LEAP 2025 Social Studies 3–8 Technical Report: 2023–2024

Prepared by DRC, Pearson, and WestEd

EXECUTIVE SUMMARY

The Louisiana Educational Assessment Program is composed of tests that are carefully constructed to fairly assess the achievement of Louisiana students. This technical report provides information on the field test administrations, scoring activities, analyses, and results of the spring 2024 administration of the LEAP 2025 Social Studies standalone field tests.

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

The chapters of this technical report outline general information about the administration and scoring activities of the LEAP 2025 assessments, classical test theory (CTT) and item response theory (IRT) analysis results.

Table of Contents

1. Introduction

The Louisiana Department of Education (LDOE) has a long and distinguished history in the development and administration of assessments that support its state accountability system and are aligned to its state content standards. Per state law, the LDOE is to administer statewide social studies assessments in grades 3–8 and high school annually. Fulfilling the directive of the Louisiana State Board of Elementary and Secondary Education (BESE), the LDOE must deliver high-quality, Louisiana-specific standards-based assessments. 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 development and field testing of the next generation of statewide summative social studies assessments for grades 3–8 as part of the Louisiana Educational Assessment Program 2025 (LEAP 2025). This report outlines the testing procedures, including forms construction, administration, and scoring and analyses. This report begins with the context of state statutes and standards that support the assessments, followed by a description of the structure and blueprints for the assessments. It then provides an outline of the item development process, including review processes involving the LDOE and Louisiana educators. It also outlines the field testing procedures, including forms construction, sampling, and administration, and provides scoring, analysis, and evaluation of the field test items.

Standards Transition

In March 2022, the BESE approved the adoption of the K-12 Louisiana Student Standards for Social Studies. The first administration of operational assessments aligned to the 2022 standards for grades 3–8 is scheduled for spring 2025. The 2022 standards reflect an increase in emphasis on critical thinking skills across the grades. The new standards create a sequence of content that is chronologically coherent and raise expectations for elementary students, balancing the acquisition of disciplinary skills and content knowledge, and integrating the historical perspectives of people from all backgrounds.

The teaching of content has shifted in some grades. Grade 3 focuses on foundations of the American experience. The teaching of world history has moved to grades 4 and 5 from grade 6, and the teaching of U.S. history and Louisiana history takes place in grades 6, 7, and 8.

LDOE Goals for the Assessments

While the process of the adoption of the new social studies standards was being implemented, the LDOE described its goals for the assessments that would align to the new standards. Initially, the LDOE planned to move away from an end-of-year summative assessment and adopt a through-year model, in which assessments would be administered in the fall, winter, and spring and align closely to the instructional frameworks at each grade. As with the previous summative assessment, the through-year assessments would focus on the students' ability to evaluate and analyze stimuli to answer questions, make inferences, and draw conclusions. With the through-year model, the LDOE intended to provide teachers and school districts with student data so that they could track students' performances throughout the year. The through-year assessments would also be shorter than the summative assessments so that they could be administered in one class period and minimize the impact on instruction. Similar to the previous summative assessment, a culminating task would be administered at the end of the year that would synthesize the content and skills students were expected to learn over the course of the year. Overall, the goal for the LEAP 2025 social studies assessments was to provide a valid and reliable assessment reflecting the content and analysis of documents as indicated in the standards, while limiting the amount of time required for testing to no more than four hours per assessment.

With these goals in mind, stimulus searching and item development was begun in March 2022. However, in December 2022, the LDOE decided to change the test designs and shift back to an end-of-year summative assessment model at grades 3–8. The LDOE still wanted to reduce both the testing time and the length of the operational summative assessments at grades 3–8. As part of this plan, it removed the extended response items from the tasks in grades 3 and 4. To support this effort, WestEd updated its item development and standalone field-testing plans.

Summary of the 2023–2024 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 2024 grades 3–8 standalone field-tests. Table 1.1 summarizes the key activities along with the months during which the activities were completed.

Table 1.1

Key Activities from December 2020 to September 2024

Date	Activity
December 2020	• BESE authorizes the review and revision of the Louisiana Student Standards for Social Studies
February/March 2022	• The LDOE creates Assessment Frameworks for grades 3–8
March 2022	• BESE approves adoption of the K-12 Louisiana Student Standards for Social Studies • Technical Advisory Committee convenes • WestEd begins topic selection and stimulus searching
May/October 2022	• The LDOE convenes Stimulus Review Committees
September 2022	• WestEd begins item writing
October 2022	• Technical Advisory Committee convenes
December 2022/January 2023	• The LDOE changes the grades 3–8 assessment model from through-year test design to end-of-year summative test design
February 2023	• Technical Advisory Committee convenes
June-July 2023	• Item Content/Bias Review Committees convenes • Item reconciliation between WestEd and the LDOE takes place

July-September 2023	• WestEd staff selects standalone field-test forms
November 2023	• Technical Advisory Committee convenes
February 2024	• Technical Advisory Committee convenes
April 2024	• Spring 2024 standalone field-tests are administered
May 2024	• Theory of Assessment document for grades 3–8 and Civics is created by the LDOE
June 2024	• Range-finding Committees convenes
August 2024	• Data Review Committees convenes
September 2024	• Data review reconciliation meetings are held between the LDOE and WestEd staff

2. Theory of Assessment

The initial assessment frameworks developed by the LDOE in February 2022 at the start of the project included information that reflected a through-year assessment model in grades 3–8. With the shift to the summative end-of-year assessment test design in December 2022, the LDOE developed a new assessment framework for grades 3–8, which it calls the Theory of Assessment. It outlines the relationship between the grade-level course frameworks, content standards, and the LEAP 2025 assessment. As such, for each grade, it describes:

- Alignment between course framework units and content standards
- Assessable content standards by reporting category
- Assessable content and excluded content for each content standard
- Test blueprints that show the range of points per reporting category and content grouping
- Test blueprints that show the range of points per skills and practices reporting category

3. Overview of the Test Development Process

This section describes the processes used to develop the field test item sets, tasks, and standalone items for the 2024 standalone assessments for grades 3–8.

Item Development Plan

WestEd's item development plan for the standalone field test in 2024 for grades 3–8 focused on developing item sets, tasks, classroom assessment tasks, and standalone items. Table 3.1 shows the item development plan for grades 3–8 in 2023–2024.

Table 3.1
Item Development Plan Grades 3-8 (2023–2024)

Grade	Classroom Assessment Task	Task	Mini Set	Item Set	Standalone Item
Grade 3	2	3	4	16	72
Grade 4	2	3	2	17	72
Grade 5	2	3	—	18	72
Grade 6	2	3	—	18	72
Grade 7	2	3	—	18	72
Grade 8	2	3	—	18	72
Total	12	18	6	105	432

Because the original intent was to have testing windows in the fall, winter, and spring for the through-year assessment, the LDOE requested that WestEd develop six item sets and 16 standalone items for each testing window for each grade. In grades 3–4, 10 items were developed for each item set and task, and six items were developed for each mini-set. In grades 5–8, 12 items were developed for each item set and task. Item types for item sets, mini-sets, and tasks included multiple choice items (MC) worth one point, multiple-select items (MS) worth one point, technology-enhanced items (TE) worth one to three points, and two-part dependent items (TPD) worth two points, or two-part independent Items independent (TPI) items, worth two to three points. The combination of item types varied from item set and task, and depended on the topic. Constructed-response items (CR) worth four points were developed for designated item sets and Extended-response items (ER) worth four points were developed for the tasks at each grade.

Following the decision to change the assessment design from a through-year assessment to an end-of-year assessment, the LDOE determined that the number of item sets proposed in the item development plan was no longer needed for the standalone field test and selected six item sets to be set aside and reserved for future item development. At the same time, it was determined that ER items would not be assessed in grades 3 and 4. However, the LDOE also requested an increase in the development of the number of standalone items. Table 3.2 shows the number of items developed by item type and grade.

Table 3.2

Item Development Plan Grades 3-8 by Item Type (2023-2024)

Grade	MC	MS	TE	TPD	TPI	CR	ER
Grade 3	234	31	71	23	60	12	10
Grade 4	219	22	81	19	20	12	10
Grade 5	241	33	85	25	26	12	7
Grade 6	239	45	77	10	17	13	6
Grade 7	248	35	67	19	27	14	13
Grade 8	249	37	66	14	29	12	7
Total	1,430	203	447	110	179	87	53

Proposal and Review of Topics and Sources

Determining Topics

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

- Which topics are in need of development based on the K-12 Louisiana Student Standards for Social Studies
- What content is eligible according to the grade-level instructional framework course documents
- Whether proposed topics will support the required item types and number of items, including overage
- How content standards 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 existing pool. The process of choosing assessable content standards for each topic is iterative and includes the identification of potential content standards 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 content standards 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. Standalone items are included in the test design to provide greater coverage of the assessable content and content standards 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 content standards that may not be sufficiently covered across the sets and tasks, with the goal of providing maximum flexibility during test construction.

Content Standard Coverage

Grade 3. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 3 except for content standard 3.18.

Grade 4. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 4 except for content standards 4.08, 4.13.d, 4.16.c, 4.16.d, and 4.16.f.

Grade 5. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 5.

Grade 6. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 6.

Grade 7. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 7 except for content standard 7.13.j.

Grade 8. By the end of the 2023–2024 development cycle, WestEd had developed at least 1 item aligned to each of the assessable content standards in grade 8 except for content standards 8.08, 8.10, 8.12.c, 8.14.h, 8.1, and 8.17.

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 lists of proposed topics also include the content standards 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. Before WestEd begins writing standalone items, it submits an item development plan to the LDOE for approval that outlines how many items will be developed, which standards the items will align to, what topics will be covered, and which item types will be created.

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.

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.

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 stimuli created internally to a maximum of 20% permissioned work. The actual percentages vary from year to year and grade to grade, depending on the needs of the content in development. Across all grades, the total percentage of permissioned work is not less than 20%. 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.

Many of the stimuli chosen as part of the 2023–2024 development cycle were found to be at grade level; however, many of the authentic historical documents were evaluated as being above grade level. In those cases, the documents were modified to improve readability and accessibility for the targeted grade levels. 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. These outline the content of the sets and tasks, identify the number and types of items to be assessed in the sets and tasks, identify the content standards and stimuli associated with the items, and provide descriptions of potential culminating items for the set.

Following the initial review by LDOE staff of the task and item set overviews and stimuli, WestEd then makes any revisions to the selection of stimuli based on feedback from the LDOE and then enters the stimuli into the Assessment Banking and Building solutions for Interoperable assessment (ABBI), Pearson's proprietary item development platform.

Stimulus Review Committees. After the stimuli are entered into ABBI, virtual stimulus review committees are held to review the quality and grade appropriateness of the proposed item set and task stimuli. 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 the LDOE and WestEd. Stimulus Review Committees were held

between May and October 2022. Table 3.3 shows the representation of educators who participated in the stimulus review committees in 2022.

Table 3.3

Representation of Educators Participating in the Stimulus Search Committees 2022

Grade	Number of Committee Participants	Classroom Teacher	Special Education	Instructional Lead or Supervisor	Visually Impaired Teacher	EL Teacher/ Supervisor	Other
3	5	3	0	0	0	2	0
4	5	1	0	2	1	0	1
5	5	3	0	2	0	0	0
6	5	4	0	1	0	0	0
7	5	3	0	2	0	0	0
8	5	3	0	1	0	0	0

Training and Security for Virtual Stimulus Committee Review. The virtual format of the stimulus review committee review allows participants to access the item development platform and vote on stimuli asynchronously before coming together in an online meeting format to discuss the stimuli as a group. Prior to accessing the platform, WestEd provides training to explain the stimulus search 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 the item set and task stimuli for content and bias considerations, explains how to use ABBI for item review, and shows educators how to individually review the stimuli and record their recommendation to accept, accept with revisions, or reject a stimulus.

Committee members are provided a review window of one week for each batch of stimuli prior to meeting as a committee, during which they access the stimuli using ABBI and vote on the stimuli. In 2022, each committee had four batches of stimuli to review. 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 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 stimuli 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 stimuli.

For standalone items, WestEd submits the items along with their corresponding stimuli to the LDOE instead of submitting the stimuli to the stimulus review committees first for review.

Item Writing and Review Process

WestEd employs item writers and editors for grades 3–8. WestEd secures the required approval from the LDOE for each writer and editor prior to beginning item development. 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. After the training, item writers are provided with approved set overviews, which identify the set topics and individual item topics, list the primary content standards 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 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 ABBI. In addition to the items and stimuli, the platform captures item metadata and allows viewers to preview items using Pearson's format viewer (TestNav 8). In this view, items appear together with their associated stimuli. The ability to examine the items and stimuli together is critical in the item review and in the evaluation of the content and cognitive demands on students.

Style Guidelines. The *LEAP Social Studies 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. Content and bias review committees were held virtually in June and July 2023. Table 3.4 provides information about the representation of educators who participated in the content and bias reviews in June and July 2023. Table 3.5 provides information about the demographic representation of the participants in the content and bias review committees.

Table 3.4

Representation of Educators Participating in the June/July 2023 Content and Bias Reviews

Grade	Number of Committee Participants	Classroom Teacher	Special Education	Instructional Lead or Supervisor	Visually Impaired Teacher	EL Teacher/ Supervisor	Other
3	6	4	1	1	0	0	0
4	10	8	0	2	0	0	0
5	9	4	1	2	0	1	1
6	10	7	1	1	0	1	0
7	9	6	0	2	0	0	1
8	9	7	1	1	0	0	0

Table 3.5

Demographic Representation of Participants in the June/July 2023 Content and Bias Reviews

Grade	Number of Committee Participants	Male	Female	White	Black or African American	Asian or Pacific Islander	American Indian or Alaska Native	Hispanic (Non-White)
3	6	0	6	3	3	0	0	0
4	10	0	10	4	4	0	0	2
5	9	2	7	6	2	0	0	1
6	10	4	6	4	3	0	1	2
7	9	2	7	8	1	0	0	0
8	9	3	6	6	3	0	0	0

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 with 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 3.6 provides the results based on the participants' individual votes following their initial review of the stimuli and the items. Table 3.7 shows the results of the group votes after discussing and reaching consensus on the disposition of the stimuli and the items.

Table 3.6

Vote Totals Based on Individual Votes Following Initial Review of Stimuli and Items (June/July 2023)

Grade	Number of Items	Accept	Accept with Edits*	No Vote	Reject	Grand Total
3	265	1,422	151	11	2	1,586
4	266	2,897	113	10	16	3,036
5	286	2,664	238	41	7	2,878
6	288	2,316	447	50	7	2,820
7	288	1,794	170	16	26	2,006
8	288	2,234	304	40	24	2,602

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

Table 3.7

Vote Totals for Items Based on Group Consensus for Stimuli and Items (June/July 2023)

Grade	Number of Items	Accept	Accept with Edits	No Vote	Reject
3	265	190	75	0	0
4	266	216	50	0	0
5	286	156	130	0	0
6	288	139	149	0	0
7	288	200	87	0	1
8	288	131	156	0	1

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

While the primary purpose of the field test was to obtain data to inform construction of the operational test forms, the field test also served as an opportunity to prepare the field for the format and rigor of the new assessments. (see Tables 4.1–4.6 for the field test form designs for each grade). To achieve this goal, a daisy chain approach was used, ensuring each item appeared as frequently as possible while adhering to a common item equating method.

In addition to content balance, test form developers were careful to avoid cueing and clanging between items. Cueing occurs when content in one item provides clues to the answer of another item. Clanging refers to overlap or similarity of content. Since content was effectively distributed across the forms, cueing and clanging was intended to have been avoided; however, developers also conducted a separate review of the forms in order to avoid inadvertent cueing or clanging.

Following the final item placement by WestEd content leads, test maps containing each item's unique identification number (UIN) were created. The test maps captured details about each proposed form, including sessions, item sequences, UINs, and associated item metadata. Item descriptions were also included for each item, to aid in the review of the selection and placement of individual items.

Table 4.1

Standalone Field Test Design for Grade 3 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	21 Items
Session 2: Standalone items 1 task 1 item set	11 Items
Session 3: Standalone items 3 item sets	21 Items
Total Number of Items Per Form	53 Items
Number of Forms	15 Forms
Totals Field Tested across Forms for Grade 3	19 item sets and tasks 326 items

Table 4.2

Standalone Field Test Design for Grade 4 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	21 Items
Session 2: Standalone items 1 task 1 item set	11 Items
Session 3: Standalone items 3 item sets	21 Items
Total Number of Items Per Form	53 Items
Number of Forms	16 Forms
Totals Field Tested across Forms for Grade 4	17 item sets and tasks 290 Items

Table 4.3

Standalone Field Test Design for Grade 5 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	24 Items
Session 2: Standalone items 1 task 1 item set	16 Items
Session 3: Standalone items 3 item sets	24 Items
Total Number of Items Per Form	64 Items
Number of Forms	16 Forms
Totals Field Tested across Forms for Grade 5	17 item sets and tasks 320 Items

Table 4.4

Standalone Field Test Design for Grade 6 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	24 Items
Session 2: Standalone items 1 task 1 item set	16 Items
Session 3: Standalone items 3 item sets	24 Items
Total Number of Items Per Form	64 Items
Number of Forms	16 Forms
Totals Field Tested across Forms for Grade 6	17 item sets and tasks 308 Items

Table 4.5

Standalone Field Test Design for Grade 7 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	24 Items
Session 2: Standalone items 1 task 1 item set	16 Items
Session 3: Standalone items 3 item sets	24 Items
Total Number of Items Per Form	64 Items
Number of Forms	16 Forms
Totals Field Tested across Forms for Grade 7	17 item sets and tasks 315 Items

Table 4.6

Standalone Field Test Design for Grade 8 Social Studies

Test Session	Numbers of Items
Session 1: Standalone items 3 item sets	24 Items
Session 2: Standalone items 1 task 1 item set	16 Items
Session 3: Standalone items 3 item sets	24 Items
Total Number of Items Per Form	64 Items
Number of Forms	16 Forms
Totals Field Tested across Forms for Grade 8	17 item sets and tasks 316 Items

Revision and Review

Psychometric Approval of Field Test Forms

Prior to submitting the field test forms to LDOE staff for review, Pearson psychometricians and WestEd content specialists participated in an iterative process of reviewing and revising the forms. The answer keys for MC items were also examined, to determine whether any forms had significantly non-uniform distributions of correct responses (A, B, C, and D). Spreadsheets were used to generate frequency tables of item types, distribution of sets across forms, and MC answer keys for each form and across forms. Pearson psychometricians also reviewed the forms to ensure that clones or enemies of items did not appear in the same form.

Deviations from expectations were identified and addressed when possible and when deemed beneficial. In this process, consideration was given to maximizing the number of field-tested items. For example, an unassigned item may have been suggested as a replacement for an item appearing on multiple forms. Moreover, small deviations from ideal reporting category representation might have been permitted, to allow for field testing of additional item sets. Additional consideration was given to minimizing cueing and clanging on a given form.

When any deviations were identified, item content was adjusted. Depending on the number of changes made, a Pearson psychometrician reviewed the forms again before they were submitted to the LDOE for review and approval. Psychometric approval was provided for all forms prior to administration.

LDOE Review

Following the psychometric reviews, the test maps and constructed item sets for each grade were delivered to the LDOE for approval. Forms were reviewed by both LDOE content and psychometric staff. Based on the LDOE review, select edits to items were made, and the sequence of answer choices and the sequence of items within sets were also evaluated and changed as requested. In light of these changes, the overall balance of answer choices and key runs were evaluated and final adjustments made to achieve the

appropriate balance. All items that had been edited were reviewed by WestEd's proofreaders before the items were transferred from ABBI to DRC.

Online and Paper Versions

At grade 3, one form was delivered on paper, and 15 forms were also delivered online. At grades 4–8, all forms were delivered online. One form in each grade was designated by the LDOE as the accommodated form, to be used with students who required accommodations; for grades 4–8, one of the online forms was offered on paper for students requiring paper testing as an accommodation. To support students with low or no vision, additional text was also provided to describe the graphic components of the assessments. The accommodated form was also rendered in Braille. Table 4.7 shows the distributions of online and paper forms for each grade.

Table 4.7
LEAP 2025 Forms for Spring 2016 Field Test

Grade	Paper Forms	Online Forms
3	1	15
4	N/A**	16
5	N/A**	16
6	N/A**	16
7	N/A**	16
8	N/A**	16

*Same form as one of the paper forms.
**One online form was also offered on paper for students requiring paper testing as an accommodation.

5. Test Administration

This chapter describes the 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 *Standards for Educational and Psychological Testing* (hereafter the *Standards*), "The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions" (AERA et al., p. 111). This chapter examines how test administration procedures implemented for the LEAP 2025 social studies assessments 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 the LEAP 2025 test administration contributed 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: the *LEAP 2025 Grade 3 Paper-Based Test Administration Manual* (TAM) and the *LEAP 2025 Grades 3–8 Computer-Based Test Administration Manual* (TAM). The TAMs provide detailed instructions for administering the LEAP assessments and include information on test security, test administrator responsibilities, test preparation, administration of tests (computer-based or paper-based), and post-test procedures. DRC also produces test coordinator manuals (TCMs) for paper- and computer-based test administrations that provide detailed instructions for district and school test coordinators' responsibilities for distributing, collecting, and returning test materials to DRC for scoring.

The LDOE assessment staff review, provide feedback, and give final approval for these manuals. The manuals are inclusive of grades 3–8 English language arts (ELA), mathematics, social studies, and science.

TAM and TCM Table of Contents

Table of Contents for *LEAP 2025 Paper-Based Test Administration Manual (TAM)*:

- 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
    - 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 Learners (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 Tests
- 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 Computer-Based Test Administration Manual (TAM)*:

    - 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)
- General Instructions
    - Reading Directions to Students
- LEAP 2025: Grades 3–8 English Language Arts (All Sessions)
- LEAP 2025: Grades 3–8 Mathematics (All Sessions)
- LEAP 2025: Grades 3–8 Science (Sessions 1–2)
- LEAP 2025: Grades 5–8 Science Session 3 Select Schools Only
- LEAP 2025: Grades 3–8 Social Studies All Sessions
- Post-Test Procedures
    - Test Administrator Post-Administration Oath of Security and Confidentiality Statement
    - Returning Test Materials to the School Test Coordinator
- Index

Table of Contents for *LEAP 2025 Paper-Based Testing Test Coordinators Manual (TCM)*:

- Key Dates
- Resources Available in DRC INSIGHT Portal
- 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
    - Test Schedule
    - Extended Time for Testing
    - Extended Breaks

    - Makeup Testing
    - Test Administration Resources
- Testing Times for Grade 3
- District Test Coordinator
    - Conduct Training Session
    - Receive Test Materials
    - Large-Print and Braille Test Materials and Communication Assistance Scripts (CAS)
    - 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: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials
    - Pickup 3: Nonscorable Test Materials
    - 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 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 District Test Coordinator
    - Pickup 1: ELA and Mathematics Scorable Test Materials
    - Pickup 2: Science and Social Studies Scorable Test Materials

    - Pickup 3: Nonscorable Test Materials
    - Final Checklist for Returning Test Materials to DTC
- Void Notification
- Index

Table of Contents for *LEAP 2025 Computer-Based Testing Test Coordinators Manual (TCM)*:

- Key Dates
- Resources Available in DRC INSIGHT Portal
- Alerts
- Pre-Administration Oath of Security and Confidentiality Statement
- Post-Administration Oath of Security and Confidentiality Statement
- General Information
    - DRC INSIGHT Portal and INSIGHT
- Test Security
    - Key Definitions
    - Violations of Test Security
- Testing Guidelines
    - Testing Eligibility
    - Testing Conditions
    - Testing Schedule
    - Extended Time for Testing
    - Extended Breaks
    - Accommodations
    - 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
    - DRC INSIGHT Portal User Guide
    - *LEAP 2025 Accommodations and Accessibility Features User Guide*
    - *INSIGHT Technology User Guide*
    - Online Tools Training (OTT)
    - Student Tutorials
- Void Notification

Standards Addressed in the TAMs and TCMs

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

Standard 4.15. The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented (AERA et al., 2014, p. 90).

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

Standard 6.1. Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user (AERA et al., 2014, p. 114).

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it is essential that the LEAP 2025 tests are administered according to the prescribed TAMs. Adhering to the test schedule is also a critical component. The TCMs include instructions for scheduling the test within the state testing window. The TAMs and TCMs also contain 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 (AERA et al., 2014, p. 115).

The LDOE staff release annual test security reports that describe a wide range of improper activities that may occur during testing, including copying and reviewing test items 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/504 Plan (IAP), or English Learner Plan (EL plan); allowing accommodations for students who do not have an IEP, IAP, or EL plan; or defining terms on the test.

Standard 6.4. The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance (AERA et al., 2014, p. 116).

The TAMs outline the steps that teachers should take to prepare the classroom testing environment for administering the LEAP 2025 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 (AERA et al., 2014, p. 116).

The TAMs 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 assessment, examiners are instructed to transcribe students' responses from the large-print or braille test book into the 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 (AERA et al., 2014, p. 117)

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

Return Material Forms and Guidelines

The paper-based TCM instructs test coordinators regarding procedures for organizing and packing materials and returning them to DRC for secure inventory purposes. The LDOE 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 the 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 district 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.

Time

Each session of each content area test is timed to provide sufficient time for students to attempt all items. Only students with an extended time accommodation were permitted to exceed the established time limits of any given session. 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 systems and school personnel set up test sessions via DRC's INSIGHT portal and print test tickets. Students enter their ticket information to access the test in INSIGHT. Students also have access to the Online Tools Training (OTT) 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, Grade 3

Schools with testers at grade 3 have the option to participate in either paper-based or computer-based testing for the spring assessment. 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/IAP/EL plan.

Accommodations may be used with students who qualify under the Individuals with Disabilities Education Act (IDEA) and have an IEP or Section 504 of the Americans with Disabilities Act and have an IAP, or who are identified as ELs and have an EL plan. Accommodations must be specified in the qualifying student's individual plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student information sheet at the time of test administration. 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 (AERA et al., 2014, p. 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 are provided by DRC:

- Braille
- Text-to-Speech
- Directions in Native Language

The following additional access and accommodation features are 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 for online testers
- Human Read Aloud
- Directions in Native Language

For more details about these accommodations, please refer to the *LEAP 2025 Accessibility and Accommodations Manual*.

Testing Windows

The computer-based testing window was available from April 15 through May 17, 2024. Paper-based testing occurred from April 17 through April 22, 2024.

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 assessments and are discussed throughout the TCMs and TAMs.

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 responsible for the security of the LEAP 2025 assessment 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. To help ensure that test scores are valid and relate to actual learning, data forensic analyses take place to assist in separating meaningful gains from spurious gains.

Multiple methods are incorporated into the forensic analysis. The following methods are applied:

- Response Change Analysis
- Score Fluctuation Analysis
- Web Monitoring
- Plagiarism Detection
- Alerts for Disturbing Content

It is important to note that although the results of the analyses may be used to identify potential problems within a school, the identification of a problem is not an accusation of misconduct.

Response Change Analysis. Students make changes to answer choices when taking the LEAP 2025 assessments, and this behavior is expected. Unfortunately, changes to student answers are sometimes influenced by school personnel who want to improve performance. Therefore, the response change analysis is conducted to identify school- and test administrator-level response change patterns that are statistically improbable when compared to the expected pattern at the state level.

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

Web Monitoring. The content of the LEAP 2025 assessments should not appear outside the boundaries of the forms administered. To protect Louisiana test content, the internet is monitored for postings that contain, or appear to contain, potentially exposed and/or copied test content. When test content is verified, steps are taken to quickly remove the infringing content.

Plagiarism Detection. The LDOE monitors for two different plagiarism situations: copying from student to student and copying from an outside source, such as Wikipedia or another internet source. Instances of plagiarism are identified by human scorers and artificial intelligence. Alerts are set to identify responses that may indicate the possibility of teacher interference or plagiarism. Alerted responses are given additional review so that the appropriate response can be taken.

Alerts for Disturbing Content

Scorers for the LEAP 2025 assessments can also apply an alert flag to student responses that may indicate disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others). All alerted responses are automatically routed to the scoring director, who reviews and forwards appropriate responses to senior project staff for review. If it is concluded that a response warrants an alert, project management will contact the LDOE to take the necessary action. At no point during this process do scorers or staff have access to demographic information for any students participating in the assessment.

6. Scoring Activities

Directory of Test Specification (DOTS) Process

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

Selected-Response (SR) Item Keycheck

SR items for social studies include multiple-choice (MC) and multiple-select (MS) items. Pearson calculates MC and MS item statistics and flags items if item statistics fall outside expected ranges. For example, items are flagged if few students select the correct response (*p*-value less than 0.25), if the item does not discriminate well between students of lower and higher ability (point-biserial correlation less than 0.20), or if many students (more than 40%) select a certain incorrect response. Lists of flagged MC and MS items, with the reasons for flagging, are provided to the LDOE and WestEd content staff for key verification. The staff reviews the list of flagged MC and MS items to confirm that the answer keys are accurate. The scoring of MC and MS items is also evaluated at data review.

Scoring of Technology-Enhanced (TE) Items

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

- A scoring rubric is created for each TE item. The 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 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 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 accommodated 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 that 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. 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. If scoring issues are identified, WestEd content staff and the LDOE recommend changes to the scoring algorithm. Any changes to the scoring algorithm are based on the LDOE's decisions. DRC, in turn, applies the approved scoring changes to any affected items.

Constructed-Response and Extended-Response Scoring

Constructed-response and extended-response items are scored by human raters trained by DRC. Ten percent of the responses are scored twice to monitor and maintain inter-rater reliability. Scoring supervisors also conduct read-behinds and review all nonscores and alerts. Operational handscoring processing rules are detailed in the *LEAP 2025 Spring 2024 Handscoring/AI Documentation* document.

Selection of Scoring Evaluators

Standard 4.20 states the following:

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

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

Recruitment and Interview Process

DRC strives to develop a highly qualified, experienced core of evaluators to appropriately maintain the integrity of all projects. All readers hired by DRC to score 2023–2024 LEAP 2025 HS test responses have at least a four-year college degree.

DRC has a human resources director dedicated solely to recruiting and retaining the handscoring staff. Applications for reader positions are screened by the handscoring project manager, the human resources director, and recruiting staff to create a large pool of potential readers. In the screening process, preference is given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, reader candidates are asked to

demonstrate their proficiency in writing by responding to a DRC writing topic and their proficiency in mathematics by solving word problems with correct work shown. These steps result in a highly qualified and diverse workforce. DRC personnel files for readers and team leaders include evaluations for each project completed. DRC uses these evaluations to place individuals on projects that best fit their professional backgrounds, their college degrees, and their performances on similar projects at DRC. Once placed, all readers go through rigorous training and qualifying procedures specific to the project on which they are placed. Any scorer who does not complete this training and does not demonstrate the ability to apply the scoring criteria by qualifying at the end of the process is not allowed to score live student responses.

Security

Whether training and scoring are conducted within a DRC facility or done remotely, security is essential to the handscoring process. When users log into DRC's secure, web-based scoring application, ScoreBoard, they are required to read and accept the security policy before they are allowed to access any project. For each project, scorers are also required to read and sign non-disclosure agreements, and during training emphasis is always given to what security means, the importance of maintaining security, and how this is accomplished.

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

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

In a remote environment, security reminders are given on a daily basis. Similar to the work that occurs within DRC scoring sites, in a remote environment, education about security expectations is the best way to maintain security of any project materials. DRC

requires scorers working remotely to work in a private environment away from other people (including family members). Restrictions are in place that define the hours during the day scorers are able to log into the system. If any type of security breach were to occur, immediate action would be taken to secure materials, and the employee would be terminated. DRC has the same policy within the scoring centers.

Handscoring Training Process

Standard 6.9 specifies:

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

Training Material Development

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

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

Training Procedures

Handscoring involves training team leaders and evaluators, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities.

Qualifying Standards

Prior to operational scoring, scorers must demonstrate their ability to apply the scoring criteria by qualifying (i.e., scoring with acceptable agreement with true scores on qualifying sets). After each qualifying set is scored, the DRC scoring director responsible for training leads the scorers in a discussion of the set. Any scorer who does not qualify during the operational scoring qualifying process is not allowed to score live student field test responses.

Monitoring the Scoring Process

Standard 6.8 states:

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

The following section explains the monitoring procedures that DRC uses to ensure that handscoring evaluators follow established scoring criteria while items are being scored. Detailed scoring rubrics, which specify the criteria for scoring, are available for all constructed- and extended-response items.

Reader Monitoring Procedures

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

Inter-Rater Reliability

Supervisors provide feedback to readers during regular read-behinds and the continuous monitoring of inter- rater reliability and score point distributions.

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

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

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

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

Reports and Reader Feedback

Reader performance and intervention information are recorded in reader feedback logs. These logs track information about actions taken with individual readers to ensure scoring consistency in regard to reliability, score point distribution, and validity performance. Due

to the brevity of the field test scoring window, DRC provides the LDOE with handscoring quality control reports for review at the end of the scoring window.

Inter-Rater Reliability

DRC and LDOE have agreed to expectations around inter-rater reliability and validity agreements as shown in Table 6.1

Table 6.1
Inter-Rater Reliability for Operational Constructed-Response Items

Agreement Rate Expectations for Validity and Inter-Rater Reliability – LEAP 2025			
Content Area/Course	**Score Point Range**	**Perfect Agreement**	**Perfect Agreement + Adjacent**
Grades 3-8 Social Studies CR and ER items	0-4 Rubric	70%	95%

A minimum of 10% of the responses for constructed- and extended-response items are scored independently by a second reader. The statistics for inter-rater reliability are calculated for all items. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the first and second scores is examined.

Tables 6.2–6.5 provide the inter-rater reliability and score point distributions for the constructed-response and extended-response items administered in the spring 2024 forms.

Table 6.2

Inter-Rater Reliability for Field Test Constructed-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
3	Item 1	≥4,000	64	29	7
	Item 2	≥4,000	71	25	4
	Item 3	≥4,000	69	29	6
	Item 4	≥4,000	67	27	6
	Item 5	≥4,000	66	27	7
	Item 6	≥3,940	68	26	5
4	Item 1	≥4,000	59	34	7
	Item 2	≥3,990	65	29	6
	Item 3	≥4,000	56	29	15
	Item 4	≥3,990	69	27	4
	Item 5	≥3,990	63	30	7
	Item 6	≥3,990	59	33	8
	Item 7	≥3,990	65	30	6
	Item 8	≥3,990	65	28	8
5	Item 1	≥4,000	62	32	7
	Item 2	≥3,990	59	34	8
	Item 3	≥4,000	55	34	11
	Item 4	≥3,990	59	33	8
	Item 5	≥4,000	63	30	8
	Item 6	≥3,990	61	31	8
6	Item 1	≥3,990	62	32	6
	Item 2	≥3,990	62	33	5
	Item 3	≥3,990	69	28	4
	Item 4	≥3,990	67	30	3
	Item 5	≥3,990	65	29	7
7	Item 1	≥3,990	58	34	8
	Item 2	≥3,990	64	33	3
	Item 3	≥3,990	64	29	7

	Item 4	≥3,990	64	32	4
8	Item 1	≥3,990	56	31	13
	Item 2	≥3,990	64	30	6
	Item 3	≥3,990	61	31	8
	Item 4	≥3,990	59	32	9
	Item 5	≥3,990	62	32	6

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

Table 6.3

Score Point Distributions for Field Test Constructed-Response Items

Grade	Item	Total	Score Point Distribution*						
			"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	Blank (%)	Foreign Language (%)**
3	Item 1	≥4,000	23	37	22	7	7	4	0
	Item 2	≥4,000	22	32	24	10	8	5	0
	Item 3	≥4,000	18	22	31	12	15	3	0
	Item 4	≥4,000	23	29	28	8	6	6	0
	Item 5	≥4,000	21	29	31	7	6	6	0
	Item 6	≥3,940	29	31	27	6	7	0	0
4	Item 1	≥4,000	11	26	37	13	12	0	0
	Item 2	≥3,990	11	27	26	15	22	0	0
	Item 3	≥4,000	18	23	37	7	15	0	0
	Item 4	≥3,990	13	28	36	13	10	0	0
	Item 5	≥3,990	16	37	32	8	6	0	0
	Item 6	≥3,990	15	32	31	11	10	0	0
	Item 7	≥3,990	12	26	35	15	13	0	0
	Item 8	≥3,990	11	17	31	17	23	0	0
5	Item 1	≥4,000	8	19	35	19	20	0	0
	Item 2	≥3,990	7	24	36	17	15	0	0
	Item 3	≥4,000	13	32	31	12	12	0	0
	Item 4	≥3,990	11	23	35	15	16	0	0
	Item 5	≥4,000	16	30	30	11	12	0	0
	Item 6	≥3,990	17	30	23	11	20	0	0
6	Item 1	≥3,990	10	25	36	16	13	0	0
	Item 2	≥3,990	13	28	33	13	12	0	0
	Item 3	≥3,990	14	25	39	12	10	0	0
	Item 4	≥3,990	11	21	29	21	18	0	0
	Item 5	≥3,990	39	27	21	9	5	0	0
7	Item 1	≥3,990	16	29	33	13	9	0	0
	Item 2	≥3,990	15	28	33	13	11	0	0

Grade	Item	Total	Score Point Distribution*						
			"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	Blank (%)	Foreign Language (%)**
	Item 3	≥3,990	18	23	29	15	15	0	0
	Item 4	≥3,990	11	28	28	15	17	0	0
8	Item 1	≥3,990	24	22	28	13	13	0	0
	Item 2	≥3,990	23	24	24	14	15	0	0
	Item 3	≥3,990	13	18	33	18	17	1	0
	Item 4	≥3,990	16	18	29	17	20	0	0
	Item 5	≥3,990	15	21	29	19	17	0	0

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

** Foreign language (F) responses cannot be assigned a score based on the rubric and count as zero points toward student scores.

Table 6.4

Inter-Rater Reliability for Field Test Extended-Response Items

Grade	Item	Inter-Rater Reliability*			
		2x	Exact Agreement (%)	Adjacent Agreement (%)	Nonadjacent (%)
5	Item 1	≥3,990	55	36	9
	Item 2	≥4,000	57	37	6
6	Item 1	≥3,990	59	35	6
	Item 2	≥3,990	68	27	5
	Item 3	≥3,990	61	35	5
7	Item 1	≥3,990	53	35	12
	Item 2	≥4,000	56	34	10
8	Item 1	≥3,990	57	36	7

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

Table 6.5

Score Point Distributions for Field Test Extended-Response Items

Grade	Score Point Distribution*								
	Item	Total	"0" (%)	"1" (%)	"2" (%)	"3" (%)	"4" (%)	Blank (%)	Foreign Language (%)**
5	Item 1	≥3,990	10	25	29	22	14	0	0
	Item 2	≥4,000	10	30	30	20	10	0	0
6	Item 1	≥3,990	9	30	33	20	7	0	0
	Item 2	≥3,990	12	41	22	17	7	0	0
	Item 3	≥3,990	7	24	32	25	12	0	0
7	Item 1	≥3,990	21	23	25	19	12	0	0
	Item 2	≥4,000	14	20	26	26	13	0	0
8	Item 1	≥3,990	11	23	32	22	12	0	0

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

** Foreign language (F) responses cannot be assigned a score based on the rubric and count as zero points toward student scores.

7. Data Analysis

Classical Item Statistics

A measure of item difficulty, *p* (or "the *p*-value"), indicates the average proportion of total points earned on an item. For example, if *p* = 0.50 on an MC item, then half of the examinees earned a score of 1. If *p* = 0.50 on a CR item, then examinees earned half of the possible points on average (e.g., 1 out of 2 possible points). The item-total correlation (point-biserial) is in general a measure of item discrimination. Items with higher item-total correlations provide better information about overall student ability (i.e., they discriminate between lower- and higher-ability students). Tables 7.1 through 7.6 provide summary item statistics by grade level–item type that were field tested.

Table 7.1

Summary of Classical Statistics for Field Test Items for Social Studies Grade 3

Grade 3							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	181	0.41	0.13	0.30	0.12	8%	0%
MS	16	0.21	0.08	0.30	0.13	19%	0%
CR	6	0.35	0.05	0.59	0.05	0%	0%
TE	55	0.40	0.14	0.35	0.12	5%	2%
TPI	47	0.39	0.11	0.37	0.12	17%	0%
TPD	15	0.32	0.11	0.35	0.13	7%	0%

Table 7.2

Summary of Classical Statistics for Field Test Items for Social Studies Grade 4

Grade 4							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	182	0.45	0.14	0.32	0.12	6%	3%
MS	9	0.31	0.12	0.37	0.06	0%	0%
CR	8	0.46	0.06	0.58	0.05	0%	0%
TE	61	0.41	0.15	0.37	0.14	7%	3%
TPI	14	0.40	0.09	0.34	0.12	21%	0%
TPD	11	0.36	0.16	0.42	0.15	18%	0%

Table 7.3

Summary of Classical Statistics for Field Test Items for Social Studies Grade 5

Grade 5							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	184	0.46	0.13	0.32	0.12	5%	3%
MS	20	0.27	0.11	0.32	0.11	0%	0%
CR	6	0.48	0.05	0.60	0.05	0%	0%
ER	2	0.50	0.03	0.63	0.03	0%	0%
TE	61	0.44	0.18	0.37	0.14	11%	2%
TPI	16	0.44	0.10	0.41	0.09	0%	0%
TPD	17	0.38	0.09	0.44	0.11	6%	0%

Table 7.4

Summary of Classical Statistics for Field Test Items for Social Studies Grade 6

Grade 6							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	186	0.44	0.14	0.36	0.13	5%	5%
MS	22	0.32	0.13	0.39	0.10	9%	0%
CR	5	0.44	0.09	0.64	0.02	0%	0%
ER	3	0.46	0.06	0.62	0.05	0%	0%
TE	59	0.42	0.15	0.43	0.16	5%	3%
TPI	12	0.42	0.11	0.42	0.17	8%	0%
TPD	7	0.40	0.15	0.47	0.10	0%	0%

Table 7.5

Summary of Classical Statistics for Field Test Items for Social Studies Grade 7

Grade 7							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	193	0.45	0.14	0.35	0.14	4%	4%
MS	22	0.25	0.09	0.34	0.16	14%	0%
CR	4	0.45	0.03	0.66	0.03	0%	0%
ER	2	0.46	0.06	0.66	0.01	0%	0%
TE	45	0.42	0.16	0.46	0.13	4%	0%
TPI	18	0.43	0.11	0.45	0.12	6%	0%
TPD	15	0.38	0.12	0.47	0.12	7%	0%

Table 7.6

Summary of Classical Statistics for Field Test Items for Social Studies Grade 8

Grade 8							
Item Type	**Number of Items**	***p*-value Mean**	***p*-value SD**	**Item-Total Correlation Mean**	**Item-Total Correlation SD**	**Percent with B-Level DIF**	**Percent with C-Level DIF**
MC	191	0.48	0.16	0.34	0.14	7%	5%
MS	24	0.32	0.18	0.36	0.17	4%	0%
CR	5	0.46	0.05	0.65	0.03	0%	0%
ER	1	0.50	0.00	0.64	0.00	0%	0%
TE	48	0.43	0.14	0.45	0.11	0%	0%
TPI	22	0.43	0.09	0.41	0.14	14%	0%
TPD	9	0.35	0.09	0.42	0.11	0%	0%

Table 7.7 summarizes the numbers of field-tested items at each grade level that were flagged. The box plots that follow illustrate the range of item *p*-values by grade level and good item–total discriminating power exhibited overall by grade level.

Table 7.7

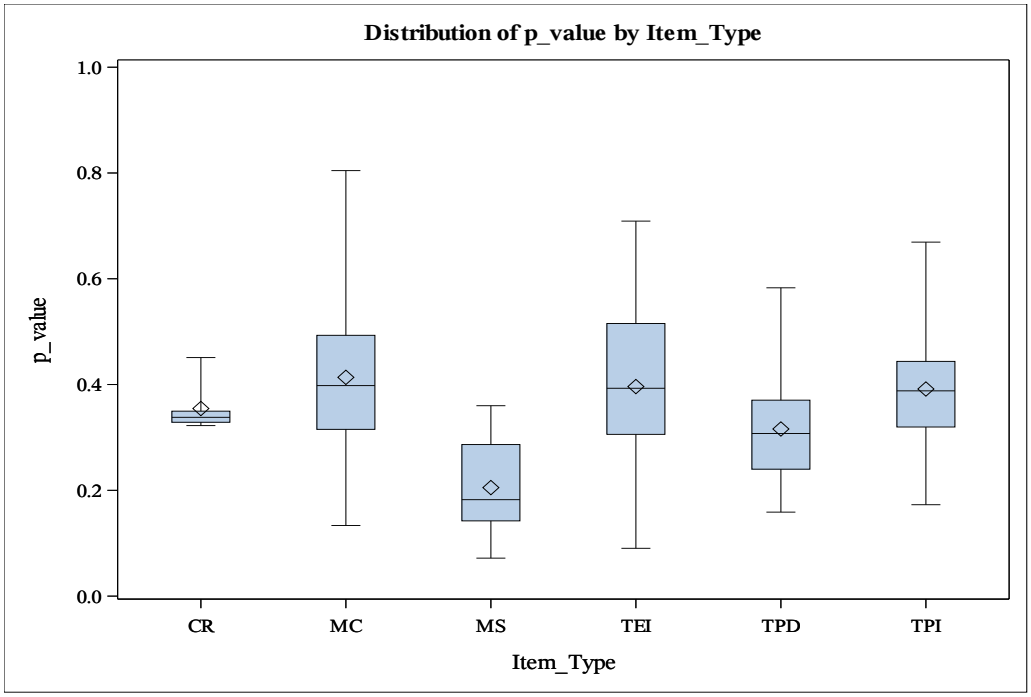
Number of Field Test Items Flagged for Item Statistics

Grade	Item Type	*N* Items	Flagged for *p*-value	Flagged for Point-Biserial Correlation	Flagged for DIF
3	CR	6	0	0	1
	MC	181	15	36	6
	MS	16	11	4	0
	TEI	55	8	6	5
	TPD	15	5	1	0
	TPI	47	4	6	0
4	CR	8	0	0	5
	MC	182	12	30	6
	MS	9	2	0	0
	TEI	61	7	7	5
	TPD	11	2	1	0
	TPI	14	0	3	0
5	CR	6	0	0	4
	ER	2	0	0	2
	MC	184	5	29	4
	MS	20	10	3	1
	TEI	61	9	7	8
	TPD	17	0	1	1
	TPI	16	0	1	1
6	CR	5	0	0	2
	ER	3	0	0	2
	MC	186	11	24	2
	MS	22	7	2	1
	TEI	59	7	6	2
	TPD	7	1	0	0
	TPI	12	1	1	1
7	CR	4	0	0	2
	ER	2	0	0	0
	MC	193	12	32	6
	MS	22	13	5	2
	TEI	45	8	1	3
	TPD	15	3	0	1
	TPI	18	1	1	0
8	CR	5	0	0	5
	ER	1	0	0	1
	MC	191	13	31	15
	MS	24	8	4	1
	TEI	48	6	1	1
	TPD	9	0	0	3
	TPI	22	1	3	0

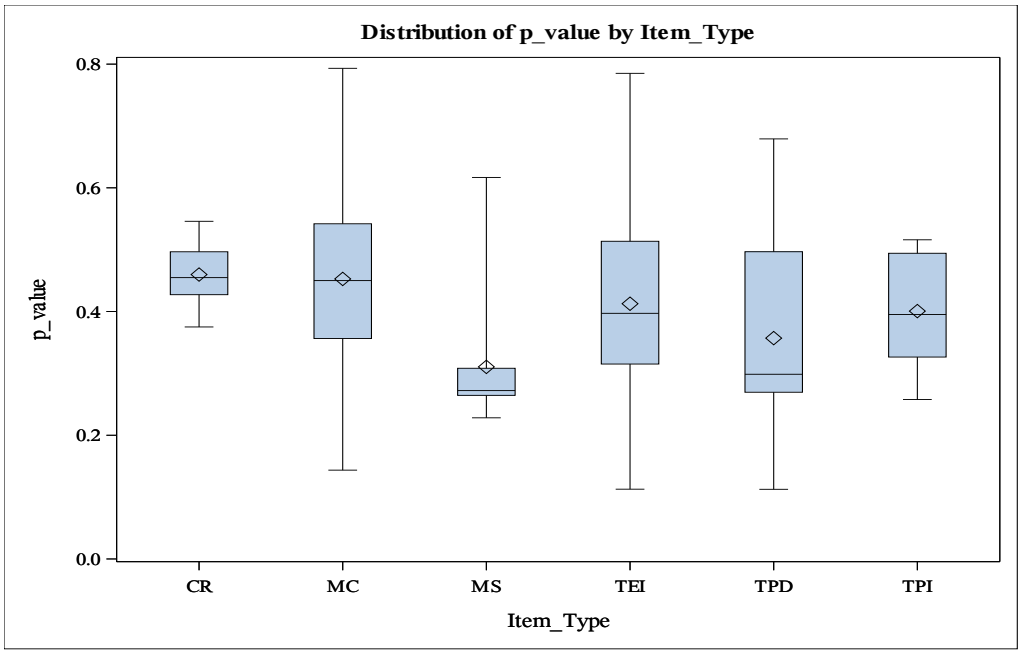
Figure 7.1

Item p-Values by Grade

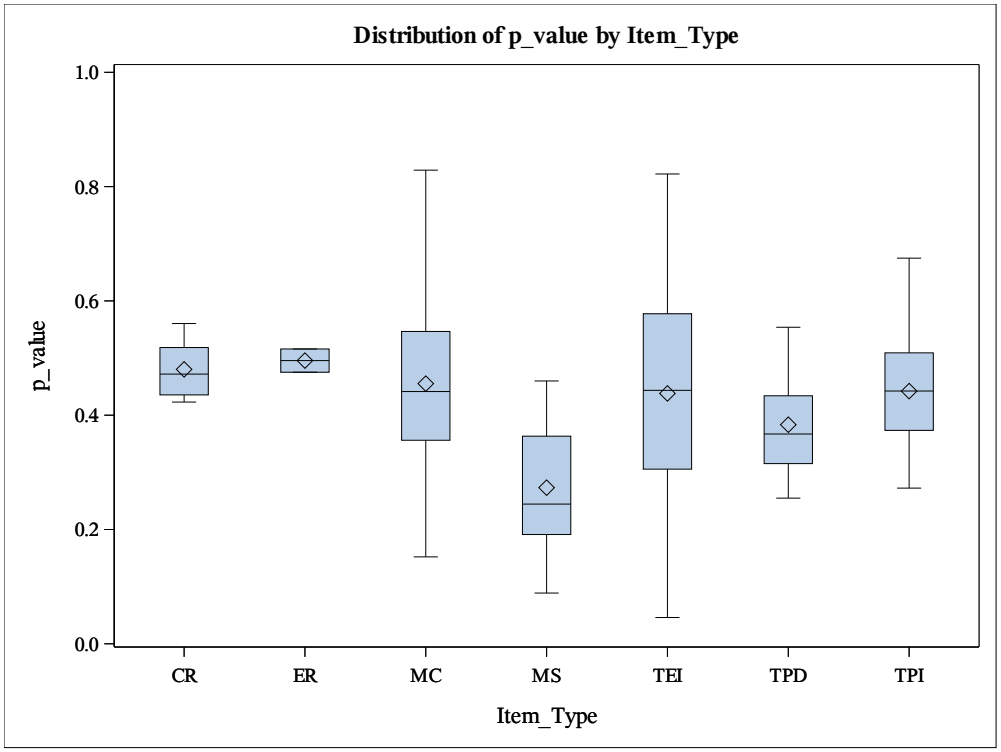
Grade 3



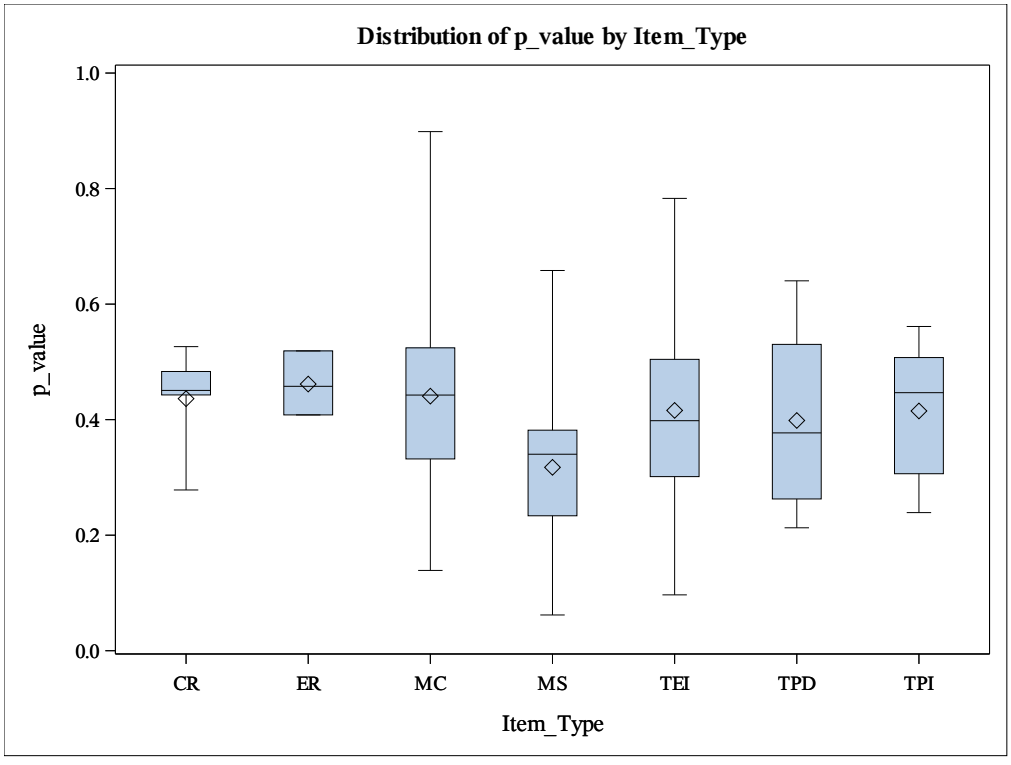
Grade 4



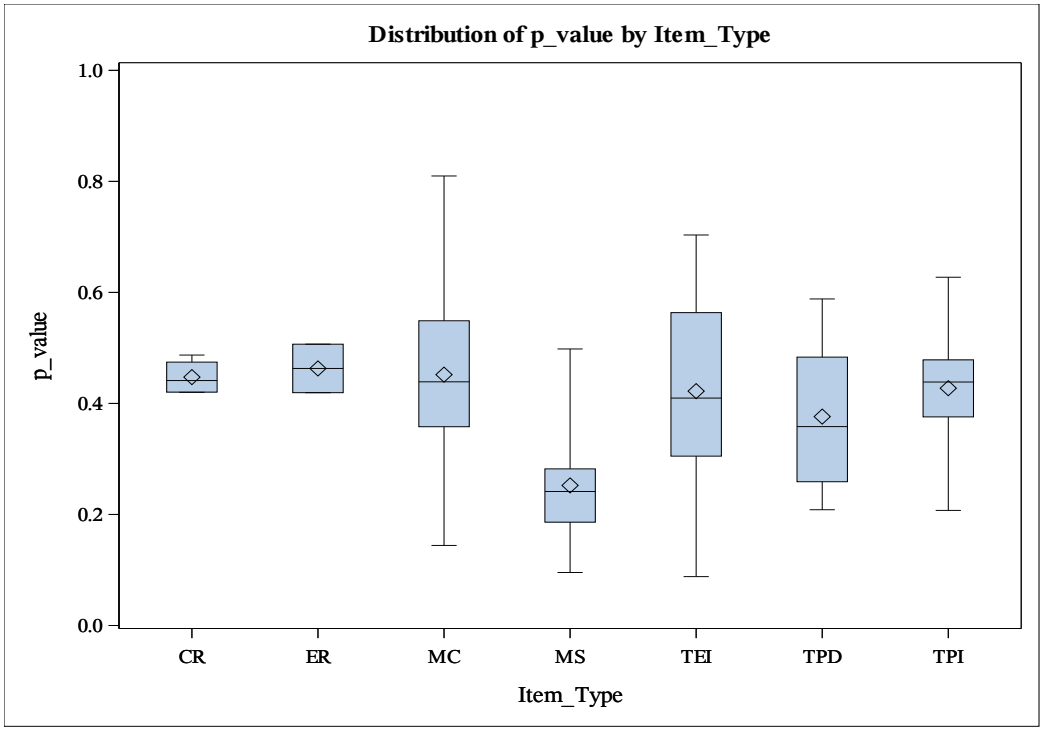
Grade 5



Grade 6



Grade 7



Grade 8

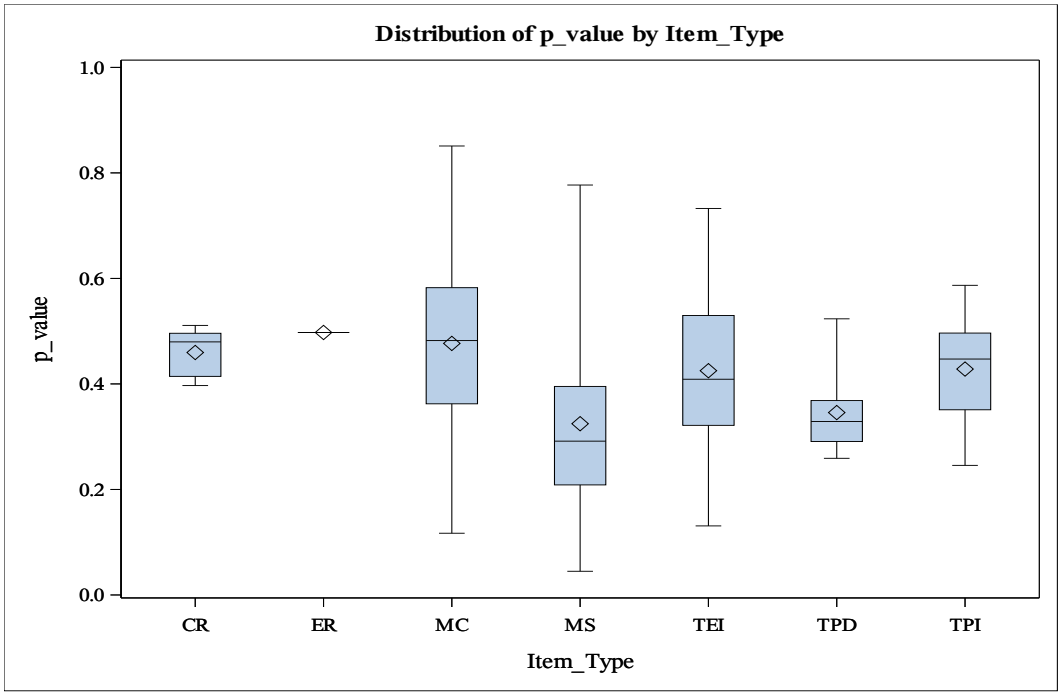
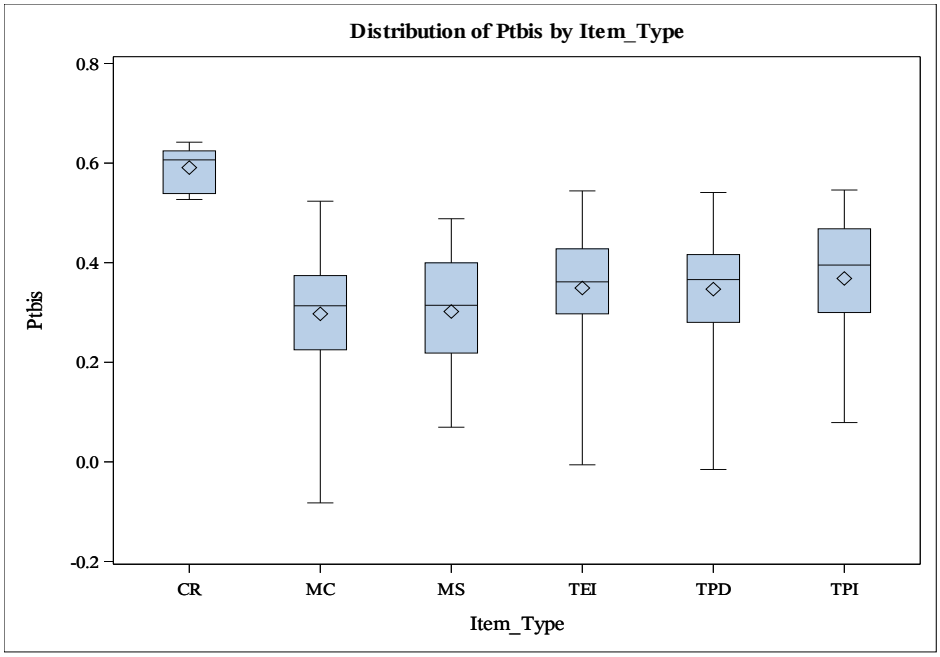


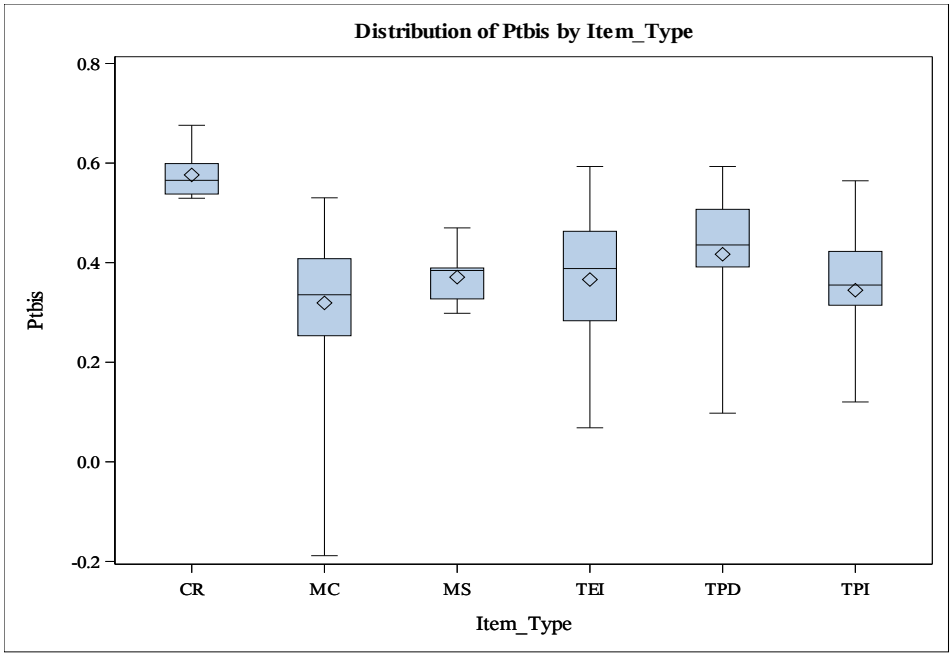
Figure 7.2

Item–Total Correlations/Point Biserial (PBIS) by Grade

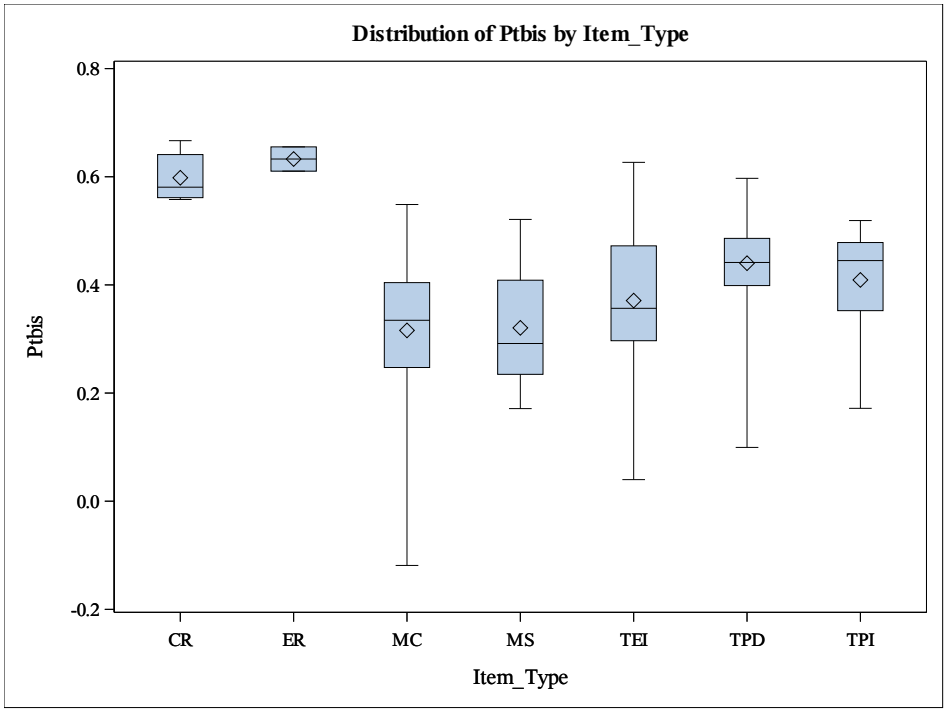
Grade 3



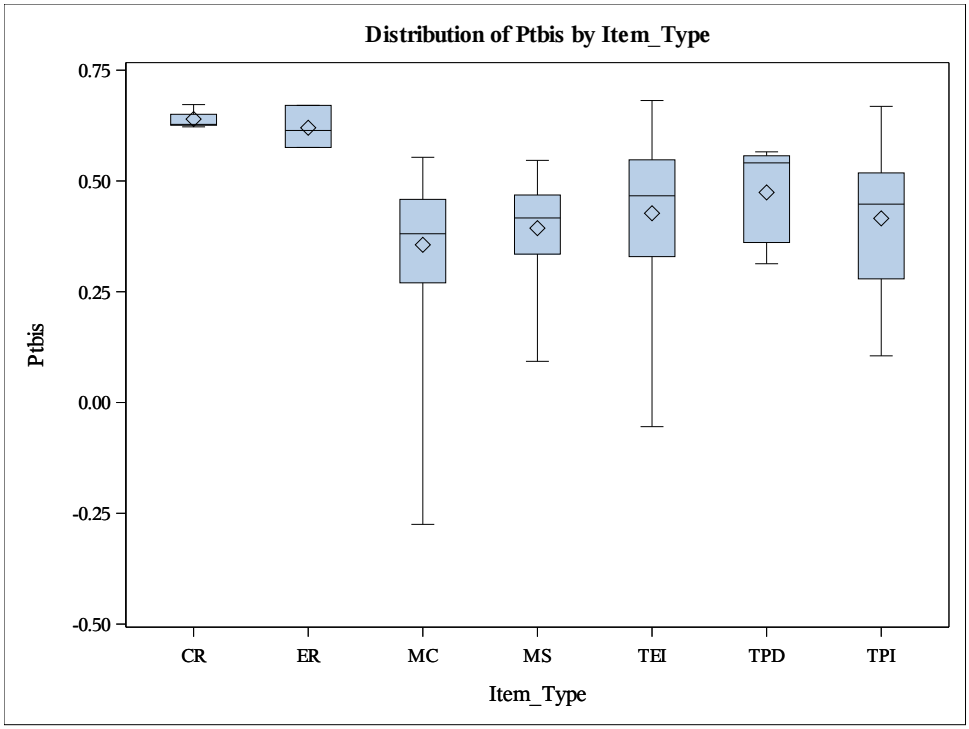
Grade 4



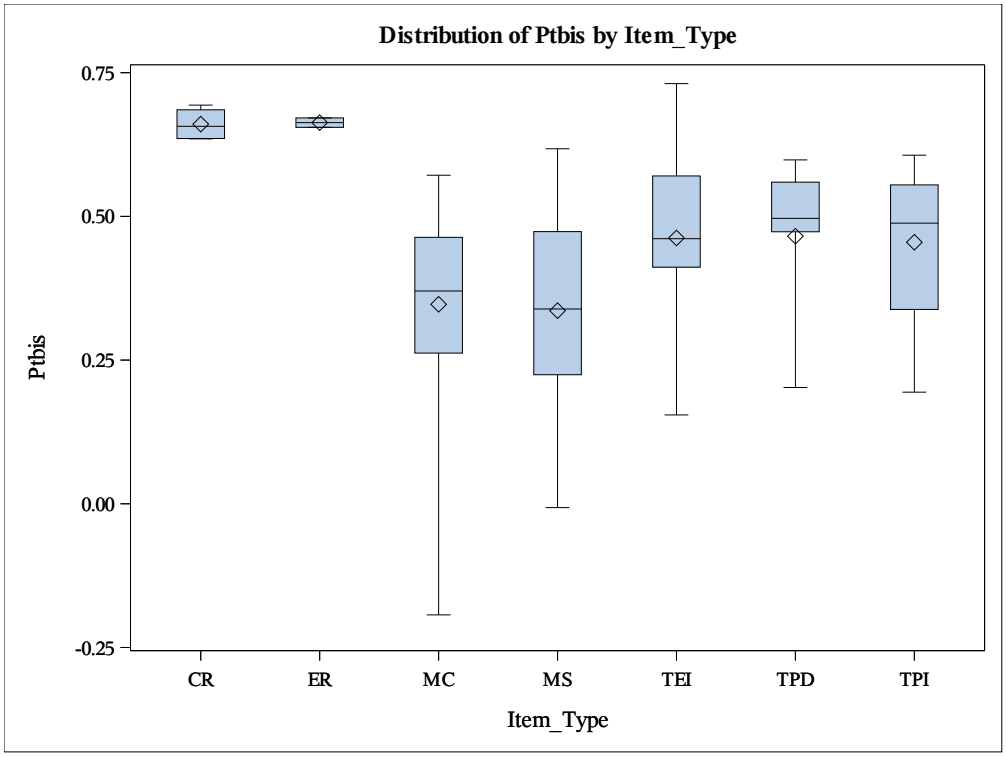
Grade 5



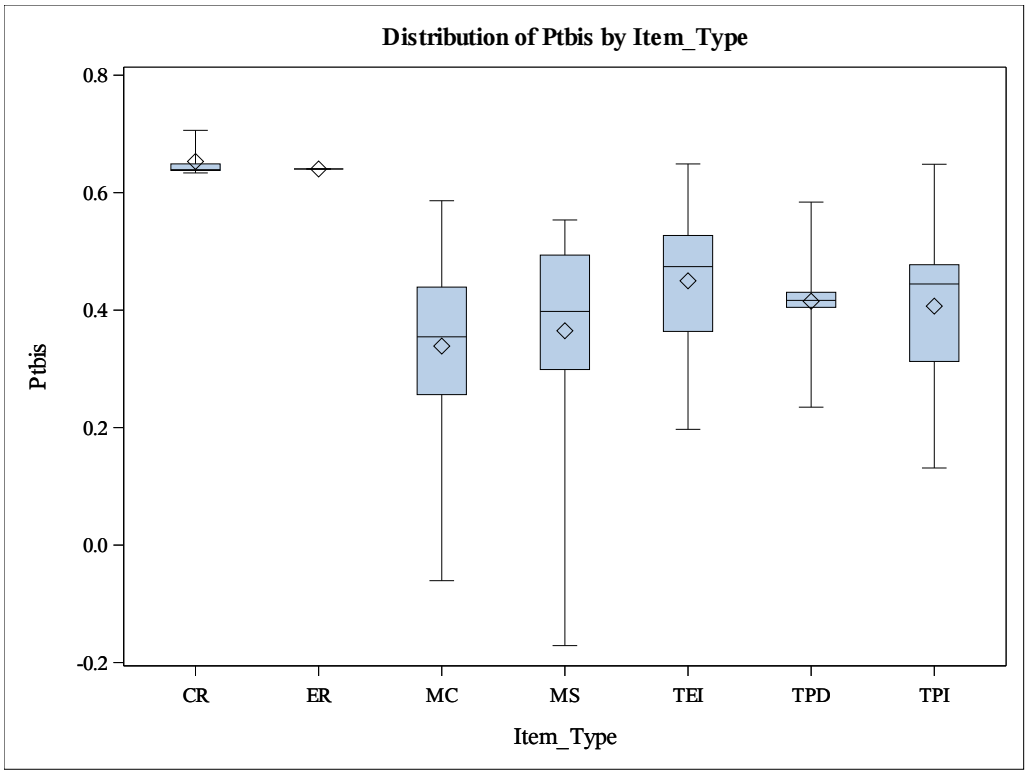
Grade 6



Grade 7



Grade 8



Differential Item Functioning

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

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

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

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

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

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

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

where N_{r1k} is the number of correct responses in the reference group at ability level *k*, N_{f0k} is the number of incorrect responses in the focal group at ability level *k*, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at

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

The *ΔMH DIF* is computed as

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

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

The MH chi-square statistic and the *ΔMH DIF* were used in combination to identify the field test items that exhibit strong, weak, or no DIF (Zieky, 1993). Table 7.8 defines the DIF categories for dichotomous items.

Table 7.8
DIF Categories for Dichotomous Items

DIF Category	Criteria
A (negligible)	| *ΔMH DIF* | is not significantly different (p <0.05) from 0.0 or is less than 1.0.
B (slight to moderate)	| *ΔMH DIF* | is significantly different (p <0.05) from 0.0 but not from 1.0, and is at least 1.0; OR | *ΔMH DIF* | is significantly different (p <0.05) from 1.0 (p <0.05) but is less than 1.5. Positive values are classified as "B+" and negative values as "B–."
C (moderate to large)	| *ΔMH DIF* | is significantly different (p <0.05) than 1.0 and is at least 1.5. Positive values are classified as "C+" and negative values as "C–."

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

which gives the greatest weight at score levels most frequently attained by students in the focal group.

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

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

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

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

The *p*-value associated with the Mantel χ^2 statistic and the *SMD* (on an effect-size metric) are used to determine DIF classifications. Table 7.9 defines the DIF categories for polytomous items.

Table 7.9
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 $0.17 < \lvert SMD/SD \rvert \leq 0.25$
C (moderate to large)	Mantel χ^2 *p*-value < 0.05 and $\lvert SMD/SD \rvert > 0.25$

Four DIF analyses were conducted for the operational test items only: Female – Male, African American – White, Hispanic/Latino – White, and Economically Disadvantaged – Not Economically Disadvantaged. 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 African American or White students, Hispanic/Latino or White students and Economically Disadvantaged or Not Economically Disadvantaged students performed unexpectedly well or unexpectedly poorly, given their

performance on the full assessment. The last two columns of Tables 7.10-7.13 provide the number of items flagged for DIF. Items flagged with A-DIF show negligible DIF, items flagged with B-DIF are said to exhibit slight to moderate DIF, and items with C-DIF are said to exhibit moderate to large DIF. Note that DIF flags for dichotomous items are based on the *MH* statistics while DIF flags for polytomous items are based on the combination of Mantel χ^2 *p*-value and *SMD* statistics. In addition, all items exhibiting DIF were reviewed by a committee of Louisiana teachers as well as LDOE and WestEd content staff.

Table 7.10

Summary of Female – Male DIF Flags for Field Test Items for Social Studies by Grade

Female – Male			
Grade	**A**	**B,[B-]**	**C,[C-]**
3	315	[11,[4]	[0],[0]
4	274	[5],[6]	[0],[0]
5	295	[5],[5]	[1],[0]
6	289	[3],[1]	[1],[0]
7	293	[4],[2]	[0],[0]
8	287	[8],[3]	[2],[0]

Table 7.11

Summary of African American – White DIF Flags for Field Test Items for Social Studies by Grade

African American – White			
Grade	A	B,[B-]	C,[C-]
3	313	[1],[5]	[0],[1]
4	279	[0],[6]	[0],[0]
5	296	[0],[8]	[0],[2]
6	292	[0],[2]	[0],[0]
7	292	[0],[5]	[2],[0]
8	294	[0],[4]	[0],[2]

Table 7.12

Summary of Hispanic –White DIF Flags for Field Test Items for Social Studies by Grade

Hispanic – White			
Grade	A	B,[B-]	C,[C-]
3	317	[11,[21	[01,[01
4	282	[11,[21	[01,[01
5	301	[21,[21	[01,[11
6	292	[01,[11	[01,[11
7	295	[11,[31	[01,[01
8	292	[21,[51	[01,[11

Table 7.13

Summary of Economically Disadvantaged – Not Economically Disadvantaged DIF Flags for Field Test Items for Social Studies by Grade

Economically Disadvantaged – Not Economically Disadvantaged			
Grade	A	B,[B-]	C,[C-]
3	318	[01,[21	[01,[01
4	284	[01,[11	[01,[01
5	303	[01,[21	[01,[11
6	292	[01,[21	[01,[01
7	299	[01,[01	[01,[01
8	297	[01,[31	[01,[01

Item Calibration

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

Item Response Theory (IRT) models were used in the item calibration for the LEAP 2025 Social Studies assessments. Scaling is the process whereby we associate student performance with some ordered value, typically a number. The most common and straightforward way to score a test is to simply use the sum of points a student earned on the test, namely, raw score. Although the raw score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Typically, a scaled metric is used, on which test forms from different years are equated.

Measurement Models

IRTPRO, a software application for item calibration and test scoring, was used to estimate item response theory (IRT) parameters from LEAP 2025 assessment data. Multiple-choice (MC), multiple-select (MS), and some technology-enhanced (TE) items were scored dichotomously (0/1), so the 3-parameter logistic model (3PL) was applied to those data:

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

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

The 2025 field test included polytomous items. Therefore, data from polytomous items were used to estimate parameters for the generalized partial credit model (GPCM) (Muraki, 1992):

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

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

Field Test Item Parameters

The distributions of item parameters are summarized by grade in Tables 7.14–7.19. Figures 7.3–7.5 provide box plot displays of the distributions of IRT parameter estimates by item type. TPI, TPD, CR, and ER items have no *c* parameters because they are polytomous items and are therefore modeled using the GPCM.

It should be noted that somewhat significant trend between classical item parameters (e.g., p-value) and IRT-based item parameters (e.g., b parameter) can be found. In addition, recommended ranges for IRT parameter estimates are functions of an assessment program and assessment results and will vary by large scale assessment programs. As each of the LEAP 2025 assessments mature, however, desired targets/ranges (e.g., point-biserial higher than 0.30) can be defined in the annual Framework documents that LDOE, Pearson and WestEd use for annual test construction.

Item Fit

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q_1 statistic (Yen, 1981) is used as an index for how well theoretical item curves match observed item responses. Q_1 is computed by first conducting an IRT item parameter estimation, then estimating students' achievement using the estimated item parameters, and, finally, using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at 10 selected equal intervals across the range of student achievement. Q_1 is computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-square (χ^2) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis

(the expected item performance). Q_1 for each item type has varying degrees of freedom because the different item types have different numbers of IRT parameters. Therefore, Q_1 is not directly comparable across item types. An adjustment or linear transformation (translation to a *Z*-score, Z_{Q_1}) is made for different numbers of item parameters and sample size to create a more comparable statistic.

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

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

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

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

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

The generalization of Q_1 for items with multiple response categories is

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

where

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

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

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

and

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

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

A summary of IRT item parameter statistics and item fit by grade is displayed in Tables 7.14 through 7.19.

Table 7.14

Summary of IRT Statistics for Field Test Items for Social Studies Grade 3

Grade 3								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	181	0.71	0.33	1.35	4.02	0.19	0.07	95%
MS	16	0.66	0.18	2.24	1.38	0.06	0.03	100%
CR	6	0.38	0.05	0.78	0.28	0.00	0.00	67%
TE	55	0.37	0.21	0.63	1.95	0.19	0.09	75%
TPI	47	0.32	0.14	1.53	2.57	0.00	0.00	72%
TPD	15	0.25	0.15	0.92	1.85	0.00	0.00	47%

*Only dichotomous items (scored 0 or 1) have *c* parameters.
*% Fit indicates % of items with no model fit issues.

Table 7.15

Summary of IRT Statistics for Field Test Items for Social Studies Grade 4

Grade 4								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	182	0.73	0.35	1.50	10.82	0.18	0.08	95%
MS	9	0.67	0.17	0.96	0.72	0.08	0.04	89%
CR	8	0.36	0.07	0.13	0.32	0.00	0.00	75%
TE	61	0.50	0.23	1.15	2.63	0.18	0.11	79%
TPI	14	0.28	0.13	1.57	2.71	0.00	0.00	86%
TPD	11	0.36	0.17	2.22	4.80	0.00	0.00	18%

*Only dichotomous items (scored 0 or 1) have *c* parameters.
*% Fit indicates % of items with no model fit issues.

Table 7.16

Summary of IRT Statistics for Field Test Items for Social Studies Grade 5

Grade 5								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	184	0.75	0.33	0.74	1.27	0.18	0.08	98%
MS	20	0.60	0.23	1.53	0.81	0.05	0.03	90%
CR	6	0.38	0.04	-0.02	0.30	0.00	0.00	67%
ER	2	0.46	0.07	-0.03	0.16	0.00	0.00	100%
TE	61	0.50	0.27	0.54	3.71	0.17	0.11	82%
TPI	16	0.37	0.14	0.48	0.86	0.00	0.00	81%
TPD	17	0.36	0.14	1.32	3.65	0.00	0.00	29%

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

*% Fit indicates % of items with no model fit issues.

*Note. TE items scored 0 and 1 have estimated c parameter.

Table 7.17

Summary of IRT Statistics for Field Test Items for Social Studies Grade 6

Grade 6								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	186	0.90	0.37	0.49	4.86	0.19	0.08	97%
MS	22	0.73	0.28	1.35	1.61	0.07	0.05	95%
CR	5	0.48	0.03	0.29	0.52	0.00	0.00	80%
ER	3	0.46	0.08	0.14	0.29	0.00	0.00	100%
TE	59	0.48	0.25	1.67	7.70	0.24	0.19	73%
TPI	12	0.41	0.21	1.38	3.04	0.00	0.00	75%
TPD	7	0.42	0.14	0.59	0.86	0.00	0.00	57%

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

*% Fit indicates % of items with no model fit issues.

*Note. TE items scored 0 and 1 have estimated c parameter.

Table 7.18

Summary of IRT Statistics for Field Test Items for Social Studies Grade 7

Grade 7								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	193	0.87	0.35	1.11	2.88	0.19	0.08	94%
MS	22	0.69	0.36	1.35	1.58	0.07	0.05	82%
CR	4	0.48	0.04	0.20	0.20	0.00	0.00	100%
ER	2	0.46	0.04	0.20	0.24	0.00	0.00	100%
TE	45	0.48	0.18	0.57	1.36	0.09	0.10	73%
TPI	18	0.46	0.17	0.67	1.10	0.00	0.00	61%
TPD	15	0.39	0.13	0.71	1.04	0.00	0.00	40%

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

*% Fit indicates % of items with no model fit issues.

*Note. TE items scored 0 and 1 have estimated c parameter.

Table 7.19

Summary of IRT Statistics for Field Test Items for Social Studies Grade 8

Grade 8								
Item Type	**Number of Items**	***a* Mean**	***a* SD**	***b* Mean**	***b* SD**	***c* Mean***	***c* SD***	**% Fit (no model fit issues)**
MC	191	0.82	0.41	0.75	4.60	0.19	0.08	96%
MS	24	0.78	0.36	0.94	1.39	0.08	0.06	96%
CR	5	0.43	0.04	0.14	0.21	0.00	0.00	80%
ER	1	0.49	0.00	-0.03	0.00	0.00	0.00	100%
TE	48	0.43	0.15	0.49	0.94	0.08	0.06	65%
TPI	22	0.40	0.20	0.84	1.64	0.00	0.00	36%
TPD	9	0.33	0.12	1.08	0.86	0.00	0.00	11%

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

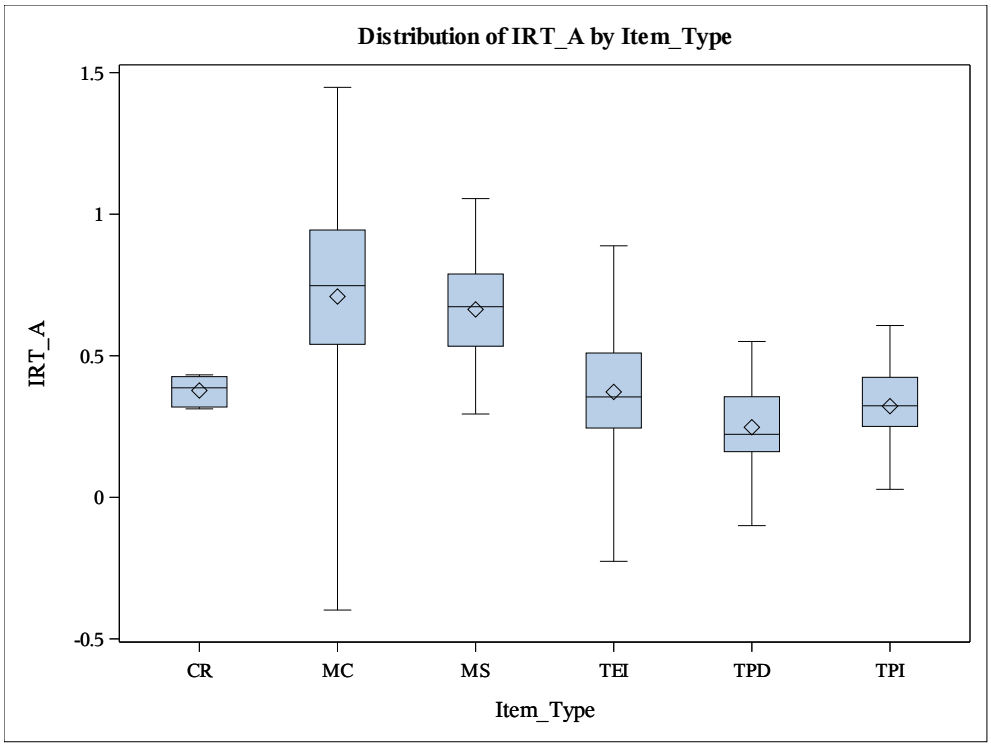
*% Fit indicates % of items with no model fit issues.

*Note. TE items scored 0 and 1 have estimated c parameter.

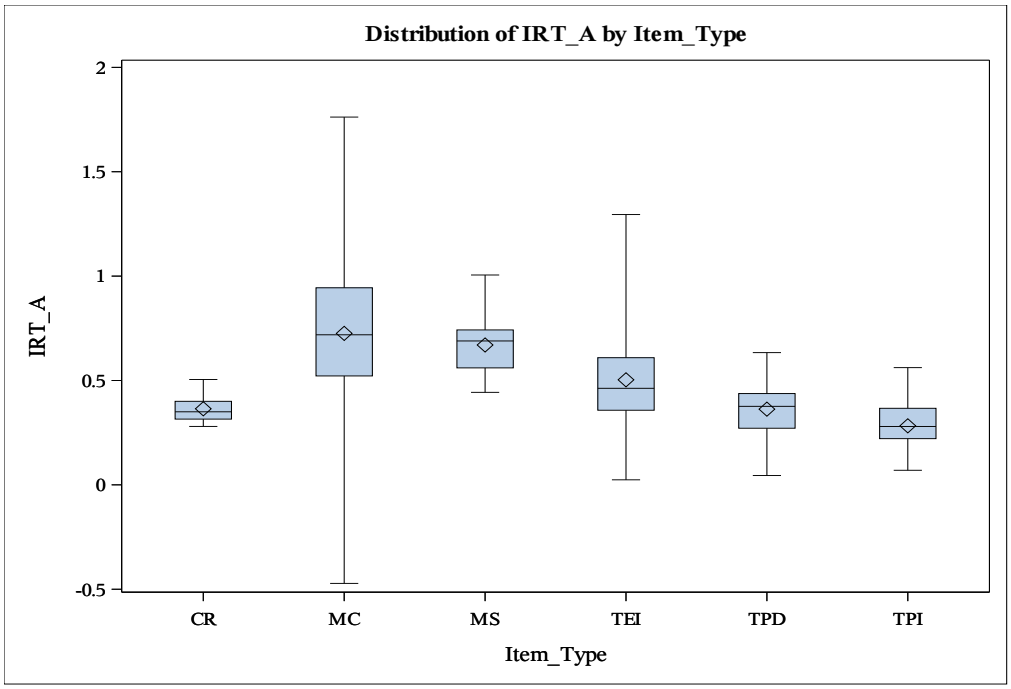
Figure 7.3

IRT A Parameters by Grade

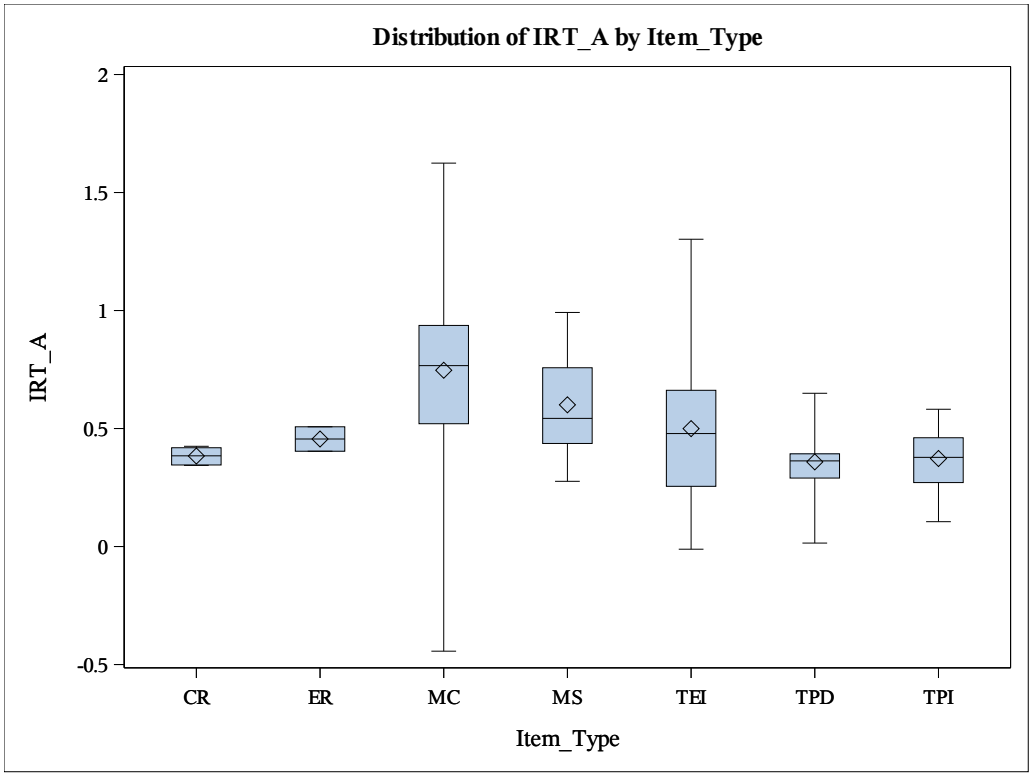
Grade 3



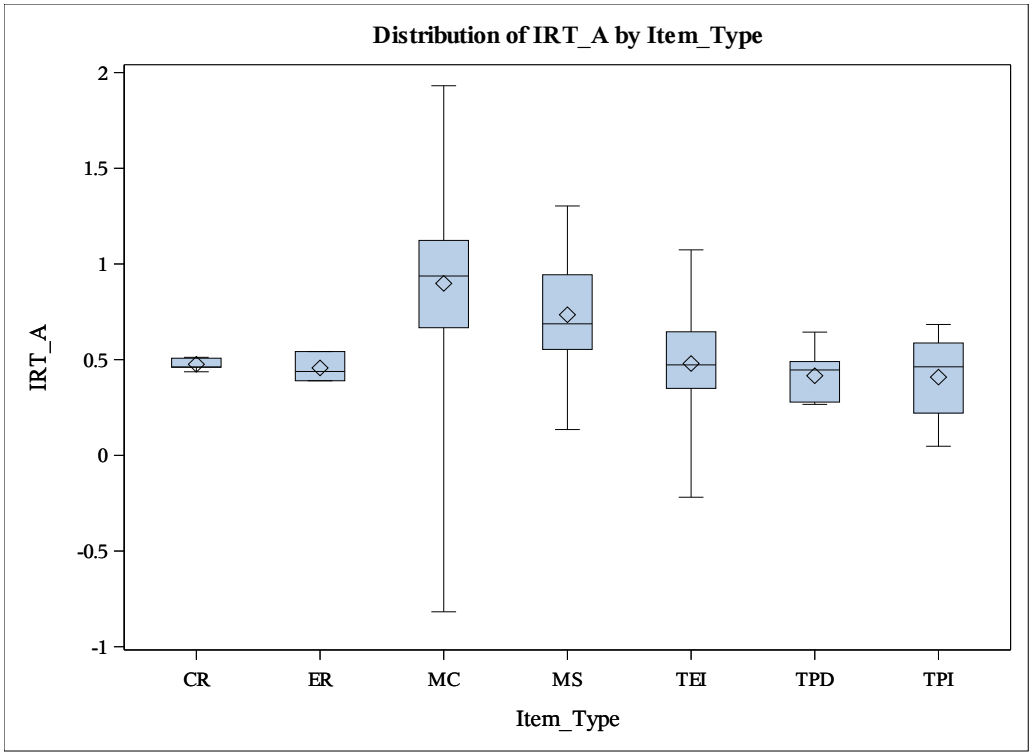
Grade 4



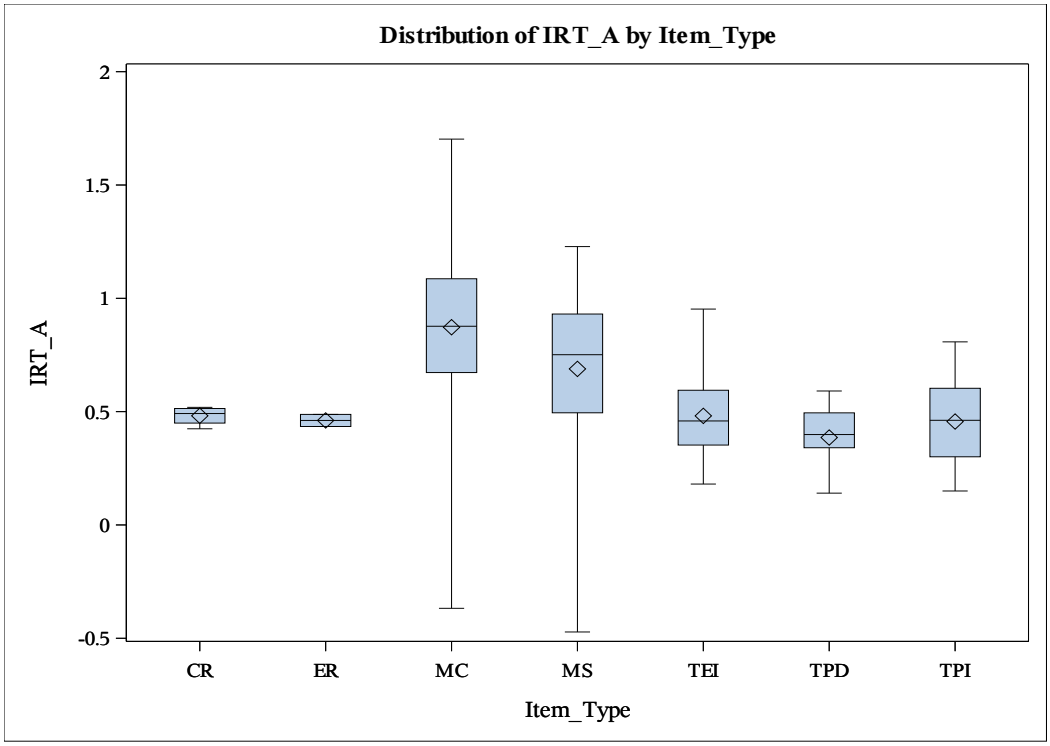
Grade 5



Grade 6



Grade 7



Grade 8

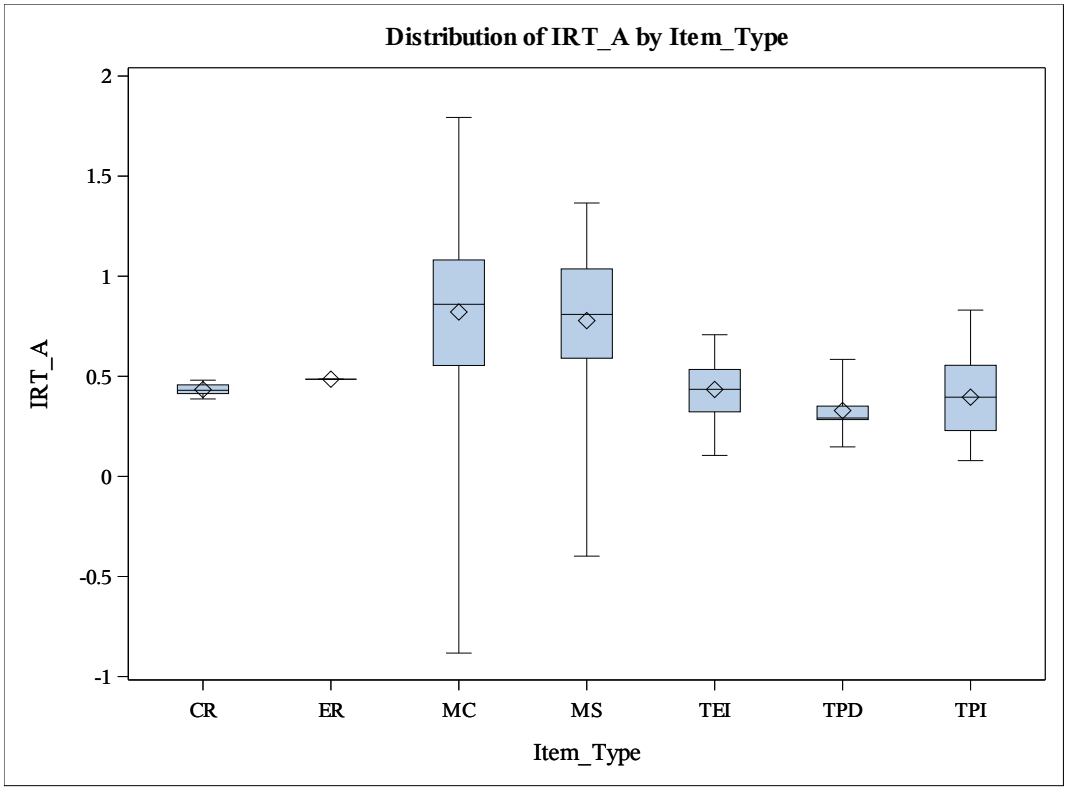
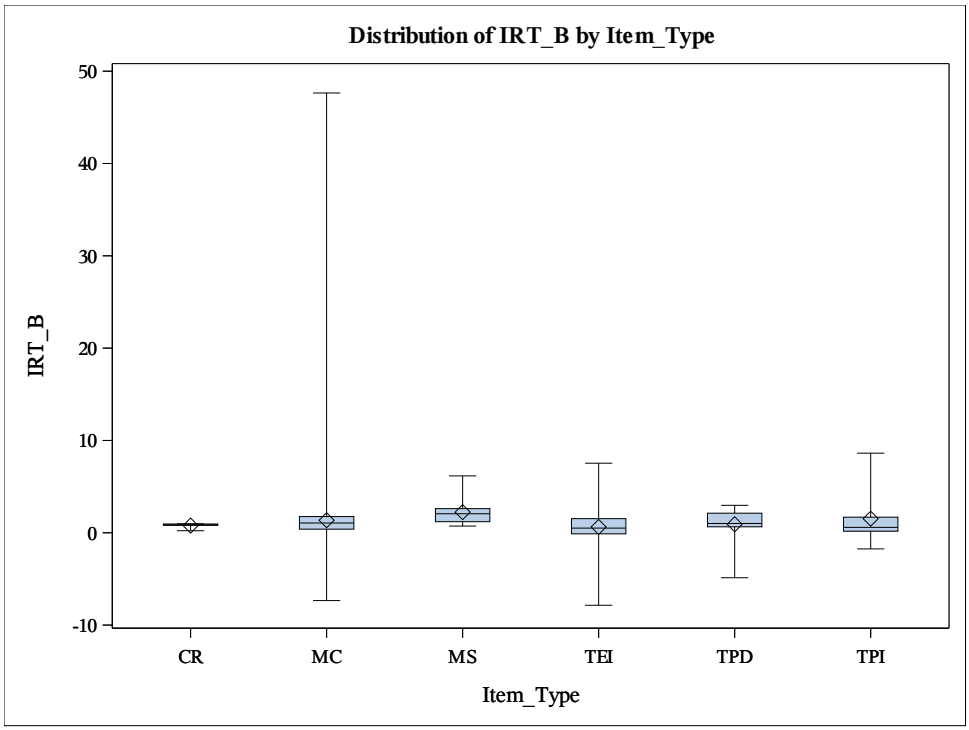


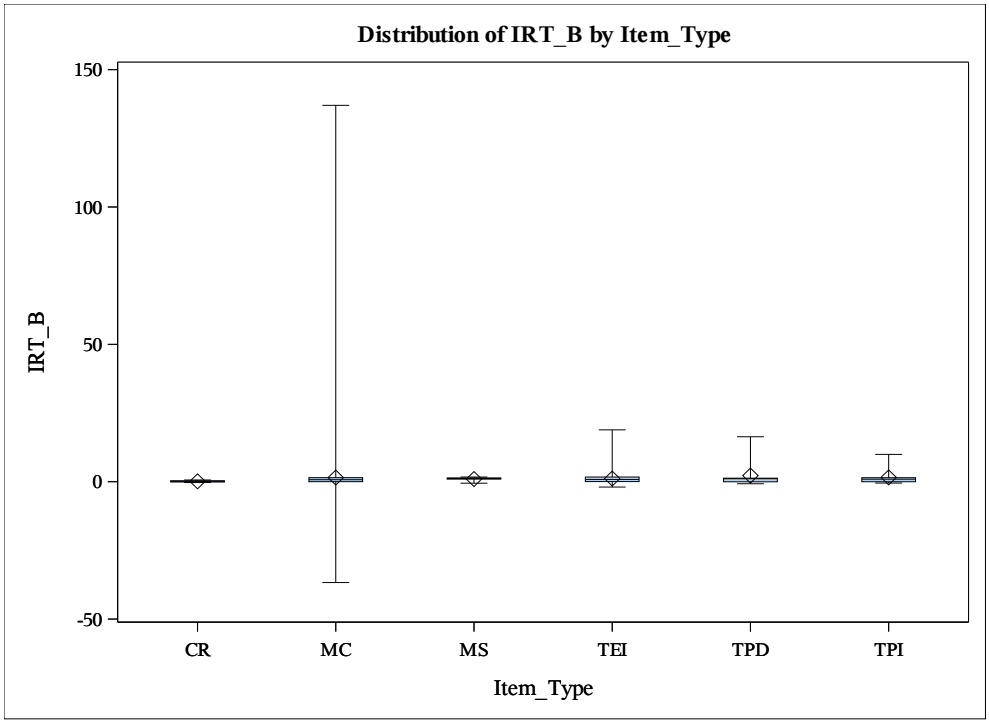
Figure 7.4

IRT B Parameters by Grade

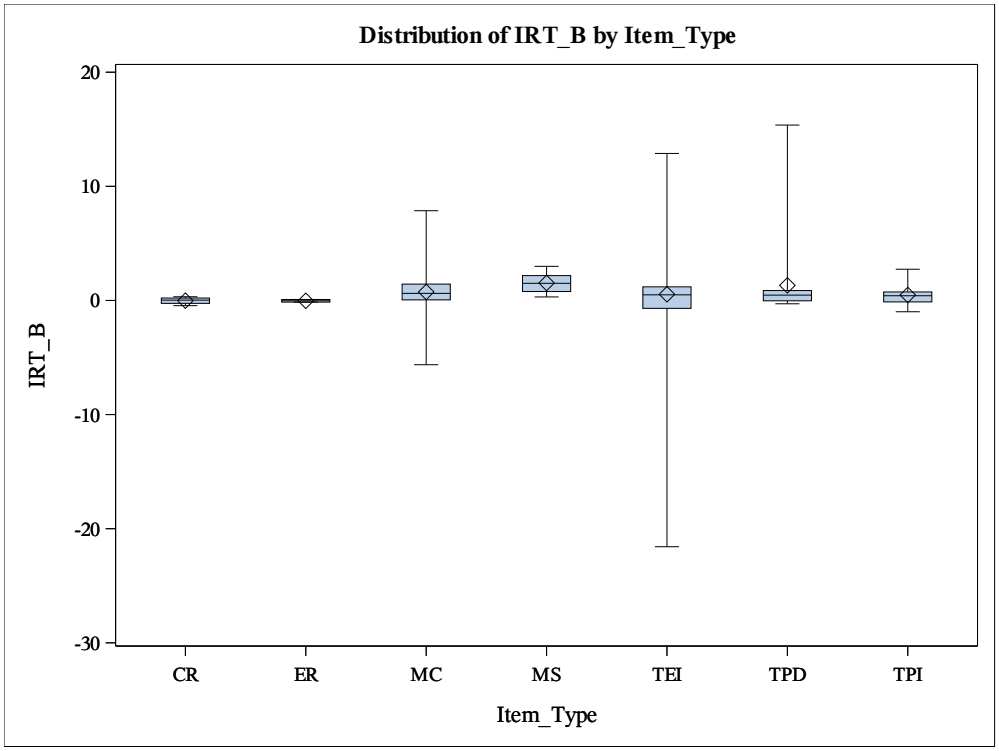
Grade 3



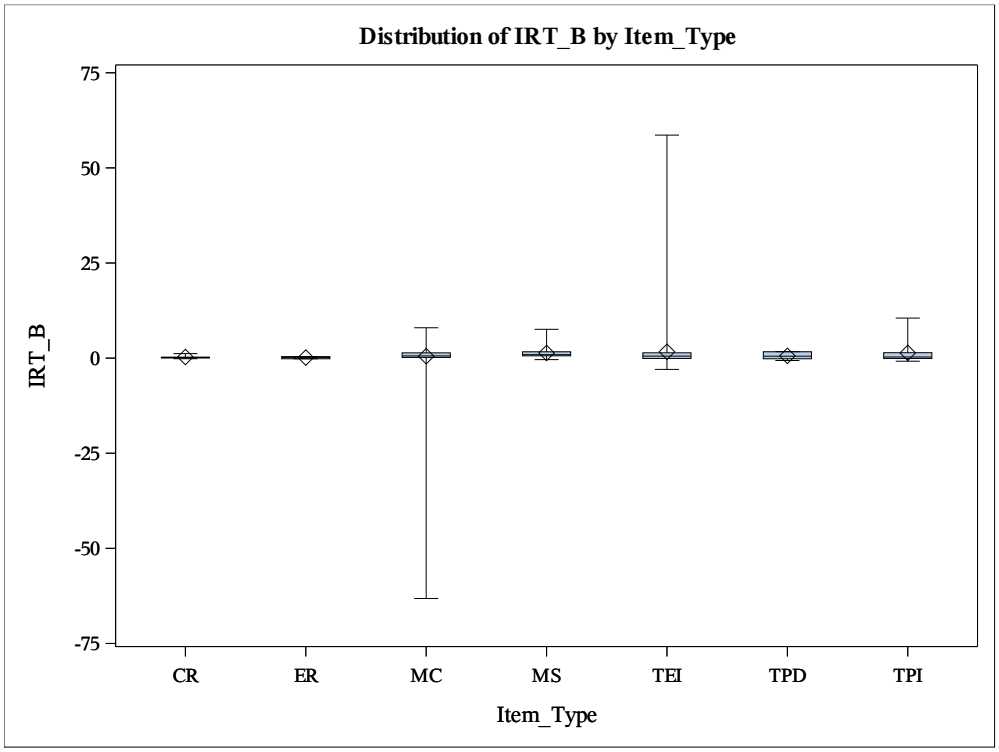
Grade 4



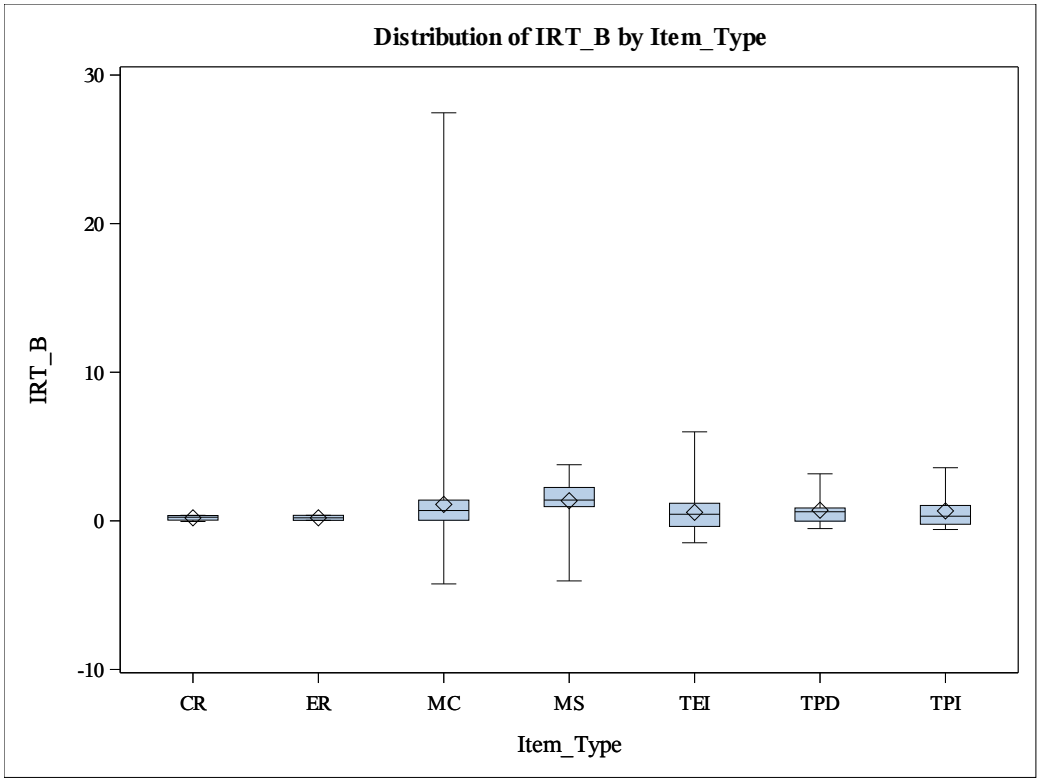
Grade 5



Grade 6



Grade 7



Grade 8

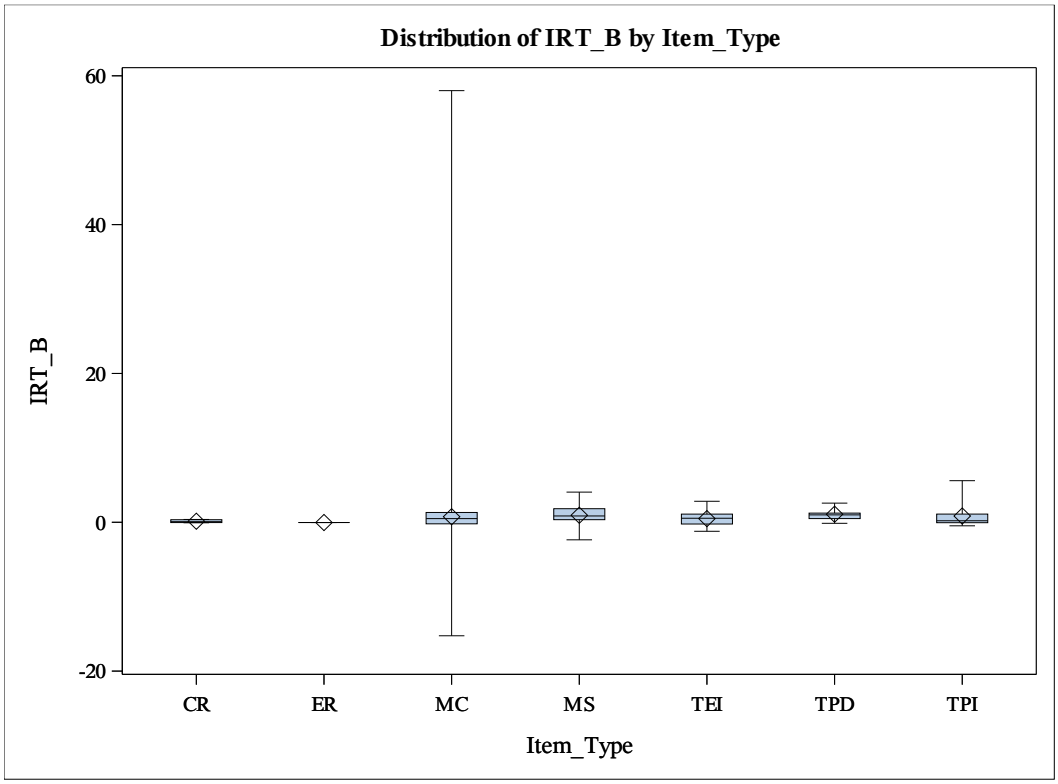
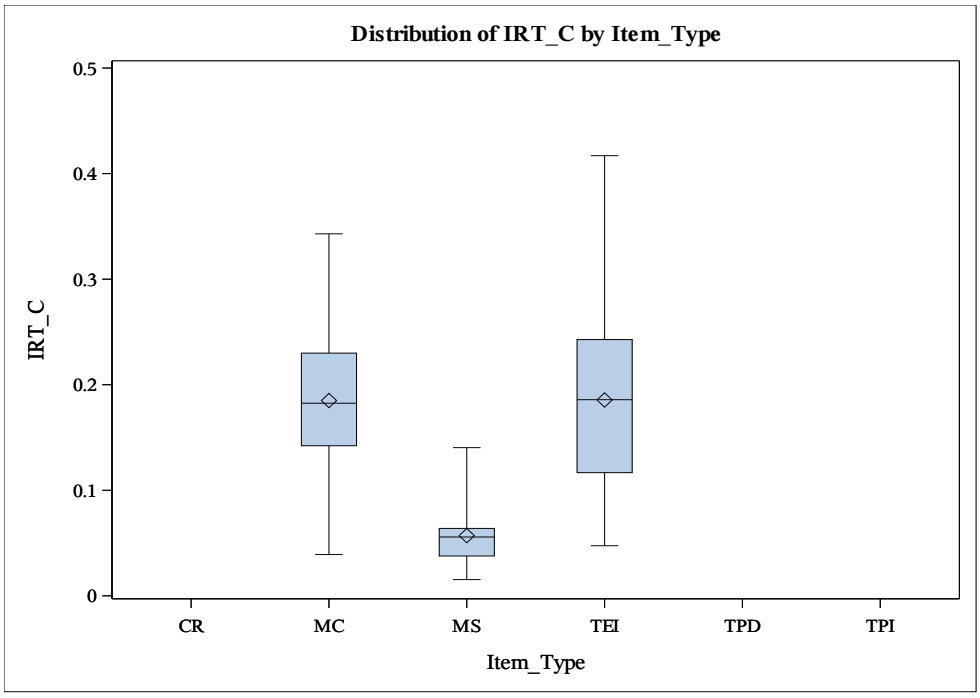


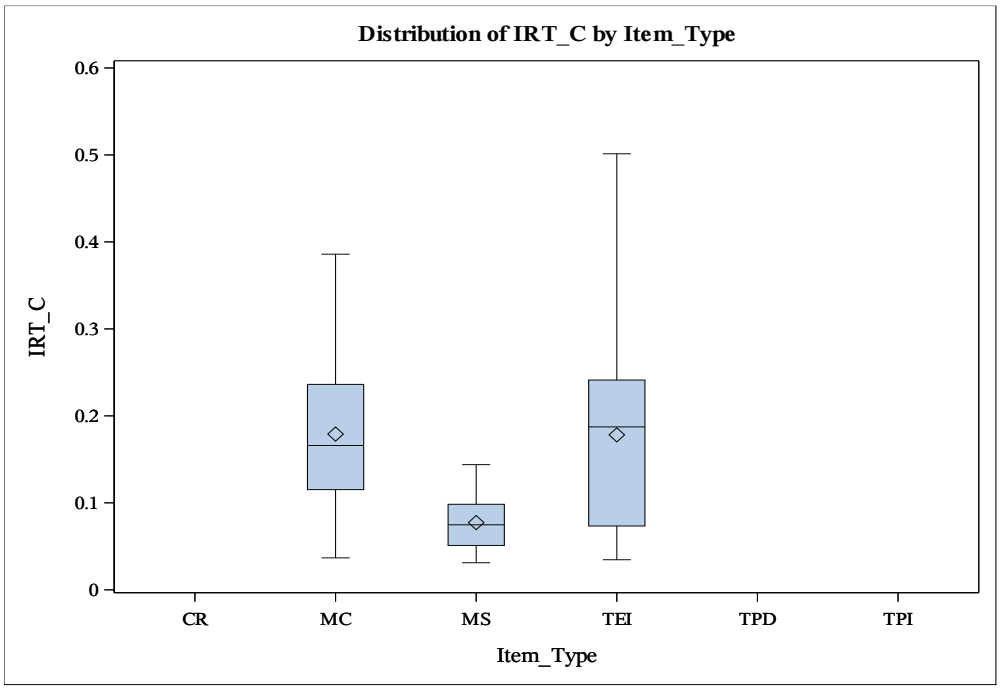
Figure 7.5

IRT C Parameters by Grade

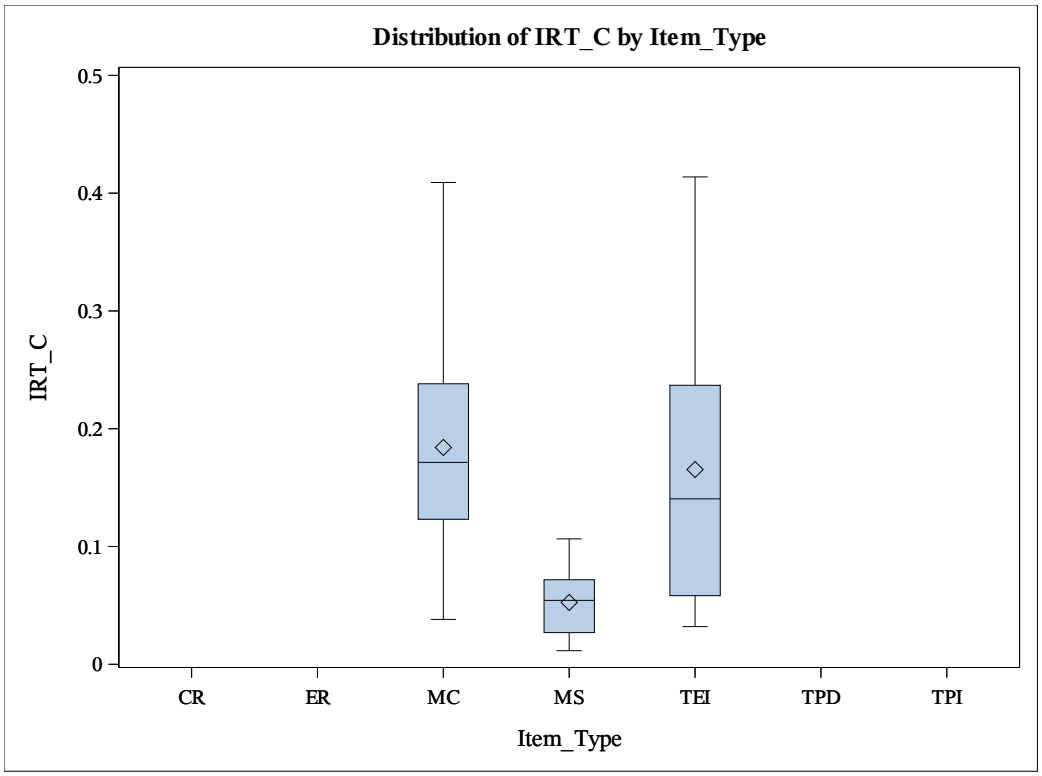
Grade 3



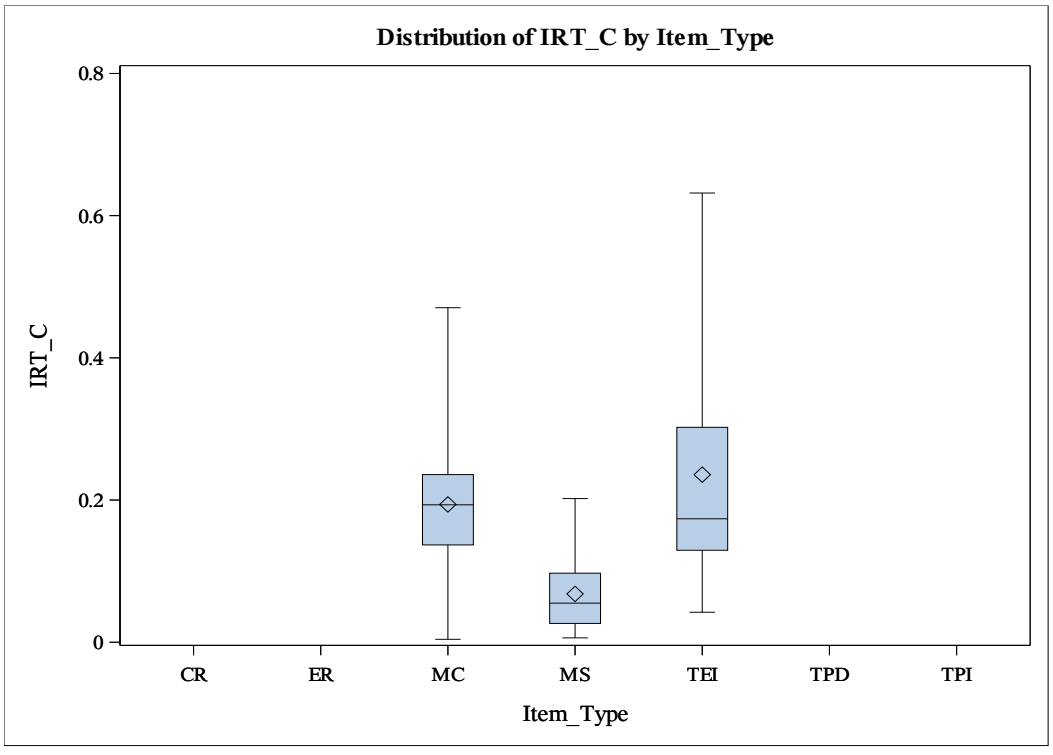
Grade 4



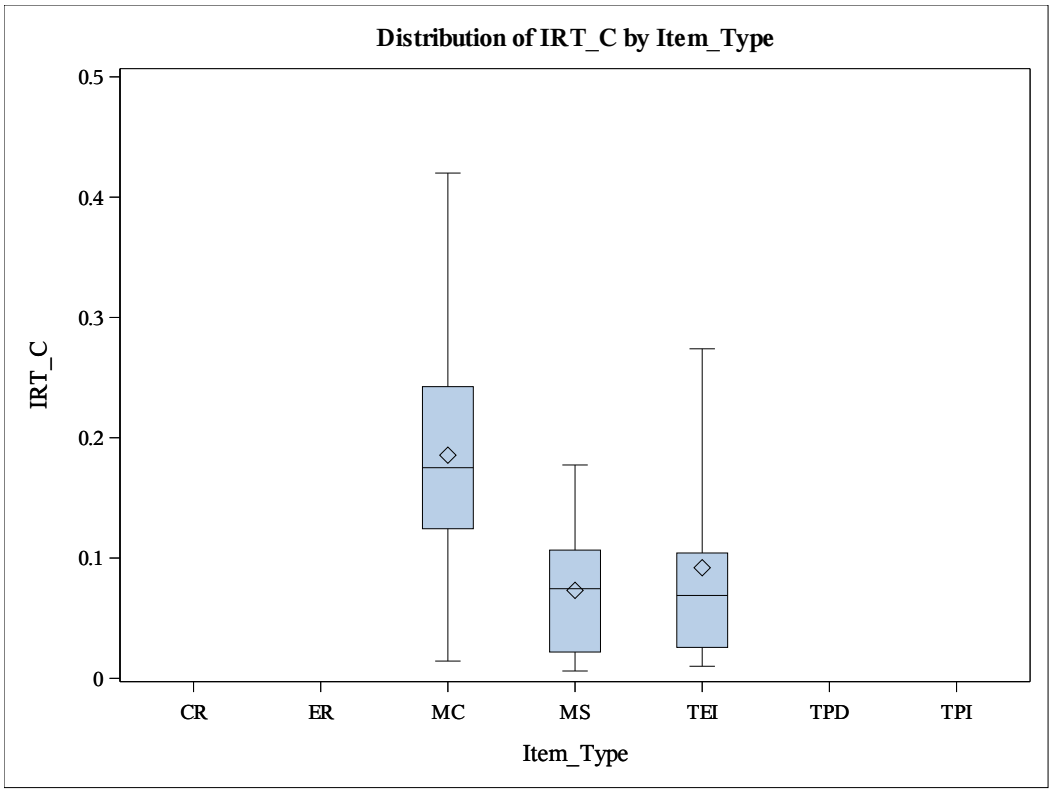
Grade 5



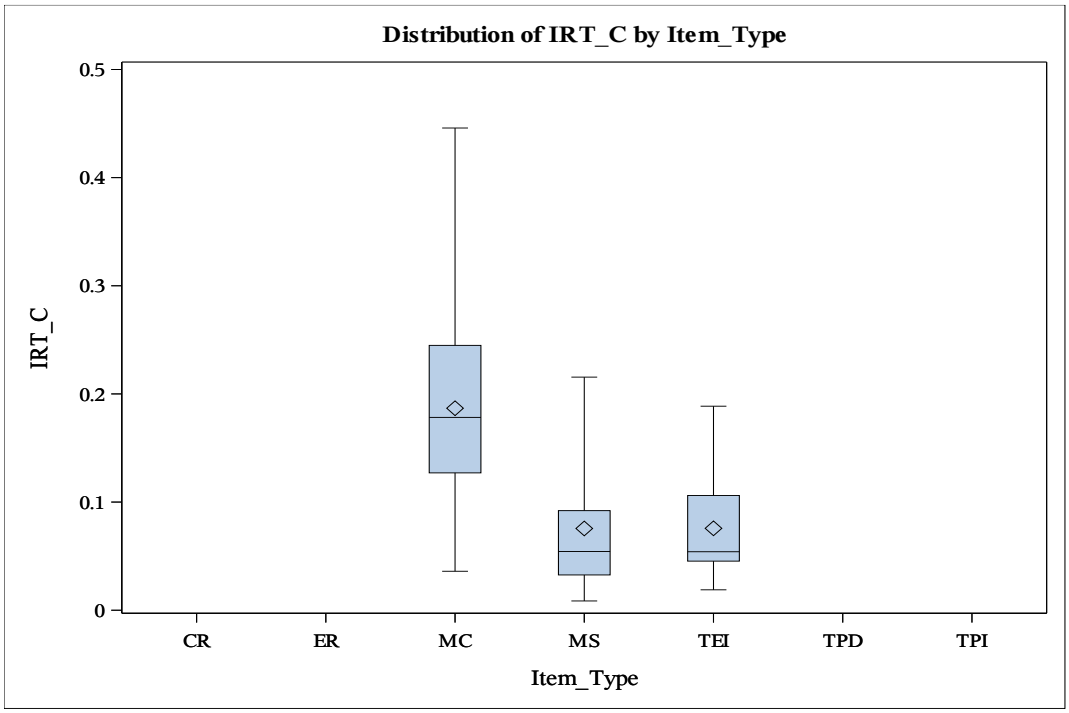
Grade 6



Grade 7



Grade 8



8. Data Review Process

During data review, invited committee members review field-tested items with accompanying data, in order to make judgments about the appropriateness of items for use on operational test forms. As part of the data review process, participants are provided with item statistics that may indicate possible problems. Items are not automatically rejected on the sole basis of statistics; only items with concrete and identifiable flaws in their content are rejected.

The data review meetings for grades 3–8 began with a presentation and introduction to data review. The introductory training included a review of appropriate interpretations on item statistics (difficulty, discrimination, DIF, score distributions), what would be considered reasonable values, and how the values might differ across item types. To reinforce the training, participants were provided with a handout defining item statistics and a checklist including statistical and content considerations to keep in mind while reviewing items.

After signing a nondisclosure agreement, each participant was provided a computer to access Pearson's ABBI platform. Participants reviewed stimuli and statistics of standalone items and item sets on the grades 3–8 field tests in ABBI. Content and psychometric representatives from the LDOE were present in the committee meetings.

Facilitators from Pearson and WestEd led the data review committees through the review of field-tested items by displaying on-screen stimuli and item statistics. Participants were instructed to evaluate the statistical information for each item and determine whether the item functioned as intended. Then, participants provided independent judgments regarding each item's suitability for future operational tests, in light of the field-test statistics. When an item exhibiting DIF was being reviewed, the facilitators specifically asked the committee members to review the DIF statistics and re-evaluate the items for any possible content problems that could lead to the item's possible differential performance. No items exhibiting DIF were identified to have flaws leading to the DIF flags. Judgments were followed by group discussion to reach consensus about each item, and consensus recommendations were then recorded. Specifically, the committees voted to accept, accept with edits (or "revise/re-field test"), or reject items. Tables 7.1–7.6 summarizes the disposition of field-tested items from data review. If the committee's decision was to edit or reject an item, additional information was captured to reflect the

reason for the committee decision. Votes were compiled by the WestEd facilitator and recorded on one main judgment form.

Table 8.1
Summary of Grade 3 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	6	0	0	6
ER	–	–	–	–
MC	143	15	6	164
MS	10	0	3	13
TE	47	6	0	53
TPD	9	1	1	11
TPI	35	6	2	43
Total	250	28	12	290

Table 8.2
Summary of Grade 4 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	8	0	0	8
ER	–	–	–	–
MC	133	1	0	134
MS	7	0	0	7
TE	35	0	0	35
TPD	6	0	0	6
TPI	6	0	0	6
Total	195	1	0	196

Table 8.3

Summary of Grade 5 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	6	0	0	6
ER	2	0	0	2
MC	129	12	2	143
MS	10	7	1	18
TE	47	3	1	51
TPD	15	0	0	15
TPI	10	0	0	10
Total	219	22	4	245

Table 8.4

Summary of Grade 6 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	5	0	0	5
ER	3	0	0	3
MC	135	35	0	170
MS	14	7	0	21
TE	43	11	0	54
TPD	5	1	1	7
TPI	5	7	0	12
Total	210	61	1	272

Table 8.5

Summary of Grade 7 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	4	0	0	4
ER	2	0	0	2
MC	147	29	4	180
MS	8	9	2	19
TE	36	7	0	43
TPD	12	2	0	14
TPI	15	2	0	17
Total	224	49	6	279

Table 8.6

Summary of Grade 8 Data Review Votes

Item Type	Number of Items			
	Accept	Accept w/Edits	Rejected	Total
CR	5	0	0	5
ER	1	0	0	1
MC	139	30	3	172
MS	16	6	0	22
TE	36	10	0	46
TPD	6	2	0	8
TPI	13	5	0	18
Total	216	53	3	272

Following the data review meetings for each grade, LDOE content specialists reviewed items again, with a focus on items that were rejected or accepted with edits. This reconciliation process provided the LDOE with an additional opportunity to review item content and consider possible revisions that would allow items to be field tested again and possibly administered operationally in the future. The reconciliation decisions were treated as the final decisions.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.

Andrich, A. (1988). *Rasch models for measurement*. Sage Publications.

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

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

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

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

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

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

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

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

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

Fleiss, J. L. (1973). Statistical methods for rates and proportions. Wiley.

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

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

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

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

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

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

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

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

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

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

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

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

Young, M. J., & Yoon, B. (1998, April). Estimating the consistency and accuracy of classifications in a standards-referenced assessment (CSE Technical Report 475). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. 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). 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. Directorate of Human Resources Research and Evaluation, Department of National Defense.

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

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

Appendix A: Accommodated Print and Braille Creation

Guidelines for Accommodated Print and Braille

Louisiana believes that all students requiring test accommodations should be presented with the same level of rigor as students taking tests without accommodations. To ensure this, Louisiana creates accommodated versions of the operational test form for each test administration, thereby enabling all students to encounter the same items, irrespective of the necessity for an accommodated presentation. Careful consideration is given to all items utilized in Louisiana assessments, evaluating their suitability to be faithfully represented in accommodated print (AP) and braille formats. Fairness across all populations, preservation of item integrity, and ensuring a consistent student-item interaction for technology-enhanced (TE) items are all factors in the item selection process for the Louisiana form. TE items are modified to ensure that students who interact with an item on an AP or braille form have a similar and equivalent experience to those engaging with that same item in the online environment. This approach maintains both the rigor and the content being assessed. Examples of the modification process are provided below.

- Drag-and-drop items in the online environment require students to place the answer options in an interactive table. For AP and braille forms, students are presented with a table containing the same information as the interactive table—complete with column or row headers, any filled cells, and blank spaces. Below the table, answer options are presented, similar to the online format where options are listed either below or to the right of the table. The directions are modified, instructing students to write the letter or number of the correct answer in its corresponding box. Additionally, students can circle the text, draw arrows to indicate placement, or add labels to the answer choices, writing only the label in the box, as long as the intended response is clear to the test administrator responsible for transcribing answers into the online system.
- Match interaction items in the online environment require students to select a checkbox in one or more columns for each of multiple rows. In the AP and braille forms, students are provided with a table and asked to mark or select the correct answer in each row.

- Highlighted-text items or item parts in the online environment require students to click on the selected text, highlighting the selected word, phrase, or sentence. In the AP and braille forms, the text is presented in the same format and students are asked to circle the answer. Words or phrases that are selectable in the online system, are underlined in the AP and braille forms indicating the words and/or phrases students should select from.
- Drop-down menu items in the online environment have answer options in a drop-down menu format, often as part of a complete sentence. The AP and braille forms display the item with a blank line in place of the drop-down menu within the sentence. Below the sentence, all answer options for the drop-down menu are displayed vertically, each lettered or numbered. The directions ask students to select the corresponding letter/number of the word/phrase that belongs in the blank.
- Short answer items in the online environment require students to type the answer in a box. AP and braille forms provide a box for students to write the response.
- Keypad input items in the online environment require students to enter a numeric response including rational and irrational numbers, as well as expressions and equations. The AP forms provide a box for students to write the response, and braille forms instruct students to answer on the provided paper.
- Graphing items, including coordinate planes, number lines, line plots, and bar graphs, in the online environment require students to complete a graph by plotting points, adding Xs to create a line plot, or raising/lowering bars to create a bar graph or histogram. The AP and braille forms provide students with the identical coordinate plane, number line, line plot, or bar graph featured in the online item. This includes completing the graph with titles, axis labels, and keys.

Displaying items consistently in both AP and braille formats, as well as in the online environment, and enabling students to interact with the items in a uniform manner, maintains item integrity by assessing a similar construct in a consistent manner regardless of how a student encounters an item. This provides students who are unable to access the assessment online with the opportunity to be evaluated at the same level of rigor as the online test.

AP forms undergo thorough review by DRC and LDOE content experts, alongside the online form. Braille forms are evaluated against the AP form by an outside third-party braille expert. Throughout the braille creation process, the vendor relies on the AP form

and consults with LDOE content experts for additional clarification, modifications, or specific items as needed.

Students' responses to the accommodated print or braille test are captured in the same online test used by the general population, either with the assistance of a scribe or by themselves if able. This ensures a valid and reliable assessment for students who are unable to participate in the online assessment. Louisiana's sample sizes are too small for traditional studies of comparability for both AP and braille forms.