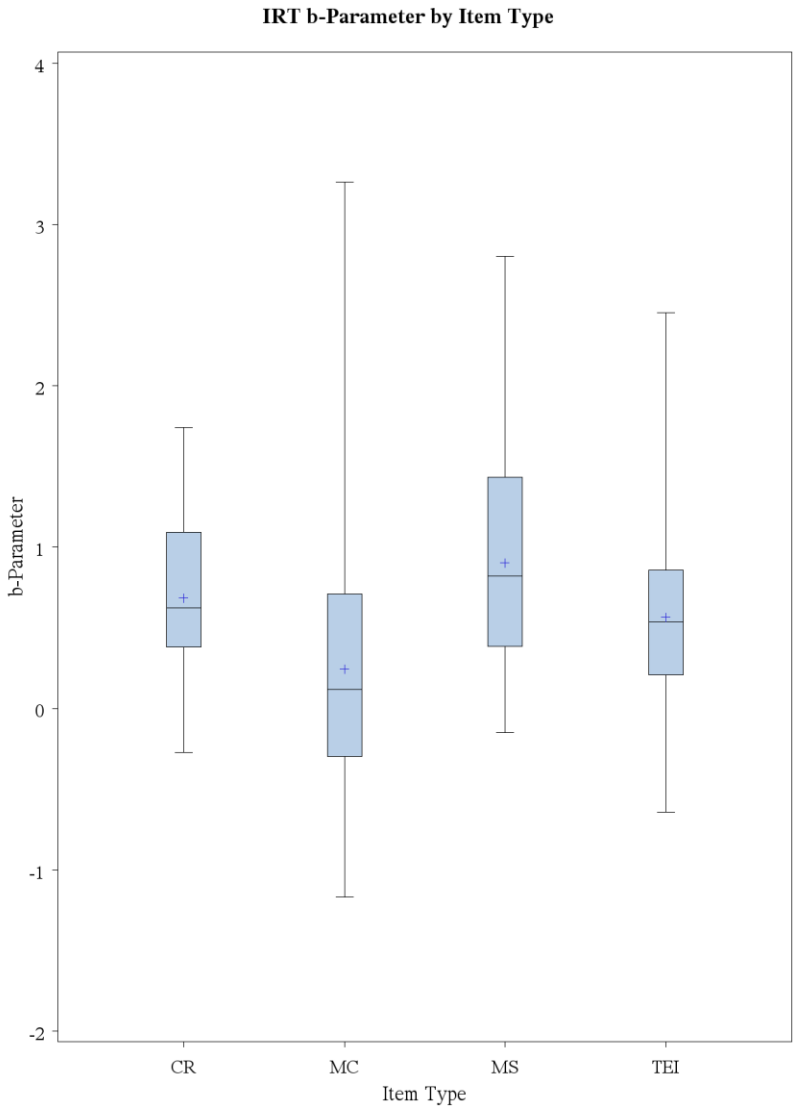
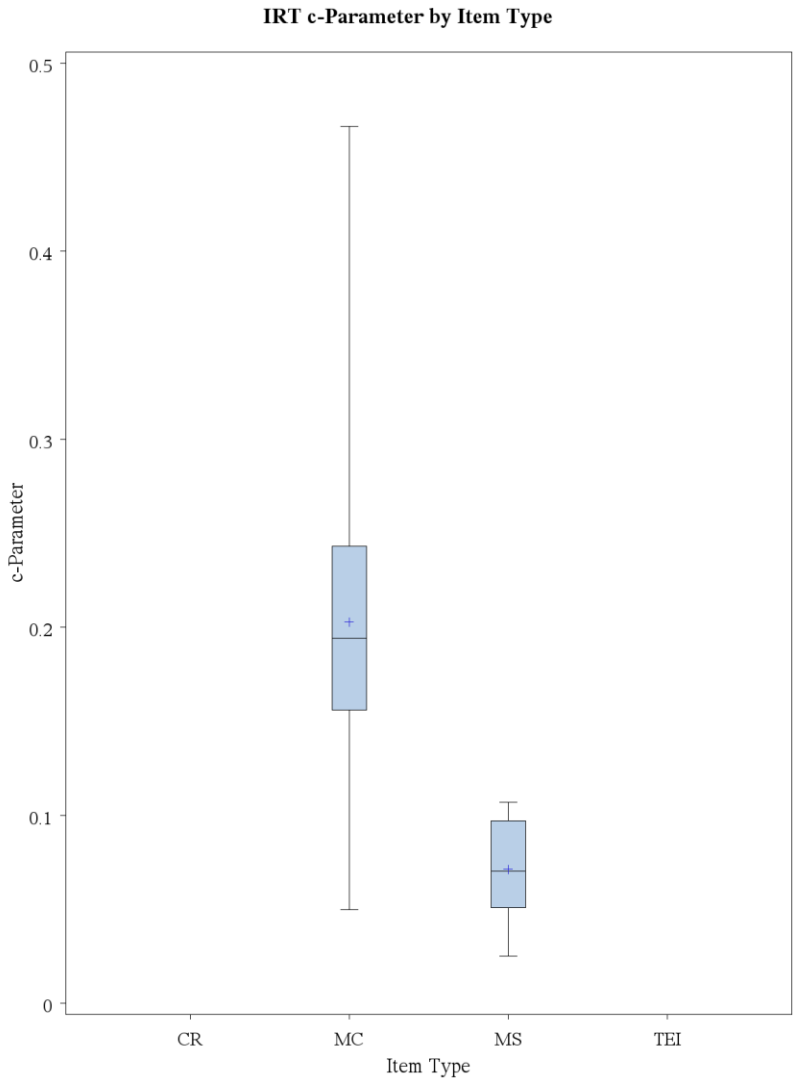


Figure 7.5 *Box Plot of IRT c-Parameters*



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

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 $ZQ_{1,\text{crit}}$ 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 intervals minus the number of independent item parameters). Items are categorized as exhibiting either Fit or Misfit.

8. Reporting for U.S. History

Additional information regarding score reporting can be found in the *Interpretive Guide English I, English II, Algebra I, Geometry, U.S. History 2017–2018* document. Detailed Information can be found at the following link: https://www.louisianabelieves.com/resources/library/assessment.
The elements of the table of contents are provided below.

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

U.S. History Standard Setting

Ancillary documentation of the standard setting process and results can be found in the *LEAP 2025 Standard Setting Meeting* document. The full report summarizes the processes and results of the standard setting. Excerpts from the Executive Report are provided in the following pages. The elements of the table of contents are listed below.

- Executive Report
- Chapter 1 – Overview of the Standard Setting Process
    - Goals of the Standard Setting Meeting
    - LEAP 2025 Achievement Levels
    - The LEAP 2025 Standard Setting Process
- Chapter 2 – Pre-Meeting Development
    - LEAP 2025 Achievement Level Descriptors
    - Development of the Participant Materials
    - Preparation of the Ordered Item Book
    - Development of the Presentation Materials
    - Facilitator Training
    - Preparation for Data Analysis during the Meetings
- Chapter 3 – Standard Setting Meetings
    - Purpose of the Standard Setting Meetings
    - Committee Participant Composition
    - Standard Setting Meeting Facilitators and Staff
    - Materials
    - Procedure
    - Standard Setting Meetings and Proceedings
    - Recommended LEAP 2025 Cut Scores from Standard Setting Committees
- Chapter 4 – Post-Standard Setting
    - Vertical Articulation Meeting
    - Standards Policy Review Committee
    - Scaling Process
- Chapter 5 – Evidence of Procedural Validity of the Standard Setting Process
    - Internal Procedures
    - Committee Representation
    - Committee Training
    - Perceived Participant Validity of the Workshop
- References
- Appendices

Standard Setting Executive Report

10 July 2018

This report summarizes the process and results of setting achievement levels for the Louisiana Educational Assessment Program (LEAP) 2025 U.S. History assessment. The Louisiana Department of Education (LDOE) and WestEd with Pearson (LEAP 2025 Social Studies assessment contractors) recommend the achievement levels shown in Table 2 of the Standard Setting Report for adoption by the Board of Elementary and Secondary Education (BESE).

LEAP 2025 U.S. History Standard Setting Process and Results

Achievement levels are used to classify student achievement on an assessment. In order to classify student achievement into the different achievement levels, the following components are required: 1) policy definitions, 2) Achievement Level Descriptors (ALDs), and 3) cut scores. Policy definitions describe the achievement levels in general terms that apply to all courses or subject areas. ALDs illustrate the achievement levels in terms that are specific to a course or subject area. Cut scores represent the lowest boundary of each achievement level on the scale.

The process of recommending achievement standards for the LEAP 2025 U.S. History test was similar to the processes followed for previous assessments in Louisiana and in line with national best practice. Results and details of the process are presented in the following sections.

Policy Definitions

Achievement level policy definitions for the LEAP 2025 U.S. History assessment are shown in Table 1. These policy definitions are also used for the social studies grades 3–8 assessments, English language arts (ELA) assessments, and mathematics assessments. The titles and descriptions of the achievement levels were defined to be part of a cohesive assessment system, and the achievement levels indicate a student's ability to demonstrate proficiency on the Louisiana student standards defined for a specific course.

Table 1
Achievement Level Policy Definitions for LEAP 2025

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

Achievement Level Descriptors (ALDs)

ALDs for the U.S. History test are shown in the appendix of the Standard Setting Executive Report. A multi-step iterative process was used in developing, reviewing, and approving the ALDs. Prior to the standard setting committee, a draft set of ALDs representing a gradual increase in expectations across the achievement levels was created by LDOE content staff in cooperation with WestEd content specialists. Panelists who participated in the standard setting committees had the opportunity to provide suggestions and edits to the draft set of ALDs based on the recommended cut score for each achievement level and the items in the ordered item book. To produce the final set of ALDs, the LDOE edited the set of draft ALDs based on suggestions generated by the panelists in the standard setting meeting.

Cut Scores

The cut scores recommended for adoption by BESE are shown in Table 2. This table shows the scale score ranges corresponding to each achievement level. The cut scores for the achievement levels are the lowest cut score within each range. There is no cut score for *Unsatisfactory*, since 650 is the lowest obtainable scale score a student can earn.

Table 2
Scale Score Ranges for LEAP 2025 Achievement Levels for U.S. History

Achievement Level	Scale Score Ranges U.S. History
Advanced	774 to 850
Mastery	750 to 773
Basic	725 to 749
Approaching Basic	711 to 724
Unsatisfactory	650 to 710

Details pertaining to the general method for obtaining the recommended cut scores are provided below.

General Method

Prior to the standard setting committee, on April 26, 2018, a policy committee was convened of teachers, school and school system leaders, and LDOE staff. The purpose of the meeting was to review information that would be useful in considering the policy implications of the cut scores for the LEAP 2025 U.S. History assessment and to provide a set of recommended ranges for the cut scores that would be presented to panelists during the standard setting meeting. The information that was shared with the committee included the impact data from the spring 2017 administration of the LEAP 2025 Social Studies assessments for grades 3–8 and the high school assessment for U.S. History, the Louisiana high school graduation rates in 2016, and the results of a contrasting groups teacher study performed for the U.S. History assessment during spring 2018. After a review and discussion of the data presented during the meeting, the policy committee members approved recommended ranges for the cut scores. The ranges, shown in Table 3, represent the maximum and minimum percentage of students that could be

reasonably expected to be classified into each achievement level or higher based on the policy considerations. These ranges helped guide the standard setting committee in understanding policy considerations as part of the standard setting process.

Table 3
Recommended Ranges from the Policy Committee

Achievement Levels	Cumulative Impact Data	
	Minimum	Maximum
Advanced	5%	15%
Mastery	25%	40%
Basic	50%	65%
Approaching Basic	70%	85%

From July 9 to July 10, 2018, after the first year of operational administration, a standard setting committee meeting was conducted to provide cut score recommendations for the LEAP 2025 U.S. History assessment. The committee was composed of 13 individuals, including teachers and non-teacher educators, who were selected for the standard setting committee to provide content expertise during the committee meeting and to be representative of the state's educators. The evidence-based bookmark method was used for the standard setting meeting (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001; Schultz & Mitzel, 2009). The key material used by the committee was a book of test items arranged in order of difficulty. Participants identified and discussed the knowledge, skills, and abilities required to respond to the test items and divided the items into two groups—items that a student who is minimally qualified for an achievement level would likely answer correctly and items too difficult for students at that same achievement level. Additionally, the participants were provided the recommended ranges from the policy committee to review and consider as part of the judgment process.

In order to create a common point of reference across the social studies assessments, cut scores and measures of student achievement on all LEAP 2025 assessments are translated to a scale that ranges from 650 to 850 points, a *Basic* cut of 725 and a *Mastery* cut of 750. The common values of 725 for the *Basic* cut score and 750 for the *Mastery* cut score across assessments do not mean that they reflect that same difficulty, or that achievement levels can be compared in difficulty through the scale values of their cut scores across grades and subjects. Similarly, the percentage of students in an achievement level is not directly comparable across grades and subjects. The population

of students tested is different for each assessment. Achievement levels from different tests are not comparable because the cut scores for these tests are criterion referenced—they are based on content-specific expectations of what students should know and be able to do.

Results for LEAP 2025 U.S. History

Table 4 shows the percent of students who took the LEAP 2025 U.S. History assessment during the spring 2018 administration that would be classified into achievement levels based on the cut score recommendations from the standard setting committee.

Table 4
Percent of Students in Achievement Levels

Achievement Level	Assessment U.S. History
Advanced	9%
Mastery	19%
Basic	32%
Approaching Basic	15%
Unsatisfactory	24%

9. Data Review Process and Results

During data review, field-tested items with accompanying data are reviewed to make judgments about the appropriateness of items for inclusion in the item pool. Items in the pool are eligible for selection and use on operational test forms. Item statistics that may signal potential concerns are reviewed during data review. Items are not rejected solely on the basis of statistics. Only items that exhibit concrete and identifiable content flaws are rejected from further consideration.

An internal data review meeting was held between the LDOE, WestEd, and Pearson for items field tested in spring 2018. Teachers did not participate in the data review. The 2018 data review for U.S. History included a refresher of appropriate interpretations about item statistics (difficulty, discrimination, DIF, score distributions), reasonable values of item statistics, and how the values might differ across different item types.

Data review materials included a booklet of item cards with an image of the item with its metadata and item statistics, and a booklet of corresponding stimuli for the items. TE items could be displayed on a screen to view items in a computer-based format. Content and psychometric representatives from the LDOE also participated in data review.

Pearson and WestEd facilitated the data review of field-tested items. Statistical information for each item was evaluated to determine whether the item functioned as intended. No items exhibiting statistical DIF were identified to have content flaws to account for the statistical DIF flags. A consensus recommendation for each field-tested item was then recorded, specifically to accept, to accept with edits (or "revise/re-field test"), or to reject an item. Collateral information to revise/re-field test was documented to aid future steps and processes.

Following data review, LDOE content specialists revisited items that were rejected or accepted with edits. This reconciliation process provides the LDOE with an additional opportunity to review item content and to consider possible revisions that would allow items to be re-field tested for possible future operational administration. The reconciliation decisions are the final decisions of record. Table 9 summarizes the disposition by item type of the field test items reviewed during 2018 data review.

Table 9

FT Item Dispositions by Item Type, 2018 Data Review

Item Type	Number of Items			
	Accept	Accept with Edits	Reject	% of Total
CR	8	–	–	4.88
MC	117	4	8	78.66
MS	6	–	1	4.27
TE	20	–	–	12.20
Total	151	4	9	100.00

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

10. Reliability and Validity

Internal Consistency Reliability Estimation

Internal consistency methods use 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 requiring multiple tests. Probably 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. Evidence for the homogeneity of LEAP 2025 U.S. History tests is obtained 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 also IRT model-based or “marginal reliability” (Thissen, Chen, & Bock, 2003) for the total test. Coefficient alpha values range from 0.92 to 0.93, and the marginal alpha value was 0.98. 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 (is 1 for standardized θ) and $E(SEM_\theta^2)$ is the average error variance (a.k.a. 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 entire population of students (see reliability and classification accuracy reports in the yearbook). 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 provided for the entire test, as well as 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). Results involving subscores must be interpreted carefully, as in some cases these measures have low reliability due to the limited number of items attached to 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. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of students.

---

[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.

Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive measures of the accuracy and consistency of the classifications. A brief description of the procedures used and the results derived from them are presented in this section.

Accuracy of Classification

According to Livingston and Lewis (1995, p. 180), the accuracy of a classification 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

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 in order to statistically model two parallel forms of the test and compare the classifications on those alternate forms.

Accuracy and Consistency Indices

There are three types of accuracy and consistency indices that can be generated from these tables: *overall*, *conditional-on-level*, and *cut point*. 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 ranges from 0.723 to 0.726 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 ranges from 0.229 to 0.232 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 ranges from 0.517 to 0.523 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.

The information mentioned above is reported in Appendix F: Reliability and Classification Accuracy.

Validity

"Validity is defined . . . as the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test" (APA, AERA, NCME, 2014).

The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process.

The 2017–2018 LEAP 2025 U.S. History test was designed and developed to provide fair and accurate scores that support appropriate, meaningful information for educational decisions. Validity evidence may be found in the following portions: Chapter 2 (Framework of Assessments), 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 2017–2018 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. Chapter 7 also describes an analysis of DIF. Complete tables of gender and ethno-racial DIF results for all 2017–2018 LEAP 2025 U.S. History operational items are presented in Appendix B.

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 2017–2018.

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

11. Statistical Summaries

For all LEAP 2025 assessments including U.S. History, the lowest obtainable scale score (LOSS) is 650 and the highest obtainable scale score (HOSS) is 850. Test results are provided in Appendix B. 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 Appendix C, scale score frequency distributions are presented in Appendix E.

References

AERA/APA/NCME. (2009/2014). *Standards for educational and psychological testing.* Washington, DC: Author.

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.

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.

Lewis, D. M., Mitzel, H.C., & Green, D. R. (1996). Standard setting: A Bookmark approach. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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.

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.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249- 281). Mahwah, NJ: Lawrence Erlbaum Assoc.

Mogilner, A. (1992). *Children's writer's word book*. Cincinnati, OH: Writer's Digest Books.

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

Schulz, E. M., & Mitzel, H. (2009). A Mapmark method of standard setting as implemented for the National Assessment Governing Board. In Smith, Jr., E.V., & Stone, G.E. (Eds.), *Applications of Rasch measurement in Criterion-Reference Testing: Practice Analysis to Score Reporting*, (pp. 194–235). Maple Grove, MN: JAM Press.

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.

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
2017–2018 Field-Test Item Development

I. **Introductions**
II. **Stimulus Set Overviews**
   a. **Item Set Topics**
      i. Themes of the 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. Recommended. The overview contains recommended stimuli that will support the item set
         2. Searchers can propose other stimuli that support the item set or task
      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. 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, 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
      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 Training Agenda
Item Development Cycle for 2017–2018 Field Test

I. **Louisiana Content Standards and GLEs**
    a. U.S. History
    b. Reporting categories
    c. Standards
    d. Grade Level Expectations

II. **Item Types and Overviews**
    a. Selected Response (Multiple Choice, Multiple Select)
        i. Rules for numbers of answer options and number correct
    b. Technology-enhanced items
        i. Templates
    c. Constructed-Response Items
    d. Item Sets
        i. Sources (Each set will have three to four sources)
    e. Standalone Items
        i. Purpose
        ii. Use of stimuli

III. **Rubrics and Scoring Guides**
    a. Constructed Response Scoring Rubrics
    b. Constructed 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. **Item Examples**

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 Distractors: Distractors should address misconceptions that the students may 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 sets 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. 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.
    ii. Sensitivity: Avoid topics that may upset or offend students in items (e.g., references to graphic violence, nudity, alcohol, drugs, natural disasters, group stereotypes, representation of ethnic groups)

VII. **ABBI Item Development Platform**

a. Functionality of the ABBI platform
b. Creating items in ABBI
c. Attaching scoring information in ABBI

VIII. **Receiving item assignments via Smartsheet**

LEAP 2025 U.S. History Editor Training Agenda
Item Development Cycle for 2017–2018 Field Test

I. **Item Set Overview**
   a. Review the item set overview to determine if the items fulfill the requirements of item set/task overview.
   b. Reporting Category alignment: Does the item align to the identified Reporting Category?
   c. GLE alignment: Does the item align to the correct GLE?
   d. DOK Level: Item aligns to the correct DOK level
   e. Source Alignment: Item aligns to the correct source(s)
   f. Item emphasis: Item meets the item emphasis
      I. Note: meeting the item emphasis could reflect meeting the expectations of the general directions in the item emphasis section or using, and possibly modifying, the stem provided. It will depend on the overview.

II. **Style Guide Requirements**
   a. General Rules
      I. Grade-level language
      II. Correct content
      III. Most items should reflect some level of analysis. Even for GLEs that would naturally lend themselves to identification items based on the wording of the GLE, try to create an item that is not DOK 1.
         - DOK 1 items are not permitted on the U.S. History assessment, but they may appear in grades 3 through 8 in very limited numbers.
      IV. Things to avoid:
         1. Idiomatic language
         2. Trivial content
         3. Cuing and clanging in a set: Clanging occurs when two or more items use the same information. For example, clanging can occur in correct answers, such as using “fertile farmland” in multiple items. Another way it can occur is when an item covers the same content in a slightly different way.
         4. Cuing within an item. Stems should not provide specific clues for the answer options.
         5. Outliers. The answer choice options should not contain grammatical or content outliers.
         6. For U.S. History, the answer choice options should come from the identified reporting category.
         7. Avoid adding a scenario to an item, if possible. It increases the reading load and typically adds unnecessary context. Context should be provided by the stimuli.
         8. Do not use the term “American Indian.” The LDOE prefers “Native American.”
         9. Do not use “slaves.” The LDOE prefers “enslaved persons/people.”

III. **Item Metadata**

a. Make sure the metadata are correctly completed on each item, including:

I. **All** the stimuli with a set have been associated with each item in the metadata

II. If you change the key in the item, update the item metadata

IV. **Stimulus Review**

a. Look over the stimuli. The stimuli have been approved by the LDOE in advance. We may not replace stimuli. If you believe a stimulus needs to be replaced, consult the Content Lead before making any changes.

I. Make sure all the stimuli have **Source X** (with the correct number) in a text box above the stimulus. (bf flush left)

II. Source all captions.

1. Captions for text passages in the upper grades follow this style:
   This excerpt is from a speech by President Franklin Delano Roosevelt, delivered on March 1, 1933. (Note: In the lower grades, the headline omits "Excerpt" and starts with "From.")
2. Captions for graphics (e.g., photographs, political cartoons), follow this style:
   This photograph by Dorothea Lange is from 1934.
   This political cartoon from Thomas Nast appeared in *Puck* in 1877. It has the caption: Graft.
3. Tables, charts, maps, and timelines do not require captions, as long as they have titles, which should be centered.

III. Sources on the stimuli:

1. Make sure text passages do not include the source and access info at the end of the passage.
2. Do not use Web links for sources for tables, maps, charts.
3. Political cartoons and photographs should include sources (e.g., Source: Library of Congress.)

IV. Make sure that this order is followed:

1. **Source X** (bf)
2. **From *Students Start Business to Learn Economics*** (title of text source, like a newspaper article, in bf. In a text box, not embedded in anything by DT.)
   by Someone's Name (by in ital. and name in roman) (text passage)

or

Source X (bf)
Louisiana Government Spending (title of map, graph, chart, quotation; all sources require a title, in bf) (for a quotation, use: Sen. Joseph McCarthy Quotation; for other things, like maps and graphs, don't use the word "map" or "graph")
(caption in roman)
(image)

Source: U.S. Geological Survey. (source line, italics, with the final period. This must be embedded flush left by the WestEd Design Team.)

V. Make sure the metadata are correctly completed on the stimulus.

V. **Item Sets**

a. Item-Set Purpose Setting Statements

I. For item sets, add a Purpose Setting Statement that appears before the stimulus/stimuli:

Example: Read and study the sources on Columbus and the New World. Then use the four sources and your knowledge of U.S. History to answer the questions.

II. (bf) Replace the first xxx with the correct number of stimuli. To add a PSS:

III. Go to the menu page in ABBI and click Create Passage.

IV. Delete the last two digits of the UIN and replace with _S (underscore capital S)

V. Mark the status as the round you are currently editing.

1. Fill out the Stimulus Title field with: PSS The Exact Title of the Set
2. Fill out Part of Stimulus Set as Yes
3. Fill out Stimulus Sequence in a Set as Scenario
4. Fill out Stimulus Group field with: The Exact Title of the Set
5. Add a text box from the Elements tab and type in the statement above
6. Save the item

b. Make sure the CRs can be answered in a fairly predictable and limited way.

c. Make sure the CRs are in the correct format. They should be in the Extended Text Interaction format. Make sure the Text Entry items score correctly in TN8. Items in the CR format will not score in TN8.

d. Make sure the CRs have answer information uploaded under the Attachments tab (select “Scoring Information” as the category) in ABBI. Use the templates available on Box.

Item-Set Constructed-Response Item Prompts:

Examples:
Stem:

Scoring for SOXXXXXXXXXXXXXX

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

Scoring Information	
Score Points	**Description**
2	Student's response correctly describes two 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 describe one way that World War II affected Louisiana or identifies an incorrect reason.
B	Blank. Student provided no response.

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.

VI. **Selected-response (multiple-choice and multiple-select Items)**

a. Reference sources in stems where possible. 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 to come at the beginning of the stem, when possible.

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.

VII. **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.1.1 Test Blueprint
Table B.1.2 Actual Percentage of Points by Reporting Category and Form
Tables B.2.1–B.2.4 Standard Coverage by Form
Table B.3 Summary of Spring 2018 EFT Item Development Field-Tested Items by Item Type
Table B.4 Summary of Spring 2018 EFT Item Development Field-Tested Items by Reporting Category
Table B.5 Spring 2018 Operational Item Summary for U.S. History
Table B.6 Raw Score Summary
Table B.7 Raw Score Summary by Reporting Category
Tables B.8.1–B.8.4 Scale Score and Raw Score Summary

Table B.1.1 *Test Blueprint*

Reporting Category	Target Percentage
Standard 2: Western Expansion to Progressivism	23%
Standard 3: Isolationism through the Great War	15%
Standard 4: Becoming World Power Through WWII	26%
Standard 5: Cold War Era	23%
Standard 6: The Modern Age	13%

Table B.1.2 *Actual Percentage of Points by Reporting Category and Form (includes Task Set Items)*

Reporting Category	Form A	Form B	Form C	Form D	Form E
Standard 2	23.2%	23.2%	20.3%	20.3%	20.3%
Standard 3	11.6%	11.6%	11.6%	11.6%	14.5%
Standard 4	21.7%	21.7%	21.7%	20.3%	31.9%
Standard 5	33.3%	33.3%	36.2%	23.2%	23.2%
Standard 6	10.1%	10.1%	10.1%	24.6%	10.1%

Table B.2 *Standard Coverage by Form*

Table B.2.1 Form B

Reporting Categories and GLEs		No. of Items					% of Test
		TE	MS	MC	ER	CR	
		N	*N*	*N*	*N*	*N*	
Standard 2	US.2.1			1			1.92
	US.2.2			1			1.92
	US.2.3			1			1.92
	US.2.4	1		2			5.77
	US.2.6			1		1	3.85
	US.2.7			1			1.92
	US.2.8		1	4			9.62
	Subtotal	**1**	**1**	**11**		**1**	**26.92**
Standard 3	US.3.1			1			1.92
	US.3.2			1			1.92
	US.3.3			2			3.85
	US.3.4		1	1			3.85
	US.3.5	1					1.92
	Subtotal	**1**	**1**	**5**			**13.46**
Standard 4	US.4.1			3			5.77
	US.4.10			1			1.92
	US.4.2	1	1	1			5.77
	US.4.6	1					1.92
	US.4.9	1		3			7.69
	Subtotal	**3**	**1**	**8**			**23.08**
Standard 5	US.5.1			2			3.85
	US.5.2	1		2			5.77
	US.5.3	1	1	2	1	1	11.54
	US.5.4		1	1			3.85
	Subtotal	**2**	**2**	**7**	**1**	**1**	**25.00**
Standard 6	US.6.2			1			1.92
	US.6.4	1		4			9.62
	Subtotal	**1**		**5**			**11.54**
Total		**8**	**5**	**36**	**1**	**2**	**100.00**

Table B.2.2 Form C

Reporting Categories and GLEs		No. of Items					% of Test
		TE	MS	MC	ER	CR	
		N	*N*	*N*	*N*	*N*	
Standard 2	US.2.1			1			1.92
	US.2.2			1			1.92
	US.2.3			1			1.92
	US.2.4	1		2			5.77
	US.2.6			1		1	3.85
	US.2.7			1			1.92
	US.2.8		1	2			5.77
	Subtotal	**1**	**1**	**9**		**1**	**23.08**
Standard 3	US.3.1			1			1.92
	US.3.2			1			1.92
	US.3.3			2			3.85
	US.3.4		1	1			3.85
	US.3.5	1					1.92
	Subtotal	**1**	**1**	**5**			**13.46**
Standard 4	US.4.1			2			3.85
	US.4.10			1			1.92
	US.4.2	1	1	1			5.77
	US.4.5			1			1.92
	US.4.6	1					1.92
	US.4.9	1		3			7.69
	Subtotal	**3**	**1**	**8**			**23.08**
Standard 5	US.5.1			2			3.85
	US.5.2	1		2			5.77
	US.5.3	1	1	5	1	1	17.31
	US.5.4			1			1.92
	Subtotal	**2**	**1**	**10**	**1**	**1**	**28.85**
Standard 6	US.6.2			1			1.92
	US.6.4	1		4			9.62
	Subtotal	**1**		**5**			**11.54**
Total		**8**	**4**	**37**	**1**	**2**	**100.00**

Table B.2.3 Form D

Reporting Categories and GLEs		No. of Items					% of Test
		TE	MS	MC	ER	CR	
		N	*N*	*N*	*N*	*N*	
Standard 2	US.2.1			1			1.92
	US.2.2			1			1.92
	US.2.3			1			1.92
	US.2.4	1		2			5.77
	US.2.6			1		1	3.85
	US.2.7			1			1.92
	US.2.8		1	2			5.77
	Subtotal	**1**	**1**	**9**		**1**	**23.08**
Standard 3	US.3.1			1			1.92
	US.3.2			1			1.92
	US.3.3			2			3.85
	US.3.4		1	1			3.85
	US.3.5	1					1.92
	Subtotal	**1**	**1**	**5**			**13.46**
Standard 4	US.4.1			2			3.85
	US.4.10			1			1.92
	US.4.2	1	1	1			5.77
	US.4.6	1					1.92
	US.4.9	1		3			7.69
	Subtotal	**3**	**1**	**7**			**21.15**
Standard 5	US.5.1			3			5.77
	US.5.2	1		2			5.77
	US.5.3	1	1	2		1	9.62
	US.5.4			1			1.92
	US.5.5			1			1.92
	Subtotal	**2**	**1**	**9**		**1**	**25.00**
Standard 6	US.6.2			1			1.92
	US.6.4	1		6	1		15.38
	Subtotal	**1**		**7**	**1**		**17.31**
Total		**8**	**4**	**37**	**1**	**2**	**100.00**

Table B.2.4 Form E

Reporting Categories and GLEs		No. of Items					% of Test
		TE	MS	MC	ER	CR	
		N	*N*	*N*	*N*	*N*	
Standard 2	US.2.1			1			1.92
	US.2.2			1			1.92
	US.2.3			1			1.92
	US.2.4	1		2			5.77
	US.2.6			1		1	3.85
	US.2.7			1			1.92
	US.2.8		1	2			5.77
	Subtotal	**1**	**1**	**9**		**1**	**23.08**
Standard 3	US.3.1			1			1.92
	US.3.2			1			1.92
	US.3.3			2			3.85
	US.3.4		1	3			7.69
	US.3.5	1					1.92
	Subtotal	**1**	**1**	**7**			**17.31**
Standard 4	US.4.1			2			3.85
	US.4.10			1			1.92
	US.4.2	1	1	2	1		9.62
	US.4.6	1					1.92
	US.4.9	1		3			7.69
	Subtotal	**3**	**1**	**8**	**1**		**25.00**
Standard 5	US.5.1			3			5.77
	US.5.2	1		2			5.77
	US.5.3	1	1	2		1	9.62
	US.5.4			1			1.92
	Subtotal	**2**	**1**	**8**		**1**	**23.08**
Standard 6	US.6.2			1			1.92
	US.6.4	1		4			9.62
	Subtotal	**1**		**5**			**11.54**
Total		**8**	**4**	**37**	**1**	**2**	**100.00**

Table B.3

Summary of Spring 2018 EFT Item Development Field-Tested Items by Item Type

Item Type	Item Count	Percent
CR	10	5%
MC	162	77%
MS	11	5%
TE	28	13%

Table B.4

Summary of Spring 2018 EFT Item Development Field-Tested Items by Reporting Category

Reporting Category	Item Count	Percent
Standard 2	40	19%
Standard 3	41	19%
Standard 4	53	25%
Standard 5	52	25%
Standard 6	25	12%

Table B.5

Spring 2018 Operational Item Summary

Form	MC	MS	TE	CR	ER
B	36	5	8	2	1
C	37	4	8	2	1
D	37	4	8	2	1
E	37	4	8	2	1

Table B.6

Raw Score Summary

Form	*N*	Mean	SD	Minimum	Maximum	Reliability	SEM
B	≥9350	34.78	13.35	2	68	0.93	3.53
C	≥8420	35.74	12.94	2	68	0.92	3.66
D	≥7300	35.18	13.04	1	68	0.93	3.45
E	≥7280	35.30	13.12	4	68	0.93	3.47

Note: Reliability is coefficient alpha.

Table B.7

Raw Score Summary by Reporting Category

Form	Reporting Category	Mean	SD	Minimum	Maximum	Reliability	SEM
B	Standard 2	7.98	3.52	0	16	0.76	1.72
B	Standard 3	4.01	1.93	0	8	0.61	1.21
B	Standard 4	7.49	3.19	0	15	0.73	1.66
B	Standard 5	11.40	4.84	0	23	0.83	2.00
B	Standard 6	3.90	1.80	0	7	0.59	1.15
C	Standard 2	6.83	3.19	0	14	0.74	1.63
C	Standard 3	4.19	1.91	0	8	0.60	1.21
C	Standard 4	7.75	3.10	0	15	0.71	1.67
C	Standard 5	12.89	4.99	0	25	0.82	2.12
C	Standard 6	4.08	1.77	0	7	0.59	1.13
D	Standard 2	6.75	3.22	0	14	0.74	1.64
D	Standard 3	4.16	1.88	0	8	0.59	1.20
D	Standard 4	7.10	2.96	0	14	0.71	1.59
D	Standard 5	9.32	3.18	0	16	0.74	1.62
D	Standard 6	7.86	3.78	0	17	0.78	1.77
E	Standard 2	6.75	3.18	0	14	0.73	1.65
E	Standard 3	5.50	2.36	0	10	0.68	1.34
E	Standard 4	9.99	4.70	0	23	0.80	2.10
E	Standard 5	9.03	3.05	0	15	0.74	1.56
E	Standard 6	4.04	1.76	0	7	0.58	1.14

Table B.8 *Scale Score and Raw Score Summary*

Table B.8.1

U.S. History Form B

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥9350	100.00	729.05	33.37	34.78	13.35
Female	4760	50.89	729.36	31.48	34.8	12.76
Male	≥4590	49.11	728.73	35.22	34.76	13.93
African American	≥4010	43.35	716.87	30.13	29.76	11.75
Asian	≥160	1.74	743.55	34.77	40.94	13.84
Hispanic/Latino	≥580	6.28	721.15	36.09	31.89	13.99
Multi-Racial	≥140	1.61	735.97	32.57	37.49	12.95
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥4350	46.98	740.53	31.57	39.45	12.88

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

Table B.8.2

U.S. History Form C

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥8420	100.00	731.63	31.77	35.74	12.94
Female	≥4370	51.92	730.51	30.11	35.2	12.38
Male	≥4040	48.08	732.85	33.44	36.33	13.5
African American	≥3460	41.39	719.1	29.65	30.57	11.74
Asian	≥170	2.08	750.31	30.65	43.21	12.53
Hispanic/Latino	≥380	4.58	729.4	32.34	34.72	13
Multi-Racial	≥130	1.59	734.89	32.41	37.1	12.86
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥4200	50.27	741.2	29.75	39.71	12.36

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

Table B.8.3

U.S. History Form D

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥7300	100.00	730.18	32.15	35.18	13.04
Female	≥3780	51.83	729.95	30.14	34.99	12.39
Male	≥3510	48.17	730.41	34.18	35.39	13.7
African American	≥3090	42.66	717.25	30.1	29.9	11.78
Asian	≥120	1.77	753.32	32.77	44.64	12.45
Hispanic/Latino	≥330	4.61	730.54	31.43	35.24	12.91
Multi-Racial	≥90	1.27	735.32	32.24	37.09	12.87
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥3590	49.59	740.23	29.8	39.31	12.41

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

Table B.8.4

U.S. History Form E

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥7280	100.00	730.43	32.35	35.3	13.12
Female	≥3790	52.11	729.11	31.07	34.65	12.67
Male	≥3480	47.89	731.87	33.64	36.01	13.56
African American	≥3000	41.41	717.27	29.56	29.84	11.66
Asian	≥100	1.43	754.12	32.9	45.09	13.05
Hispanic/Latino	≥340	4.81	729.85	33.47	35.29	13.38
Multi-Racial	≥110	1.52	736.44	30.67	37.62	12.74
Native Hawaiian or Other Pacific Islander	<10	NR	NR	NR	NR	NR
White	≥3670	50.75	740.32	30.43	39.39	12.54

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

Contents
Table C.1 P-Value Summary by Item Type
Plot C.1 P-Value by Item Type
Table C.2 Item-Total Correlation: Summary of Point-Biserial Correlation
Plot C.2 Item-Total Correlation by Item Type
Table C.3 Summary of Corrected Point-Biserial Correlation
Plot C.3 Corrected Point-Biserial Correlation
Table C.4 Item-Total Correlation Summary by Reporting Category
Table C.5 IRT Parameter Summary
Plot C.4 IRT a-Parameter
Plot C.5 IRT b-Parameter
Plot C.6 IRT c-Parameter

Table C.1

P-Value Summary by Item Type

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.313	0.313	0.406	0.498	0.498
ER	4	0.269	0.276	0.325	0.346	0.364
MC	48	0.277	0.550	0.616	0.721	0.871
MS	5	0.341	0.385	0.640	0.682	0.690
TE	8	0.281	0.307	0.385	0.469	0.599

Plot C.1
P-Value by Item Type

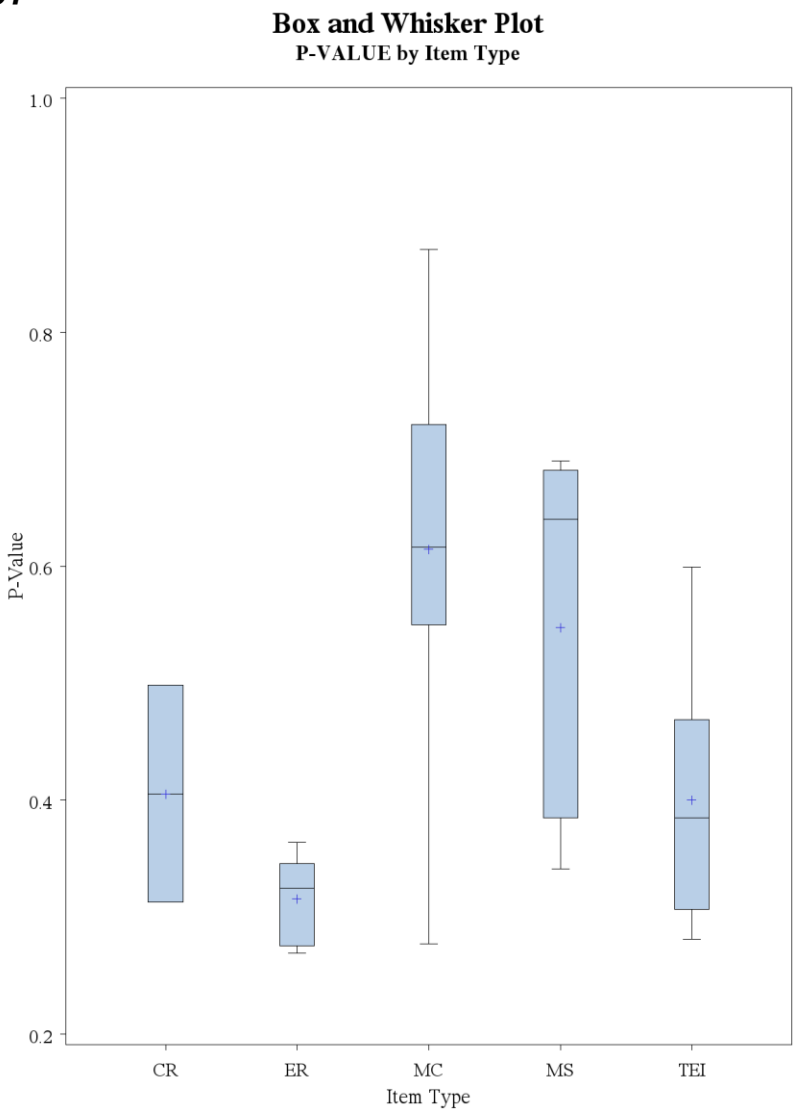


Table C.2

Item-Total Correlation Summary: Point-Biserial Correlation

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.573	0.573	0.631	0.689	0.689
ER	4	0.733	0.735	0.747	0.764	0.775
MC	48	0.237	0.358	0.407	0.456	0.550
MS	5	0.355	0.492	0.519	0.563	0.584
TE	8	0.407	0.463	0.494	0.555	0.598

Plot C.2

Item-Total Correlation by Item Type

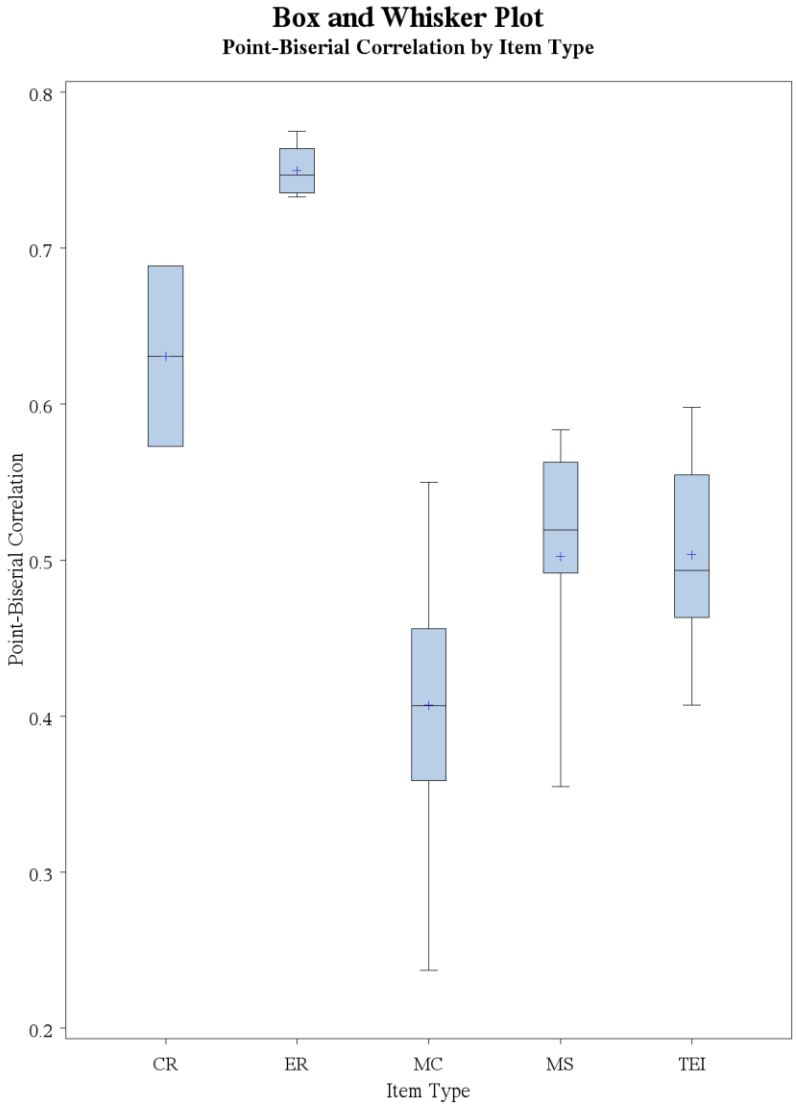


Table C.3

Summary of Corrected Point-Biserial Correlation*

Item Type	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	2	0.537	0.537	0.596	0.655	0.655
ER	8	0.690	0.694	0.708	0.726	0.741
MC	48	0.201	0.325	0.377	0.428	0.523
MS	5	0.322	0.464	0.493	0.537	0.559
TE	8	0.363	0.428	0.457	0.519	0.558

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

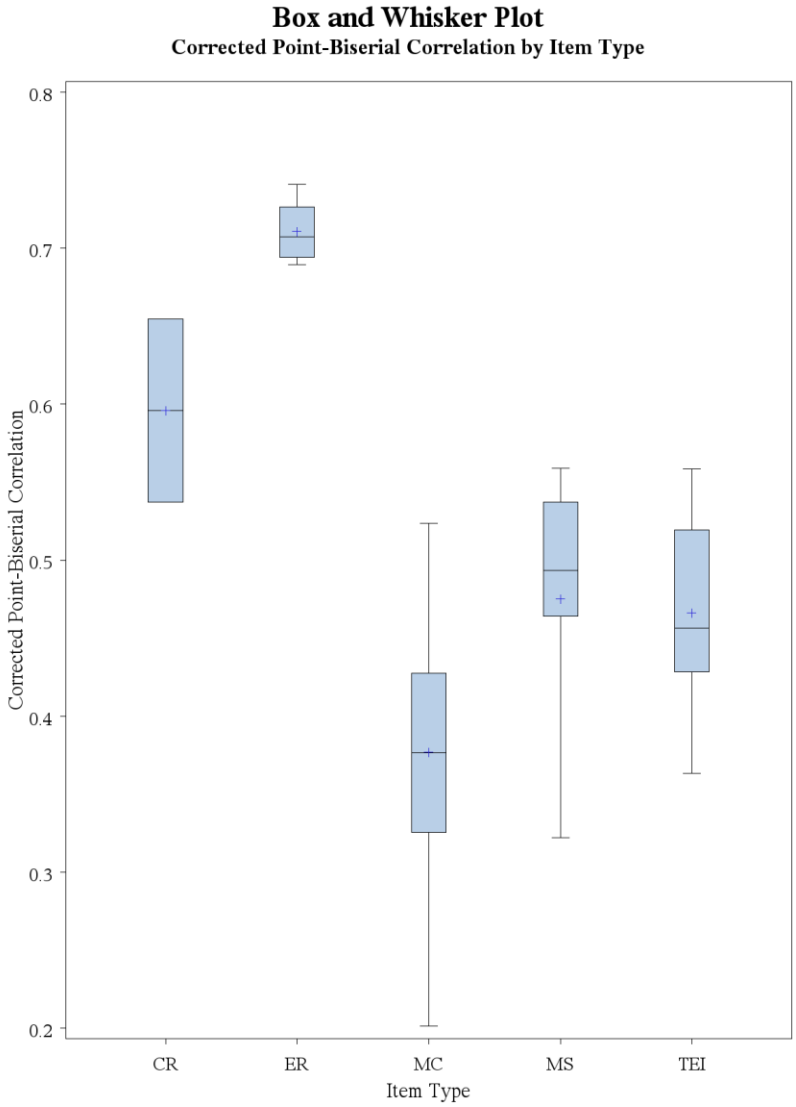


Table C.4

Item-Total Correlation Summary by Reporting Category

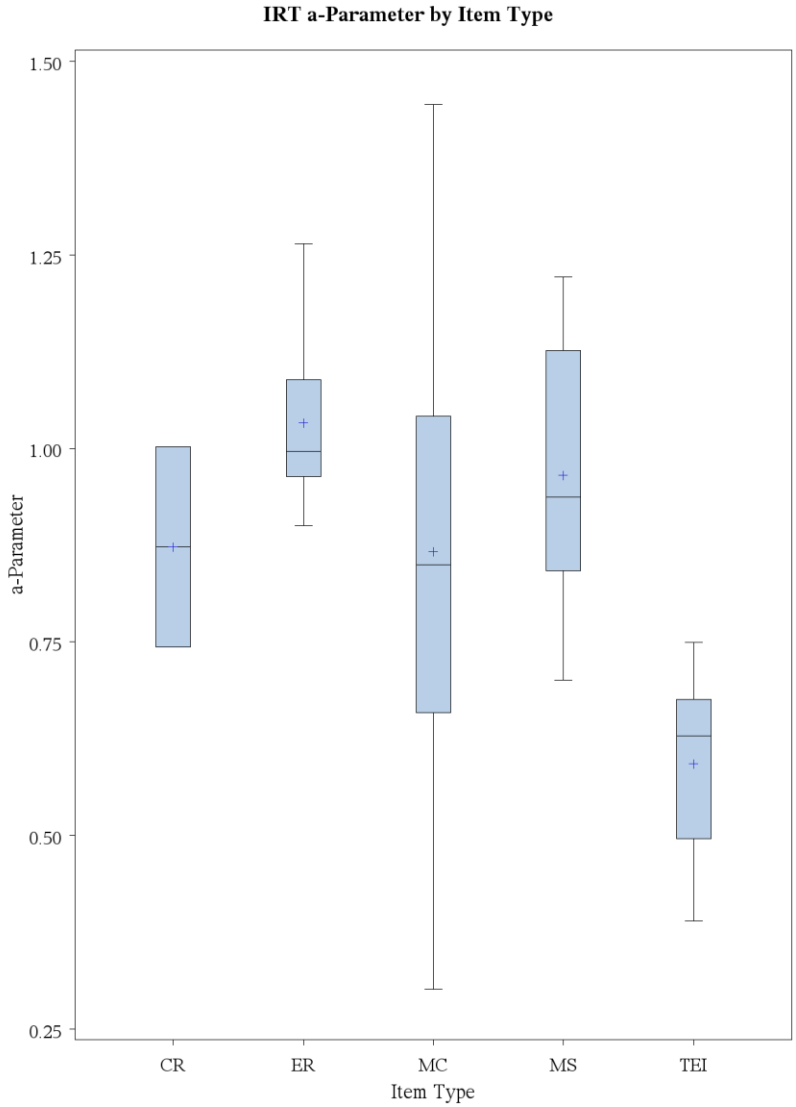
Item Type	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	Standard 2	1	0.689	0.689	0.689	0.689	0.689
CR	Standard 5	1	0.573	0.573	0.573	0.573	0.573
ER	Standard 4	1	0.734	0.734	0.735	0.737	0.737
ER	Standard 5	2	0.733	0.736	0.747	0.755	0.756
ER	Standard 6	1	0.772	0.772	0.773	0.775	0.775
MC	Standard 2	11	0.272	0.381	0.401	0.425	0.470
MC	Standard 3	7	0.355	0.412	0.489	0.504	0.538
MC	Standard 4	10	0.237	0.325	0.394	0.433	0.550
MC	Standard 5	13	0.253	0.356	0.392	0.452	0.548
MC	Standard 6	7	0.328	0.345	0.432	0.463	0.489
MS	Standard 2	1	0.563	0.563	0.563	0.563	0.563
MS	Standard 3	1	0.355	0.355	0.355	0.355	0.355
MS	Standard 4	1	0.492	0.492	0.492	0.492	0.492
MS	Standard 5	2	0.519	0.519	0.552	0.584	0.584
TE	Standard 2	1	0.513	0.513	0.513	0.513	0.513
TE	Standard 3	1	0.474	0.474	0.474	0.474	0.474
TE	Standard 4	3	0.474	0.474	0.523	0.598	0.598
TE	Standard 5	2	0.407	0.407	0.430	0.453	0.453
TE	Standard 6	1	0.586	0.586	0.586	0.586	0.586

Table C.5

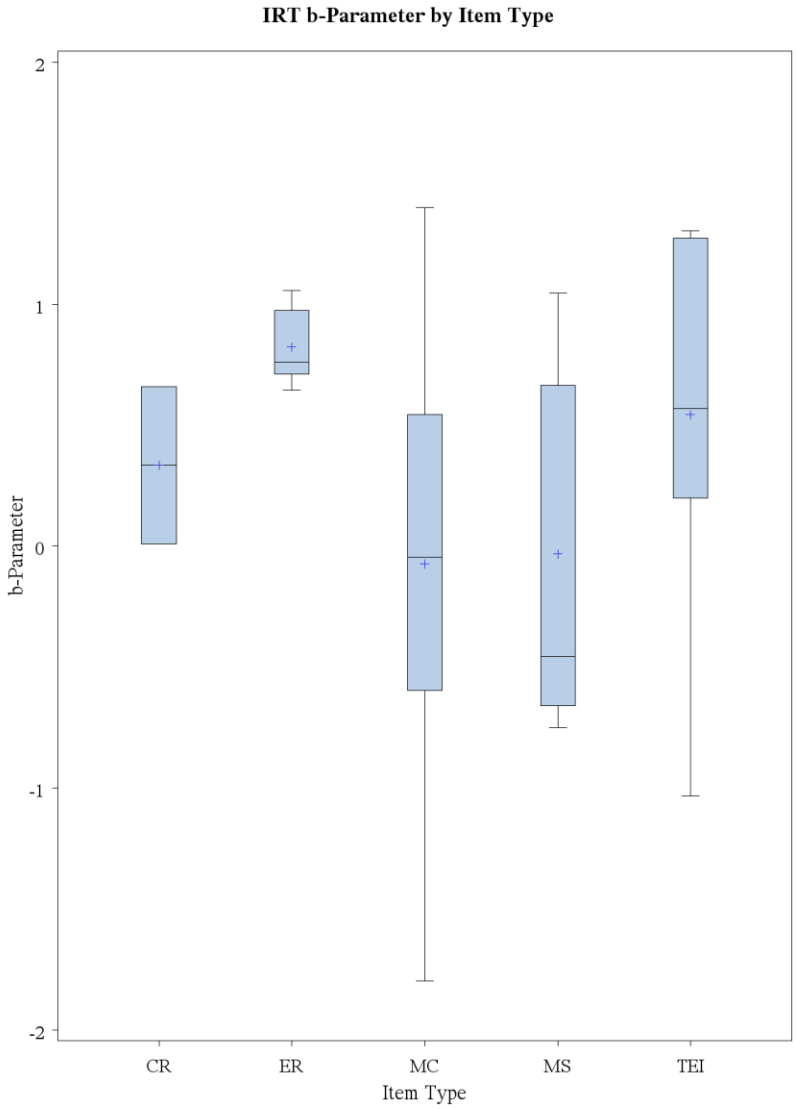
IRT Parameter Summary

Item Type	Parameter	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
CR	*a*	2	0.743	0.743	0.873	1.002	1.002
CR	*b*	2	0.010	0.010	0.335	0.660	0.660
ER	*a*	4	0.900	0.964	0.996	1.089	1.265
ER	*b*	4	0.646	0.712	0.759	0.975	1.058
MC	*a*	48	0.301	0.659	0.850	1.042	1.444
MC	*b*	48	–1.799	–0.596	–0.045	0.545	1.401
MC	*c*	48	0.054	0.131	0.183	0.265	0.380
MS	*a*	5	0.700	0.842	0.937	1.126	1.222
MS	*b*	5	–0.751	–0.659	–0.455	0.666	1.047
MS	*c*	5	0.015	0.046	0.048	0.051	0.155
TE	*a*	8	0.390	0.495	0.629	0.675	0.750
TE	*b*	8	–1.033	0.199	0.570	1.273	1.304

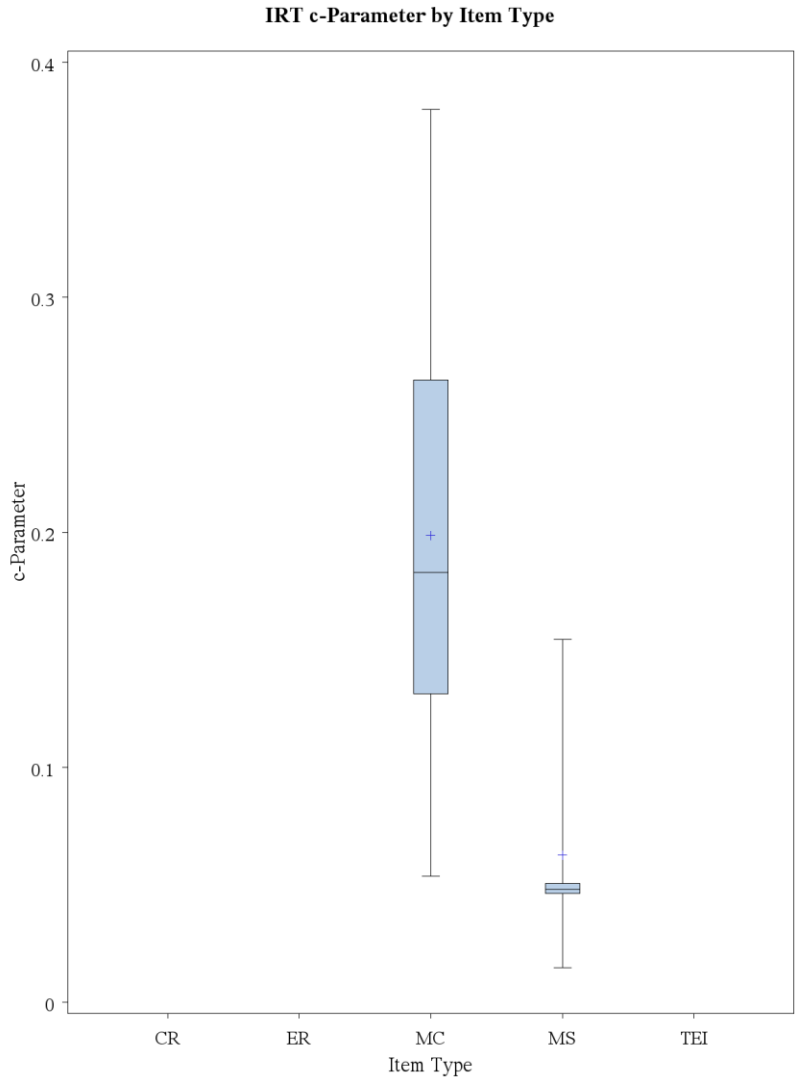
Plot C.4
IRT a-Parameter



Plot C.5
IRT b-Parameter



Plot C.6
IRT c-Parameter



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

Appendix D: Dimensionality

Dimensionality Reports

U.S. History

Contents
Table D.1 Zq1 Statistics and Summary Data
Table D.2 Q3 Statistics and Summary Data
Table D.3.1–D.3.4 Reporting Category Intercorrelation Coefficients

Table D.1

Zq1 Statistics and Summary Data

Form	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
B	CR	20.59	20.59	37.80	55.02	55.02	0
B	ER	19.40	19.40	19.66	19.91	19.91	0
B	MC	1.34	4.97	8.06	13.37	61.20	0
B	MS	7.46	7.56	10.43	12.15	22.18	0
B	TE	9.75	27.86	32.74	60.90	87.66	1
C	CR	20.59	20.59	37.80	55.02	55.02	0
C	ER	8.31	8.31	9.71	11.11	11.11	0
C	MC	1.29	4.65	7.81	12.51	61.20	0
C	MS	7.46	8.95	11.29	17.17	22.18	0
C	TE	9.75	27.86	32.74	60.90	87.66	1
D	CR	20.59	20.59	37.80	55.02	55.02	0
D	ER	9.27	9.27	10.11	10.94	10.94	0
D	MC	0.22	4.65	7.81	12.51	61.20	0
D	MS	7.46	8.95	11.29	17.17	22.18	0
D	TE	9.75	27.86	32.74	60.90	87.66	1
E	CR	20.59	20.59	37.80	55.02	55.02	0
E	ER	3.73	3.73	4.15	4.56	4.56	0
E	MC	1.50	4.65	7.81	12.51	61.20	0
E	MS	7.46	8.95	11.29	17.17	22.18	0
E	TE	9.75	27.86	32.74	60.90	87.66	1

Table D.2

Q3 Statistics and Summary Data

Form	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
B	0.196	–0.095	–0.048	–0.016	0.026	0.916
C	0.183	–0.083	–0.049	–0.016	0.028	0.883
D	0.185	–0.076	–0.049	–0.017	0.030	0.906
E	0.190	–0.089	–0.048	–0.017	0.025	0.930

Table D.3 ***Reporting Category Intercorrelation Coefficients***

Table D.3.1

U.S. History Form B

Reporting Category	Standard 2	Standard 3	Standard 4	Standard 5	Standard 6
Standard 2	1.00				
Standard 3	0.67	1.00			
Standard 4	0.74	0.67	1.00		
Standard 5	0.75	0.64	0.73	1.00	
Standard 6	0.66	0.58	0.63	0.67	1.00

Table D.3.2

U.S. History Form C

Reporting Category	Standard 2	Standard 3	Standard 4	Standard 5	Standard 6
Standard 2	1.00				
Standard 3	0.66	1.00			
Standard 4	0.72	0.65	1.00		
Standard 5	0.72	0.64	0.71	1.00	
Standard 6	0.64	0.58	0.62	0.64	1.00

Table D.3.3

U.S. History Form D

Reporting Category	Standard 2	Standard 3	Standard 4	Standard 5	Standard 6
Standard 2	1.00				
Standard 3	0.65	1.00			
Standard 4	0.71	0.65	1.00		
Standard 5	0.68	0.60	0.68	1.00	
Standard 6	0.73	0.64	0.70	0.74	1.00

Table D.3.4

U.S. History Form E

Reporting Category	Standard 2	Standard 3	Standard 4	Standard 5	Standard 6
Standard 2	1.00				
Standard 3	0.69	1.00			
Standard 4	0.75	0.72	1.00		
Standard 5	0.68	0.65	0.74	1.00	
Standard 6	0.64	0.60	0.65	0.63	1.00

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 ALL

N	≥32360		
Mean	730.29	Median	731.00
Std deviation	32.47	Variance	1054.15
Skewness	-0.0899	Kurtosis	0.0241
Mode	719.00	Std Error Mean	0.1805
Range	200.00	Interquartile Range	43.00

Quantile	Estimate
100% Max	850
99%	805
95%	781
90%	770
75% Q3	752
50% Median	731
25% Q1	709
10%	689
5%	676
1%	650
0% Min	650

```
                 Histogram                          #  Boxplot
855+*                                             < 10     0
   .*                                             ≥ 20     0
835+
   .*                                             ≥ 40     0
815+**                                           ≥ 160     0
   .****                                         ≥ 280     |
795+******                                       ≥ 440     |
   .***********                                  ≥ 960     |
775+********************                       ≥ 1740     |
   .**************************                  ≥ 2240     |
755+*************************************       ≥ 3320  +-----+
   .********************************************  ≥ 3840  |     |
735+*************************************       ≥ 3270  *--+--*
   .************************************************ ≥ 4200  |     |
715+****************************************      ≥ 3380  |     |
   .**********************************           ≥ 2920  +-----+
695+**********************                      ≥ 1920     |
   .******************                          ≥ 1570     |
675+**********                                   ≥ 790     |
   .****                                         ≥ 320     |
655+**********                                   ≥ 870     |
    ----+----+----+----+----+----+----+----+----+---
    * may represent ≥ 80 counts
```

```
                 Normal Probability Plot
855+                                                  *
   |                                                  *
835+
   |                                                  *
815+                                                 **
   |                                              ****
795+                                            ***
   |                                         ****
775+                                      *****
   |                                   ****
755+                                ****
   |                            *****
735+                          ***
   |                       ****
715+                    ****
   |                *****
695+             ***
   |          ****
675+       ****
   |    ++**
655+*******
    +----+----+----+----+----+----+----+----+----+----+
        -2        -1         0        +1        +2
```

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

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

FREQUENCY DISTRIBUTION - SCALE SCORES

ScaleScore		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	≥ 600	≥ 600	1.86	1.86
657	**********	≥ 260	≥ 870	0.83	2.69
665	**********	≥ 320	≥ 1190	1.00	3.70
671	**********	≥ 370	≥ 1570	1.16	4.85
676	**********	≥ 410	≥ 1980	1.29	6.15
681	**********	≥ 470	≥ 2460	1.48	7.63
685	**********	≥ 480	≥ 2950	1.50	9.13
689	**********	≥ 600	≥ 3560	1.87	11.00
692	**********	≥ 610	≥ 4170	1.90	12.90
696	**********	≥ 660	≥ 4830	2.04	14.94
699	**********	≥ 650	≥ 5480	2.01	16.95
701	**********	≥ 700	≥ 6180	2.17	19.12
704	**********	≥ 750	≥ 6940	2.33	21.46
707	**********	≥ 710	≥ 7650	2.20	23.66
709	**********	≥ 750	≥ 8410	2.33	25.99
712	**********	≥ 810	≥ 9230	2.53	28.52
714	**********	≥ 820	≥ 10050	2.54	31.06
716	**********	≥ 850	≥ 10900	2.63	33.69
719	**********	≥ 890	≥ 11790	2.76	36.45
721	**********	≥ 800	≥ 12600	2.49	38.94
723	**********	≥ 810	≥ 13420	2.53	41.47
725	**********	≥ 850	≥ 14270	2.65	44.12
727	**********	≥ 870	≥ 15150	2.70	46.82
729	**********	≥ 840	≥ 15990	2.61	49.43
731	**********	≥ 880	≥ 16870	2.72	52.15
733	**********	≥ 820	≥ 17690	2.54	54.69
735	**********	≥ 760	≥ 18460	2.37	57.06
738	**********	≥ 800	≥ 19270	2.49	59.55
740	**********	≥ 750	≥ 20020	2.32	61.88
742	**********	≥ 800	≥ 20830	2.49	64.37
744	**********	≥ 760	≥ 21590	2.37	66.74
746	**********	≥ 750	≥ 22350	2.33	69.07
748	**********	≥ 760	≥ 23110	2.37	71.44
750	**********	≥ 690	≥ 23810	2.15	73.59
752	**********	≥ 690	≥ 24500	2.14	75.72
754	**********	≥ 700	≥ 25200	2.17	77.89
756	**********	≥ 640	≥ 25850	1.99	79.88
758	**********	≥ 590	≥ 26440	1.84	81.72
761	**********	≥ 580	≥ 27020	1.80	83.52
763	**********	≥ 590	≥ 27620	1.85	85.37
765	**********	≥ 520	≥ 28150	1.63	87.00
768	**********	≥ 530	≥ 28690	1.65	88.65
770	**********	≥ 470	≥ 29160	1.48	90.13
773	**********	≥ 430	≥ 29600	1.35	91.48
775	**********	≥ 440	≥ 30050	1.38	92.87
778	**********	≥ 380	≥ 30430	1.19	94.05
781	**********	≥ 380	≥ 30820	1.20	95.25
784	**********	≥ 290	≥ 31110	0.91	96.16
788	**********	≥ 270	≥ 31390	0.86	97.02
791	**********	≥ 220	≥ 31620	0.70	97.72
795	**********	≥ 210	≥ 31840	0.67	98.39
800	********	≥ 160	≥ 32000	0.51	98.90
805	******	≥ 110	≥ 32120	0.36	99.26
811	****	≥ 80	≥ 32210	0.27	99.53
818	****	≥ 70	≥ 32280	0.23	99.76
828	**	≥ 40	≥ 32330	0.14	99.90
841	*	≥ 20	≥ 32350	0.07	99.97
850		< 10	≥ 32360	NR	100.00

100 200 300 400 500 600 700 800 900

Table E.3 ***Scale Score Descriptive Statistics and Plots***

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

N	≥ 9350		
Mean	729.05	Median	729.00
Std deviation	33.37	Variance	1113.45
Skewness	-0.0800	Kurtosis	-0.0175
Mode	719.00	Std Error Mean	0.3450
Range	200.00	Interquartile Range	45.00

Quantile	Estimate
100% Max	850
99%	805
95%	781
90%	770
75% Q3	752
50% Median	729
25% Q1	707
10%	685
5%	671
1%	650
0% Min	650

```
                 Histogram                          #  Boxplot
855+*                                             < 10     0
   .*                                             < 10     0
835+
   .*                                              ≥ 10     0
815+**                                             ≥ 40     |
   .****                                           ≥ 80     |
795+*****                                         ≥ 120     |
   .***********                                   ≥ 280     |
775+********************                          ≥ 500     |
   .************************                      ≥ 610     |
755+************************************          ≥ 910  +-----+
   .****************************************     ≥ 1030  |     |
735+************************************          ≥ 930  |     |
   .************************************************ ≥ 1240  *--+--*
715+************************************          ≥ 980  |     |
   .********************************             ≥ 830  +-----+
695+**********************                        ≥ 570     |
   .*****************                             ≥ 450     |
675+***********                                   ≥ 270     |
   .*****                                          ≥ 110     |
655+************                                  ≥ 300     |
    ----+----+----+----+----+----+----+----+----+---
    * may represent ≥ 20 counts
```

```
                 Normal Probability Plot
855+                                                  *
   |                                                  *
835+
   |                                                  *
815+                                                 **
   |                                              ****
795+                                            ***
   |                                         ****
775+                                     *****
   |                                  ****
755+                              ****
   |                           ****
735+                        ****
   |                     ****
715+                  ****
   |               ****
695+            ****
   |          ***
675+       ****
   |    ++**
655+********
    +----+----+----+----+----+----+----+----+----+----+
       -2        -1         0        +1        +2
```

Table E.4 *Frequency Distribution of Scale Scores*

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

ScaleScore		Freq	Cum. Freq	Percent	Cum. Percent
650	**********	⩾ 210	⩾ 210	2.25	2.25
657	**********	⩾ 90	⩾ 300	1.04	3.28
665	**********	⩾ 110	⩾ 420	1.22	4.50
671	**********	⩾ 120	⩾ 540	1.29	5.79
676	**********	⩾ 150	⩾ 690	1.61	7.41
681	**********	⩾ 160	⩾ 850	1.72	9.13
685	**********	⩾ 120	⩾ 980	1.37	10.50
689	**********	⩾ 170	⩾ 1150	1.82	12.32
692	**********	⩾ 180	⩾ 1330	1.95	14.26
696	**********	⩾ 190	⩾ 1530	2.13	16.39
699	**********	⩾ 180	⩾ 1720	2.02	18.41
701	**********	⩾ 200	⩾ 1920	2.19	20.60
704	**********	⩾ 210	⩾ 2140	2.29	22.89
707	**********	⩾ 200	⩾ 2340	2.16	25.05
709	**********	⩾ 210	⩾ 2560	2.33	27.38
712	**********	⩾ 230	⩾ 2790	2.46	29.84
714	**********	⩾ 250	⩾ 3040	2.71	32.55
716	**********	⩾ 230	⩾ 3280	2.53	35.08
719	**********	⩾ 260	⩾ 3540	2.81	37.89
721	**********	⩾ 230	⩾ 3770	2.47	40.36
723	**********	⩾ 250	⩾ 4030	2.73	43.09
725	**********	⩾ 250	⩾ 4280	2.68	45.77
727	**********	⩾ 250	⩾ 4530	2.73	48.50
729	**********	⩾ 240	⩾ 4780	2.66	51.16
731	**********	⩾ 240	⩾ 5020	2.58	53.74
733	**********	⩾ 230	⩾ 5260	2.54	56.28
735	**********	⩾ 230	⩾ 5490	2.50	58.78
738	**********	⩾ 220	⩾ 5720	2.41	61.19
740	**********	⩾ 200	⩾ 5930	2.21	63.40
742	**********	⩾ 200	⩾ 6130	2.21	65.62
744	**********	⩾ 190	⩾ 6330	2.07	67.69
746	**********	⩾ 220	⩾ 6550	2.36	70.05
748	**********	⩾ 200	⩾ 6750	2.20	72.25
750	**********	⩾ 180	⩾ 6940	1.95	74.20
752	**********	⩾ 190	⩾ 7130	2.11	76.31
754	**********	⩾ 200	⩾ 7340	2.18	78.49
756	**********	⩾ 170	⩾ 7510	1.88	80.37
758	**********	⩾ 150	⩾ 7670	1.67	82.04
761	**********	⩾ 150	⩾ 7820	1.65	83.68
763	**********	⩾ 150	⩾ 7980	1.70	85.38
765	**********	⩾ 140	⩾ 8130	1.59	86.98
768	**********	⩾ 150	⩾ 8280	1.61	88.59
770	**********	⩾ 130	⩾ 8420	1.44	90.04
773	**********	⩾ 120	⩾ 8540	1.34	91.37
775	**********	⩾ 120	⩾ 8670	1.35	92.72
778	**********	⩾ 110	⩾ 8790	1.26	93.98
781	**********	⩾ 100	⩾ 8890	1.15	95.14
784	**********	⩾ 90	⩾ 8990	0.98	96.12
788	**********	⩾ 80	⩾ 9070	0.89	97.01
791	**********	⩾ 60	⩾ 9130	0.67	97.68
795	**********	⩾ 60	⩾ 9190	0.64	98.32
800	**********	⩾ 50	⩾ 9240	0.56	98.88
805	******	⩾ 20	⩾ 9270	0.31	99.19
811	*****	⩾ 20	⩾ 9300	0.27	99.45
818	****	⩾ 20	⩾ 9320	0.24	99.69
828	****	⩾ 10	⩾ 9340	0.20	99.89
841	**	< 10	⩾ 9350	NR	99.99
850		< 10	⩾ 9350	NR	100.00

20 40 60 80 100 120 140 160 180 200 220 240 260

Table E.5 ***Scale Score Descriptive Statistics and Plots***

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

N	≥ 8420		
Mean	731.63	Median	731.00
Std deviation	31.77	Variance	1009.48
Skewness	-0.0605	Kurtosis	0.1019
Mode	733.00	Std Error Mean	0.3462
Range	200.00	Interquartile Range	42.00

Quantile	Estimate
100% Max	850
99%	805
95%	781
90%	770
75% Q3	754
50% Median	731
25% Q1	712
10%	692
5%	676
1%	650
0% Min	650

```
                    Histogram                          #   Boxplot
855+*                                                < 10      0
   .*                                                < 10      0
835+
   .*                                                ≥ 10      0
815+**                                               ≥ 40      0
   .****                                             ≥ 80      |
795+*****                                           ≥ 110      |
   .**********                                      ≥ 230      |
775+*******************                             ≥ 450      |
   .***************************                     ≥ 630      |
755+***************************************          ≥ 910   +-----+
   .*********************************************   ≥ 1040   |     |
735+************************************            ≥ 860   *--+--*
   .************************************************ ≥ 1110   |     |
715+************************************             ≥ 830   +-----+
   .********************************                ≥ 750      |
695+*********************                           ≥ 490      |
   .*****************                               ≥ 390      |
675+********                                        ≥ 170      |
   .***                                              ≥ 70      |
655+********                                        ≥ 170      |
    ----+----+----+----+----+----+----+----+----+--
    * may represent ≥ 20 counts
```

```
                    Normal Probability Plot
855+                                                  *
   |                                                  *
835+
   |                                                  *
815+                                                 **
   |                                               ****
795+                                             ***
   |                                          ****
775+                                       *****
   |                                     ****
755+                                 *****
   |                             ****
735+                          ****
   |                       ****
715+                    ****
   |                 ****
695+              ****
   |           ****
675+        ****
   |     ++**
655+******
    +----+----+----+----+----+----+----+----+----+----+
       -2        -1         0        +1        +2
```

Table E.6 *Frequency Distribution of Scale Scores*

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

ScaleScore			Cum. Freq	Freq	Percent	Cum. Percent
650	**********	≥ 110	≥ 110	1.37	1.37
657	**********	≥ 60	≥ 170	0.76	2.13
665	**********	≥ 70	≥ 250	0.86	2.98
671	**********	≥ 80	≥ 330	0.97	3.95
676	**********	≥ 90	≥ 420	1.14	5.09
681	**********	≥ 110	≥ 540	1.34	6.44
685	**********	≥ 140	≥ 680	1.66	8.10
689	**********	≥ 130	≥ 820	1.65	9.75
692	**********	≥ 160	≥ 980	1.92	11.67
696	**********	≥ 170	≥ 1150	2.08	13.75
699	**********	≥ 160	≥ 1310	1.91	15.66
701	**********	≥ 160	≥ 1480	1.97	17.63
704	**********	≥ 180	≥ 1670	2.24	19.88
707	**********	≥ 200	≥ 1870	2.39	22.27
709	**********	≥ 200	≥ 2070	2.41	24.68
712	**********	≥ 200	≥ 2280	2.41	27.09
714	**********	≥ 170	≥ 2450	2.10	29.19
716	**********	≥ 220	≥ 2680	2.72	31.91
719	**********	≥ 220	≥ 2910	2.67	34.58
721	**********	≥ 210	≥ 3120	2.58	37.16
723	**********	≥ 200	≥ 3330	2.40	39.56
725	**********	≥ 230	≥ 3560	2.74	42.30
727	**********	≥ 220	≥ 3780	2.66	44.96
729	**********	≥ 230	≥ 4020	2.81	47.77
731	**********	≥ 230	≥ 4260	2.81	50.59
733	**********	≥ 230	≥ 4490	2.83	53.41
735	**********	≥ 180	≥ 4680	2.20	55.61
738	**********	≥ 200	≥ 4880	2.41	58.02
740	**********	≥ 210	≥ 5100	2.55	60.57
742	**********	≥ 210	≥ 5310	2.52	63.09
744	**********	≥ 200	≥ 5520	2.48	65.57
746	**********	≥ 200	≥ 5720	2.41	67.98
748	**********	≥ 200	≥ 5920	2.39	70.37
750	**********	≥ 190	≥ 6120	2.32	72.69
752	**********	≥ 170	≥ 6290	2.03	74.72
754	**********	≥ 180	≥ 6480	2.23	76.95
756	**********	≥ 180	≥ 6660	2.24	79.19
758	**********	≥ 160	≥ 6830	2.01	81.20
761	**********	≥ 150	≥ 6990	1.86	83.07
763	**********	≥ 170	≥ 7170	2.11	85.18
765	**********	≥ 150	≥ 7320	1.81	86.98
768	**********	≥ 140	≥ 7470	1.72	88.71
770	**********	≥ 120	≥ 7590	1.44	90.14
773	**********	≥ 110	≥ 7700	1.32	91.46
775	**********	≥ 120	≥ 7820	1.51	92.97
778	**********	≥ 90	≥ 7920	1.14	94.11
781	**********	≥ 90	≥ 8010	1.09	95.20
784	**********	≥ 70	≥ 8080	0.86	96.06
788	**********	≥ 60	≥ 8150	0.78	96.84
791	**********	≥ 60	≥ 8220	0.77	97.61
795	**********	≥ 50	≥ 8270	0.62	98.23
800	*********	≥ 40	≥ 8310	0.51	98.74
805	*******	≥ 30	≥ 8350	0.44	99.18
811	*****	≥ 20	≥ 8370	0.30	99.48
818	****	≥ 20	≥ 8390	0.25	99.73
828	**	≥ 10	≥ 8410	0.14	99.87
841	*	< 10	≥ 8410	NR	99.95
850	*	< 10	≥ 8420	NR	100.00

```
----+---+---+---+---+---+---+---+---+---+---+---+
20  40  60  80 100 120 140 160 180 200 220 240
Frequency
```

Table E.7 ***Scale Score Descriptive Statistics and Plots***

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

N	≥ 7300		
Mean	730.18	Median	731.00
Std deviation	32.15	Variance	1033.60
Skewness	-0.1027	Kurtosis	0.0107
Mode	725.00	Std Error Mean	0.3762
Range	200.00	Interquartile Range	43.00

Quantile	Estimate
100% Max	850
99%	805
95%	781
90%	770
75% Q3	752
50% Median	731
25% Q1	709
10%	689
5%	676
1%	650
0% Min	650

```
                    Histogram                                  #  Boxplot
855+*                                                       < 10     0
   .*                                                       < 10     0
835+
   .*                                                       < 10     0
815+**                                                      ≥ 30     0
   .***                                                     ≥ 50     |
795+*****                                                   ≥ 90     |
   .************                                           ≥ 230     |
775+*******************                                    ≥ 370     |
   .**************************                               ≥ 510     |
755+*************************************                   ≥ 730  +-----+
   .*********************************************            ≥ 880  |     |
735+*************************************                   ≥ 730  *--+--*
   .************************************************         ≥ 940  |     |
715+*****************************************                ≥ 780  |     |
   .***********************************                      ≥ 680  +-----+
695+*********************                                   ≥ 410     |
   .*******************                                     ≥ 360     |
675+*********                                               ≥ 160     |
   .****                                                     ≥ 80     |
655+**********                                              ≥ 180     |
    ----+----+----+----+----+----+----+----+----+---
    * may represent ≥ 20 counts
```

```
                    Normal Probability Plot
855+                                                    *
   |                                                    *
835+
   |                                                    *
815+                                                    *
   |                                                 ****
795+                                              ****
   |                                           ****
775+                                        *****
   |                                     ****
755+                                  ****
   |                              *****
735+                            ***+
   |                         ****
715+                      ****
   |                   *****
695+                ***
   |            *****
675+          ***
   |       ++**
655+*******
    +----+----+----+----+----+----+----+----+----+----+
       -2        -1         0        +1        +2
```

Table E.8 *Frequency Distribution of Scale Scores*

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

ScaleScore	Freq	Cum. Freq	Percent	Cum. Percent
650	≥ 130	≥ 130	1.86	1.86
657	≥ 50	≥ 180	0.73	2.59
665	≥ 80	≥ 260	1.10	3.68
671	≥ 70	≥ 340	1.08	4.77
676	≥ 80	≥ 430	1.16	5.93
681	≥ 100	≥ 530	1.40	7.33
685	≥ 110	≥ 640	1.55	8.87
689	≥ 140	≥ 790	2.04	10.91
692	≥ 130	≥ 930	1.83	12.75
696	≥ 130	≥ 1070	1.90	14.65
699	≥ 130	≥ 1200	1.89	16.54
701	≥ 160	≥ 1370	2.26	18.80
704	≥ 180	≥ 1550	2.49	21.29
707	≥ 160	≥ 1720	2.31	23.61
709	≥ 160	≥ 1890	2.29	25.89
712	≥ 190	≥ 2080	2.67	28.56
714	≥ 190	≥ 2280	2.70	31.26
716	≥ 190	≥ 2470	2.66	33.92
719	≥ 190	≥ 2670	2.71	36.63
721	≥ 180	≥ 2860	2.59	39.22
723	≥ 170	≥ 3040	2.44	41.65
725	≥ 200	≥ 3240	2.79	44.45
727	≥ 190	≥ 3440	2.72	47.17
729	≥ 170	≥ 3620	2.40	49.57
731	≥ 190	≥ 3810	2.68	52.25
733	≥ 180	≥ 3990	2.46	54.72
735	≥ 170	≥ 4160	2.34	57.06
738	≥ 190	≥ 4350	2.62	59.67
740	≥ 150	≥ 4510	2.15	61.82
742	≥ 190	≥ 4710	2.71	64.54
744	≥ 190	≥ 4910	2.70	67.23
746	≥ 150	≥ 5060	2.18	69.41
748	≥ 170	≥ 5240	2.42	71.83
750	≥ 160	≥ 5410	2.29	74.12
752	≥ 140	≥ 5550	1.96	76.08
754	≥ 150	≥ 5700	2.05	78.13
756	≥ 140	≥ 5840	1.92	80.05
758	≥ 130	≥ 5980	1.83	81.88
761	≥ 130	≥ 6110	1.89	83.77
763	≥ 140	≥ 6260	1.96	85.73
765	≥ 110	≥ 6370	1.52	87.25
768	≥ 110	≥ 6490	1.62	88.87
770	≥ 100	≥ 6590	1.38	90.25
773	≥ 110	≥ 6700	1.59	91.84
775	≥ 90	≥ 6790	1.25	93.09
778	≥ 70	≥ 6860	0.97	94.06
781	≥ 100	≥ 6970	1.47	95.52
784	≥ 50	≥ 7030	0.74	96.26
788	≥ 70	≥ 7100	0.97	97.23
791	≥ 40	≥ 7140	0.60	97.84
795	≥ 50	≥ 7190	0.71	98.55
800	≥ 30	≥ 7220	0.42	98.97
805	≥ 20	≥ 7250	0.36	99.33
811	≥ 10	≥ 7270	0.26	99.59
818	≥ 10	≥ 7290	0.26	99.85
828	< 10	≥ 7290	NR	99.93
841	< 10	≥ 7300	NR	99.97
850	< 10	≥ 7300	NR	100.00

Frequency

Table E.9 ***Scale Score Descriptive Statistics and Plots***

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

N	≥ 7280		
Mean	730.43	Median	731.00
Std deviation	32.35	Variance	1046.62
Skewness	-0.1094	Kurtosis	-0.0093
Mode	719.00	Std Error Mean	0.3790
Range	200.00	Interquartile Range	43.00

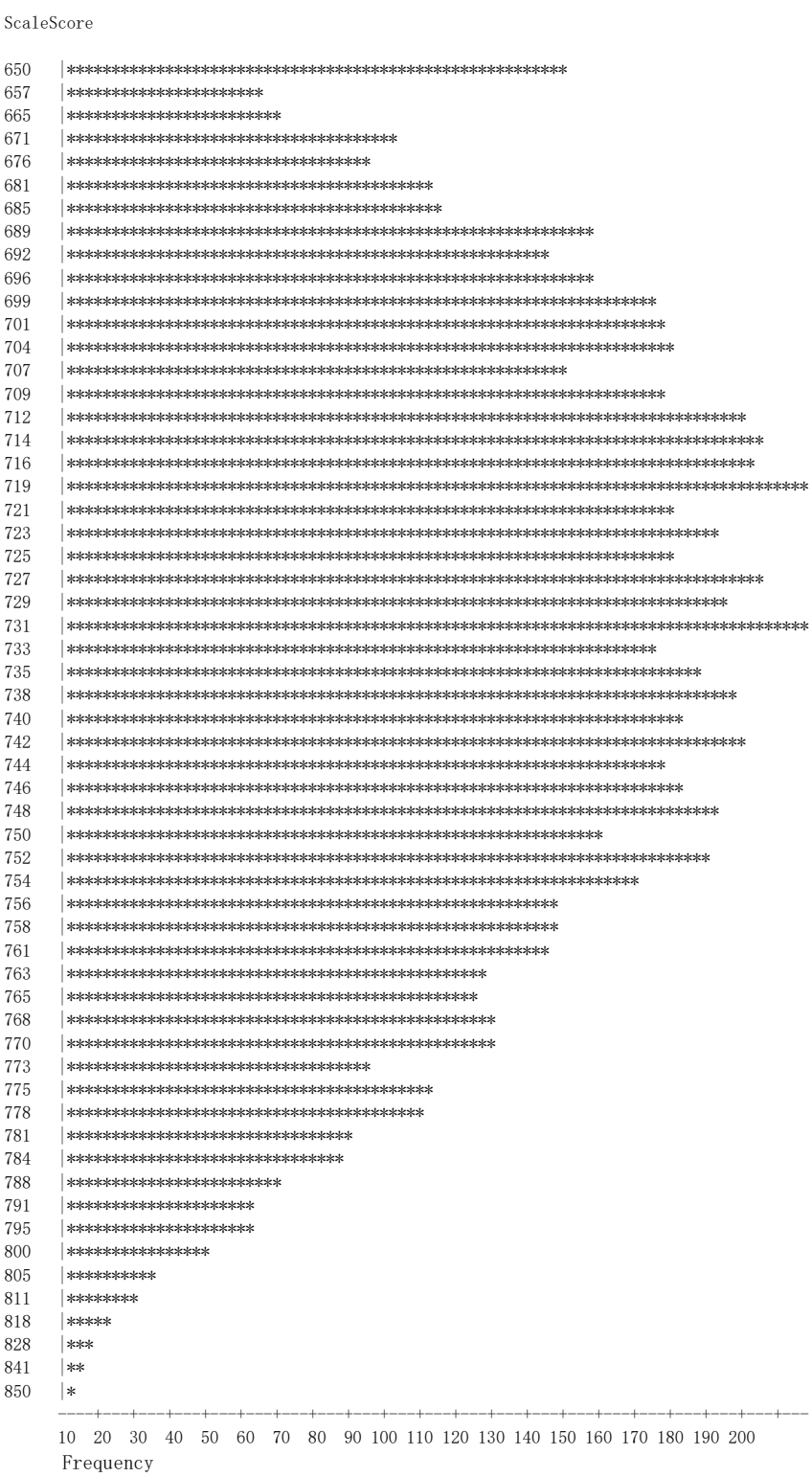
Quantile	Estimate
100% Max	850
99%	800
95%	781
90%	770
75% Q3	752
50% Median	731
25% Q1	709
10%	689
5%	676
1%	650
0% Min	650

```
                    Histogram                          #  Boxplot
855+*                                                 < 10     0
   .*                                                 < 10     0
835+
   .*                                                 < 10     0
815+**                                                ≥ 30     0
   .****                                              ≥ 60     |
795+******                                           ≥ 100     |
   .************                                     ≥ 210     |
775+**********************                           ≥ 400     |
   .**************************                        ≥ 480     |
755+*****************************************         ≥ 760  +-----+
   .**********************************************    ≥ 880  |     |
735+****************************************           ≥ 730  *--+--*
   .************************************************  ≥ 900  |     |
715+*****************************************         ≥ 780  |     |
   .**********************************                ≥ 640  +-----+
695+************************                          ≥ 440     |
   .******************                                ≥ 350     |
675+**********                                       ≥ 170     |
   .****                                              ≥ 50     |
655+***********                                      ≥ 190     |
    ----+----+----+----+----+----+----+----+----+---
    * may represent ≥ 10 counts
```

```
                    Normal Probability Plot
855+                                                  *
   |                                                  *
835+
   |                                                  *
815+                                                  *
   |                                                ****
795+                                             ****
   |                                          ****
775+                                       *****
   |                                    ****
755+                                 ****
   |                             *****
735+                           ***
   |                        ****
715+                     ****
   |                  *****
695+               ***
   |            *****
675+         ***
   |      ++**
655+*******
      +----+----+----+----+----+----+----+----+----+----+
         -2        -1         0        +1        +2
```

Table E.10 *Frequency Distribution of Scale Scores*

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



ScaleScore	Freq	Cum. Freq	Percent	Cum. Percent
650	≥ 140	≥ 140	1.94	1.94
657	≥ 50	≥ 190	0.75	2.69
665	≥ 50	≥ 250	0.81	3.50
671	≥ 90	≥ 340	1.28	4.78
676	≥ 80	≥ 430	1.18	5.96
681	≥ 100	≥ 530	1.41	7.37
685	≥ 100	≥ 640	1.44	8.81
689	≥ 140	≥ 790	2.03	10.84
692	≥ 130	≥ 920	1.87	12.71
696	≥ 140	≥ 1070	2.02	14.73
699	≥ 160	≥ 1230	2.25	16.98
701	≥ 160	≥ 1400	2.29	19.27
704	≥ 170	≥ 1570	2.33	21.61
707	≥ 140	≥ 1710	1.94	23.54
709	≥ 160	≥ 1880	2.29	25.83
712	≥ 190	≥ 2070	2.61	28.44
714	≥ 190	≥ 2260	2.68	31.12
716	≥ 190	≥ 2450	2.64	33.75
719	≥ 200	≥ 2660	2.84	36.60
721	≥ 160	≥ 2830	2.32	38.92
723	≥ 180	≥ 3010	2.51	41.43
725	≥ 170	≥ 3180	2.35	43.77
727	≥ 190	≥ 3380	2.68	46.45
729	≥ 180	≥ 3560	2.54	48.99
731	≥ 200	≥ 3770	2.84	51.83
733	≥ 160	≥ 3940	2.26	54.10
735	≥ 170	≥ 4110	2.43	56.53
738	≥ 180	≥ 4300	2.58	59.11
740	≥ 170	≥ 4470	2.37	61.48
742	≥ 180	≥ 4660	2.59	64.08
744	≥ 160	≥ 4830	2.29	66.37
746	≥ 170	≥ 5000	2.36	68.73
748	≥ 180	≥ 5180	2.50	71.23
750	≥ 150	≥ 5340	2.07	73.30
752	≥ 180	≥ 5520	2.47	75.77
754	≥ 160	≥ 5680	2.20	77.97
756	≥ 130	≥ 5810	1.89	79.86
758	≥ 130	≥ 5950	1.88	81.74
761	≥ 130	≥ 6080	1.84	83.58
763	≥ 110	≥ 6200	1.62	85.20
765	≥ 110	≥ 6320	1.59	86.79
768	≥ 120	≥ 6440	1.66	88.46
770	≥ 120	≥ 6560	1.66	90.12
773	≥ 80	≥ 6650	1.18	91.30
775	≥ 100	≥ 6750	1.41	92.71
778	≥ 90	≥ 6850	1.36	94.07
781	≥ 80	≥ 6930	1.10	95.17
784	≥ 70	≥ 7010	1.06	96.23
788	≥ 50	≥ 7060	0.81	97.04
791	≥ 50	≥ 7120	0.73	97.76
795	≥ 50	≥ 7170	0.73	98.49
800	≥ 30	≥ 7210	0.54	99.03
805	≥ 20	≥ 7230	0.34	99.37
811	≥ 10	≥ 7250	0.26	99.63
818	≥ 10	≥ 7270	0.18	99.81
828	< 10	≥ 7270	NR	99.92
841	< 10	≥ 7280	NR	99.97
850	< 10	≥ 7280	NR	100.00

10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
Frequency

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

U.S. History

Contents
Table F.1 Reliability for Overall and Subgroups
Table F.2 Cronbach Alpha and Marginal Reliability
Table F.3.1–F.3.7 Classification Accuracy and Decision Consistency

Table F.1

Reliability for Overall and Subgroups for U.S. History

Subgroup	Form B	Form C	Form D	Form E
All Students	0.929	0.923	0.925	0.926
Female	0.922	0.916	0.917	0.920
Male	0.936	0.931	0.933	0.933
African American	0.910	0.909	0.909	0.909
Asian	0.936	0.917	0.922	0.925
Hispanic/Latino	0.935	0.925	0.923	0.929
Multi-Racial	0.926	0.922	0.925	0.922
Native Hawaiian or Other Pacific Islander	0.935	0.848	0.954	0.955
White	0.925	0.916	0.919	0.920
Ethnicity Unknown	0.909	0.896	0.902	0.935
English Learners	0.868	0.885	0.914	0.907

Table F.2

Cronbach Alpha and Marginal Reliability for U.S. History

Form	Cronbach Alpha	Marginal Reliability
B	0.929	0.98
C	0.923	0.98
D	0.925	0.98
E	0.926	0.98

Table F.3 ***Classification Accuracy and Decision Consistency for U.S. History***

Table F.3.1

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

Form	Accuracy	Consistency	PChance	Kappa
B	0.726	0.633	0.232	0.523
C	0.726	0.631	0.229	0.521
D	0.723	0.629	0.232	0.517
E	0.724	0.630	0.230	0.519

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)
B	0.868	0.542	0.714	0.663	0.771
C	0.865	0.522	0.722	0.691	0.772
D	0.864	0.541	0.717	0.672	0.761
E	0.865	0.542	0.716	0.669	0.769

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
B	0.932	0.911	0.922	0.954
C	0.934	0.912	0.917	0.956
D	0.933	0.910	0.920	0.955
E	0.933	0.911	0.920	0.953

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
B	0.903	0.877	0.889	0.936
C	0.905	0.877	0.883	0.937
D	0.904	0.876	0.886	0.936
E	0.905	0.877	0.887	0.934

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
B	0.765	0.750	0.732	0.586
C	0.753	0.745	0.725	0.607
D	0.758	0.745	0.725	0.569
E	0.760	0.748	0.730	0.574

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
B	0.037	0.041	0.036	0.029
C	0.033	0.042	0.041	0.027
D	0.036	0.041	0.038	0.029
E	0.035	0.041	0.038	0.030

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
B	0.031	0.048	0.042	0.017
C	0.033	0.046	0.042	0.017
D	0.032	0.048	0.042	0.017
E	0.031	0.048	0.042	0.017