LEAP 2025 Social Studies Grades 3–8 Technical Report: 2018–2019

Prepared by DRC, Pearson, and WestEd

Foreword

Improving student achievement is a primary goal of any educational assessment program such as the Louisiana Educational Assessment Program 2025 (LEAP 2025). This technical report and its associated materials have been produced in a way that can help educators understand the technical characteristics of the assessment used to measure student achievement.

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2009) and in the new edition, *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Table of Contents

1. Introduction

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

The purpose of this Technical Report is to describe the process for the operational administration of the statewide summative social studies assessments for grades 3–8. This report outlines the testing procedures, including forms construction, administration, scoring and analyses, and reporting of scores.

Summary of the 2018–2019 Activities

WestEd and Pearson, in partnership with the LDOE and Data Recognition Corporation (DRC), the administration vendor, developed a timeline to capture the major activities necessary to produce the spring 2019 grades 3–8 operational forms with embedded field tests (EFT). Table 1.1 summarizes those key activities along with the months during which the activities were completed.

Table 1.1
Key Activities from November 2018 to September 2019

Date	Activity
November 2018	• Started item development planning for spring 2019 EFT
January–March 2018	• Item development plans approved • Content development specifications and style guide updated • LDOE conducted Source Review committees in February 2018 • WestEd began item writing and development • WestEd updated 2018–2019 Assessment Framework document
March 2018	• 2018–2019 Assessment Framework document proposed
March–June 2018	• LDOE staff reviewed proposed content
June 2018	• Item Content/Bias Review Committees convened • 2018–2019 Assessment Framework document approved; preliminary test construction activities began
July 2018	• Reconciliation meeting held between LDOE and WestEd staff
August–October 2018	• Data reviewed for spring field test and operational items 2018 results • LDOE staff reviewed proposed spring 2019 operational test selections • Biannual planning meeting held
October–November 2018	• LDOE staff reviewed proposed spring 2018 EFT selections • Initial batch of content delivered to administration vendor
November 2018	• Technical Advisory Committee Meeting convened
December 2018	• Remaining batches of content delivered to administration vendor
January 2019	• Biannual planning meeting held
March 2019	• Technical Advisory Committee Meeting convened
April–May 2019	• Spring 2019 tests administered, including EFT
September 2019	• Data reviewed to verify accuracy of spring 2019 results

2. Assessment Frameworks

The initial assessment frameworks developed at the start of the project included

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

Before the spring 2019 operational test forms were constructed, the Assessment Frameworks were updated to reflect changes to the design and field test plan, as well as to clarify the criteria used to guide item and form selection.

3. Overview of the Development Process

This section describes the processes used to develop field test item sets, tasks, and standalone items to embed within the LEAP 2025 Social Studies assessments.

Item Development Plan

WestEd's proposed item development plans may include tasks, item sets, and standalone items. In grade 4, WestEd proposed to develop two TE items as part of an item set and one standalone TE item. The period 2018–2019 is the first development cycle in which TE items have been developed in grade 4. The LDOE wanted to field test the items to observe student performance with this item type before deciding if TE items should be included as operational test items on the grade 4 assessment in subsequent administrations. Table 3.1 shows the item development plan for grades 3–8 in 2018–2019.

Table 3.1

Item Development Plan Grades 3–8

Grades 3–8 Item Development Plan		Total Tasks	Total Item Sets	Total Items per Set	MC/MS	CR	TE*	ER	Total Items
2019	Tasks	1		10	8	0	0	2	10
	Item sets		18	13-14	184	18	24	0	226
	Standalone items (MC/MS/TE)			10–15	72	0	1	0	73
	TOTALS				264	18	25	2	309

Key

MC: multiple choice MS: multiple select CR: constructed response

TE: technology enhanced ER: extended response

*The period 2018–2019 was the first development cycle in which TE items were developed in grade 4. Development included 2 one-point TE items within an item set and 1 one-point standalone TE item.

Proposal and Review of Topics and Sources

Determining Topics

When identifying possible topics, WestEd content leads consider the following:

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

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

Tasks and Item Sets. Tasks and item sets contain multiple, related stimuli that provide the content from which students answer groups of questions. Sets allow students to delve deeply into a topic, and may include items aligned to GLEs across reporting categories—allowing a set to highlight the interrelated nature of history, geography, civics, and economics—or from a subset of those categories.

Standalone Items. Standalone items assess content that may or may not be connected to a stimulus. A goal in standalone item development is to have a stimulus for 80% of the standalone items to best support students in answering the questions. With the exception of grade 4, which has standalone TE items, all standalone items are selected response (SR) items (multiple choice, multiple select). Standalone items are included in the test design to provide greater coverage of the assessable content and GLEs and to provide flexibility in meeting the blueprints and test characteristic curve targets across test administrations.

Content leads select topics for standalone items based on content and GLEs that may not be sufficiently covered across the sets, with the goal of providing maximum flexibility during test construction. Consequently, the standalone items are typically developed last.

GLE Coverage

Grade 3. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 40 assessable GLEs in grade 3. (GLEs 3.1.3, 3.1.4, 3.1.5, 3.3.5, and 3.3.6 have been identified as those that should not be addressed directly in assessment items.)

Grade 4. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 39 assessable GLEs in grade 4. (GLEs 4.1.3, 4.1.4, 4.1.7, and 4.4.7 have been identified as those that should not be addressed directly in assessment items.)

Grade 5. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 24 assessable GLEs in grade 5. (GLEs 5.1.2, 5.1.4, and 5.4.2 have been identified as those that should not be addressed directly in assessment items.)

Grade 6. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 23 assessable GLEs in grade 6. (GLEs 6.1.1, 6.1.2, 6.1.3, and 6.1.4 have been identified as those that are unlikely to be addressed directly in assessment items.)

Grade 7. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 39 assessable GLEs in grade 7. (GLEs 7.1.1, 7.1.2, 7.1.4, 7.1.5, 7.5.2, and 7.9.1 have been identified as those that are unlikely to be addressed directly in assessment items.)

Grade 8. By the end of the 2018–2019 development cycle, WestEd had developed at least 1 item aligned to each of the 30 assessable GLEs in grade 8. (GLEs 8.1.1, 8.1.2, 8.2.10, 8.3.3, and 8.10.5 have been identified as those that are unlikely to be addressed directly in assessment items.)

While some GLEs at each grade level may have only 1 item aligned to them, others have several. This variation is a result of differences in the importance and scope of content covered by individual GLEs.

Obtaining LDOE Approval for Topics

For tasks and item sets, WestEd submits lists of proposed topics at each grade level to the LDOE for review prior to item development. These lists describe the topics and possible related stimuli so that the LDOE can review and approve them simultaneously. The proposed topic lists also include the GLEs and reporting categories that might be assessed by the tasks and item sets. Once the LDOE approves the topics to be developed for the development cycle, stimulus-searching and development of the task and item set overviews begin.

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

Identifying Stimuli

The LEAP 2025 Social Studies assessments focus on the use of authentic historical and contemporary documents, including maps, letters, journal entries, speeches, photographs, paintings, reports, and other primary source documents. The assessments also include secondary source documents, such as authentic newspaper articles and book excerpts. These documents are supplemented by timelines, tables, charts, and graphic organizers created by WestEd's Design Team. On rare occasions, a stimulus, such as a newspaper article or a scenario, is written by WestEd editors to meet a specific assessment purpose.

Both internal and external editors locate appropriate stimuli for tasks, item sets, and standalone items. Before the stimuli searchers begin, WestEd trains them on the search process, on the LDOE's objectives, and on best practices, including bias and sensitivity training. For an outline of the training, see LEAP 2025 Social Studies Grades 3–8 Stimulus Search Training Agenda (2018–2019) in Appendix A.

All stimuli are submitted to WestEd for evaluation for alignment and appropriateness for the approved topics. Based on this evaluation, the WestEd content leads select the final sources to propose to the LDOE.

Public Domain versus Permissioned Work. WestEd endeavors to maintain a ratio of 80% royalty-free stimuli from the public domain or created internally to a maximum of 20% permissioned work. The actual percentages of permissioned work for the 2018–2019 development cycle exceeded the target of 20% at grades 3 and 8 where that value was 43% and 30%, respectively. Across all grades, the total percentage of permissioned work was 23%. Before administration of the assessment, WestEd's permissions coordinator obtains permissions from the rights holders for five years of use of any work that was not in the public domain or created internally.

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

Most of the stimuli chosen as part of the 2018–2019 development cycle were found to be at grade level; however, some of the authentic historical documents were evaluated as being above grade level. In those cases, footnotes were added for words that were above grade level and for words or phrases that were thought to be a potential source of confusion for students. If an authentic historical document required many footnotes or if the language was considered too arcane or incomprehensible for students at the given grade level, the document was modified to improve readability and accessibility. These

modifications were made evident by use of the phrase “Adapted from” in the title of the document. After modification, the stimuli were re-evaluated to ensure that the changes resulted in the desired outcomes.

Obtaining LDOE Approval for Tasks, Item Sets, and Stimuli

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

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

Item Writing and Review Process

WestEd employs item writers and editors for grades 3–8. Some of the WestEd writers have been part of item development since the first development cycle in 2015–2016. WestEd secures the required approval from the LDOE for each writer during their first development cycle. Writers and editors receive training from WestEd that outlines lessons learned from previous development cycles, LDOE expectations, and best practices for item development, including consideration of bias and sensitivity. For an outline of the information covered at the 2018–2019 training, see Appendix A for the LEAP 2025 Social Studies Grades 3–8 Item Writer and Editor Training Agenda. After the training, item writers are provided with approved set overviews, which identify the set topics, list the GLEs to be addressed, specify the number and type of items to be written, and offer specific guidance to the item writer about how the content for each item within a set should be assessed. The use of the set overviews allows WestEd to control the quality of the task and item sets.

Once written, items go through two rounds of content editing, one round of proofreading, and a final round of review before being submitted to the LDOE for their first round of review. The LDOE has two rounds of review prior to content and bias review committee meetings.

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

Style Guidelines. The *LEAP 2025 Social Studies Content Style Guide* is updated immediately following test construction to reflect final formatting decisions made by the LDOE. Throughout the development and review process, when questions of style arise that are unanswered by existing documentation, WestEd consults the LDOE, and approved changes are added to the Style Guide.

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

Content and Bias Review Committees. After the completion of item development and the initial rounds of LDOE review, virtual content and bias review meetings are held. The LDOE recruits educators from different parts of Louisiana, who represent all Louisiana students, to serve on the committees. The meetings are led jointly by facilitators from the LDOE and WestEd. Table 3.2 provides information about the representation of educators who participated in the content and bias reviews in June 2018.

Table 3.2

Representation of Educators Participating in June 2018 Content and Bias Reviews

Grade	Number of Committee Participants‡	Classroom Teacher	Special Education	Instructional Lead or Supervisor	Visually Impaired Teacher	EL Teacher/ Supervisor
3	8	6	1	3	1	1
4	11	7	-	5	2	1
5	7	5	1	3	1	2
6	9	6	1	3	1	1
7	7	7	2	-	1	-
8	8	5	1	4	1	1

‡The number of committee participants is lower than the sum of the educational roles of participants at each grade because several participants listed multiple roles.

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

Committee members are provided a pre-review day during which they access the items using ABBI and vote on the items. Comments are compiled and shared with LDOE and WestEd facilitators prior to the joint virtual committee review. When the committee convenes as a group, the committee members revisit and discuss items and stimuli. A WestEd recorder takes detailed notes about discussions and records the final committee recommendations. These notes are compiled for reconciliation with the LDOE and post-review implementation. Access to the items is tightly controlled by WestEd, with password

access shutting off immediately following the close of each pre-review and review section. At the close of each session, committee members are instructed to clear their internet browser history. In addition, all participants complete a nondisclosure agreement prior to accessing any items.

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

Table 3.3

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

Grade	Number of Items and Stimuli	Accept	Accept with Edits	No Vote	Reject	Grand Total
3	55	361	65*	0	11	437
4	62	468	71	0	16	555
5	61	305	42*	9	8	364
6	61	324	27*	5	16	372
7	63	363	68*	1	7	439
8	63	324	27*	5	16	372

*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 Items and Stimuli	Accept	Accept with Edits	No Vote	Reject
3	55	28	27	0	0
4	62	42	20	0	0
5	61	34	27	0	0
6	61	32	29	0	0
7	63	25	38	0	0
8	63	35	28	0	0

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

4. Construction of Test Forms

Initial Construction

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

Spring 2019 Operational Forms

Items approved during data review from the spring 2018 embedded field tests were available for use on the spring 2019 operational assessments. (See the 2017–2018 Technical Report for results from the data review and reconciliation of the spring 2018 field test items.) Items approved during data review from previous years were also available for use on the spring 2019 operational assessments.

Prior to test form construction for spring 2019, WestEd, Pearson, and the LDOE discussed the test designs for grades 3–8. In grades 3 and 4, a decision was reached to eliminate the task from the assessments and reduce the number of sessions from three to two sessions. In grades 5–8, the overall number of points was reduced by two points in each grade. Table 4.1 provides a comparison of the 2018 and 2019 tests for each grade.

Table 4.1

LEAP Social Studies Grades 3–8 2018 and 2019 Test Information

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	2018/2019	2018/2019	2018/2019	2018/2019	2018/2019	2018/2019
Overall OP Points	52/45	57/45	60/58	68/66	68/66	68/66
OP Points Excluding Task	45/45*	46/45*	49/46	56/54	56/54	56/54
Overall OP Items	47/43	48/43	47/46	55/53	55/53	55/53
FT Items per Form	5/7	5/7	5/6	6/5-8	6/5-8	6/5-8
OP Sessions	3/2	3/2	3/3	3/3	3/3	3/3

*Does not have a task.

WestEd completed item selection for one operational form per grade for consideration by Pearson psychometricians before submission to the LDOE. The operational forms were designed to adhere to the blueprint for the tested grade and exhibit the broadest possible balance of content and breadth of GLE coverage. The task was selected first for grades 5–8, followed by item sets with CRs, other item sets, and standalone items. For grades 3–4, item sets with CRs were selected first, followed by other item sets and standalone items. Test-form developers worked to avoid cueing and clanging between items. Cueing occurs when content in one item provides clues to the answer of another item. Clanging refers to overlap or similarity of content. Because content was purposely distributed across sessions, cueing and clanging were intended to have been avoided; however, developers also conducted a separate review of the forms to check for inadvertent cueing and clanging. During item selection, test maps were created to capture details of the forms, including each item's unique identification number (UIN), test session, item sequence, item descriptions, and associated item metadata. Tables 4.2–4.7 provide the operational test composition for the grades 3–8 spring 2019 forms.

Table 4.2

Grade 3 Social Studies Operational Test Composition for 2019

Item Sets/Item Types	Total Sets	Total Points per Set	SR	CR	Total Items	Total Points
7-Item Set	1	7	7		7	7
5-Item Set with CR	2	6	8	2	10	12
5-Item Set	1	5	5		5	5
4-Item Set	1	4	4		4	4
Standalone Items	0	0	17		17	17
Total Items			41	2	43	45

Table 4.3

Grade 4 Social Studies Operational Test Composition for 2019

Item Sets/Item Types	Total Sets	Total Points per Set	SR	CR	Total Items	Total Points
6-Item Set	3	6	18		18	18
5-Item Set with CR	2	6	8	2	10	12
Standalone Items	0	0	15		15	15
Total Items			41	2	43	45

Table 4.4

Grade 5 Social Studies Operational Test Composition for 2019

Item Sets/Item Types	Total Sets	Total Points per Set	SR	CR	TE	ER	Total Items	Total Points
6-Item Set with TE	2	7	10		1		12	14
5-Item Set with CR	2	6	8	2			10	12
5-Item Set with TE	1	6	4		1		5	6
Standalone Items	0	0	14				14	14
5-Item Task	1	12	4			1	5	12
Total Items			35	2	2	1	46	58

Table 4.5

Grade 6 Social Studies Operational Test Composition for 2019

Sets and Standalone Items	Total Sets	Total Points per Set	SR	CR	TE	ER	Total Items	Total Points
6-Item Set with TE	2	7	10		2		12	14
6-Item Set with CR	2	7	10	2			12	14
5-Item Set with TE	2	6	8		2		10	12
Standalone Items	0	0	14				14	14
5-Item Task	1	12	4			1	5	12
Total Items			46	2	4	1	53	66

Table 4.6

Grade 7 Social Studies Operational Test Composition for 2019

Sets and Standalone Items	Total Sets	Total Points per Set	SR	CR	TE	ER	Total Items	Total Points
6-Item Set with TE	3	7	15		3		18	21
6-Item Set with CR	1	7	5	1			6	7
5-Item Set with TE	1	6	4		1		5	6
5-Item Set with CR	1	6	4	1			5	6
Standalone Items	0	0	14				14	14
5-Item Task	1	12	4			1	5	12
Total Items			46	2	4	1	53	66

Table 4.7

Grade 8 Social Studies Operational Test Composition for 2019

Sets and Standalone Items	Total Sets	Total Points per Set	SR	CR	TE	ER	Total Items	Total Points
6-Item Set with TE	2	7	10		2		12	14
6-Item Set with CR	2	7	10	2			12	14
5-Item Set with TE	2	6	8		2		10	12
Standalone Items	0	0	14				14	14
5-Item Task	1	12	4			1	5	12
Total Items			42	2	4	1	53	66

Spring 2019 Field Test Forms

For grades 3 and 4, item sets and standalone items developed in 2018 were field tested in spring 2019. For grades 5–8, a combination of task sets developed in 2017 and item sets and standalone items developed in 2018 were field tested in spring 2019. Sets were placed on multiple field test forms, with a different combination of items on each form, to ensure field testing of the maximum number of items in each set. Nine field test forms were administered in each grade for grades 3–8. During item placement by WestEd content leads, test maps also captured details about the field test forms.

Field test items were embedded in Session 2 for grades 3 and 4 and part of a fourth session for grades 5–8. All students in grades 3 and 4 took field test items, and a sample of students in grades 5–8 participated in the field test session.

Field test forms for the spring 2019 administration were organized as follows.

Grade 3

- Nine forms had one 5-item set with CR and two standalone items (in Session 2)

Grade 4

- Nine forms had one 5-item set with CR and two standalone items (in Session 2)
- One form had one 5-item set with two TEI and one standalone SR item and one standalone TE item (in Session 2)

Grade 5

- Six forms had one task set with four SR and 1 ER item (in Session 4)
- Two forms had one 5-item set with TEI and three standalone items (in Session 4)
- One form had one 5-item set with CR and three standalone items (in Session 4)

Grade 6

- Six forms had one task set with four SR and 1 ER (in Session 4)
- Two forms had one 6-item set with TE and two standalone items (in Session 4)
- One form had one 6-item set with CR and two standalone items (in Session 4)

Grade 7

- Six forms had one task set with four SR items and 1 ER (in Session 4)
- Two forms had one 6-item set with TE and two standalone items (in Session 4)
- One form had one 6-item set with CR and two standalone items (in Session 4)

Grade 8

- Six forms had one task set with four SR items and 1 ER (in Session 4)
- Two forms had one 6-item set with TE and two standalone items (in Session 4)
- One form had one 6-item set with CR and two standalone items (in Session 4)

In grades 3 and 4, standalone items were repeated on field test forms as necessary to fill all available positions.

Revision and Review

Psychometric Approval of Operational Forms

Prior to submitting the forms to LDOE staff for review, Pearson psychometricians and WestEd content specialists participate in an iterative process of reviewing and revising the forms. The psychometric review consists of comparisons of the expected representation and the actual representation of reporting categories (History, Geography, Civics, Economics) and item types—selected response (SR), constructed response (CR), technology enhanced (TE), and extended response (ER)—on the operational forms.

The answer keys for multiple-choice (MC) items also are examined, to determine whether any forms have significantly non-uniform distributions of correct responses (A, B, C, and D). Spreadsheets are used to generate frequency tables of reporting categories, item types, and MC answer keys for each form. They are also used to compare to operational forms from previous years for each grade. Deviations from the blueprint are identified and addressed. Test characteristic curves (TCC) based on item response theoretic models are applied to the data, and conditional standard errors of measurement are computed for each iteration during the test construction process to evaluate how well a proposed test form matches psychometric targets. Psychometric approval from Pearson is provided

for all forms prior to submission to the LDOE for their review. Please refer to the following table for criteria to flag items based on scoring point.

Table 4.8
Summary of Flagging Criteria to Select/Flag Items: Classical Analysis and IRT

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

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

LDOE Review

Following the psychometric reviews, the test maps and constructed sets for each grade are delivered to the LDOE for approval. Forms are reviewed by both LDOE content and psychometric staff. Based on the LDOE review, sets or items are replaced and resequenced as requested. After these changes, the overall balance of answer choices and key runs is re-evaluated, and final adjustments are made to achieve the appropriate balance. Pearson also updates the TCC and SEM curves to be sure that the selection meets psychometric targets.

Finalized test maps are used to create PDF versions, or constructed sets, of the forms, which are reviewed by WestEd's proofreaders before the items are transferred from ABBI to DRC.

Online and Paper Versions

At grades 3 and 4, forms are mainly delivered on paper, and at least one of the forms at each grade is identified for delivery online. At grades 5–8, all forms are delivered online.

One form in each grade is designated by the LDOE as the accommodated version, to be used with students who have accommodations. For grades 5–8, the accommodated version is available in print form to students who require paper testing. The accommodated version is also rendered in braille. To support students with low or no vision, additional text (alternate text) is provided to describe the graphic components of the assessments. Content specialists evaluate the graphics and draft the alternate text. Table 4.8 shows the number of online and paper forms for each grade in spring 2019.

Table 4.8
Numbers of LEAP 2025 Forms for Spring 2019 Operational and Embedded Field Test

Grade	Paper Forms	Online Forms
3	9	1*
4	9	1**
5	1	9
6	1	9
7	1	9
8	1	9

*Same form as one of the paper forms.
**A form 1A was created as an online form with TE companion forms as the regular and online accommodated form.

Whenever possible, the comparability between the primary test mode at a grade level (paper for grades 3 and 4; online for grades 5–8) and the secondary mode (online for grades 3 and 4; accommodated forms for grades 5–8) is evaluated empirically. At grades 3 and 4, for example, assessment results are separately analyzed by paper and by online versions. At grades 5–8, a historical limitation for the same types of analysis exists because of the lack of examinees who take an accommodated version of the test. Comparability between online and accommodated versions of tests at grades 5–8 is defined by comprehensive content evaluations.

Since two modes were administered for grades 3 and 4, the following techniques (i.e., mode effect analysis and equating) were applied to operational test data to investigate item mode effect. It should be noted that the CBT sample size for grade 3 is less than 10% of the population, Pearson did not conduct mode-effect analysis.

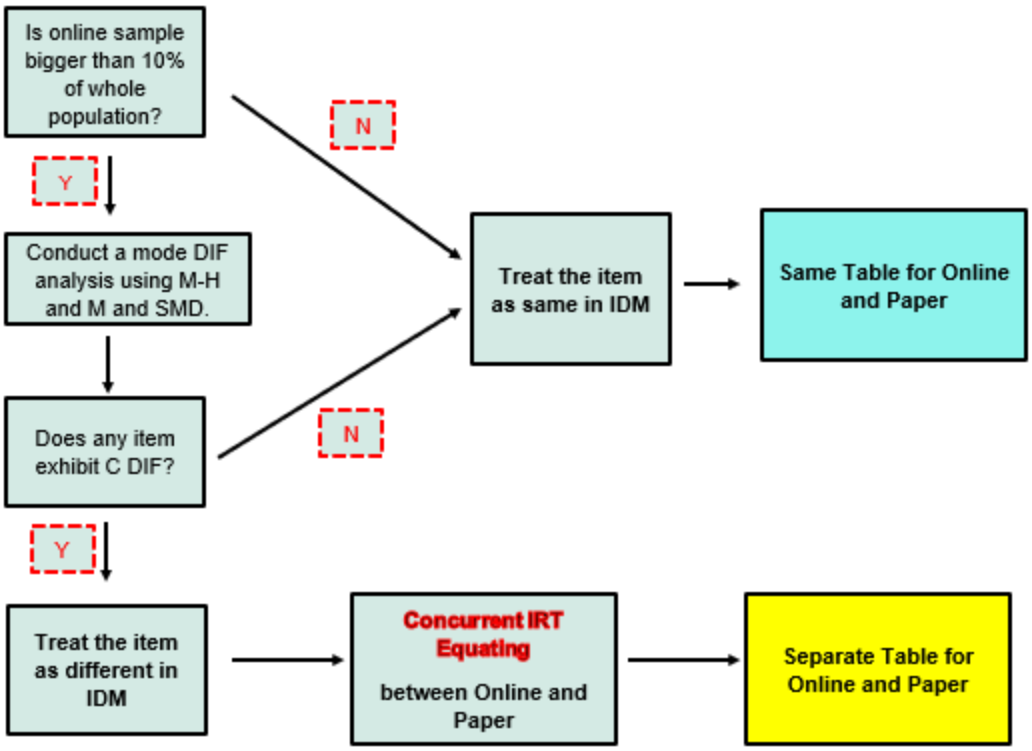


Figure 4.1 General overview of equating, including mode-effect analysis

5. Test Administration

This chapter describes processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) *Standards for Educational and Psychological Testing*, "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 Education Assessment Program 2025 (LEAP 2025) 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 the LEAP 2025 assessments are administered and scored in accordance with the department's policies, the LDOE takes a primary role in communicating with and training school system personnel. The LDOE provides train-the-trainer opportunities for the district test coordinators, who in turn convey test administration training to schools within their school systems. The LDOE conducts quality-assurance visits during testing to ensure adherence to the standardized administration of the tests.

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

Ancillary Materials

Ancillary materials for LEAP 2025 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 test administration, DRC produces two administration manuals:

1. *LEAP 2025 Grades 3–4 Paper-Based Test Administration Manual*
2. *LEAP 2025 Grades 3–8 Computer-Based Test Administration Manual*

DRC also produces Test Coordinators Manuals for paper-based test administration and for computer-based test administration. LDOE assessment staff review, provide feedback, and give final approval for these manuals. The Test Coordinators Manuals are inclusive of grades 3–8 English Language Arts (ELA), Mathematics, Social Studies, and Science. They provide detailed instructions for district and school test coordinators' responsibilities for distributing, collecting, and returning test materials to DRC for scoring.

Table of Contents for Paper-Based Testing Test Coordinators Manual

- Key Dates
- Spring 2019 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
- Test Security
    - Key Definitions
    - Violations of Test Security
    - Answer Change Analysis
    - Voiding Student Tests
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions

    - Testing in Class-sized Groups
    - Test Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Makeup Testing
    - Test Administration Resources
    - Testing Times for Grades 3 and 4
- District Test Coordinator
    - Conduct Training Session
    - Receive Test Materials
    - Large-print Braille and CAS Test Materials
    - Accommodated Materials
    - Verify and Distribute Test Materials to School Test Coordinators
    - Request Additional Test Materials and Bar-code Labels
    - Collect Materials from Schools After Testing
    - Used and Unused Consumable Test Booklets (Defined)
    - Unscorable Documents and Unscorable Document Labels
- Directions for Returning Test Materials to DRC in May
    - Pickup 1
    - Pickup 2
    - Pickup 3
    - Final Checklist for Returning Test Materials to DRC
- School Test Coordinator
    - Receive and Verify Test Materials
    - Conduct Test Administration and Security Training Session
    - Supervise Application of Bar-code Labels and Coding of Consumable Test Booklets
    - Soiled, Damaged, and Other Unscorable Answer Documents and Consumable Test Booklets

    - Verify and Distribute Materials to Test Administrators
    - Supervise Test Administration
    - Collect Test Materials
    - Used and Unused Consumable Test Booklets (Defined)
    - Coding Responsibilities of Principals—Before Testing
    - Coding Responsibilities of Principals—Before or After Testing
    - Coding Responsibilities of Principals—After Testing
- Directions for Returning Test Materials to the DTC
    - Pickup 1
    - Pickup 2
    - Pickup 3: Nonscorable Test Materials
    - Final Checklist for Returning Test Materials to the DTC
- Void Form
- Index

Table of Contents for Computer-Based Testing Test Coordinators Manual

- Key Dates Spring 2019
- Resources Available in eDIRECT Spring 2019
- Spring 2019 Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - eDIRECT and INSIGHT
- LEAP 2025
- Test Security
    - Key Definitions
    - Violations of Test Security
- Testing Guidelines
    - Testing Eligibility

    - Testing Conditions
    - Testing in Class-sized Groups
    - Testing Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grades 3 through 8
- Roles and Responsibilities
    - District Test Coordinator
    - School Test Coordinator
    - Technology Coordinator
- Managing Test Tickets
    - Student Transfers
    - Locked Test Tickets
    - Technical Issues
    - Invalidating Test Tickets
- Resources for Online Testing
    - Test Administration Manuals
    - *eDIRECT User Guides*
    - *LEAP 2025 Accommodations and Accessibility Features User Guide*
    - *INSIGHT Technology User Guide*
    - Online Tools Training (OTT)
    - Student Tutorials

The test administration manuals provide detailed instructions for administering the LEAP 2025 assessments. The manuals include instructions for test security, test administrator responsibilities, test preparation, administration of tests (online or paper), and post-test procedures. Information included in the test administration manuals is listed below.

Table of Contents for LEAP 2025 Test Administration Manual (PBT)

- Spring 2019 Notes and Reminders
- Test Administrator Pre-Administration Oath of Security and Confidentiality Statement
- Test Administrator Post-Administration Oath of Security and Confidentiality Statement
- Overview
- Test Security
    - Secure Test Materials
    - Testing Irregularities and Security Breaches
    - Testing Environment
    - Violations of Test Security
    - Answer Change Analysis
    - Voiding Student Tests
- Test Administrator Responsibilities
- Test Administration Checklists
    - Before Testing
    - During Testing
    - After Testing (Daily)
    - After Testing (Last Day)
- Test Administrators' Frequently Asked Questions
- Test Materials
    - Receipt of Test Materials
- Testing Guidelines
    - Testing Eligibility
    - Test Schedule
    - Extended Time for Testing
- Testing Times for Grades 3 and 4
    - Makeup Testing

    - Testing Conditions
- Special Populations and Accommodations
    - IDEA Special Education Students
    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard-of-Hearing Students
    - English Learner (ELs)
- Hand-coded Consumable Test Booklets
- Students Absent from Testing
- Consumable Test Booklet Coding
    - Coding the Demographic Section
- Sample Grade 3 English Language Arts Consumable Test Booklet
- General Instructions for LEAP 2025
    - Student Marking/Erasing on Consumable Test Booklet
    - Reading Directions to Students
    - Special Instructions
- Directions for Administering LEAP 2025
- Post-Test Procedures
    - Test Administrator Oath of Security and Confidentiality Statement
    - Used and Unused Consumable Test Booklets (Defined)
    - Transferring Student Responses
    - Returning Test Materials to the School Test Coordinator
- Index

Table of Contents for LEAP 2025 Test Administration Manual (CBT)

- Spring 2019 Notes and Reminders
- Test Administrator Pre-Administration Oath of Security and Confidentiality Statement (CBT)

- Test Administrator Post-Administration Oath of Security and Confidentiality Statement (CBT)
- Overview
- Test Security
    - Secure Test Materials
    - Testing Irregularities and Security Breaches
    - Testing Environment
    - Violations of Test Security
    - Voiding Student Tests
- Test Administrator Responsibilities
    - Software Tools and Features for Test Administrators
- Test Administration Checklists
    - Before Testing
    - During Testing
    - After Testing (Daily)
    - After Testing (Last Day)
- Test Administrators' Frequently Asked Questions
- Testing Guidelines
    - Testing Eligibility
    - Testing Schedule
    - Extended Time for Testing
- Testing Times for Grades 3 through 8
    - Makeup Testing
    - Testing Conditions
- Online Tools Training
- Student Tutorials
- Directions for Administering the Grades 3–8 LEAP 2025 Tests
- Special Populations and Accommodations
    - IDEA Special Education Students

    - Students with One or More Disabilities According to Section 504
    - Gifted and Talented Special Education Students
    - Test Accommodations for Special Education and Section 504 Students
    - Special Considerations for Deaf and Hard-of-Hearing Students
    - English Learners (ELs)
- Students Absent from Testing
- Test Materials
    - Receipt of Test Materials
- General Instructions
    - Reading Directions to Students
- Post-Test Procedures
    - Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    - Returning Test Materials to the School Test Coordinator
- Index

The *Standards* contain multiple references relevant to test administration. Information in the LEAP 2025 test administration manuals addresses these in the following manner.

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

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

The LEAP 2025 test administration manuals provide instructions for activities that happen before, during, and after testing with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the test administration manuals describe

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

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

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

Test administration manuals outline the steps that teachers should take to prepare the classroom testing environment for administering the LEAP 2025 online test. These include the following.

- Determine the layout of the classroom environment.

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

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

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

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

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

Return Material Forms and Guidelines

The Test Coordinators Manual instructs test coordinators regarding procedures for organizing and packing materials and returning them to DRC for secure inventory purposes. LDOE assessment staff have opportunities to review, provide feedback, and give final approval of the guidelines. The purpose of the instructions is to ensure that

secure test materials are properly accounted for and organized appropriately for return shipment.

Security Checklists

As soon as printed test materials are received by a school system, the district test coordinator ensures that the first and last security barcodes on the tests match the packing list they received. The district test coordinator then packages the tests to be sent to schools. Upon returning test books to DRC, school and system test coordinators are required to complete and submit an accountability form that details the number of test books or printed test forms returned. This form also requires that systems/schools document nonstandard situations, including lost, damaged, destroyed, extra, or missing test books.

Interpretive Guides

Essential to making valid interpretations of test scores is an understanding of what the test scores mean and how to interpret score reports. The Interpretive Guide is written for Louisiana teachers and administrators who receive the LEAP 2025 score reports. https://www.louisianabelieves.com/resources/library/assessment-guidance

Time

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

Online Forms Administration, Grades 3–8

The online forms are administered via DRC's INSIGHT online assessment system. School system and school personnel set up test sessions via DRC's online testing portal, eDIRECT, and print test tickets. Students enter their ticket information to access the test in INSIGHT. In addition, students have access to Online Tools Training before the testing window, which allows them to practice using tools and features within INSIGHT. Tutorials with

online video clips that demonstrate features of the system are also available to students before testing.

Paper-Based Forms Administration, Grades 3 and 4

Schools with testers at grades 3 and 4 have the option to participate in either paper-based or computer-based testing. DRC prints and ships paper materials to the sites that choose paper-based testing. These materials are returned to DRC after testing, for processing and scoring with the online tests.

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

Online testing for grades 3–8 was available from Monday, April 1, through Friday, May 3, 2019. Paper-based testing occurred from Monday, April 29, through Friday, May 3, 2019.

Test Security Procedures

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors, such as unauthorized exposure of test items that would affect the valid interpretation of test scores. Several test security measures are implemented for the LEAP 2025. Test security procedures are discussed throughout the test coordinators manual and test administration manuals.

Test coordinators and administrators are instructed to keep all test materials in locked storage, except during actual test administration, and access to secure materials must be restricted to authorized individuals only (e.g., test administrators and the school test coordinator). During the testing sessions, test administrators are directly responsible for the security of the LEAP 2025 and must account for all test materials and supervise the test administrators at all times.

Data Forensic Analyses

Due to the importance of the LEAP 2025 assessment, it is prudent to ensure that the results from the assessments are based on effective instruction and true student achievement. While there are many ways to achieve meaningful understanding of student knowledge via test scores, there are also ways to obtain higher test scores that are not related to actual learning. To assist ensuring that assessment results are valid, data forensic analyses are conducted to help separate meaningful gains from spurious gains. It is important to note that although the results may be used to identify potential problems within a school, the identification of a problem is not an accusation of misconduct. Multiple methods were incorporated into the forensic analysis. The following methods were applied:

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

Response Change Analysis

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

Score Fluctuation Analysis

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

Web Monitoring

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

Plagiarism Detection

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

6. Scoring Activities

Answer Key Verification

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

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

Selected-Response (SR) Item Keycheck. Scoring of SR items is evaluated with TRIAN, a standardized Pearson program that calculates MC item statistics, to verify that MC items are keyed correctly (i.e., that the true correct response is applied during scoring). Items are flagged if item statistics fall outside expected ranges. For example, items are flagged if few students select the correct response (*p*-value less than 0.15), if the item does not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) select a certain incorrect response. Lists of flagged items, with the reasons for flagging, are provided to WestEd content staff for key verification. WestEd staff review the list of flagged MC and MS items to confirm that the answer keys are accurate. Scoring of MS items is also evaluated at data review.

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

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

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

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

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

Constructed-response items were scored by human raters who were trained by DRC. Handscoring and Artificial Intelligence (AI) processing rules are detailed in the *LEAP 2025 Spring 2019 Handscoring/AI Documentation* document.

Selection of Scoring Evaluators. Standard 4.20 states the following:

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

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

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

DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, or recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to demonstrate their proficiency in writing by responding to a DRC writing topic and their proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all

readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and demonstrate the ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

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

Handscoring Training Process. Standard 6.9 specifies:

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

Training Material Development. DRC scoring supervisors trained scorers using training materials developed by DRC in conjunction with and approved by the LDOE. These materials were made for use with WestEd-developed items.

- Prompts
- Rubrics
- Anchor sets
- Practice sets
- Qualifying sets

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

Qualifying Standards. Scorers demonstrated their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with true scores on qualifying sets). After each qualifying set was scored, the DRC scoring director responsible for training led the scorers in a discussion of the set.
Any scorer who did not qualify by the end of the qualifying process for an item was not allowed to score live student responses.

Monitoring the Scoring Process. Standard 6.8 states:

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

Reader Monitoring Procedures. Throughout the handscoring process, DRC project managers, scoring directors, and team leaders reviewed the statistics that were generated daily. DRC used one team leader for every 10 to 12 readers. If scoring concerns were apparent among individual scorers, team leaders dealt with those issues on an individual basis. If a scorer appeared to need clarification of the scoring rules, DRC supervisors typically monitored one out of five of the scorer's readings, making adjustments to that ratio as needed. If a supervisor disagreed with a reader's scores during monitoring, the supervisor provided retraining in the form of direct feedback to the reader, using rubric language and applicable training responses.

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

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

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

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

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

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

Calibration Sets. DRC pulled calibration responses for items. DRC used these sets to perform calibration across the entire scorer population for an item if trends were detected (e.g., low agreement between certain score points if a certain type of response was missing from initial training). These calibrations were designed to help refocus scorers on how to properly use the scoring guidelines. They were selected to help

illustrate particular points and familiarize scorers with the types of responses commonly seen during operational scoring. After readers scored a calibration set, the scoring director reviewed it from the front of the room, using rubric language and scoring concepts exemplified by the anchor responses to explain the reasoning behind each response's score.

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

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

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

Table 6.1

Operational Constructed-Response Inter-Rater Reliability

Grade	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
3	Grade3_Item1	≥12,480	86	14	0
	Grade3_Item2	≥12,800	93	7	0
4	Grade4_Item1	≥12,630	81	19	0
	Grade4_Item2	≥10,790	92	8	0
5	Grade5_Item1	≥23,720	92	8	0
	Grade5_Item2	≥17,220	88	12	0
6	Grade6_Item1	≥12,130	90	10	0
	Grade6_Item2	≥16,480	88	12	0
7	Grade7_Item1	≥14,880	87	13	0
	Grade7_Item2	≥17,540	88	11	0
8	Grade8_Item1	≥14,750	83	17	0
	Grade8_Item2	≥12,610	80	20	0

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

Table 6.2

Operational Constructed-Response Score Point Distributions

Grade	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
3	Grade3_Item1	≥59,270	56	29	10	3
	Grade3_Item2	≥59,450	62	12	19	2
4	Grade4_Item1	≥61,160	34	35	26	2
	Grade4_Item2	≥60,280	79	15	3	1
5	Grade5_Item1	≥66,340	41	34	7	0
	Grade5_Item2	≥63,110	37	41	10	0
6	Grade6_Item1	≥60,630	55	35	5	0
	Grade6_Item2	≥62,760	27	51	11	0
7	Grade7_Item1	≥59,200	46	35	9	0
	Grade7_Item2	≥60,560	39	28	19	0
8	Grade8_Item1	≥57,430	35	39	19	0
	Grade8_Item2	≥56,380	24	53	17	0

Table 6.3

Operational Extended-Response Inter-Rater Reliability

Grade		Inter-Rater Reliability			
	2x	Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
5	≥53,370	Content	92	8	0
		Claim	92	7	0
6	≥39,810	Content	91	8	0
		Claim	92	8	0
7	≥54,930	Content	94	5	0
		Claim	94	6	0
8	≥50,970	Content	92	8	0
		Claim	92	8	0

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

Table 6.4

Operational Extended-Response Score Point Distributions

Grade	Score Point Distribution							
	Total	Dimension	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent "3" Rating	Percent "4" Rating	Percent Blank
5	≥81,270	Content	39	31	15	3	0	0
		Claim	44	28	13	3	0	0
6	≥74,480	Content	36	36	13	3	0	0
		Claim	46	31	9	2	0	0
7	≥79,240	Content	34	35	12	4	1	0
		Claim	40	29	11	4	1	0
8	≥75,540	Content	17	32	27	10	4	0
		Claim	17	29	28	10	5	0

Table 6.5

Field Test Constructed-Response Inter-Rater Reliability

Grade	Item	Inter-Rater Reliability			
		2x	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
3	Grade3_Item1	≥400	90	10	0
	Grade3_Item2	≥430	95	5	0
	Grade3_Item3	≥470	99	1	0
4	Grade4_Item1	≥580	98	2	0
	Grade4_Item2	≥490	100	0	0
	Grade4_Item3	≥370	95	5	0
5	Grade5_Item1	≥440	93	7	0
6	Grade6_Item1	≥420	90	10	0
7	Grade7_Item1	≥490	92	8	0
8	Grade8_Item1	≥450	85	15	0

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

Table 6.6

Field Test Constructed-Response Score Point Distributions

Grade	Item	Score Point Distribution				
		Total	Percent "0" Rating	Percent "1" Rating	Percent "2" Rating	Percent Blank
3	Grade3_Item1	≥1,700	49	34	11	2
	Grade3_Item2	≥1,710	50	25	16	3
	Grade3_Item3	≥1,730	71	1	2	2
4	Grade4_Item1	≥1,790	73	9	1	2
	Grade4_Item2	≥1,740	77	7	2	2
	Grade4_Item3	≥1,680	61	26	6	3
5	Grade5_Item1	≥1,720	61	19	2	0
6	Grade6_Item1	≥1,710	62	25	5	0
7	Grade7_Item1	≥1,740	41	37	10	0
8	Grade8_Item1	≥1,720	26	48	16	0

Table 6.7

Field Test Extended-Response Inter-Rater Reliability

Grade	Item	2x	Inter-Rater Reliability			
			Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
5	Grade5_Item1	≥3,210	Content	83	17	0
			Claim	82	18	0
	Grade5_Item2	≥3,230	Content	80	19	1
			Claim	80	19	1
	Grade5_Item3	≥3,210	Content	88	12	0
			Claim	85	15	0
	Grade5_Item4	≥3,250	Content	80	19	1
			Claim	81	19	1
	Grade5_Item5	≥3,230	Content	80	20	0
			Claim	80	20	0

Grade	Item	2x	Inter-Rater Reliability			
			Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
6	Grade6_Item1	≥3,370	Content	77	23	0
			Claim	77	23	0
	Grade6_Item2	≥3,360	Content	77	23	0
			Claim	77	23	0
	Grade6_Item3	≥3,380	Content	79	21	0
			Claim	77	23	0
	Grade6_Item4	≥3,360	Content	81	19	0
			Claim	84	16	0
	Grade6_Item5	≥3,370	Content	84	16	0
			Claim	89	11	0
	Grade6_Item6	≥3,350	Content	81	19	0
			Claim	87	12	0
7	Grade7_Item1	≥3,150	Content	80	19	0
			Claim	79	20	1
	Grade7_Item2	≥3,180	Content	81	19	0
			Claim	81	19	1
	Grade7_Item3	≥3,180	Content	81	19	0
			Claim	79	20	0
	Grade7_Item4	≥3,150	Content	80	20	0
			Claim	79	21	0
	Grade7_Item5	≥3,170	Content	82	18	0
			Claim	81	19	1
	Grade7_Item6	≥3,180	Content	81	19	0
			Claim	83	17	0
8	Grade8_Item1	≥3,310	Content	80	20	0
			Claim	77	23	0
	Grade8_Item2	≥3,310	Content	81	19	0
			Claim	79	21	0

Grade	Item	2x	Inter-Rater Reliability			
			Dimension	Percent Exact Agreement	Percent Adjacent Agreement	Percent Non-Adjacent
	Grade8_Item3	≥3,300	Content	78	22	0
			Claim	79	21	0
	Grade8_Item4	≥3,280	Content	79	21	0
			Claim	78	22	0
	Grade8_Item5	≥3,320	Content	81	19	0
			Claim	80	20	0
	Grade8_Item6	≥3,290	Content	79	21	0
			Claim	79	21	0

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

Table 6.8

Field Test Extended-Response Score Point Distributions

Grade/ Item	Score Point Distribution							
	Total	Dimension	% "0" Rating	% "1" Rating	% "2" Rating	% "3" Rating	% "4" Rating	% Blank
Grade 5 Item1	≥3,210	Content	41	45	7	1	0	0
		Claim	45	42	7	1	0	0
Grade 5 Item2	≥3,230	Content	44	39	6	1	0	0
		Claim	46	37	6	1	0	0
Grade 5 Item3	≥3,210	Content	50	38	3	0	0	0
		Claim	55	33	3	0	0	0
Grade 5 Item4	≥3,250	Content	56	33	4	1	0	0
		Claim	57	32	4	1	0	0
Grade 5 Item5	≥3,230	Content	48	40	3	1	0	0
		Claim	49	39	6	0	0	0
Grade 6 Item1	≥3,370	Content	38	44	11	1	0	0
		Claim	37	48	9	1	0	0
Grade 6 Item2	≥3,360	Content	27	52	15	2	0	0
		Claim	22	61	12	1	0	0

Grade/ Item	Score Point Distribution							
	Total	Dimension	% "0" Rating	% "1" Rating	% "2" Rating	% "3" Rating	% "4" Rating	% Blank
Grade 6 Item3	≥3,380	Content	30	48	12	1	0	0
		Claim	26	54	12	2	0	0
Grade 6 Item4	≥3,360	Content	34	50	9	1	0	0
		Claim	29	56	8	1	0	0
Grade 6 Item5	≥3,370	Content	19	68	10	1	0	0
		Claim	10	78	9	1	0	0
Grade 6 Item6	≥3,350	Content	32	54	8	1	0	0
		Claim	19	68	7	1	0	0
Grade 7 Item1	≥3,150	Content	33	51	9	1	0	0
		Claim	43	41	9	1	0	0
Grade 7 Item2	≥3,180	Content	32	50	9	2	0	0
		Claim	39	43	9	2	0	0
Grade 7 Item3	≥3,180	Content	35	53	5	0	0	0
		Claim	49	39	5	0	0	0
Grade 7 Item4	≥3,150	Content	30	56	6	1	0	0
		Claim	49	37	6	0	0	0
Grade 7 Item5	≥3,170	Content	50	37	5	0	0	0
		Claim	58	29	5	0	0	0
Grade 7 Item6	≥3,180	Content	51	35	4	1	0	0
		Claim	61	25	4	1	0	0
Grade 8 Item1	≥3,310	Content	23	43	22	4	1	0
		Claim	26	41	21	4	1	0
Grade 8 Item2	≥3,310	Content	19	47	23	3	0	0
		Claim	18	47	23	3	0	0
Grade 8 Item3	≥3,300	Content	36	42	13	3	0	0
		Claim	34	44	13	3	0	0
Grade 8 Item4	≥3,280	Content	40	37	12	3	0	0
		Claim	45	33	11	3	0	0
Grade 8 Item5	≥3,320	Content	27	43	19	4	1	0
		Claim	23	47	19	4	1	0
Grade 8 Item6	≥3,290	Content	25	45	20	3	1	0
		Claim	23	47	20	3	1	0

7. Data Analysis

Classical Item Statistics

This section shows the results of the classical item analysis for data obtained from the LEAP operational tests. These item analysis results serve two purposes: 1) to inform item performance and 2) to provide item statistics for the item bank. LEAP classical item analysis consists of the following types of items: key/multiple option-based items, rule-based machine-scored items such as technology-embedded items, and hand-scored constructed response items. For each operational item, the analysis produces item difficulty (i.e., *p*-value) and item discrimination (p-b serial).

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

The point biserial correlation of any MC item should be greater than 0.20. Any item with negative point-biserial correlation should not be selected. However, there may be cases in which items required to meet content guidelines do not meet the point-biserial correlation guideline. The corrected point-biserial correlation is a measure of item discrimination. Items with higher item-total correlations provide better information about how well items discriminate between lower- and higher-performing students. In addition, the following flagging criteria was also used to review any field test items for data review:

- Correct Response *p*-value < 0.25
- Correct Response point-biserial < 0.20
- Distractor *p*-value > 0.40

Please note that statistical results of FT items can be found at Pearson using ABBI.

Differential Item Functioning

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

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

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

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

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

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

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

The *ΔMH DIF* is then computed as

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

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

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

Table 7.1
DIF Categories for Dichotomous Items

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

For polytomous items, the standardized mean difference (*SMD*) (Dorans & Schmitt, 1991; Zwick, Thayer, & Mazzeo, 1997) and the Mantel χ^2 statistic (Mantel, 1963) are used to identify items with DIF. *SMD* estimates the average difference in performance between the

reference group and the focal group while controlling for student ability. To calculate *SMD*, let *M* represent the matching variable (total test score). For all $M = m$, identify the students with raw score *m* and calculate the expected item score for the reference group (E_{rm}) and the focal group (E_{fm}). DIF is defined as $D_m = E_{fm} - E_{rm}$, and *SMD* is a weighted average of D_m using the weights $w_m = N_{fm}$ (the number of students in the focal group with raw score *m*), which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

The *MH DIF* statistic is based on a $2\times(T+1)\times M$ (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 7.2 defines the DIF categories for polytomous items.

Table 7.2

DIF Categories for Polytomous Items

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

Three DIF analyses were conducted for operational test items: female/male, black/white, and Hispanic/white. That is, item score data were used to detect items on which female or male students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The same methods were used to detect items on which black or white students performed unexpectedly well or unexpectedly poorly, given their performance on the full assessment. The last two columns of Tables 7.3–7.4 provide the number of items flagged for DIF. Items flagged with B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Very few operational test items were flagged for C-DIF by either analysis.

Note that DIF flags for dichotomous items are based on the Mantel–Haenszel statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 and *SMD* statistics. Table 7.3 and Table 7.5 summarize the operational-test DIF statistics for the operational items on the 2019 spring test forms.

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

Table 7.3
Summary of DIF Flags (Female – Male) for Operational Items by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	43	[0],[0]	[0],[0]
4	43	[0],[0]	[0],[0]
5	46	[0],[0]	[0],[0]
6	51	[2],[1]	[0],[0]
7	52	[1],[1]	[0],[0]
8	49	[2],[2]	[1],[0]

Table 7.4

Summary of DIF Flags (Black – White) for Operational Items by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	43	[0],[0]	[0],[0]
4	42	[0],[1]	[0],[0]
5	46	[0],[0]	[0],[0]
6	52	[0],[1]	[0],[0]
7	52	[0],[0]	[0],[1]
8	52	[0],[1]	[0],[0]

Table 7.5

Summary of Hispanic – White DIF Flags for Operational Items by Grade

Grade	A	[B+],[B-]	[C+],[C-]
3	43	[0],[0]	[0],[0]
4	43	[0],[0]	[0],[0]
5	46	[0],[0]	[0],[0]
6	53	[0],[0]	[0],[0]
7	51	[0],[1]	[0],[1]
8	53	[0],[0]	[0],[0]

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

Item Calibration and Scaling

LEAP 2025 Social Studies assessments are standards-based assessments that have been constructed to align rigorously to the social studies student standards, as defined by the LDOE and Louisiana educators. For each grade level, the content standards specify the subject matter students should know and the skills they should be able to perform. In addition, performance standards specify what students need to master in order to achieve proficiency. Constructing tests aligned to content standards enables the tests to assess the same constructs from one year to the next.

Item Response Theory (IRT) models were used in the item calibration for all LEAP 2025 Social Studies tests. Each grade-level test was calibrated separately. All calibration activities were independently replicated by Pearson staff as an added quality-control check.

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

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

In that model, $\boldsymbol{p_i(\theta_j)}$ 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):

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

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

Operational and Field Test Item Parameters

The distributions of item parameters are summarized in Table C.6. Figures in Appendix C provide graphical displays of the distributions of IRT parameter estimates for each grade. TE, CR, and ER items have no *c* parameters because they are polytomous items and are therefore modeled using the GPCM, and the number of ER tasks reflects the item parameter estimates for both trait scores (e.g., 12 actually represents 6 items × 2 traits).

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 test items by comparing observed and expected item performance. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. For dichotomous items, Q_1 is computed as

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

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

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

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

The generalization of Q_1 for items with multiple response categories is

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

where

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

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

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

and

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

where *df* is the degrees of freedom (the number of intervals minus the number of independent item parameters).

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

Dimensionality and Local Item Dependence

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

Computation of the Q_3 statistic starts with expected student performance on each item, which is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference can be thought of as what is left in performance after accounting for underlying achievement. If performance on an item is driven by a single achievement construct, then not only will the residual be small (as tested by the Q_1 statistic), but the correlation between residuals of the pair of items also will be small. These correlations are analogous to partial correlations, which can be interpreted as the relationship between two variables (items) after the effects of a third variable (underlying achievement) are held constant or "accounted for." The correlation among IRT residuals is the Q_3 statistic. When calculating the level of local item dependence for two items (*i* and *j*), the Q_3 statistic is

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

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

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

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

With *n* items, there are $n(n - 1)/2$ Q_3 statistics. For example, LEAP 2025 Social Studies grade 5 has 48 items and 1128 Q_3 values. The Q_3 values should all be small. Summaries of the distributions of Q_3 are provided in Appendix D: Dimensionality. Specifically, Q_3 data are summarized by minimum, 5th percentile, median, 95th percentile, and maximum values for LEAP 2025 Social Studies grades 3 through 8. To add perspective to the meaning of Q_3 distributions, the average zero-order correlation (simple intercorrelation) among item responses is also shown. If the achievement construct is "accounting for" the relationships among the items, Q_3 values should be much smaller than the zero-order correlations. The Q_3 summary tables in the dimensionality reports in Appendix D indicate that, for all grades and subjects, at least 90% (between the 5th and 95th percentiles) of the items are expectedly small. These data, coupled with the Q_1 data above, indicate that the unidimensional IRT model provides a very reasonable solution for capturing the essence of student achievement defined by the carefully selected set of items for each grade and subject.

Unidimensionality and Principal Component Analysis

It should be noted that Appendix D provides information about principal component analysis of grades 3-9 science. Measurement implies order and magnitude along a single dimension (Andrich, 2004). Consequently, in the case of scholastic achievement, one-dimensional scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 2004; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant

factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983). To check the unidimensionality of the 2017 ISTEP Grade 10 ELA, Math, and Science assessments, the relative sizes of the eigenvalues associated with a principal component analysis of the item set were examined using SAS program. The first and the second principal component eigenvalues were compared *without rotation*. Table 1 and Figures 1, 2, and 3 summarize the results of the first and second principal component eigenvalues of the assessments.

A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the table and figures, the first component is substantially larger than the second eigenvalue across the assessments: the first eigenvalue was at least 5 times as big as the second eigenvalue for each test except for grades 3 and 4. In addition, the figures indicate that the second component sharply drops from the first and gets flat. As a result, we could conclude that the unidimensionality assumption of 2019 assessment was met.

Scaling

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

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

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

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

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

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

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

which are rearranged as

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

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

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

The scaling constants *A* and *B* are calculated, and the Advanced cut score and the Approaching Basic cut score on the θ scale are transformed to the reporting scale, rounded to the nearest integer. At this point, the score ranges associated with the five achievement levels are determined. The same scaling constants *A* and *B* are used to convert student ability estimates to the reporting scale until new achievement level standards are set.

Descriptive Statistics and Frequency Distribution of LEAP 2025 Social Studies Scale Scores can be found in Appendix E: Scale Distribution and Statistics Report.

8. Reporting for 3–8 Social Studies

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

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

School Roster Report

A School Roster Report, which provides summary information about student performance on the LEAP 2025 ELA and Mathematics tests, is available to school systems and schools through eDIRECT. Total test scores and achievement-level indicators are shown for the content area of interest. Reporting category and subcategory performance ratings are also reported for students. At the school level, the percentage of students at each achievement level and rating by category and subcategory are summarized. More details can be found in the *LEAP 2025 Interpretive Guide*.

Individual Student-Level Report (ISR)

The ISR is another type of report available through the eDIRECT system. ISRs may be downloaded and printed by schools to be sent home to parents. At the top of the page, overall student performance is reported by scale scores and achievement level. To give context to the student score, the student's school system and state averages are presented to the right of the student information. In the middle of the page, category and subcategory performance indicators are reported. Achievement-level descriptors and the percentage of students in each achievement level by school, school system, and the state, which allows comparisons of the student's overall achievement level to those of their peers, are found at the bottom of the page. When a student does not receive a scale score, their achievement level will be left blank. ISRs for students whose scores were invalidated will display a blank scale score for a given content area.

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

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

The *LEAP 2025 Interpretive Guide* has three sections. The first section presents an introduction and an overview of key terms and test-related concepts. The second section discusses assessment terms and types of scores that are presented on the ISRs. Sample ISRs are included in the guide. The third section discusses information that is presented on the School Roster Report and an example of the report.

In summary, the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information developed by DRC address multiple best practices of the testing industry but are particularly related to the following standards:

> **Standard 5.1** Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (102)
>
> **Standard 6.10** When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (119)
>
> **Standard 7.0** Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (125)

Standard 12.18 In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (200)

9. Data Review Process and Results

During data review of the spring 2019 EFT items, content experts and psychometric support staff review field test items with accompanying data to make judgments about the appropriateness of items for use on operational test forms. Statistically flagged items are not rejected on the sole basis of statistics; only items with identifiable flaws based on content are rejected.

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

Facilitators from WestEd and Pearson lead the data review. Statistical information is evaluated for each item to determine whether the item functions as intended. Each item's suitability for future operational tests is then evaluated in the context of field test statistics. Judgments to accept, accept with edits (or "revise/re-field test"), or reject are then recorded. If the decision is to edit or to reject an item, additional information is captured to document, giving the reason for the decision. Table 9 summarizes the decisions by item type for the field test items reviewed during 2019 data review.

Table 9

FT Item Decisions by Item Type, 2019 Data Review

Grade	Item Type	Number of Items			
		Accept	Accept with Edits	Reject	% of Total
3	CR	2	1	–	6.12
	MC	30	3	6	79.59
	MS	5	–	2	14.29
	TE	–	–	–	0.00
	Total	37	4	8	100.00
4	CR	1	1	1	5.77
	MC	33	2	6	78.85
	MS	4	–	1	9.61
	TE	3	–	–	5.77
	Total	41	3	8	100.00
5	CR	1	–	–	2.00
	ER	3		2	10.00
	MC	30	3	2	70.00
	MS	6	1	–	14.00
	TE	2	–	–	4.00
	Total	42	4	4	100.00
6	CR	1	–	–	1.96
	ER	6	–	–	11.76
	MC	30	3	2	68.63
	MS	6	1	–	13.73
	TE	2	–	–	3.92
	Total	45	4	2	100.00
7	CR	1	–	–	2.04
	ER	6	–	–	12.24
	MC	31	–	4	71.43
	MS	5	–	–	10.20
	TE	1	–	1	4.08
	Total	44	–	5	100.00
8	CR	1	–	–	1.96
	ER	6	–	–	11.76
	MC	33	–	7	78.43
	MS	1	–	1	3.92
	TE	2	–	–	3.92
	Total	43	–	8	100.00

10. Reliability and Validity

Internal Consistency Reliability Estimation

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

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

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the ith item or component, and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogeneous. The homogeneity of LEAP 2025 Social Studies tests is evidenced through a dimensionality analysis. Dimensionality analyses results are discussed in "Chapter 7. Data Analysis."

The reliability and classification accuracy reports in Appendix F: Reliability and Classification Accuracy provide coefficient alpha and IRT model-based or "marginal reliability" (Thissen, Chen, & Bock, 2003) for the total tests. Coefficient alpha values range from 0.85 to 0.91, and marginal alpha values range from 0.87 to 0.92 across grades. Marginal reliability is described as "an average reliability over levels of θ or theta" (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 "posterior standard deviations" (SEMs) in the IRTPRO output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31 θs. Marginal reliability is the average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals.

The formula for marginal reliability is

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

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

Additional reliabilities were calculated on various demographic subgroups[1] using the population of students (see reliability and classification accuracy reports in the yearbook). Included with coefficient alpha in the tables are the number of students responding to the test, the mean score obtained by this group of students, and the standard deviation of the scores obtained for this group.

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

Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their scale scores. It is important to know the reliability of student scores in any examination, but assessing the reliability of the classification decisions based on these scores is of even greater importance. Classification decision reliability is estimated by the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive accuracy and consistency classification measures.

---

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

Accuracy of Classification. According to Livingston and Lewis (1995, p. 180), the classification accuracy is "the extent to which the actual classifications of the test takers... agree with those that would be made based on 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 somehow be obtained (Young & Yoon, 1998).

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

Accuracy and Consistency Indices. Three types of accuracy and consistency indices were generated: *overall*, *conditional-on-level*, and *cut point*, provided in Appendix F: Reliability and Classification Accuracy. The *overall accuracy* of performance-level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels. It is a proportion (or percentage) of correct classification across all the levels. The overall accuracy index ranges from 0.640 to 0.715 for grades of LEAP 2025 Social Studies assessments.

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 et al., 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 0.211 to 0.245 across grades of LEAP 2025 Social Studies assessments.

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 0.381 to 0.501 across grades.

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

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

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

Validity

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (AERA/APA/NCME, 2009, 2014). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process.

The spring 2019 LEAP 2025 Social Studies tests were designed and developed to provide fair and accurate achievement scores that support appropriate, meaningful, and useful educational decisions. Validity evidence may be found in the following parts: Chapter 2 (Assessment Frameworks), Chapter 3 (Overview of the Test Development Process), Chapter 4 (Construction of Test Forms), Chapter 5 (Test Administration), Chapter 6 (Scoring Activities), Chapter 7 (Data Analysis), Chapter 8 (Reporting for 3–8 Social Studies),

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 Social Studies assessments is an adequate and representative sample of appropriate content and that the content forms 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 Social Studies assessments.

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 spring 2019 LEAP 2025 Social Studies tests. 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 the spring 2019 LEAP 2025 Social Studies operational items by grade 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 Social Studies assessments for spring 2019.

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

11. Statistical Summaries

The LEAP 2025 test results for Social Studies for grades 3–8 are not on a vertical scale, and therefore the scale scores across grades cannot be compared. For each grade, the lowest obtainable scale score on the Social Studies tests is 650 and the highest obtainable scale score is 850. Test results are presented in Tables 11.1 through 11.6. For each grade, scale-score means and standard deviations as well as the percentages of students in each performance level are reported for the state and select demographic groups. In addition to the descriptive statistics presented in Tables 11.1 through 11.6, scale score frequency distributions are presented in Appendix E. The information for each grade is provided within separate tables.

It should be noted that the grades 3 and 4 operational test became shorter compared to previous years. Therefore, we directly compared a reliability index (i.e., Cronbach's Alpha) between the two years (i.e., Year 2018 and Year 2019). The results indicated that there was not much change between the two years (i.e., 0.85 to 0.84 for grade 3 and 0.86 to 0.85 for grade 4).

The current years' unidimensionality results can be found in Appendix D. We continue to conduct a principal component analysis. Measurement implies order and magnitude along a single dimension (Andrich, 1989). Consequently, in the case of scholastic achievement, a one-dimensional scale is required to reflect this idea of measurement (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 1988; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983).

To check the unidimensionality, the relative sizes of the eigenvalues associated with a principal component analysis of the item set will be examined using the SAS program. The first and the second principal component eigenvalues will be then compared without rotation.

Table 11.1

Spring 2019 LEAP 2025 State Test Results Grade 3

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥46,540	723.39	39.18	21	29	24	18	7
Gender								
Female	≥23,660	724.76	38.37	20	30	25	18	7
Male	≥22,870	721.98	39.94	23	29	23	17	7
Gender Unknown	<10	NR	NR	NR	NR	NR	NR	NR
Ethnicity								
Hispanic/Latino	≥4,290	717.23	38.22	25	30	24	16	5
American Indian or Alaska Native	≥340	721	37	23	25	27	21	4
Asian	≥780	745.23	41.84	11	19	22	27	21
Black	≥19,720	711.61	36.62	29	34	22	11	3
Native Hawaiian or Other Pacific Islander	≥30	731.34	33.64	13	29	29	16	13
White	≥19,890	735.06	38.06	13	24	26	24	12
Multi-Racial	≥1,470	728.15	37.18	16	29	26	21	7
Economically Disadvantaged								
No	≥13,520	741.44	37.74	10	21	26	27	15
Yes	≥33,010	715.99	37.31	26	33	23	14	4
LEP Status								
Fully English Proficient	≥44,050	724.37	39.11	21	29	24	18	8
English Learner	≥2,490	705.91	36.25	35	35	19	9	2

Table 11.2

Spring 2019 LEAP 2025 State Test Results Grade 4

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥48,290	726.99	37.56	19	26	25	23	6
Gender								
Female	≥24,770	726.41	35.87	19	28	26	22	5
Male	≥23,510	727.61	39.25	20	25	24	24	7
Gender Unknown	≥10	685.60	25.53	60	30	10		
Ethnicity								
Hispanic/Latino	≥4,200	721.17	37.38	23	28	25	19	4
American Indian or Alaska Native	≥370	722.00	32	17	33	29	19	2
Asian	≥720	751.19	729	9	14	22	34	21
Black	≥20,830	713.91	20831	28	33	23	14	2
Native Hawaiian or Other Pacific Islander	≥40	723.19	42	19	31	29	17	5
White	≥20,610	740.10	20615	11	20	27	33	10
Multi-Racial	≥1,550	732.82	1550	15	23	28	27	7
Economically Disadvantaged								
No	≥14,290	745.28	35.89	9	17	25	35	14
Yes	≥34,000	719.30	35.53	24	30	25	18	3
LEP Status								
Fully English Proficient	≥46,080	728.06	37.41	19	26	25	24	6
English Learner	≥2,210	704.78	33.65	36	35	20	8	1

Table 11.3

Spring 2019 LEAP 2025 State Test Results Grade 5

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥48,430	728.87	34.70	19	23	31	22	6
Gender								
Female	≥24,810	728.64	33.76	18	24	32	21	5
Male	≥23,610	729.11	35.66	20	22	30	23	6
Ethnicity								
Hispanic/Latino	≥4,060	724.23	36.25	24	23	29	20	5
American Indian or Alaska Native	≥320	730.96	33.46	16	23	30	25	6
Asian	≥800	751.78	37.11	9	9	25	36	21
Black	≥20,600	716.98	32.33	27	29	29	13	2
Native Hawaiian or Other Pacific Islander	≥40	747.63	29.22	2	25	31	29	13
White	≥21,060	740.13	32.41	10	18	33	30	9
Multi-Racial	≥1,530	733.24	33.11	14	22	32	24	7
Economically Disadvantaged								
No	≥14,750	745.59	32.38	8	14	31	34	12
Yes	≥33,670	721.55	33.11	24	27	30	16	3
LEP Status								
Fully English Proficient	≥46,610	729.93	34.38	18	23	31	22	6
English Learner	≥1,810	701.60	31.65	45	29	20	5	0

Table 11.4

Spring 2019 LEAP 2025 State Test Results Grade 6

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥48,960	726.62	35.86	20	24	29	18	9
Gender								
Female	≥25,000	726.64	34.72	19	25	30	17	8
Male	≥23,960	726.61	37.01	22	23	28	18	9
Ethnicity								
Hispanic/Latino	≥3,820	719.22	37.78	28	23	26	16	6
American Indian or Alaska Native	≥320	727.87	34.04	19	22	32	19	7
Asian	≥760	754.74	37.16	8	9	23	29	31
Black	≥20,760	713.73	33.13	30	30	26	11	3
Native Hawaiian or Other Pacific Islander	≥40	737.22	38.42	15	13	35	22	15
White	≥21,800	738.75	33.15	11	19	32	24	14
Multi-Racial	≥1,440	733.02	34.60	14	23	32	20	11
Economically Disadvantaged								
No	≥15,360	744.04	33.25	8	16	31	27	18
Yes	≥33,600	718.67	34.15	26	28	28	13	5
LEP Status								
Fully English Proficient	≥47,410	727.75	35.44	19	24	30	18	9
English Learner	≥1,550	692.18	31.04	56	26	14	3	1

Table 11.5

Spring 2019 LEAP 2025 State Test Results Grade 7

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥46,910	732.95	39.02	23	17	25	22	13
Gender								
Female	≥23,710	733.94	37.87	21	17	26	23	13
Male	≥23,200	731.93	40.14	25	16	23	22	14
Ethnicity								
Hispanic/Latino	≥3,500	726.32	41.46	30	16	22	20	12
American Indian or Alaska Native	≥310	735.65	36.84	18	20	23	25	14
Asian	≥750	762.42	41.74	10	7	16	26	40
Black	≥20,110	719.86	36.05	33	21	25	15	6
Native Hawaiian or Other Pacific Islander	≥40	739.50	39.16	19	21	13	31	17
White	≥20,950	745.25	36.75	13	13	25	29	20
Multi-Racial	≥1,230	737.34	37.52	18	15	28	23	15
Economically Disadvantaged								
No	≥15,150	751.70	36.70	10	11	23	31	25
Yes	≥31,760	724.01	36.86	29	19	26	18	8
LEP Status								
Fully English Proficient	≥45,500	734.10	38.65	22	17	25	23	14
English Learner	≥1,400	695.80	32.27	61	19	14	5	1

Table 11.6

Spring 2019 LEAP 2025 State Test Results Grade 8

	Scale Score			% at Performance Level				
	Number	Mean	Standard Deviation	Unsatisfactory	Approaching Basic	Basic	Mastery	Advanced
TOTAL	≥45,730	740.27	37.46	15	17	25	30	13
Gender								
Female	≥23,080	742.03	35.71	13	17	26	31	13
Male	≥22,650	738.47	39.09	18	16	24	29	13
Ethnicity								
Hispanic/Latino	≥3,400	730.13	43.08	27	15	21	26	11
American Indian or Alaska Native	≥310	743.05	35.89	13	17	25	31	14
Asian	≥760	766.01	41.70	9	6	14	33	38
Black	≥19,330	726.91	34.75	23	23	27	22	5
Native Hawaiian or Other Pacific Islander	≥20	751.18	37.92	9	9	23	45	14
White	≥20,850	752.93	33.73	7	12	24	39	19
Multi-Racial	≥1,030	747.78	35.65	10	16	24	35	16
Economically Disadvantaged								
No	≥15,510	758.30	33.56	5	9	21	41	24
Yes	≥30,210	731.01	35.96	20	21	27	25	7
LEP Status								
Fully English Proficient	≥44,230	741.78	36.66	14	17	25	31	13
English Learner	≥1,500	695.81	33.30	59	20	14	6	1

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.

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.

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.

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

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

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

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

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

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

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

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

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

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

Appendix A: Training Agendas

LEAP 2025 Social Studies Grades 3–8 Stimulus Search Training Agenda
Test Item Development Cycle for 2018-2019

I. Introductions

II. Stimulus Set Overviews
  a. Item Set and Task Topics
    i. Themes of the item set or task that will need to be developed and supported by stimuli and items
    ii. Reporting Categories
    iii. Potential Assessable GLEs
      1. Stimuli should support these GLEs
    iv. Potential Types of Stimuli
      1. The overview contains recommended stimuli that will support the item set or task
      2. Searchers can propose other stimuli that will 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 and standards. The focus on content aligned to the GLEs reduces the potential for bias that can occur by including content that is not aligned to instruction. This could give an advantage to one student group over other student groups.
    ii. Sensitivity: Avoid topics in stimuli that may upset or offend students in items (e.g., references to graphic violence, nudity, alcohol, drugs, recent natural disasters, caricature representation of ethnic groups)
    iii. Universal design and visual impairments

III. Receiving stimulus search assignments

IV. Submitting stimuli for assignments
  a. Text-based stimuli
    i. Readability measurements
      1. Lexile
        a. Lexile bands
      2. ATOS

ii. Originals and marked-up copies of texts
iii. Text Complexity
iv. Range of Textual Evidence
v. Levels of Inference

b. Graphic-based stimuli
i. PDFs with source of graphic and location
ii. Word document with caption
iii. Gifs and JPEGs

V. Completing Webforms

VI. Using Box

VII. Additional Resources

LEAP 2025 Grade 3–8 Item Writer and Editor Training Agenda
Item Development Cycle for 2018-2019

I. Louisiana Student Standards and GLEs
   a. Grades 3–8
      i. Reporting Categories (History, Geography, Civics, Economics)
      ii. Grade-Level Expectations (GLEs)

II. Item Types and Set Overviews
   a. Selected-Response Items (Multiple Choice, Multiple Select)
      i. Rules for numbers of answer options and number correct
   b. Constructed-Response Items (Item Sets Only)
   c. Extended-Response Items (Tasks Only)
   d. Technology-Enhanced Items (Item Sets Only)
   e. Item Sets
      i. Sources (Each set will have multiple sources)
      ii. Item Set Overviews
         1. Item stems provided for each item
         2. Metadata associated with each item
         3. Answer options and the nature of distractors
   f. Task Sets
      i. Sources (Each set will have multiple sources)
      ii. Task Set Overviews
         1. Item stems provided for each item
         2. Metadata associated with each item
         3. Answer options and nature of distractors
   g. Standalone Items
      i. Purpose
      ii. Stimuli

III. Writing and Editing Rubrics and Scoring Guides
   a. Constructed-Response Item Scoring Rubrics
   b. Constructed-Response Item Scoring Information
   c. Extended-Response Scoring Rubrics
      i. Content
      ii. Claims
   d. Extended-Response Scoring Information

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

V. Item Writing and Editing 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 (e.g., information that is not part of the curriculum, standards)
    ii. Sensitivity: Avoid topics that may upset or offend students in items (e.g., references to graphic violence, nudity, alcohol, drugs, recent natural disasters, group stereotypes, representation of ethnic groups)

VI. ABBI Item Development Platform
a. Functionality of the ABBI platform
b. Writing Items in ABBI
c. Editing Items in ABBI
d. Attaching Scoring Information in ABBI
e. Checking Scoring of Technology-Enhanced Items

VII. Receiving Item Assignments via Smartsheet

VIII. Graphic Art Requests (Editing only)
a. Using the Smartsheet Form
b. Attaching Marked-Up Graphics in ABBI
c. Confirming graphic edits have been made

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

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

Scoring for SOXXXXXXXXXXXXXX

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

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

Scoring Notes:

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

Accept other reasonable answers.

XI. Selected-response (multiple-choice, multiple-select Items)

a. Reference sources in stems where appropriate. Use the language Sources 1 and 2 rather than Source 1 and Source 2. When referring to all of the sources, say "all of the sources." Refer to the source in the stem, where it is most appropriate.
b. Make sure MS items are in the correct format:
   Which natural resources inspired Americans to migrate westward?
   Select the two correct answers.
c. Make sure the item scores correctly.

XII. 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

Test Summary Reports
Social Studies

Contents
Table B.1.1 Test Blueprint Distribution by Reporting Category for Spring 2019 Operational Social Studies Table B.1.2 Actual Percentage of Points by Reporting Category in Spring 2019 (includes Task Items)
Tables B.2.1–B.2.6 GLE Coverage by Grade for Spring 2019 Operational Social Studies
Table B.3 Summary of Spring 2019 EFT Item Development
Table B.4 Spring 2019 Operational Item Summary for Social Studies
Table B.5 Raw Score Summary for Spring 2019 Operational Social Studies
Tables B.6.1–B.6.6 Raw Score Summary by Reporting Category: Spring 2019 Operational Social Studies
Tables B.7.1–B.7.6 Scale Score and Raw Score Summary: Spring 2019 Operational Social Studies

Table B.1.1

Test Blueprint Distribution by Reporting Category for Spring 2019 Operational Social Studies

Reporting Category	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
History	25%	25%	50%	52%	50%	55%
Geography	25%	25%	15%	22%	13%	15%
Civics	25%	25%	15%	13%	24%	15%
Economics	25%	25%	20%	13%	13%	15%

Table B.1.2

Actual Percentage of Points by Reporting Category in Spring 2019 (includes Task Items)

Reporting Category	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
History	27%	26%	57%	46%	57%	59%
Geography	23%	26%	15%	32%	11%	15%
Civics	28%	26%	13%	9%	20%	13%
Economics	23%	23%	15%	13%	11%	13%

Table B.2

GLE Coverage by Grade for Spring 2019 Operational Social Studies

Table B.2.1

Grade 3

Reporting Categories and GLE		No. of Items				% of Test
		MS	MC	ER	CR	
History	3.1.1		1			2.33
	3.1.2		1			2.33
	3.2.1		3	0		6.98
	3.2.2	1	1			4.65
	3.2.3		3			6.98
	3.2.4		1			2.33
	Subtotal	1	10			25.58
Geography	3.4.1		2			4.65
	3.4.2		1			2.33
	3.4.3		1		1	4.65
	3.4.5	1	3			9.30
	3.4.6		1			2.33
	Subtotal	1	8		1	23.26
Civics	3.5.1		2			4.65
	3.5.2		2			4.65
	3.5.4		1			2.33
	3.5.5		1			2.33
	3.5.6	1				2.33
	3.6.1		2			4.65
	3.6.2		2			4.65
	3.6.3		1			2.33
	Subtotal	1	11			27.91
Economics	3.7.3		2			4.26
	3.8.2		1		1	2.13
	3.8.3		2			4.26
	3.8.4		2			4.26
	3.9.2		1			4.26
	3.10.1		1			2.13
	Subtotal		9		1	23.26
Total		3	38		2	100.00

Table B.2.2
Grade 4

Reporting Categories and GLE		No. of Items				% of Test
		MS	MC	ER	CR	
History	4.1.2	1				2.33
	4.2.1		1		1	4.65
	4.2.3		2			4.65
	4.2.4		1			2.33
	4.3.1	1	4			11.63
	Subtotal	2	8		1	25.58
Geography	4.4.1		1			2.33
	4.4.3		1			2.33
	4.4.5		2			4.65
	4.5.1	1				2.33
	4.5.2		1			2.33
	4.5.3		1			2.33
	4.6.1		2			4.65
	4.6.2		2			4.65
	Subtotal	1	10			25.58
Civics	4.7.1		3			6.98
	4.7.2		1			2.33
	4.7.3		2			4.65
	4.7.4		2			4.65
	4.8.2		1			2.33
	4.8.3		1			2.33
	4.8.4		1			2.33
	Subtotal		11			20.83
Economics	4.9.1		2			4.65
	4.9.2	1				2.33
	4.9.3		2			4.65
	4.9.4		1			2.33
	4.9.5		1			2.33
	4.9.6				1	2.33
	4.9.8		1			2.33
	4.9.10		1			2.33
	Subtotal	1	8		1	23.26
Total		4	37		2	100.00

Table B.2.3
Grade 5

Reporting Categories and GLE		No. of Items					% of Test
		TE	MS	MC	ER	CR	
History	5.1.1	1					2.13
	5.2.1			1			2.13
	5.2.4			2			4.26
	5.3.2			1			2.13
	5.3.3			1			2.13
	5.3.4	1		7			17.02
	5.3.5			4	1		12.77
	5.3.6		1	5			12.77
	5.3.7			1			2.13
	Subtotal	2	1	22	1		57.45
Geography	5.4.3			1			2.13
	5.5.1		1	4			10.64
	5.5.2			1			2.13
	Subtotal			7			14.89
Civics	5.6.1			1			2.13
	5.6.2			3		1	8.51
	5.7.1			1			2.13
	Subtotal			5		1	12.77
Economics	5.8.1			1			2.13
	5.9.1	1		2		1	8.51
	5.10.1			2			4.26
	Subtotal	1		5		1	14.89
Total		3	2	38	2	2	100.00

Table B.2.4
Grade 6

Reporting Categories and GLE		No. of Items					% of Test
		TE	MS	MC	ER	CR	
History	6.2.2			1			1.85
	6.2.3		1	3			7.41
	6.2.4		1	3			7.41
	6.2.5		1	2			5.56
	6.2.6	1		3			7.41
	6.2.7			1			1.85
	6.2.8	1		1			3.70
	6.2.9			1			1.85
	6.2.10	1		4			9.26
	Subtotal	3	3	19			46.30
Geography	6.3.2			1			1.85
	6.3.3			1			1.85
	6.4.1			2			3.70
	6.4.2			2			3.70
	6.4.3		1	7	1	1	20.37
	Subtotal		1	13	1	1	31.48
Civics	6.5.1	1		1			3.70
	6.5.2			2		1	5.56
	Subtotal	1		3		1	9.26
Economics	6.6.1			1			1.85
	6.6.2			1			1.85
	6.6.3			3			5.56
	6.6.4			2			3.70
	Subtotal			7			12.96
Total		4	4	42	1	2	100.00

Table B.2.5
Grade 7

Reporting Categories and GLE		No. of Items					% of Test
		TE	MS	MC	ER	CR	
History	7.2.1			4	1		11.11
	7.2.2	1	1	4			11.11
	7.2.3			1			1.85
	7.2.4			2			3.70
	7.3.1			1			1.85
	7.3.2			2			3.70
	7.3.3			5			9.26
	7.3.4			1			1.85
	7.3.5			1			1.85
	7.4.1			2			3.70
	7.4.2	1		2			5.56
	7.4.3			1			1.85
	Subtotal	2	1	26	1		57.41
Geography	7.5.3			1			1.85
	7.6.1			1			1.85
	7.6.2			1		1	3.70
	7.6.3			1			1.85
	7.6.4			1			1.85
	Subtotal			5		1	11.11
Civics	7.8.2			1			1.85
	7.8.4			2			3.70
	7.8.5	1		1			3.70
	7.8.6			2			3.70
	7.9.2			1			1.85
	7.10.2			1		1	3.70
	7.10.3		1				1.85
	Subtotal	1	1	8		1	20.37
Economics	7.11.1	1		2			5.56
	7.11.2			1			1.85
	7.11.3			2			3.70
	Subtotal	1		5			11.11
Total		4	2	44	1	2	100.00

Table B.2.6
Grade 8

Reporting Categories and GLE		No. of Items					% of Test
		TE	MS	MC	ER	CR	
History	8.2.1			1			1.85
	8.2.2			1			1.85
	8.2.4		1	3	1		11.11
	8.2.5			4		1	9.26
	8.2.6	1	2	4			12.96
	8.2.7	1		4			9.26
	8.2.8			1			1.85
	8.2.9		1	4		1	11.11
	Subtotal	2	4	22	1	2	59.26
Geography	8.3.1			1			1.85
	8.3.2			1			1.85
	8.4.1			1			1.85
	8.4.2			1			1.85
	8.5.1		1	1			3.70
	8.5.2			2			3.70
	Subtotal		1	7			14.81
Civics	8.6.3		1	1			3.70
	8.7.1			1			1.85
	8.7.2			1			1.85
	8.8.1	1		1			3.70
	8.8.2			1			1.85
	Subtotal	1	1	5			12.96
Economics	8.9.2			1			1.85
	8.9.3		1				1.85
	8.10.2			1			1.85
	8.10.3	1	1	2			7.41
	Subtotal	1	2	4			12.96
Total		4	8	38	1	2	100.00

Table B.3

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

Grade	MC	MS	TE	CR	ER
3	39	7		3	-
4	41	5	3	3	-
5	35	7	2	1	6
6	35	5	2	1	6
7	37	3	2	1	6
8	37	3	2	1	6

Table B.4

Item Type Summary by Grade: Spring 2019 Operational Social Studies

Grade	MC	MS	TE	CR	ER
3	38	3	-	2	-
4	37	4	-	2	-
5	38	2	3	2	1
6	42	4	4	2	1
7	44	2	4	2	1
8	38	8	4	2	1

Table B.5

Raw Score Summary by Grade: Spring 2019 Operational Social Studies

Grade	*N*	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
3	≥46,540	20	8	0	44	0.45	0.35	0.84	3.05
4	≥48,290	22	8	0	45	0.50	0.37	0.85	3.02
5	≥48,430	22	9	1	55	0.43	0.38	0.88	3.27
6	≥48,960	31	11	2	63	0.51	0.39	0.90	3.43
7	≥46,910	29	12	0	65	0.47	0.40	0.91	3.53
8	≥45,730	35	13	1	66	0.56	0.43	0.92	3.59

Note: Reliability is coefficient alpha.

Table B.6

Raw Score Summary by Reporting Category: Spring 2019 Operational Social Studies

Table B.6.1

Grade 3

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	4.49	2.13	0	11	0.40	0.31	0.49	1.52
Geography	4.60	2.18	0	11	0.43	0.34	0.51	1.53
Civics	6.56	2.49	0	12	0.54	0.36	0.61	1.56
Economics	4.57	2.61	0	11	0.42	0.40	0.64	1.57

Note: Reliability is coefficient alpha.

Table B.6.2

Social Studies Grade 4

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	5.81	2.52	0	12	0.49	0.37	0.59	1.61
Geography	5.54	2.29	0	11	0.50	0.35	0.55	1.54
Civics	5.75	2.22	0	11	0.52	0.34	0.54	1.51
Economics	4.81	2.45	0	11	0.47	0.43	0.68	1.39

Note: Reliability is coefficient alpha.

Table B.6.3

Social Studies Grade 5

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	13.08	5.92	0	34	0.42	0.38	0.81	2.58
Geography	3.24	1.64	0	7	0.46	0.35	0.45	1.22
Civics	2.36	1.52	0	7	0.35	0.33	0.38	1.20
Economics	4.33	2.03	0	9	0.53	0.44	0.61	1.27

Note: Reliability is coefficient alpha.

Table B.6.4

Social Studies Grade 6

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	14.20	4.82	1	28	0.51	0.36	0.77	2.31
Geography	10.37	4.27	0	24	0.53	0.43	0.79	1.96
Civics	3.42	1.57	0	7	0.50	0.42	0.49	1.12
Economics	3.65	1.71	0	7	0.52	0.36	0.49	1.22

Note: Reliability is coefficient alpha.

Table B.6.5

Social Studies Grade 7

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	15.56	6.8	0	39	0.45	0.39	0.84	2.72
Geography	2.63	1.53	0	6	0.45	0.39	0.50	1.08
Civics	6.90	3.01	0	13	0.55	0.42	0.69	1.68
Economics	3.40	1.76	0	7	0.47	0.42	0.55	1.18

Note: Reliability is coefficient alpha.

Table B.6.6

Social Studies Grade 8

Reporting Category	Mean	SD	Min	Max	Mean_Pval	Mean_Pbis	Reliability	SEM
History	21.51	8.31	1	42	0.54	0.44	0.88	2.88
Geography	4.66	2.06	0	8	0.58	0.45	0.65	1.22
Civics	5.08	1.87	0	8	0.63	0.39	0.54	1.27
Economics	4.51	1.80	0	8	0.56	0.38	0.52	1.25

Note: Reliability is coefficient alpha.

Tables B.7

Scale Score and Raw Score Summary: Spring 2019 Operational Social Studies

Table B.7.1

Grade 3

Subgroup	*N*-Count	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥46,540	100.00	723.39	39.18	20	8
Female	≥23,660	50.83	724.76	38.37	20	8
Male	≥22,870	49.15	721.98	39.94	20	8
Gender Unknown	<10	NR	NR	NR	NR	NR
African American	≥19,720	42.69	711.61	36.62	18	7
Asian	≥780	1.71	745.23	41.84	25	8
Hispanic/Latino	≥4,290	9.29	717.85	38.22	19	7
Multi-Racial	≥1,470	3.18	728.15	37.18	21	7
Native Hawaiian or Other Pacific Islander	≥30	0.08	731.34	33.64	21	8
White	≥19,890	43.05	735.06	38.06	23	8
Economically Disadvantaged	≥33,010	70.94	715.99	37.31	19	7
English Learners	≥2,490	5.35	705.91	36.25	17	7

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

Table B.7.2

Grade 4

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥48,290	100.00	726.99	37.56	22	8
Female	≥24,770	51.29	726.41	35.87	22	7
Male	≥23,510	48.69	727.61	39.25	22	8
Gender Unknown	≥10	0.02	685.60	25.53	14	7
African American	≥20,830	43.42	713.91	34.68	19	7
Asian	≥720	1.52	751.19	39.94	27	8
Hispanic/Latino	≥4,200	8.77	721.18	37.38	21	8
Multi-Racial	≥1,550	3.23	732.82	36.27	23	8
Native Hawaiian or Other Pacific Islander	≥40	0.09	723.19	38.69	21	8
White	≥20,610	42.97	740.10	35.41	25	8
Economically Disadvantaged	≥34,000	70.40	719.30	35.53	20	7
English Learners	≥2,210	4.58	704.78	33.65	17	6

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

Table B.7.3

Grade 5

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥48,430	100.00	728.87	34.70	22	9
Female	≥24,810	51.23	728.64	33.76	22	9
Male	≥23,610	48.77	729.11	35.66	22	10
African American	≥20,600	42.53	716.98	32.33	19	8
American Indian or Alaska Native	≥320	0.66	730.96	33.46	23	9
Asian	≥800	1.66	751.78	37.11	29	11
Hispanic/Latino	≥4,060	8.40	724.23	36.25	21	9
Multi-Racial	≥1,530	3.16	733.24	33.11	23	9
Native Hawaiian or Other Pacific Islander	≥40	0.10	747.63	29.22	27	10
White	≥21,060	43.49	740.13	32.41	25	10
Economically Disadvantaged	≥33,670	69.54	721.55	33.11	20	8
English Language Learners	≥1,810	3.75	701.60	31.65	15	6

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

Table B.7.4

Grade 6

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥48,960	100.00	726.62	35.86	31	11
Female	≥25,000	51.06	726.64	34.72	31	11
Male	≥23,960	48.94	726.61	37.01	31	11
African American	≥20,760	42.40	713.73	33.13	27	10
American Indian or Alaska Native	≥320	0.66	727.87	34.04	31	10
Asian	≥760	1.56	754.74	37.16	40	11
Hispanic/Latino	≥3,820	7.82	719.22	37.78	29	11
Multi-Racial	≥1,440	2.94	733.02	34.60	33	11
Native Hawaiian or Other Pacific Islander	≥40	0.09	737.22	38.42	34	12
White	≥21,800	44.52	738.75	33.15	35	10
Economically Disadvantaged	≥33,600	68.63	718.67	34.15	28	10
English Language Learners	≥1,550	3.17	692.18	31.04	21	8

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

Table B.7.5

Grade 7

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥46,910	100.00	732.95	39.02	29	12
Female	≥23,710	50.54	733.94	37.87	29	12
Male	≥23,200	49.46	731.93	40.14	28	12
African American	≥20,110	42.87	719.86	36.05	25	10
American Indian or Alaska Native	≥310	0.66	735.65	36.84	29	11
Asian	≥750	1.60	762.42	41.74	38	13
Hispanic/Latino	≥3,500	7.47	726.32	41.46	27	12
Multi-Racial	≥1,230	2.62	737.34	37.52	30	12
Native Hawaiian or Other Pacific Islander	≥40	0.10	739.50	39.16	31	12
White	≥20,950	44.67	745.25	36.75	32	12
Economically Disadvantaged	≥31,760	67.70	724.01	36.86	26	11
English Language Learners	≥1,400	3.00	695.80	32.27	18	7

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

Table B.7.6

Grade 8

Subgroup	*N*	Percent	Scale Score Mean	Scale Score SD	Raw Score Mean	Raw Score SD
Total	≥45,730	100.00	740.27	37.46	35	13
Female	≥23,080	50.47	742.03	35.71	36	12
Male	≥22,650	49.53	738.47	39.09	35	13
African American	≥19,330	42.27	726.91	34.75	31	12
American Indian or Alaska Native	≥310	0.70	743.05	35.89	36	12
Asian	≥760	1.67	766.01	41.70	44	13
Hispanic/Latino	≥3,400	7.45	730.13	43.08	32	14
Multi-Racial	≥1,030	2.27	747.78	35.65	38	12
Native Hawaiian or Other Pacific Islander	≥20	0.05	751.18	37.92	39	12
White	≥20,850	45.59	752.93	33.73	40	12
Economically Disadvantaged	≥30,210	66.07	731.01	35.96	32	12
English Language Learners	≥1,500	3.28	695.81	33.30	21	10

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

Appendix C: Item Analysis Summary Report

Summary Statistics Reports

Social Studies

Contents
Table C.1 *P*-Value by Grade: Spring 2019 Operational Social Studies
Plot C.1 *P*-Value by Grade: Spring 2019 Operational Social Studies
Table C.2 Item-Total Correlation, Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies
Plot C.2 Item-Total Correlation, Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies
Table C.3 Corrected* Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies
Plot C.3 Corrected* Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies
Table C.4 Item-Total Correlation by Reporting Category: Spring 2019 Operational Social Studies
Table C.5 Statistically Flagged Operational Items: Spring 2019 Operational Social Studies
Table C.6 IRT Parameters: Spring 2019 Operational Social Studies
Plot C.4 IRT a-Parameter: Spring 2019 Operational Social Studies
Plot C.5 IRT b-Parameter: Spring 2019 Operational Social Studies
Plot C.6 IRT c-Parameter: Spring 2019 Operational Social Studies

Table C.1

P-Value by Grade: Spring 2019 Operational Social Studies

Grade	No. of OP Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	43	0.17	0.36	0.44	0.54	0.76
4	43	0.10	0.40	0.47	0.85	0.81
5	46	0.19	0.35	0.42	0.52	0.79
6	53	0.17	0.42	0.51	0.63	0.85
7	53	0.17	0.40	0.45	0.56	0.76
8	53	0.29	0.47	0.55	0.66	0.80

Plot C.1

P-Value: Spring 2019 Operational Social Studies

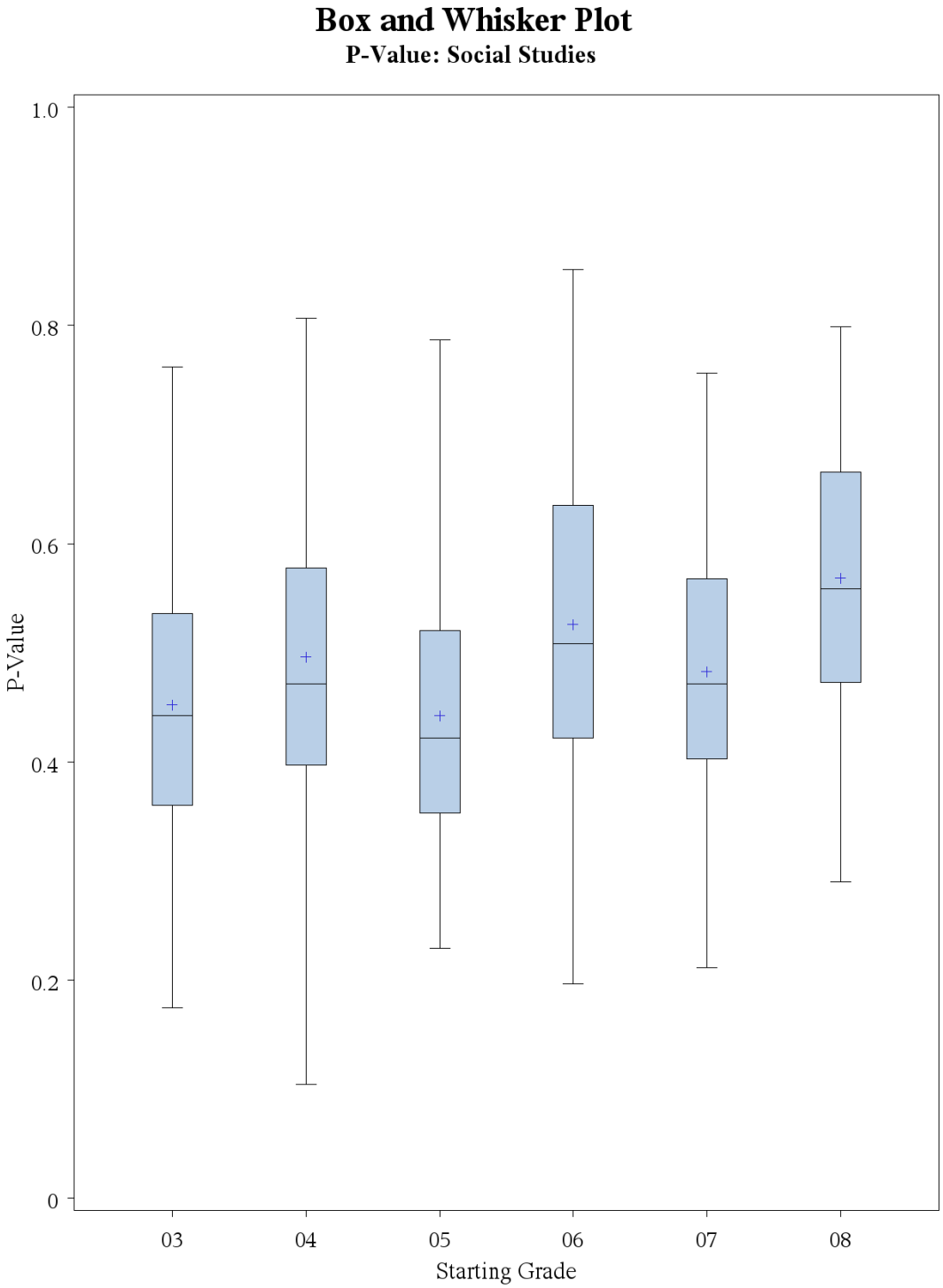


Table C.2

Item-Total Correlation, Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies

Grade	No. of OP Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	43	0.13	0.28	0.37	0.42	0.59
4	43	0.20	0.29	0.38	0.43	0.55
5	46	0.12	0.32	0.37	0.44	0.74
6	53	0.18	0.30	0.37	0.46	0.70
7	53	0.19	0.32	0.40	0.46	0.72
8	53	0.09	0.35	0.43	0.50	0.76

Plot C.2

Item-Total Correlation, Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies

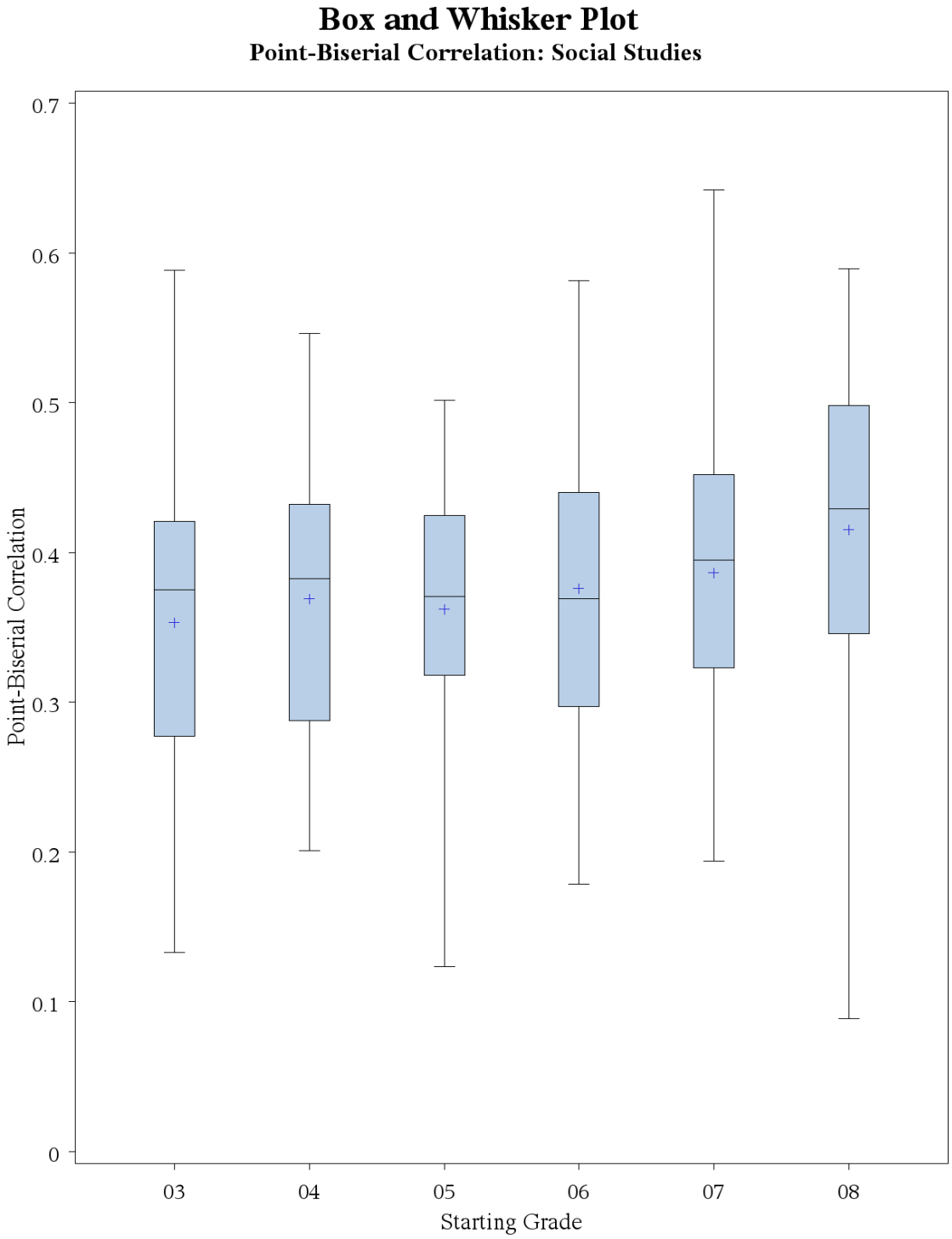


Table C.3

Corrected Point-Biserial Correlation by Grade: Spring 2019 Operational Social Studies*

Grade	No. of OP Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	43	0.07	0.22	0.32	0.37	0.51
4	43	0.14	0.23	0.33	0.38	0.47
5	46	0.07	0.27	0.33	0.40	0.69
6	53	0.13	0.26	0.33	0.42	0.66
7	53	0.15	0.29	0.36	0.42	0.68
8	53	0.05	0.32	0.40	0.47	0.72

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 Grade: Spring 2019 Operational Social Studies*

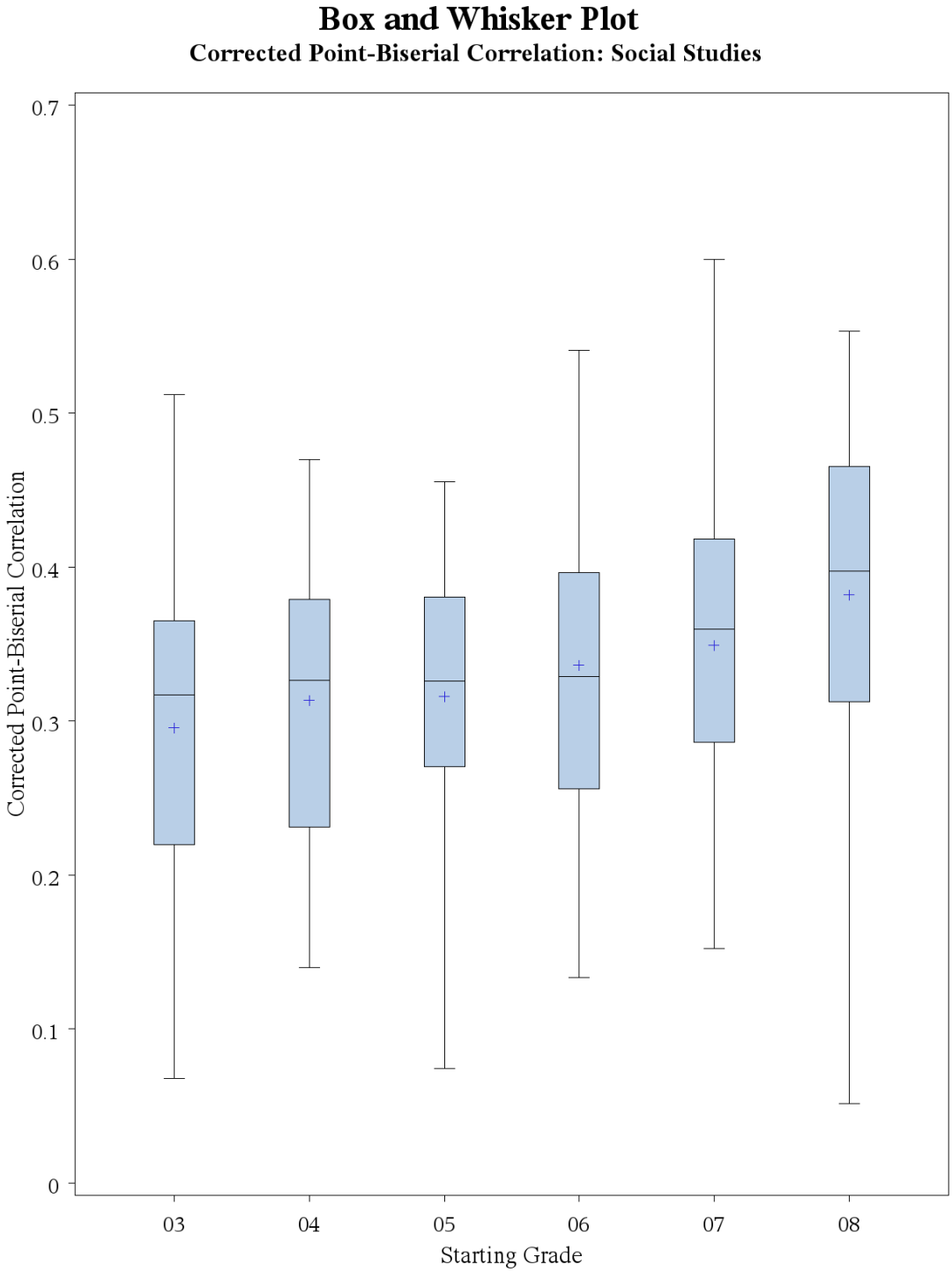


Table C.4

Item-Total Correlation by Reporting Category: Spring 2019 Operational Social Studies

Grade	Reporting Category	No. of OP Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	History	11	0.16	0.21	0.36	0.40	0.42
	Geography	10	0.18	0.29	0.35	0.42	0.45
	Civics	12	0.13	0.26	0.38	0.44	0.53
	Economics	10	0.28	0.36	0.39	0.43	0.59
4	History	11	0.23	0.31	0.35	0.46	0.55
	Geography	11	0.20	0.23	0.37	0.43	0.52
	Civics	11	0.22	0.24	0.33	0.41	0.50
	Economics	10	0.33	0.41	0.43	0.46	0.49
5	History	26	0.18	0.30	0.36	0.39	0.50
	Geography	7	0.20	0.25	0.39	0.41	0.45
	Civics	6	0.12	0.18	0.37	0.46	0.50
	Economics	7	0.35	0.40	0.47	0.50	0.50
6	History	25	0.18	0.29	0.35	0.43	0.54
	Geography	16	0.23	0.30	0.38	0.46	0.57
	Civics	5	0.28	0.33	0.45	0.48	0.58
	Economics	7	0.25	0.34	0.36	0.42	0.42
7	History	30	0.19	0.30	0.38	0.43	0.54
	Geography	6	0.24	0.26	0.40	0.46	0.60
	Civics	11	0.24	0.35	0.40	0.50	0.64
	Economics	6	0.31	0.32	0.41	0.48	0.58
8	History	31	0.27	0.35	0.43	0.50	0.59
	Geography	8	0.30	0.40	0.44	0.51	0.56
	Civics	7	0.09	0.30	0.41	0.52	0.58
	Economics	7	0.26	0.28	0.41	0.45	0.51

Table C.5

Statistically Flagged Operational Items by Grade: Spring 2019 Operational Social Studies

Grade	Item Type	*N* OP Items	*N* Items Flagged for *P*-Value	*N* Items Flagged for Mean	*N* Items Flagged for Point-Biserial Correlation	*N* Items Flagged for DIF	*N* Items Flagged for Omitting
3	CR	2	1	1	0	0	0
	MC	38	0	0	3	0	0
	MS	3	1	0	0	0	0
4	CR	2	1	1	0	0	0
	MC	37	0	0	0	1	0
	MS	4	1	0	0	0	0
5	CR	2	0	0	0	0	0
	ER	1	1	1	0	0	0
	MC	38	2	0	4	0	0
	MS	2	0	0	0	0	0
	TE	3	1	1	0	0	0
6	CR	2	0	0	0	0	0
	ER	1	1	1	0	1	0
	MC	42	0	0	2	1	0
	MS	4	1	0	0	0	0
	TE	4	0	0	0	0	0
7	CR	2	0	0	0	1	0
	ER	1	1	1	0	1	0
	MC	44	0	0	1	0	0
	MS	2	0	0	0	1	0
	TE	4	1	1	0	2	0
8	CR	2	0	0	0	1	0
	ER	1	0	0	0	1	0
	MC	38	0	0	0	3	0
	MS	8	0	0	1	0	0
	TE	4	0	0	0	0	0

Table C.6

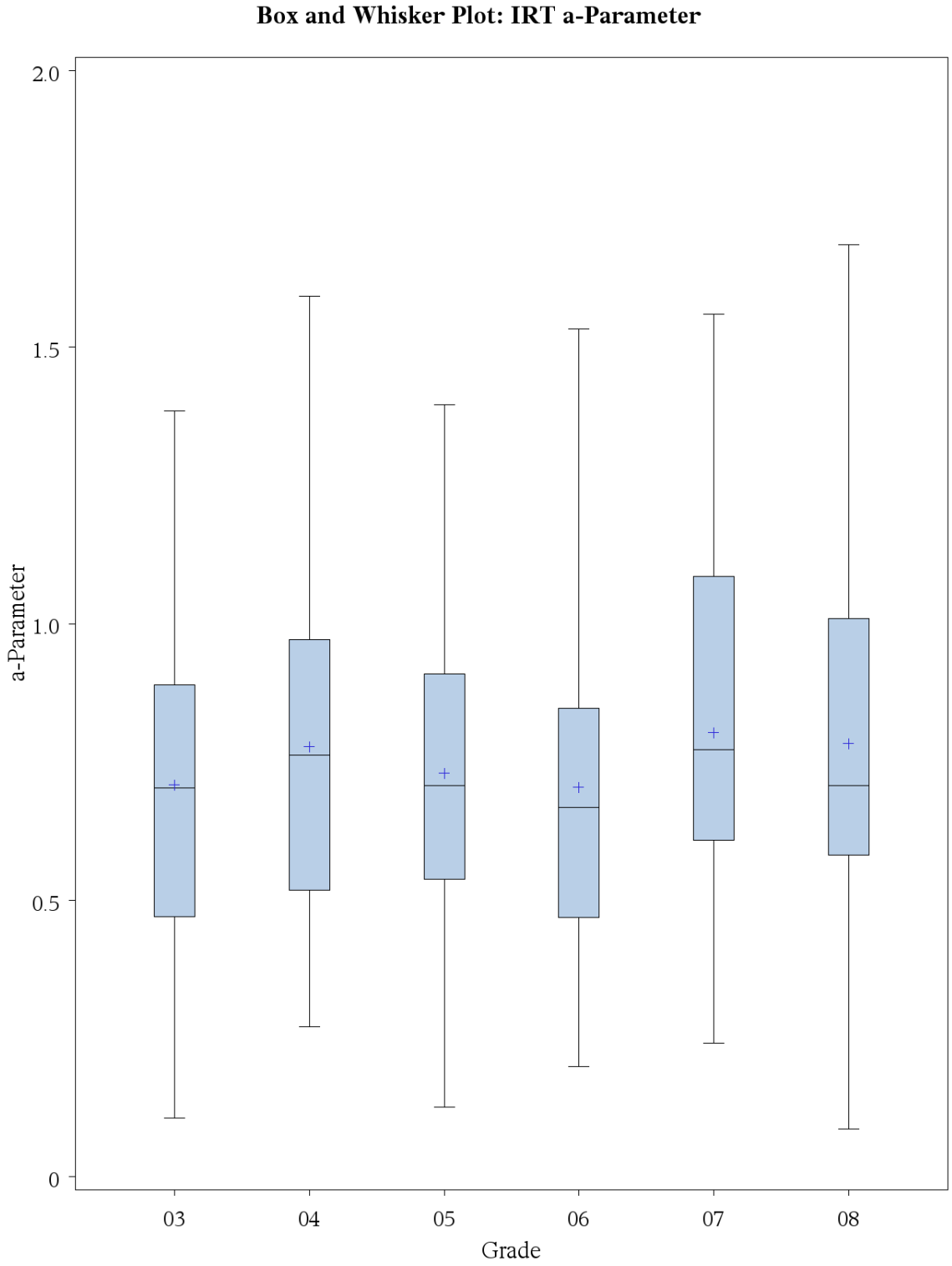
IRT Item Parameters by Grade: Spring 2019 Operational Social Studies

Grade	Parameter	No. of OP Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	a	43	0.105	0.471	0.703	0.889	1.384
	b	43	-1.201	0.357	0.917	1.207	2.639
	c	43	0.011	0.107	0.171	0.222	0.407
4	a	43	0.271	0.518	0.762	0.971	1.593
	b	43	-1.934	-0.246	0.610	1.234	2.826
	c	43	0.015	0.096	0.196	0.225	0.358
5	a	46	0.126	0.538	0.708	0.909	1.397
	b	46	-0.896	0.657	0.912	1.452	4.714
	c	46	0.073	0.124	0.187	0.250	0.381
6	a	53	0.199	0.469	0.667	0.847	1.533
	b	53	-1.540	-0.374	0.456	0.960	2.540
	c	53	0.019	0.119	0.160	0.200	0.529
7	a	53	0.241	0.609	0.773	1.086	1.559
	b	53	-1.047	0.079	0.584	1.135	1.860
	c	53	0.031	0.122	0.196	0.266	0.352
8	a	53	0.087	0.582	0.707	1.009	1.685
	b	53	-1.162	-0.303	0.179	0.578	7.132
	c	53	0.018	0.118	0.166	0.234	0.376

Note: c-Parameter summaries include MC and MS items only.

Plot C.4

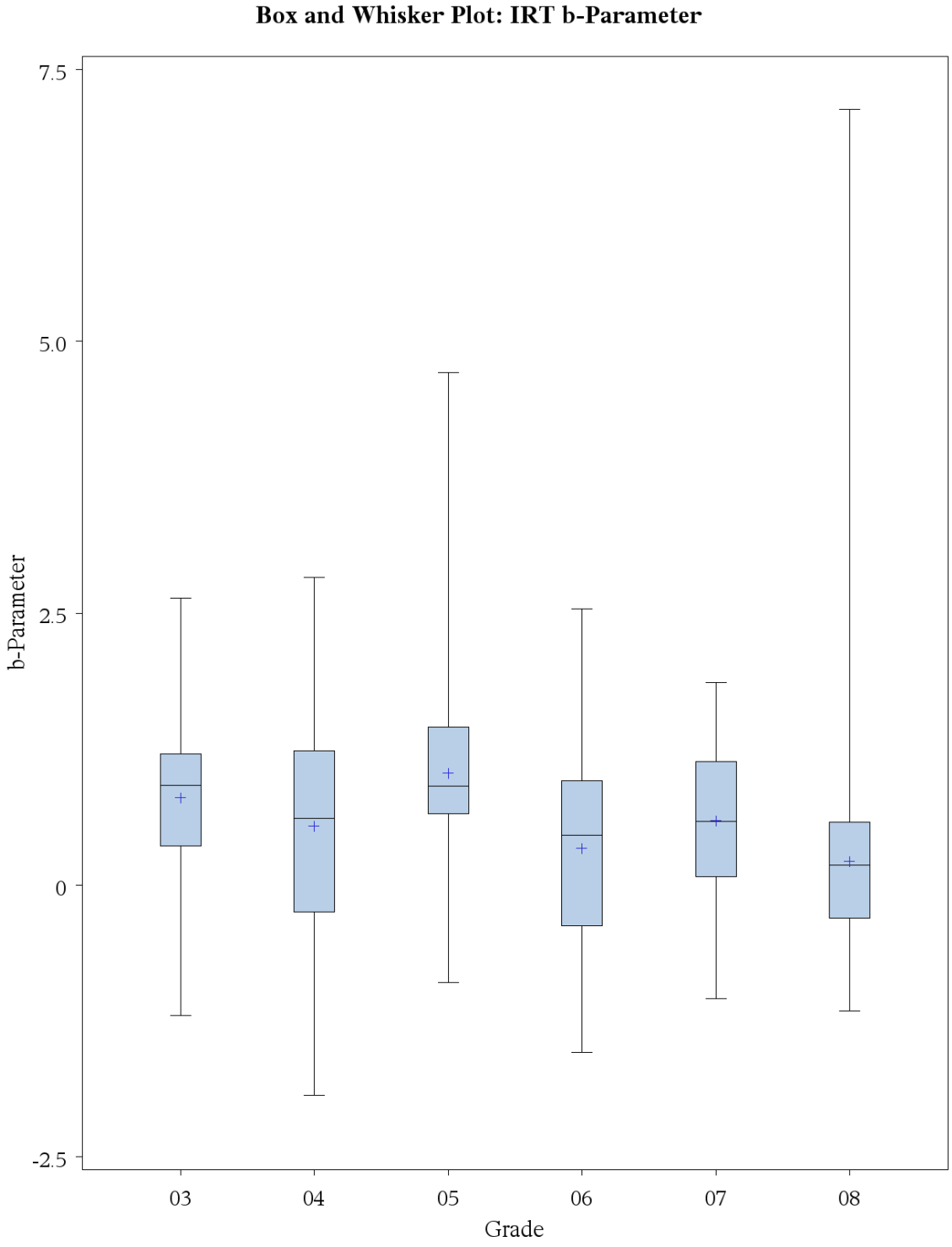
IRT a-Parameter: Spring 2019 Operational Social Studies



Note: All item types are included in a-Parameter summaries.

Plot C.5

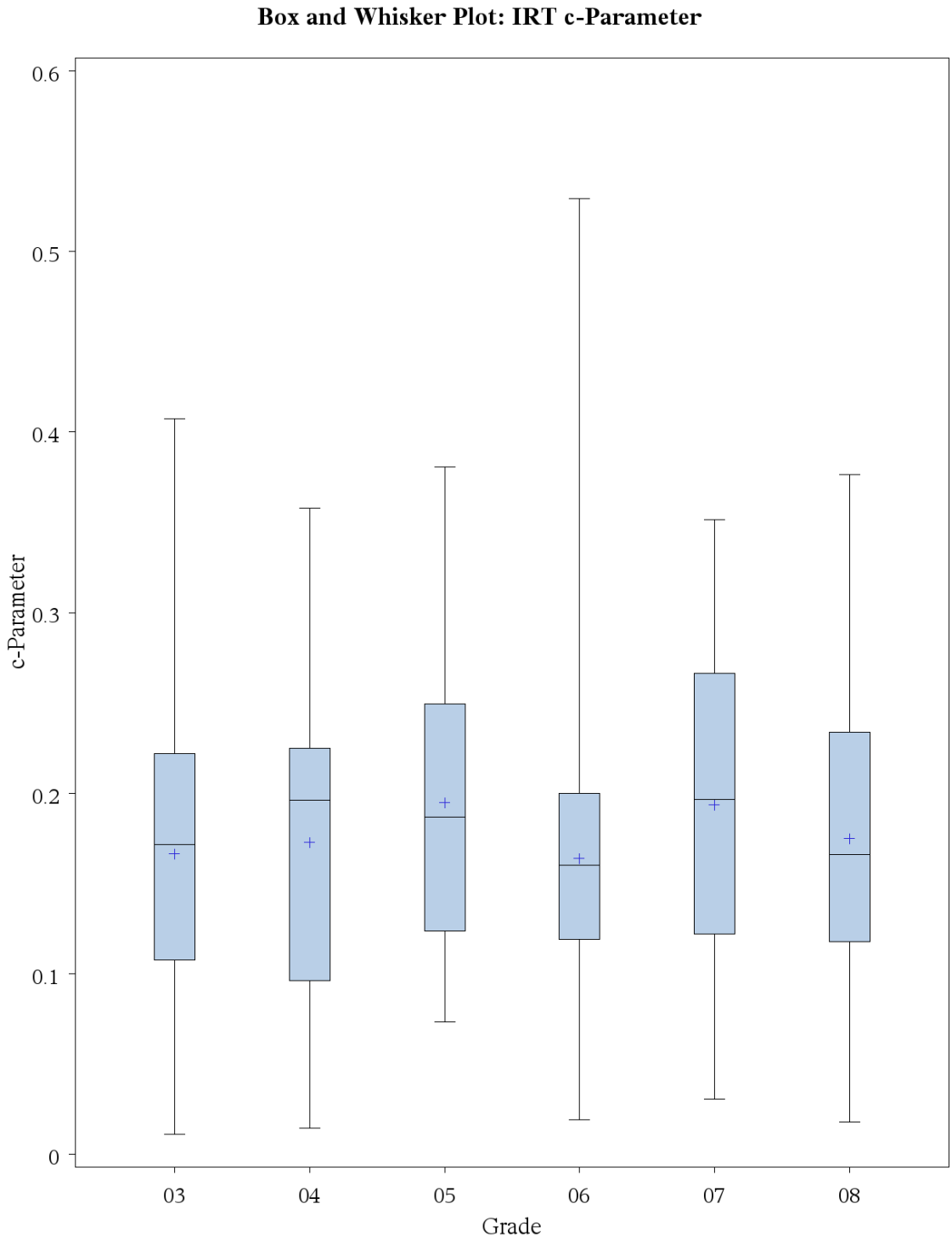
IRT b-Parameter: Spring 2019 Operational Social Studies



Note: All item types are included in b-Parameter summaries.

Plot C.6

IRT c-Parameter: Spring 2019 Operational Social Studies



Note: Only MC and MS items are included in c-Parameter summaries.

Appendix D: Dimensionality

Dimensionality Reports

Social Studies

Contents
Table D.1 Zq1 Statistics and Summary Data by Item Type and Grade: Spring 2019 Operational Social Studies
Table D.2 Q3 Statistics and Summary Data by Grade: Spring 2019 Operational Social Studies
Tables D.3.1–D.3.6 Reporting Category Intercorrelation Coefficients: Spring 2019 Operational Social Studies
Table D.4.1 First and Second Eigenvalues by Grade: Spring 2019 Operational Social Studies
Figures D.4.1-D.4.2 Principal Component Analysis Plot: Spring 2019 Operational Social Studies

Table D.1

Zq1 Statistics and Summary Data by Item Type and Grade: Spring 2019 Operational Social Studies

Grade	Type	Minimum	25th Percentile	Median	75th Percentile	Maximum	Num. of Items with Poor Fit
3	CR	35.76	35.76	42.33	48.90	48.90	0
	MC	5.51	14.78	20.48	34.77	87.36	2
	MS	14.82	14.82	18.75	45.21	45.21	0
4	CR	6.19	7.15	20.25	36.69	40.99	0
	MC	0.86	6.05	12.94	32.33	315.46	7
	MS	2.33	7.36	10.61	22.64	29.83	0
5	CR	24.88	24.88	27.73	30.58	30.58	0
	ER	44.14	44.14	46.47	48.79	48.79	1
	MC	-0.35	2.72	4.49	7.94	45.55	1
	MS	4.29	4.29	17.08	29.86	29.86	0
	TE	12.96	12.96	27.10	29.98	29.98	0
6	CR	12.14	12.14	19.26	26.39	26.39	0
	ER	19.17	19.17	20.63	22.09	22.09	0
	MC	2.34	4.53	6.42	9.40	46.66	1
	MS	4.79	8.47	12.60	15.02	16.98	0
	TE	15.50	15.55	17.28	49.58	80.22	1
7	CR	16.97	16.97	19.50	22.02	22.02	0
	ER	15.97	15.97	16.00	16.03	16.03	0
	MC	0.63	3.29	5.30	9.83	37.12	0
	MS	12.69	12.69	43.67	74.66	74.66	1
	TE	15.18	15.62	25.08	59.56	85.03	1
8	CR	13.48	13.48	16.25	19.02	19.02	0
	ER	25.94	25.94	27.66	29.38	29.38	0
	MC	0.14	2.47	3.58	4.92	14.39	0
	MS	3.79	6.35	8.84	19.32	63.23	1
	TE	6.50	10.02	19.14	26.72	28.72	0

Table D.2

Q3 Statistics and Summary Data by Grade: Spring 2019 Operational Social Studies

Grade	Average Zero-Order Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.102	-0.068	-0.046	-0.024	0.004	0.092
4	0.112	-0.086	-0.047	-0.025	0.008	0.077
5	0.125	-0.097	-0.047	-0.020	0.017	0.871
6	0.135	-0.085	-0.043	-0.018	0.011	0.798
7	0.144	-0.078	-0.041	-0.017	0.014	0.148
8	0.169	-0.123	-0.041	-0.014	0.018	0.908

Tables D.3

Reporting Category Intercorrelation Coefficients: Spring 2019 Operational Social Studies

Table D.3.1

Grade 3

Reporting Category	History	Geography	Civics	Economics
History	1.00	0.77*	0.71*	0.76*
Geography	0.51	1.00	0.76*	0.77*
Civics	0.51	0.54	1.00	0.78*
Economics	0.56	0.56	0.57	1.00

*Correlation coefficients below the main diagonal are Pearson correlations. Since the correlation coefficient between the two assessments could have a value lower (attenuated) than it actually would have if there were no measurement error for the two assessments, the disattenuated correlation was calculated.

Table D.3.2

Grade 4

Reporting Category	History	Geography	Civics	Economics
History	1.00	0.83*	0.84*	0.81*
Geography	0.57	1.00	0.76*	0.75*
Civics	0.57	0.51	1.00	0.74*
Economics	0.62	0.57	0.56	1.00

*Denotes disattenuated correlations. Correlations below the main diagonal are Pearson correlations.

Table D.3.3

Grade 5

Reporting Category	History	Geography	Civics	Economics
History	History	1.00	0.95*	1.00*
Geography	Geography	0.60	1.00	0.76*
Civics	Civics	0.57	0.43	1.00
Economics	Economics	0.69	0.53	0.50

*Denotes disattenuated correlations. Correlations below the main diagonal are Pearson correlations.

Table D.3.4

Grade 6

Reporting Category	History	Geography	Civics	Economics
History	1.00	0.89*	0.93*	0.90*
Geography	0.75	1.00	0.93*	0.89*
Civics	0.62	0.62	1.00	0.73*
Economics	0.60	0.59	0.48	1.00

*Correlation coefficients below the main diagonal are Pearson correlations. Since the correlation coefficient between the two assessments could have a value lower (attenuated) than it actually would have if there were no measurement error for the two assessments, the disattenuated correlation was calculated.

Table D.3.5

Grade 7

Reporting Category	History	Geography	Civics	Economics
History	1.00	0.98*	0.94*	0.95*
Geography	0.66	1.00	0.75*	0.75*
Civics	0.74	0.59	1.00	0.86*
Economics	0.67	0.53	0.61	1.00

*Denotes disattenuated correlations. Correlations below the main diagonal are Pearson correlations.

Table D.3.6

Grade 8

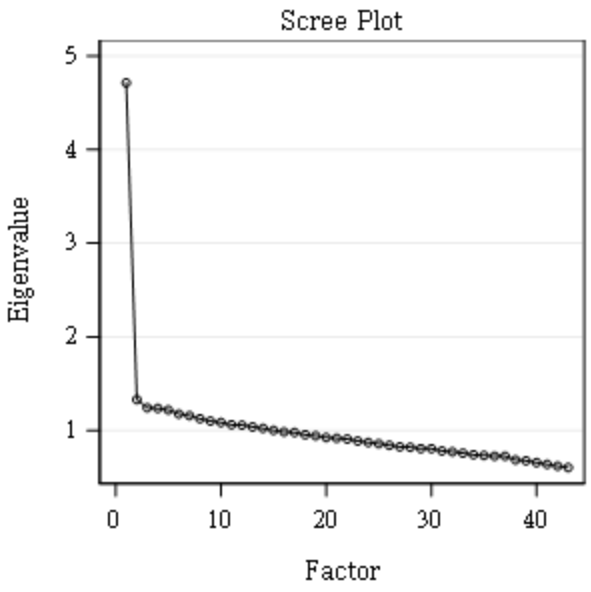
Reporting Category	History	Geography	Civics	Economics
History	1.00	0.96*	0.97*	0.95*
Geography	0.74	1.00	0.82*	0.83*
Civics	0.68	0.58	1.00	0.74*
Economics	0.66	0.58	0.51	1.00

*Denotes disattenuated correlations. Correlations below the main diagonal are Pearson correlations.

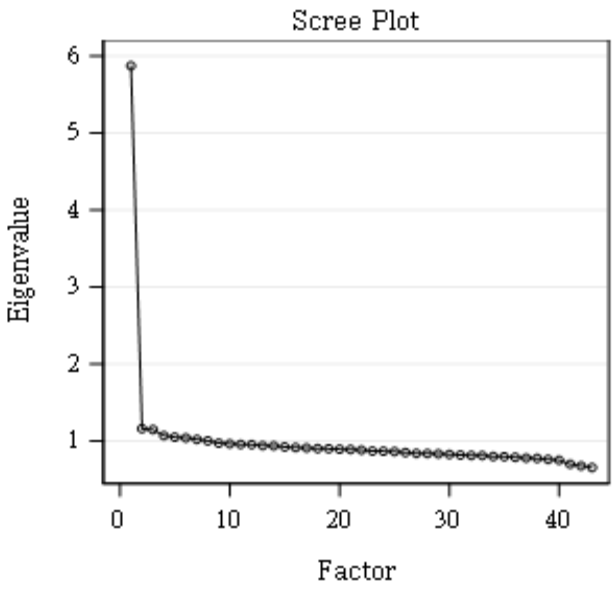
Table D.4.1

First and Second Eigenvalues by Grade: Spring 2019 Operational Social Studies

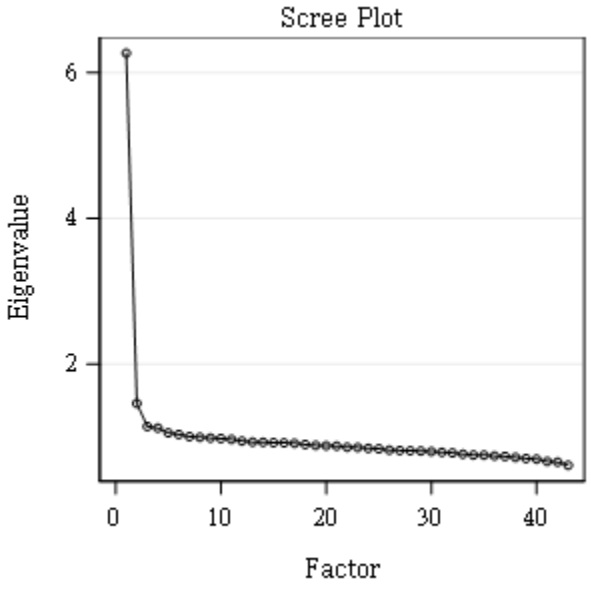
Content	Form	First Eigenvalue	Second Eigenvalue
3	Online	4.713	1.329
	Paper	5.872	1.160
4	Online	6.266	1.462
	Paper	6.292	1.515
5	Online	7.333	1.215
6	Online	8.879	1.224
7	Online	9.512	1.416
8	Online	10.725	1.349



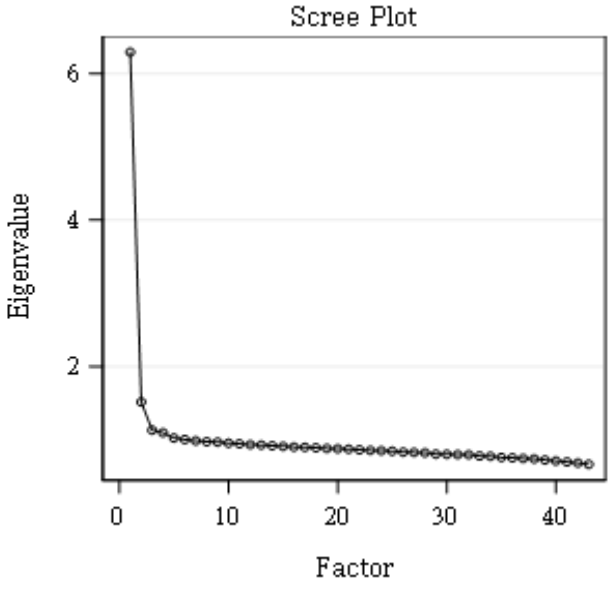
LEAP Social Studies Online: Grade 3



LEAP Social Studies Paper: Grade 3

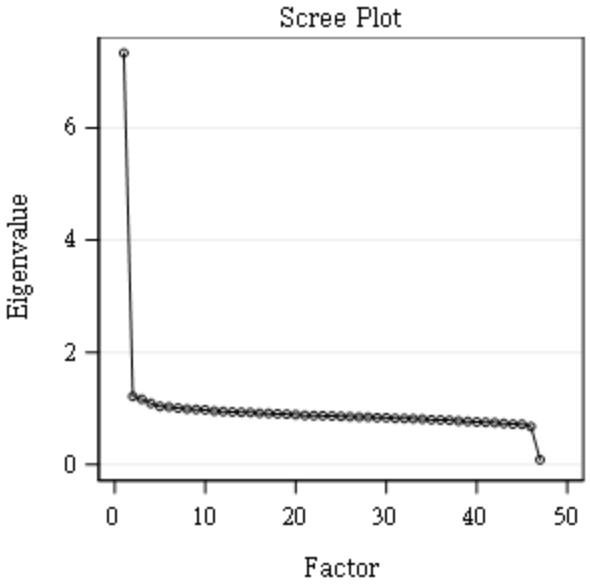


LEAP Social Studies Online: Grade 4

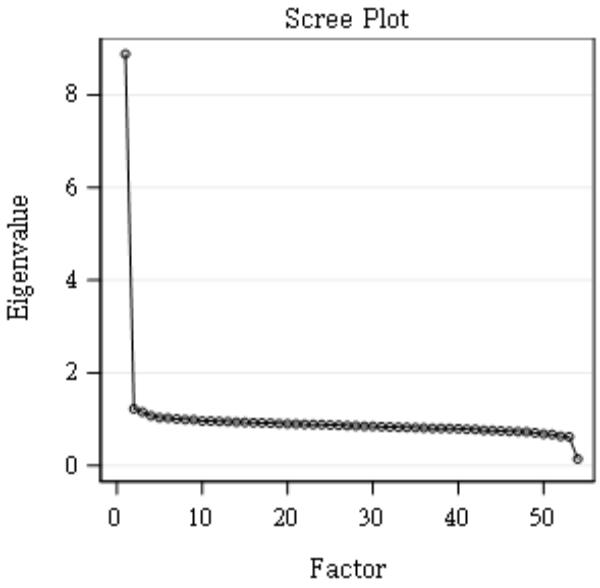


LEAP Social Studies Paper: Grade 4

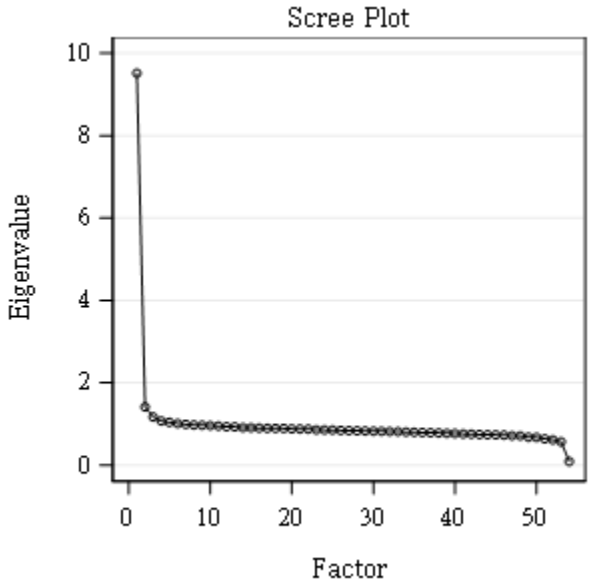
Figure D.4.1
Principal Component Analysis Plot for Spring 2019 Operational Social Studies: Grades 3 and 4



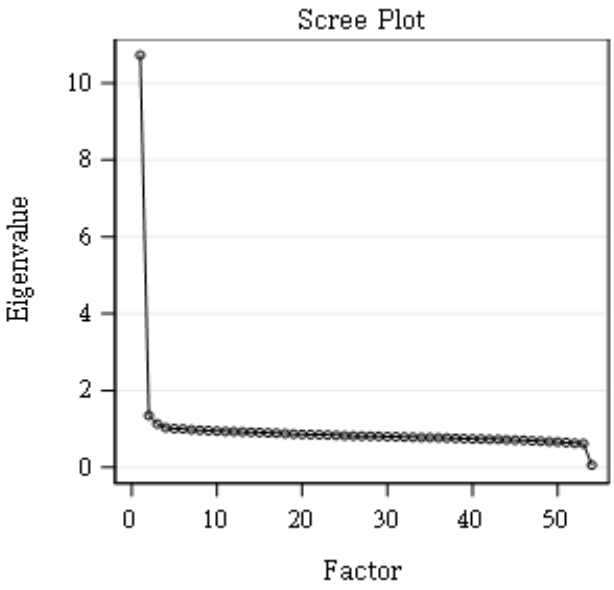
LEAP Social Studies Online: Grade 5



LEAP Social Studies Online: Grade 6



LEAP Social Studies Online: Grade 7



LEAP Social Studies Online: Grade 8

Figure D.4.2
Principal Component Analysis Plot for Spring 2019 Operational Social Studies: Grades 5-8

Appendix E: Scale Distribution and Statistics Report

Table E.1 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 03

N	≥46540		
Mean	723.39	Median	724.00
Std deviation	39.18	Variance	1534.95
Skewness	-0.0423	Kurtosis	-0.3519
Mode	650.00	Std Error Mean	0.1816
Range	200.00	Interquartile Range	56.00

Quantile	Estimate
100% Max	850
99%	810
95%	786
90%	772
75% Q3	752
50% Median	724
25% Q1	696
10%	668
5%	650
1%	650
0% Min	650

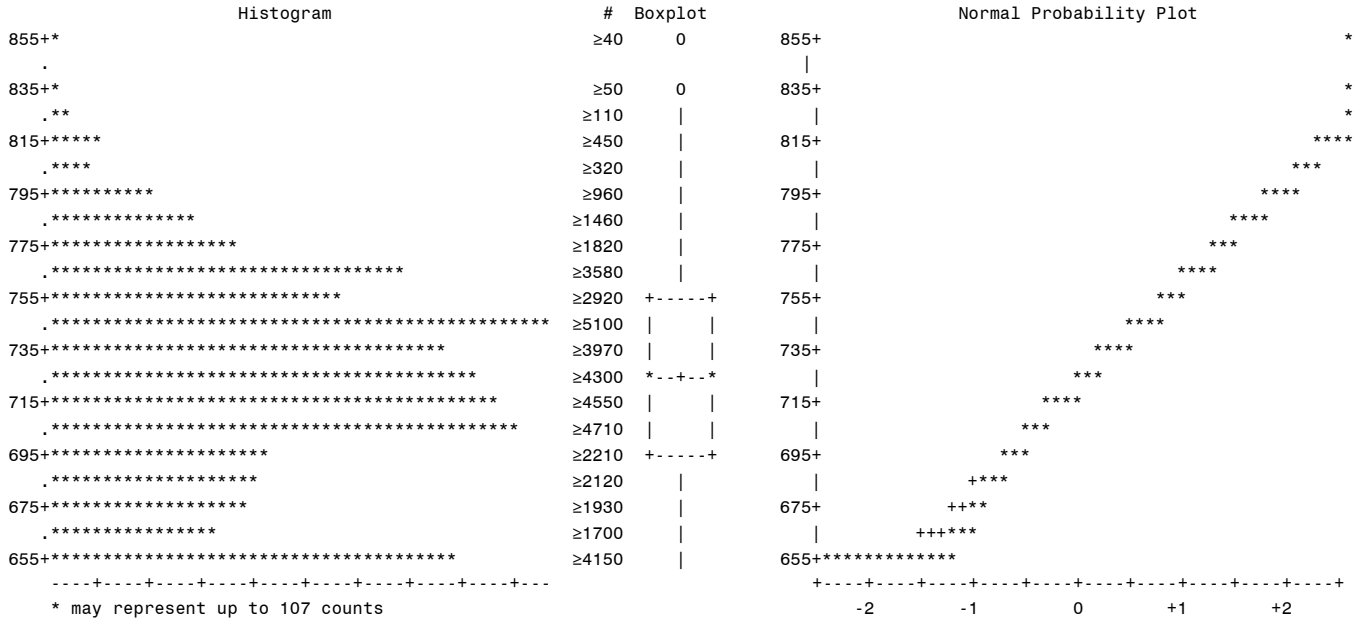


Table E.2 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 03

SCALE_SCORE		Freq	Freq	Cum. Percent	Cum. Percent
650	**	≥2660	≥2660	5.72	5.72
653	*****************************	≥1490	≥4150	3.21	8.93
668	*******************************	≥1700	≥5850	3.65	12.58
679	************************************	≥1930	≥7780	4.15	16.73
688	***	≥2120	≥9900	4.55	21.28
696	**	≥2210	≥12120	4.76	26.04
703	**	≥2300	≥14430	4.96	31.00
708	**	≥2400	≥16830	5.17	36.17
714	**	≥2310	≥19140	4.96	41.13
719	***	≥2240	≥21390	4.83	45.96
724	***	≥2170	≥23560	4.66	50.63
728	***	≥2130	≥25690	4.58	55.21
732	*************************************	≥2010	≥27700	4.32	59.53
736	**************************************	≥1960	≥29670	4.22	63.74
740	**********************************	≥1720	≥31390	3.71	67.45
744	**********************************	≥1760	≥33150	3.79	71.24
748	********************************	≥1610	≥34770	3.47	74.71
752	*****************************	≥1520	≥36290	3.27	77.98
756	**************************	≥1390	≥37690	3.01	80.98
760	**************************	≥1320	≥39020	2.85	83.83
764	**********************	≥1190	≥40210	2.57	86.40
768	********************	≥1060	≥41280	2.29	88.69
772	********************	≥990	≥42280	2.14	90.83
776	*****************	≥830	≥43110	1.78	92.62
781	****************	≥770	≥43880	1.66	94.28
786	**************	≥690	≥44570	1.48	95.76
791	**********	≥520	≥45090	1.12	96.89
797	*********	≥440	≥45540	0.96	97.84
803	*******	≥320	≥45860	0.70	98.54
810	******	≥280	≥46140	0.60	99.14
818	***	≥170	≥46320	0.37	99.52
827	**	≥110	≥46440	0.25	99.77
839	*	≥50	≥46490	0.13	99.90
850	*	≥40	≥46540	0.10	100.00

400 800 1200 1600 2000 2400
Frequency

Table E.3 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 04

N	≥48290		
Mean	726.99	Median	727.00
Std deviation	37.56	Variance	1410.72
Skewness	-0.1198	Kurtosis	-0.1628
Mode	650.00	Std Error Mean	0.1709
Range	200.00	Interquartile Range	50.00

Quantile	Estimate
100% Max	850
99%	810
95%	784
90%	774
75% Q3	754
50% Median	727
25% Q1	704
10%	676
5%	650
1%	650
0% Min	650

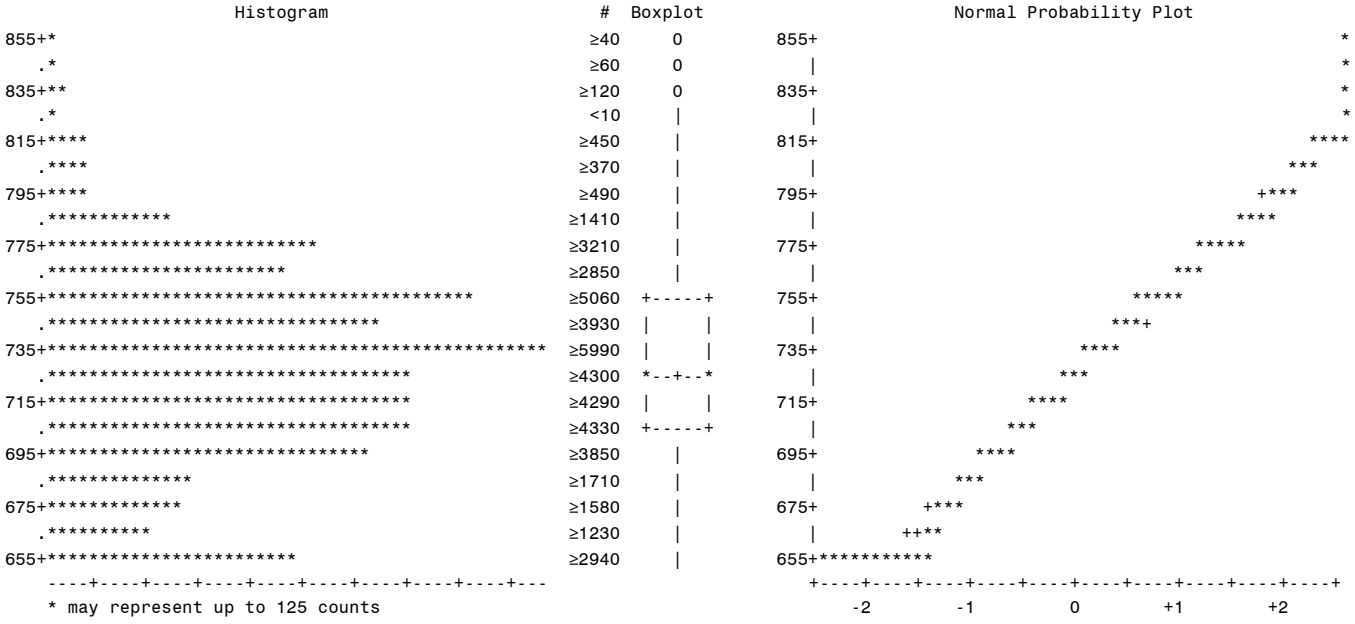


Table E.4 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 04

SCALE_SCORE		Freq	Freq	Cum. Percent	Cum. Percent
650	**********	≥2940	≥2940	6.10	6.10
664	**********	≥1230	≥4170	2.55	8.65
676	**********	≥1580	≥5760	3.28	11.93
685	**********	≥1700	≥7470	3.54	15.47
692	**********	≥1940	≥9410	4.02	19.49
698	**********	≥1910	≥11320	3.96	23.45
704	**********	≥2170	≥13500	4.50	27.96
709	**********	≥2150	≥15650	4.46	32.42
714	**********	≥2140	≥17800	4.44	36.86
719	**********	≥2140	≥19950	4.44	41.31
723	**********	≥2160	≥22120	4.49	45.80
727	**********	≥2130	≥24250	4.42	50.22
731	**********	≥2040	≥26300	4.24	54.46
735	**********	≥2030	≥28330	4.21	58.68
739	**********	≥1910	≥30250	3.95	62.63
743	**********	≥1960	≥32210	4.07	66.70
747	**********	≥1960	≥34180	4.07	70.78
750	**********	≥1750	≥35940	3.64	74.42
754	**********	≥1680	≥37620	3.48	77.91
758	**********	≥1620	≥39250	3.37	81.28
762	**********	≥1490	≥40750	3.10	84.38
766	**********	≥1350	≥42100	2.80	87.18
770	**********	≥1150	≥43260	2.39	89.57
774	**********	≥1080	≥44340	2.24	91.81
779	**********	≥970	≥45320	2.02	93.84
784	**********	≥770	≥46100	1.61	95.45
789	**********	≥630	≥46730	1.31	96.76
795	**********	≥490	≥47220	1.01	97.78
802	**********	≥370	≥47600	0.78	98.56
810	**********	≥260	≥47860	0.54	99.10
819	**********	≥190	≥48050	0.40	99.51
830	**********	≥120	≥48180	0.27	99.77
844	**********	≥60	≥48250	0.13	99.91
850	**********	≥40	≥48290	0.09	100.00

400 800 1200 1600 2000 2400 2800
Frequency

Table E.5 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 05

N	≥48430		
Mean	728.87	Median	728.00
Std deviation	34.70	Variance	1204.13
Skewness	-0.2098	Kurtosis	-0.0948
Mode	709.00	Std Error Mean	0.1577
Range	200.00	Interquartile Range	45.00

Quantile	Estimate
100% Max	850
99%	803
95%	783
90%	772
75% Q3	754
50% Median	728
25% Q1	709
10%	680
5%	667
1%	650
0% Min	650

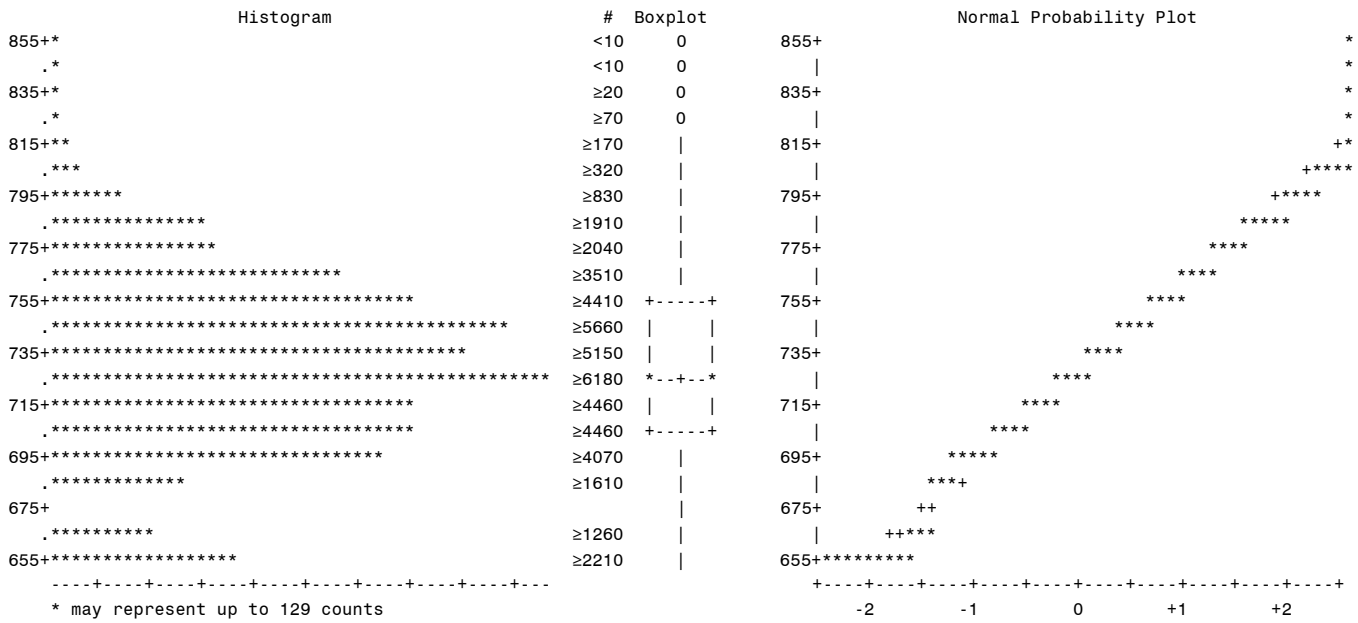


Table E.6 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 05

SCALE_SCORE		Freq	Freq	Cum. Percent	Cum. Percent
650	**	≥2210	≥2210	4.56	4.56
667	************************	≥1260	≥3470	2.61	7.17
680	*******************************	≥1610	≥5090	3.34	10.51
690	**********************************	≥1920	≥7010	3.97	14.48
697	**	≥2150	≥9160	4.44	18.92
703	**	≥2180	≥11340	4.52	23.43
709	**	≥2270	≥13620	4.70	28.13
713	**	≥2230	≥15860	4.61	32.75
717	**	≥2230	≥18090	4.61	37.36
721	**	≥2150	≥20240	4.44	41.80
725	*************************************	≥2070	≥22310	4.28	46.08
728	**********************************	≥1960	≥24280	4.06	50.13
731	*********************************	≥1810	≥26090	3.74	53.88
735	********************************	≥1660	≥27750	3.44	57.32
738	*********************************	≥1670	≥29430	3.46	60.78
740	******************************	≥1520	≥30950	3.14	63.92
743	******************************	≥1480	≥32440	3.07	66.99
746	**************************	≥1410	≥33850	2.92	69.91
749	************************	≥1240	≥35090	2.56	72.47
751	**********************	≥1160	≥36250	2.40	74.86
754	**********************	≥1150	≥37410	2.39	77.26
757	**********************	≥1080	≥38500	2.24	79.50
759	********************	≥1010	≥39510	2.09	81.59
762	********************	≥990	≥40510	2.05	83.64
764	******************	≥900	≥41410	1.86	85.51
767	*****************	≥830	≥42240	1.72	87.23
769	****************	≥780	≥43020	1.61	88.84
772	***************	≥760	≥43790	1.58	90.42
775	**************	≥690	≥44480	1.44	91.86
778	************	≥580	≥45070	1.21	93.07
780	************	≥580	≥45650	1.20	94.27
783	**********	≥490	≥46150	1.03	95.29
786	*********	≥440	≥46590	0.92	96.21
789	********	≥390	≥46980	0.81	97.02
793	*******	≥330	≥47320	0.70	97.72
796	******	≥270	≥47600	0.57	98.29
799	****	≥220	≥47820	0.46	98.75
803	****	≥180	≥48010	0.38	99.13
807	***	≥130	≥48140	0.28	99.41
811	**	≥90	≥48240	0.20	99.62
815	*	≥70	≥48320	0.15	99.77
820	*	≥50	≥48370	0.11	99.88
825		≥20	≥48400	0.05	99.93
831		≥10	≥48410	0.04	99.97
838		<10	≥48420	0.01	99.98
847		<10	≥48420	0.01	99.99
850		<10	≥48430	0.01	100.00

400 800 1200 1600 2000
Frequency

Table E.7 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 06

N	≥48960		
Mean	726.62	Median	727.00
Std deviation	35.86	Variance	1285.92
Skewness	-0.0989	Kurtosis	-0.3204
Mode	650.00	Std Error Mean	0.1621
Range	200.00	Interquartile Range	50.00

Quantile	Estimate
100% Max	850
99%	807
95%	783
90%	772
75% Q3	752
50% Median	727
25% Q1	702
10%	679
5%	658
1%	650
0% Min	650

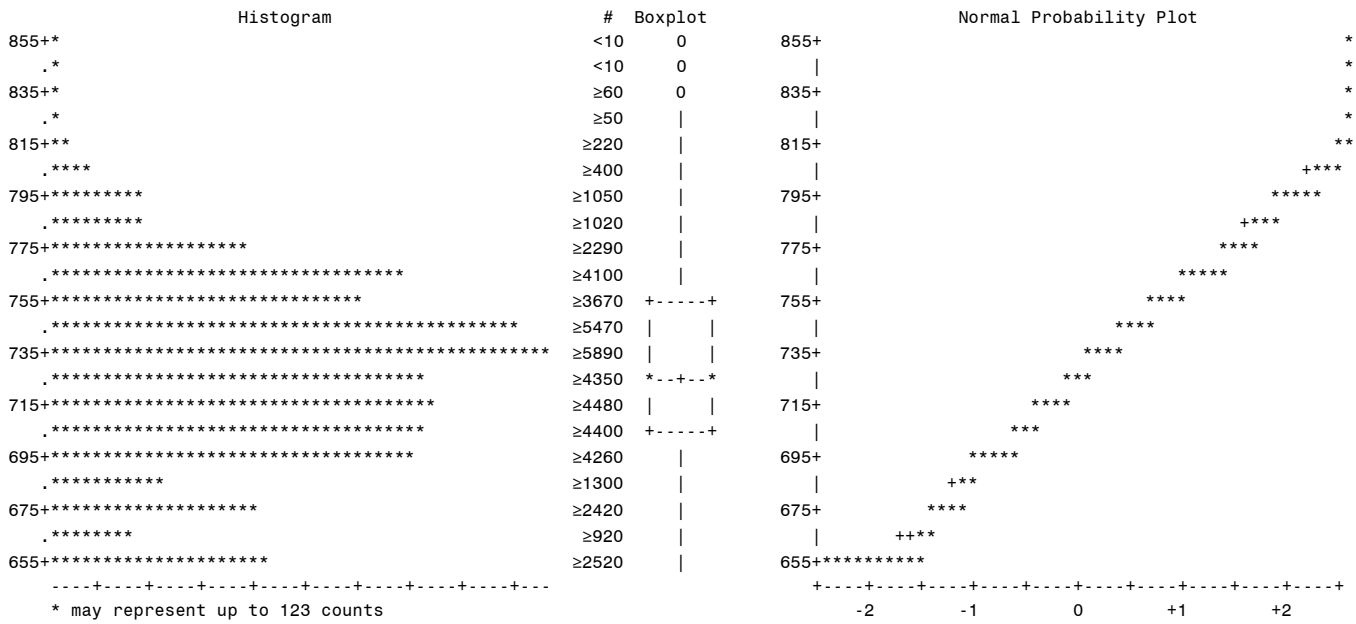


Table E.8 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 06

SCALE_SCORE	Freq	Freq	Cum. Percent	Cum. Percent
650	≥1670	≥1670	3.43	3.43
658	≥840	≥2520	1.73	5.16
666	≥920	≥3440	1.88	7.04
673	≥1150	≥4590	2.35	9.39
679	≥1270	≥5870	2.60	11.99
685	≥1300	≥7170	2.66	14.64
690	≥1370	≥8550	2.82	17.46
694	≥1460	≥10010	3.00	20.46
698	≥1410	≥11430	2.89	23.35
702	≥1450	≥12890	2.98	26.32
706	≥1430	≥14320	2.93	29.25
709	≥1510	≥15830	3.09	32.34
713	≥1530	≥17360	3.13	35.47
716	≥1450	≥18820	2.97	38.44
719	≥1500	≥20320	3.06	41.50
722	≥1450	≥21770	2.96	44.47
725	≥1460	≥23240	2.99	47.46
727	≥1440	≥24680	2.94	50.40
730	≥1460	≥26140	2.99	53.39
733	≥1470	≥27610	3.01	56.40
735	≥1460	≥29080	2.99	59.39
738	≥1490	≥30570	3.05	62.44
741	≥1430	≥32010	2.94	65.38
744	≥1310	≥33330	2.69	68.07
746	≥1450	≥34780	2.97	71.04
749	≥1260	≥36050	2.58	73.62
752	≥1260	≥37310	2.58	76.20
754	≥1250	≥38560	2.56	78.76
757	≥1150	≥39720	2.36	81.12
760	≥1100	≥40830	2.26	83.38
763	≥1070	≥41900	2.20	85.58
766	≥970	≥42880	1.99	87.57
769	≥940	≥43820	1.93	89.50
772	≥860	≥44690	1.77	91.27
776	≥760	≥45450	1.56	92.83
779	≥660	≥46120	1.36	94.19
783	≥560	≥46690	1.16	95.35
786	≥450	≥47140	0.92	96.28
790	≥430	≥47580	0.89	97.17
794	≥350	≥47940	0.73	97.90
798	≥260	≥48200	0.53	98.43
803	≥220	≥48420	0.47	98.90
807	≥170	≥48600	0.36	99.26
812	≥130	≥48740	0.28	99.54
818	≥90	≥48830	0.18	99.73
823	≥50	≥48890	0.12	99.85
830	≥30	≥48930	0.08	99.92
837	≥20	≥48950	0.05	99.98
845	<10	≥48960	0.02	99.99
850	<10	≥48960	0.01	100.00

100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600
Frequency

Table E.9 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 07

N	≥46910		
Mean	732.95	Median	734.00
Std deviation	39.02	Variance	1522.82
Skewness	-0.0957	Kurtosis	-0.3544
Mode	702.00	Std Error Mean	0.1802
Range	200.00	Interquartile Range	54.00

Quantile	Estimate
100% Max	850
99%	816
95%	795
90%	783
75% Q3	760
50% Median	734
25% Q1	706
10%	677
5%	656
1%	650
0% Min	650

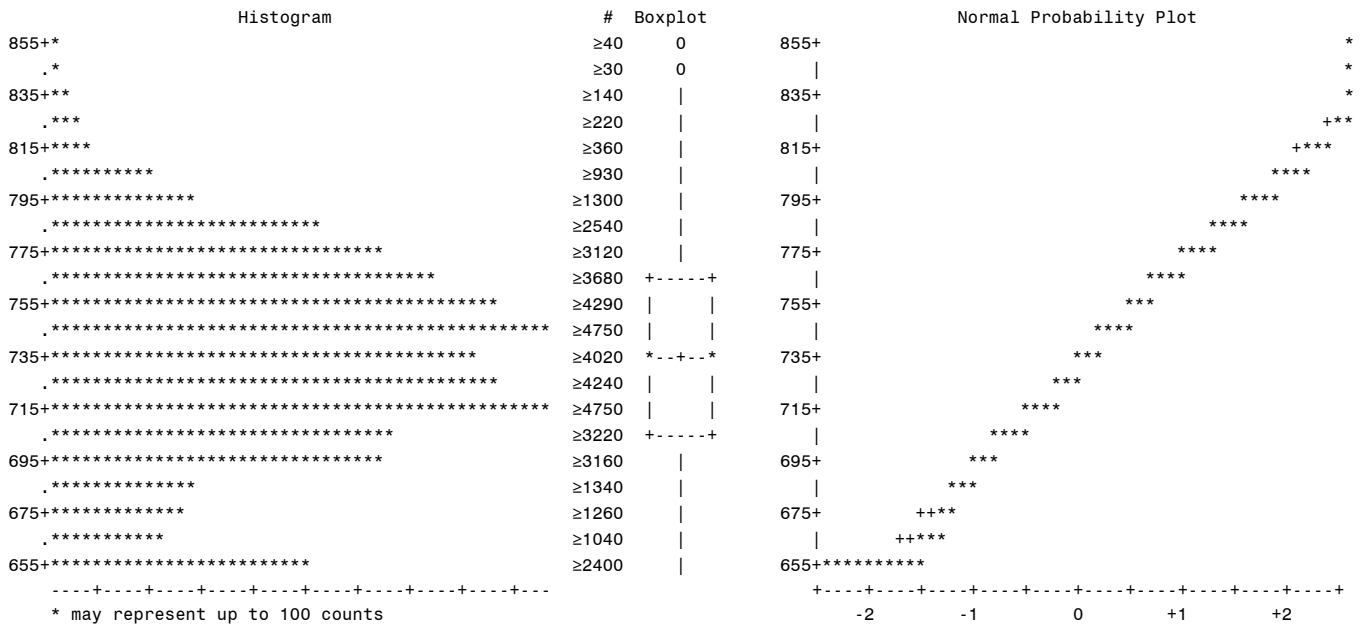


Table E.10 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 07

SCALE_SCORE	Freq	Freq	Cum. Percent	Cum. Percent
650	≥1570	≥1570	3.35	3.35
656	≥830	≥2400	1.78	5.13
667	≥1040	≥3450	2.22	7.36
677	≥1260	≥4710	2.69	10.05
684	≥1340	≥6060	2.87	12.92
691	≥1550	≥7610	3.31	16.23
697	≥1610	≥9220	3.44	19.67
702	≥1630	≥10860	3.49	23.16
706	≥1580	≥12450	3.37	26.54
711	≥1600	≥14050	3.42	29.96
715	≥1630	≥15680	3.47	33.43
718	≥1520	≥17200	3.24	36.68
722	≥1440	≥18650	3.09	39.77
725	≥1410	≥20060	3.01	42.77
728	≥1380	≥21450	2.96	45.73
731	≥1370	≥22820	2.92	48.65
734	≥1340	≥24160	2.86	51.51
737	≥1300	≥25470	2.79	54.30
740	≥1260	≥26740	2.69	57.00
743	≥1170	≥27910	2.51	59.50
745	≥1160	≥29080	2.48	61.98
748	≥1140	≥30220	2.45	64.43
750	≥1140	≥31370	2.44	66.87
753	≥1110	≥32490	2.39	69.25
755	≥1060	≥33550	2.27	71.52
758	≥960	≥34520	2.05	73.58
760	≥940	≥35460	2.01	75.59
763	≥990	≥36450	2.12	77.71
765	≥860	≥37320	1.85	79.56
768	≥880	≥38200	1.88	81.44
770	≥850	≥39050	1.81	83.25
773	≥810	≥39870	1.75	85.00
775	≥750	≥40630	1.61	86.60
778	≥700	≥41330	1.50	88.10
781	≥660	≥41990	1.41	89.51
783	≥690	≥42690	1.48	90.99
786	≥640	≥43330	1.37	92.37
789	≥540	≥43870	1.15	93.52
792	≥480	≥44360	1.03	94.55
795	≥420	≥44790	0.91	95.46
798	≥390	≥45180	0.83	96.30
801	≥350	≥45530	0.76	97.06
804	≥320	≥45860	0.69	97.75
808	≥250	≥46110	0.54	98.29
812	≥180	≥46300	0.40	98.70
816	≥170	≥46470	0.36	99.06
821	≥110	≥46590	0.25	99.32
826	≥100	≥46700	0.22	99.54
831	≥80	≥46780	0.18	99.72
838	≥50	≥46840	0.13	99.84
847	≥30	≥46870	0.07	99.91
850	≥40	≥46910	0.09	100.00

100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600
Frequency

Table E.11 Scale Score Descriptive Statistics and Plots

DESCRIPTIVE STATISTICS - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 08

N	≥45730		
Mean	740.27	Median	743.00
Std deviation	37.46	Variance	1403.52
Skewness	-0.1965	Kurtosis	-0.1342
Mode	758.00	Std Error Mean	0.1752
Range	200.00	Interquartile Range	50.00

Quantile	Estimate
100% Max	850
99%	821
95%	796
90%	786
75% Q3	765
50% Median	743
25% Q1	715
10%	691
5%	676
1%	650
0% Min	650

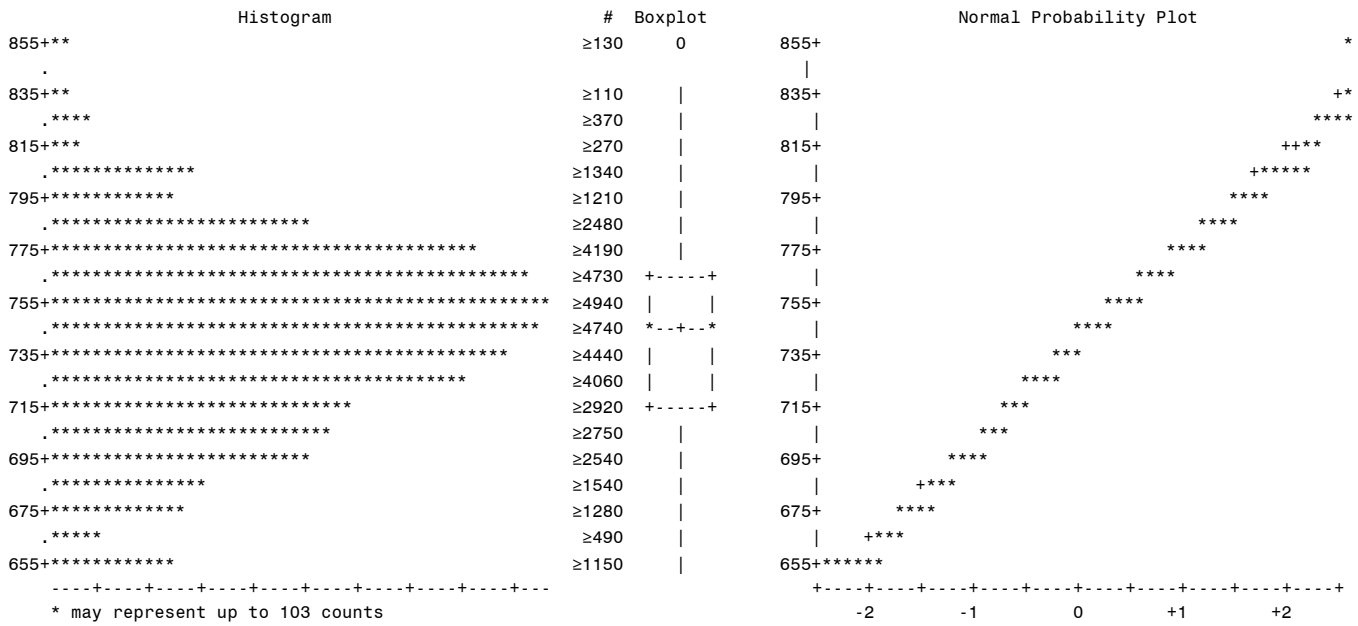


Table E.12 Frequency Distribution of Scale Scores

FREQUENCY DISTRIBUTION - SCALE SCORES
Social Studies
ALL STUDENTS
GRADE 08

SCALE_SCORE		Freq	Freq	Cum. Percent	Cum. Percent
650	**********	≥780	≥780	1.71	1.71
653	*******	≥370	≥1150	0.81	2.51
662	********	≥490	≥1640	1.08	3.59
670	*********	≥600	≥2240	1.32	4.91
676	*********	≥680	≥2920	1.49	6.40
682	*********	≥720	≥3640	1.58	7.98
687	**********	≥820	≥4460	1.79	9.77
691	**********	≥810	≥5280	1.78	11.55
695	**********	≥860	≥6140	1.90	13.44
699	**********	≥860	≥7000	1.88	15.32
703	**********	≥880	≥7880	1.92	17.25
706	**********	≥950	≥8830	2.08	19.33
709	**********	≥920	≥9760	2.02	21.34
712	**********	≥920	≥10680	2.03	23.37
715	**********	≥1000	≥11690	2.21	25.58
718	**********	≥980	≥12680	2.15	27.73
721	**********	≥1010	≥13690	2.21	29.94
723	**********	≥940	≥14630	2.07	32.01
726	**********	≥1060	≥15700	2.32	34.33
728	**********	≥1040	≥16740	2.29	36.62
731	**********	≥1080	≥17830	2.37	38.99
733	**********	≥1140	≥18970	2.49	41.49
736	**********	≥1110	≥20090	2.44	43.93
738	**********	≥1090	≥21180	2.40	46.33
740	**********	≥1120	≥22300	2.45	48.78
743	**********	≥1170	≥23480	2.58	51.35
745	**********	≥1270	≥24750	2.78	54.13
748	**********	≥1170	≥25930	2.58	56.71
750	**********	≥1200	≥27140	2.64	59.34
753	**********	≥1270	≥28410	2.79	62.13
755	**********	≥1170	≥29590	2.58	64.71
758	**********	≥1280	≥30880	2.81	67.51
760	**********	≥1200	≥32080	2.63	70.15
763	**********	≥1190	≥33280	2.62	72.77
765	**********	≥1150	≥34440	2.53	75.30
768	**********	≥1170	≥35610	2.56	77.86
771	**********	≥1130	≥36740	2.47	80.33
774	**********	≥1120	≥37860	2.45	82.78
776	**********	≥970	≥38830	2.13	84.91
779	**********	≥960	≥39800	2.11	87.02
783	**********	≥900	≥40700	1.98	89.00
786	**********	≥810	≥41510	1.77	90.77
789	**********	≥770	≥42290	1.69	92.46
793	**********	≥640	≥42930	1.40	93.86
796	**********	≥570	≥43500	1.26	95.12
800	**********	≥490	≥44000	1.09	96.21
805	**********	≥460	≥44460	1.01	97.21
809	**********	≥380	≥44840	0.84	98.05
815	**********	≥270	≥45120	0.60	98.65
821	**********	≥200	≥45320	0.44	99.10
828	**********	≥170	≥45490	0.38	99.48
838	**********	≥110	≥45600	0.24	99.72
850	**********	≥130	≥45730	0.28	100.00

100 200 300 400 500 600 700 800 900 1000 1100 1200
Frequency

Appendix F: Reliability and Classification Accuracy

Reliability and Classification Accuracy Reports

Social Studies

Contents
Table F.1 Reliability for All Students and for Subgroups: Spring 2019 Operational Social Studies
Table F.2 Cronbach's Alpha and Marginal Reliability: Spring 2019 Operational Social Studies
Tables F.3.1–F.3.7 Classification Accuracy and Decision Consistency: Spring 2019 Operational Social Studies

Table F.1

Reliability for All Students and for Subgroups: Spring 2019 Operational Social Studies

Subgroup	3	4	5	6	7	8
All Students	0.836	0.848	0.877	0.896	0.905	0.918
Female	0.830	0.835	0.867	0.889	0.900	0.912
Male	0.842	0.861	0.886	0.904	0.910	0.925
African American	0.795	0.805	0.830	0.870	0.871	0.902
Asian	0.870	0.876	0.905	0.899	0.930	0.927
Hispanic/Latino	0.819	0.841	0.878	0.906	0.907	0.929
Multi-Racial	0.823	0.843	0.868	0.897	0.904	0.915
Native Hawaiian or Other Pacific Islander	0.824	0.843	0.896	0.933	0.889	0.934
White	0.839	0.843	0.876	0.884	0.904	0.908
Ethnicity Unknown	0.815	0.816	0.887	0.898	0.899	0.913
English Learners	0.782	0.772	0.759	0.834	0.763	0.863

Table F.2

Cronbach's Alpha and Marginal Reliability: Spring 2019 Operational Social Studies

Grade	Cronbach's Alpha	Marginal Reliability
3	0.84	0.83
4	0.85	0.91
5	0.88	0.88
6	0.90	0.91
7	0.91	0.90
8	0.92	0.93

Table F.3

Classification Accuracy and Decision Consistency: Spring 2019 Operational Social Studies

Table F.3.1

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

Grade	Accuracy	Consistency	PChance	Kappa
3	0.590	0.489	0.224	0.342
4	0.620	0.512	0.228	0.368
5	0.647	0.543	0.235	0.402
6	0.685	0.578	0.219	0.460
7	0.667	0.565	0.211	0.449
8	0.724	0.622	0.222	0.515

Table F.3.2

Accuracy of Classification at Each Achievement Level by Grade

Grade	Unsatisfactory (1)	Below Basic (2)	Basic (3)	Mastery (4)	Advanced (5)
3	0.800	0.637	0.492	0.462	0.000*
4	0.815	0.629	0.526	0.566	0.000*
5	0.799	0.615	0.647	0.588	0.625
6	0.848	0.676	0.688	0.556	0.696
7	0.852	0.536	0.621	0.596	0.728
8	0.862	0.637	0.671	0.735	0.753

*Inestimable, default output values due to restricted sample size.

Table F.3.3

Accuracy of Dichotomous Categorizations by Grade

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.916	0.861	0.865	0.926
4	0.925	0.873	0.869	0.938
5	0.936	0.886	0.884	0.935
6	0.944	0.904	0.902	0.931
7	0.932	0.902	0.900	0.924
8	0.957	0.926	0.909	0.931

Table F.3.4

Consistency of Dichotomous Categorizations by Grade

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.876	0.810	0.811	0.901
4	0.890	0.825	0.816	0.916
5	0.905	0.843	0.837	0.921
6	0.918	0.865	0.862	0.907
7	0.902	0.865	0.859	0.893
8	0.938	0.895	0.872	0.903

Table F.3.5

Kappa of Dichotomous Categorizations by Grade

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.660	0.620	0.529	0.196
4	0.672	0.649	0.573	0.158
5	0.688	0.670	0.631	0.187
6	0.734	0.724	0.673	0.438
7	0.736	0.719	0.696	0.529
8	0.758	0.758	0.742	0.593

Table F.3.6

Accuracy of Dichotomous Categorizations by Grade: False Positive Rates

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.043	0.062	0.057	0.074
4	0.035	0.057	0.056	0.062
5	0.034	0.052	0.047	0.062
6	0.027	0.047	0.043	0.046
7	0.034	0.046	0.044	0.044
8	0.020	0.037	0.042	0.038

Table F.3.7

Accuracy of Dichotomous Categorizations by Grade: False Negative Rates

Grade	1/ 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	0.041	0.077	0.078	0.000*
4	0.040	0.069	0.075	0.000*
5	0.030	0.062	0.068	0.002
6	0.030	0.049	0.055	0.022
7	0.034	0.051	0.056	0.032
8	0.023	0.037	0.049	0.032

*Inestimable, default output values due to restricted sample size.