West Virginia General Summative Assessment

2021–2022

Volume 1
Annual Technical Report

West Virginia DEPARTMENT OF
EDUCATION

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

The West Virginia General Summative Assessment (WVGSA) is a series of assessments for English language arts (ELA) and mathematics in grades 3–8 and for science in grades 5 and 8. The *WVGSA 2021–2022 Annual Technical Report* is provided to document and make transparent all methods used in item development, test construction, psychometrics, standard setting, test administration, and score reporting, including summaries of student results, and evidence and support for intended uses and interpretations of the test scores. The technical report is provided as seven separate, self-contained volumes, which are updated as changes occur. The volumes include the following:

1) ***Annual Technical Report.*** This annually updated volume provides a global overview of the tests administered to students each school year.

2) ***Test Development.*** This volume summarizes the procedures used to construct test forms and provides summaries of the item bank and item development process.

3) ***Setting Performance Standards.*** This volume documents the methods and results of the WVGSA standard-setting process held in 2018.

4) ***Reliability and Validity.*** This volume provides technical summaries of the test quality and special studies to support the intended uses and interpretations of the test scores.

5) ***Test Administration.*** This volume describes the methods used to administer all tests, security protocols, and available modifications or accommodations.

6) ***Score Interpretation Guide.*** This volume describes the score types reported and details the appropriate inferences that can be drawn from each score.

7) ***Special Studies.*** This volume consists of any special studies conducted for the WVGSA. It is updated each year to reflect studies relevant to the respective administration.

Volume 3, Setting Performance Standards, is the only static volume, as standard setting occurs only when a new testing system is put into place. The West Virginia Department of Education (WVDE) communicates the quality of the WVGSA by making these technical reports accessible to the public on the state's website.

1.1 TEST BACKGROUND AND HISTORICAL CONTEXT

The WVGSA for students in grades 3–8 is an online summative test given toward the end of the school year to measure student performance on the state's content standards. These standards provide clear, consistent guidelines for what students should know and be able to do at each grade level. The WVGSA was first administered to students during spring 2018, replacing the Smarter Balanced Assessment Consortium (SBAC) tests in ELA and mathematics and replacing the West Virginia Educational Standards Test 2 (WESTEST2) in science. Students in grades 3–8 are assessed in ELA and mathematics. Students in grades 5 and 8 are assessed in science, as well. The

ELA assessment consists of two segments: reading and writing. In this document, the term *ELA* is used when referring to the combination of reading and writing, and *reading* is used when referring to only the reading portion of the test.

1.2 PURPOSE AND INTENDED USES OF THE WVGSA

The WVGSA is a criterion-referenced test established using principles of evidence-centered design to yield overall and reporting category-level test scores at the student level and other levels of aggregation that reflect student achievement of the West Virginia College- and Career-Readiness (WVCCR) Standards for ELA and mathematics. It reflects student achievement of the West Virginia Next Generation Standards and Objectives for Science in West Virginia Schools (WV NxGen Science Standards), which were built on the Next Generation Science Standards (NGSS). The WVCCR Standards and WV NxGen Science Standards establish a set of knowledge and skills that all students need to pursue a wide range of high-quality post-secondary opportunities, including higher education and the workplace. The WVGSA supports instruction and student learning by providing valuable feedback to educators and parents, which can be used to form instructional strategies to remediate or enrich instruction. An array of reporting metrics is provided so that achievement can be evaluated at the student level and at aggregate levels and to monitor growth at the student and group levels over time.

The WVGSA ELA and mathematics tests draw all items from the Independent College and Career Readiness (ICCR) item bank (refer to Volume 2, Test Development), which is a rigorously developed bank of items aligned to nationally recognized career and college readiness standards.

For WVGSA science, the three-dimensional NGSS reflect the latest research and advances in modern science and differ from previous science standards in multiple ways. First, rather than describe general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The NGSS refers to such performed knowledge and skills as *performance expectations* (PEs). Second, while unidimensionality is a typical goal of standards (and the items that measure them), the NGSS are intentionally multi-dimensional. Each PE incorporates all three dimensions from the NGSS Framework—a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Third, while traditional standards do not consider other subject areas, the NGSS connects to other subjects like the Common Core mathematics and ELA standards.

Another unique feature of the NGSS is the assumption that students should learn all science disciplines rather than a select few, as is traditionally done in many high schools, where students may elect, for example, to take biology and chemistry but not physics or astronomy. Items are drawn from an item bank that consists of Independent College and Career Readiness (ICCR) items, items owned by West Virginia, and items owned by several other states that share a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods. Full members of the MOU include Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. Cambium Assessment, Inc. (CAI) has a supporting and coordinating role. New Hampshire, North Dakota, South Dakota, and U.S. Virgin Islands observe and participate in some activities. CAI and the WVDE worked together to ensure that the items in the test forms constructed for all grades within the state uniquely measure the WVCCRs for ELA and mathematics and measure the WV NxGen Science Standards for science.

Table 1 outlines the required uses and citations for the WVGSA based on Section 18-2E-5-(d)(3) of the West Virginia Statutes and the federal *Every Student Succeeds Act (ESSA)* plan. The WVGSA fulfills all the requirements described in Table 1.

Table 1. Required Uses and Citations for the WVGSA

Required Use	Citation
Indicator of academic achievement and progress	ESSA Plan Section 1 A. i.; ESSA Plan Section 4 4.1 A
Administer end-of-course mathematics assessments to high school students to meet the requirements under Section 1111(b)(2)(B)(v)(I)(bb) of the Elementary and Secondary Education Act (ESEA)	ESSA Plan Section 3 A
Test administration frequency and grade levels	15.1-21-08.1
Compilation of test scores	15.1-21-09
Publication of test scores	15.1-21-10
Requirement for alignment of test to academic content standards	15.1-21-11

1.3 Participants in the Development and Analysis of the WVGSA

The WVDE manages the West Virginia state assessment program with the assistance of several participants, including West Virginia educators, a Technical Advisory Committee (TAC), and various vendors. WVDE fulfills the diverse requirements of implementing West Virginia's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

1.3.1 West Virginia Department of Education

The WVDE Office of Assessment manages test development, administration, scoring, and reporting of results for the statewide comprehensive assessment programs, including coordinating with other WVDE offices, West Virginia public schools, and vendors.

1.3.2 West Virginia Educators

West Virginia educators participate in most aspects of the conceptualization and development of the WVGSA. Educators participate in the development of the academic standards, the clarification of how these standards are assessed, the test design, and the review of test questions and passages.

1.3.3 Technical Advisory Committee

The WVDE convenes an advisory committee panel multiple times each year to discuss psychometric, test development, administrative, and policy issues of relevance to current and future West Virginia assessments. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple West Virginia school districts.

1.3.4 Cambium Assessment, Inc.

Cambium Assessment, Inc. (CAI) is the vendor selected through the state-mandated competitive procurement process. CAI is responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the WVGSA described in this report. Additionally, CAI is responsible for developing and maintaining the ICCR item bank.

1.3.5 Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2022 test administration to ensure that secure testing materials such as items and prompts were not leaked.

1.4 Available Test Formats and Special Versions

The WVGSA for ELA and mathematics are administered as online assessments using an adaptive item selection algorithm (refer to Volume 2, Appendix K) and making use of technology-enhanced item types. For science, the test is administered online using an adaptive test design. Science items are centered on a scientific phenomenon. They can consist of shorter items (stand-alone) or items with several parts (item clusters) requiring the student to interact with them in various ways. The science test was an operational field test in 2018, the first year of the new science assessment, and was an operational test with embedded field-test slots in spring 2019, 2021, and 2022. In every administration, new items are field tested to build out the item bank.

Students unable to participate in the online administration have the option to use print-on-demand— a feature that provides the same items administered to students online in a paper format. Spanish versions of mathematics tests and science tests (developed to meet the same content standards as the English versions) are available for all tested grades. Students participating in the computer-based WVGSA can use standard online testing features in the Test Delivery System (TDS), including a selection of font color and size and the ability to zoom in and zoom out or highlight text. In addition to the resources available to all students, options are available to accommodate students with an Individualized Education Program (IEP) or Section 504 Plan. These include braille, American Sign Language (ASL), closed captioning, and large print. Students with disabilities have the option to take the WVGSA with or without accommodations or to take an alternate assessment. For additional information about the testing features and testing accommodations, refer to Volume 5, Test Administration.

1.5 Student Participation

All students in West Virginia public schools are required to participate in the statewide assessments. The WVGSA for ELA, mathematics, and science are administered in the spring.

Table 2 shows the number of students who were tested (Number Tested) and the number of students whose scores were included for the analyses in this technical report (Number Reported). Students who took a print-on-demand or braille form in ELA and mathematics, as well as those who took a Spanish version of the mathematics test, were excluded in all summary results in this report, unless otherwise noted (21 students took a large print Data Entry Interface [DEI] form for

ELA and mathematics, and 47 students took a Spanish language version of the mathematics test). Table 3 through Table 5 show the demographic characteristics of the student population, by counts and percentages, in the spring administration of the 2021–2022 assessments. The subgroups reported here are gender, ethnicity, and students with limited English proficiency (LEP). Also included in Table 3 are those students who declined to report their ethnicity. These students were not included in the remaining demographic tables.

Table 2. Number of Students Participating in WVGSA, Spring 2022

Grade	ELA		Mathematics		Science	
	Number Tested	Number Reported	Number Tested	Number Reported	Number Tested	Number Reported
3	17,547	17,526	17,572	17,548	-	-
4	17,355	17,323	17,345	17,342	-	-
5	17,737	17,684	17,730	17,727	17,699	17,698
6	17,750	17,699	17,731	17,721	-	-
7	18,340	18,242	18,308	18,302	-	-
8	18,776	18,698	18,749	18,742	18,698	18,694

Table 3. Demographic Distribution of Tested Population, ELA

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
3	*N*	17,526	8,538	8,988	685	6	102	370	800	10	15,252	301	185
	%	100	48.72	51.28	3.91	0.03	0.58	2.11	4.56	0.06	87.02	1.72	1.06
4	*N*	17,323	8,454	8,869	630	8	98	370	794	8	15,129	286	131
	%	100	48.8	51.2	3.64	0.05	0.57	2.14	4.58	0.05	87.33	1.65	0.76
5	*N*	17,683	8,611	9,072	649	7	113	372	804	8	15,439	291	131
	%	100	48.7	51.3	3.67	0.04	0.64	2.1	4.55	0.05	87.31	1.65	0.74
6	*N*	17,697	8,678	9,019	683	13	100	410	723	9	15,282	477	100
	%	100	49.04	50.96	3.86	0.07	0.57	2.32	4.09	0.05	86.35	2.70	0.57
7	*N*	18,242	8,979	9,263	783	28	117	390	696	3	15,946	279	122
	%	100	49.22	50.78	4.29	0.15	0.64	2.14	3.82	0.02	87.41	1.53	0.67
8	*N*	18,698	9,012	9,686	717	20	120	440	729	9	16,371	292	133
	%	100	48.2	51.8	3.83	0.11	0.64	2.35	3.9	0.05	87.55	1.56	0.71

Table 4. Demographic Distribution of Tested Population, Mathematics

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
3	*N*	17,548	8,551	8,997	686	6	102	370	802	10	15,270	302	184
	%	100	49	51	3.91	0.03	0.58	2.11	4.57	0.06	87.02	1.72	1.05
4	*N*	17,342	8,463	8,879	631	8	98	369	797	8	15,143	288	131
	%	100	49	51	3.64	0.05	0.57	2.13	4.6	0.05	87.32	1.66	0.76

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
5	*N*	17,727	8,637	9,090	651	7	113	372	806	8	15,483	287	131
	%	100	49	51	3.67	0.04	0.64	2.1	4.55	0.05	87.34	1.62	0.74
6	*N*	17,721	8,687	9,034	708	13	100	411	741	9	15,437	302	100
	%	100	49	51	4	0.07	0.56	2.32	4.18	0.05	87.11	1.7	0.56
7	*N*	18,302	9,006	9,296	788	29	118	392	698	3	15,995	279	124
	%	100	49	51	4.31	0.16	0.64	2.14	3.81	0.02	87.39	1.52	0.68
8	*N*	18,742	9,031	9,711	718	20	120	440	732	9	16,409	294	133
	%	100	48	52	3.83	0.11	0.64	2.35	3.91	0.05	87.55	1.57	0.71

Table 5. Distribution of Demographic Characteristics of Tested Population, Science

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multiple race	Pacific Islander	White	Declined to Report	LEP
5	*N*	17,698	8,621	9,077	647	7	112	371	800	8	15,305	448	131
	%	100.00	48.71	51.29	3.66	0.04	0.63	2.10	4.52	0.05	86.48	2.53	0.74
8	*N*	18,694	9,013	9,681	711	20	120	438	723	9	16,265	408	134
	%	100.00	48.21	51.79	3.80	0.11	0.64	2.34	3.87	0.05	87.01	2.18	0.72

2. Operational Practices and Procedures

2.1 Test Administration

Table 6 shows the testing window for the 2021–2022 WVGSA by subject. As a part of the statewide assessment, interim assessments for English language arts (ELA) and mathematics were administered with multiple opportunities prior to the WVGSA for local school districts and staff to monitor students' progress. The interim assessment results related to the WVGSA ELA and mathematics are presented in Volume 4, Evidence of Reliability and Validity, of this technical report.

Table 6. WVGSA Testing Windows by Subject Area

Subject	Grade(s)	Testing Window
ELA (Reading and Writing)	3–8	April 4–May 27, 2022
Mathematics	3–8	April 4–May 27, 2022
Science	5 & 8	April 4–May 27, 2022

The key personnel involved with the West Virginia test administration included the district test coordinators (DCs), school test coordinators (SCs), and test administrators (TAs), who proctored the test. Test administration manuals were provided so that personnel involved with the statewide assessment administrations could maintain both standardized administration conditions and test security.

A secure browser developed by Cambium Assessment, Inc. (CAI) was required to access the online WVGSA tests. The online browser provided a secure environment for student testing by disabling the hot keys, copy, and screen capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). During the online assessment, students could pause a test, review previously answered questions, and modify their response if the test had not been paused for more than 20 minutes. Students do not have a required time limit for each test session, but for test administration planning purposes, schools are given approximate time estimates for how long each test may take for most students.

2.2 Simulations

Prior to the operational testing window, CAI employs a simulation approach. Simulations are performed for all WVGSA assessments, including ELA, mathematics, and science.

For ELA and mathematics, simulations are used to configure the adaptive algorithm (described further in Volume 2, Part 1: ELA and Mathematics, Appendix K: ICCR Adaptive Algorithm Design), seeking to maximize test score precision while meeting blueprint specifications based on the available pool of test items. Psychometricians review ELA and mathematics simulation results for the following key diagnostic factors:

- The **match-to-test blueprint** determines that the tests have the correct number of test items overall and the appropriate proportion by content strands, as specified in the test blueprints for every grade and subject.
- **Precision** determines whether the size of the standard error of measurement (SEM) is within the acceptable range and whether there is any possible bias in the estimates of student ability.
- The **item exposure rate** evaluates the utility of item pools and identifies overexposed and underexposed items.

These diagnostics are interrelated. For example, if the test pool for a particular content strand is limited (i.e., if there are only a few items available), achieving a 100% match to the blueprint for this content strand will lead to a high item-exposure rate, which means that a large number of students will see the same items. A high item-exposure rate results in decreased benefits from adaptive testing relative to using a fixed form, such as the usual increased security caused by a larger pool of items. The software system that performs the simulation allows the adjustment of test configuration to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews initial results, adjusts these system parameters, runs new simulations, reviews new results, and repeats the exercise until an optimal balance is achieved. The final setting is then applied for operational tests. The ELA and mathematics simulation reports in Appendix A: Simulation Summary Report describe in detail the simulation approach and results evaluated based on blueprint, precision, and item exposure rate.

For science, administered under an adaptive test design, the test is delivered using an item-selection algorithm in which operational items are selected on the fly based on a student's performance on past items while ensuring that the test blueprint is followed for each individual student. Simulations were carried out to configure the settings of the algorithm and to evaluate whether individual tests adhered to the test blueprint and monitor item exposure rates. The simulation approaches and results for science are discussed in Volume 2, Test Development, of this technical report.

2.3 ACCOMMODATIONS

The accessibility supports discussed in this volume include embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that are available to all students as they access instructional or assessment content; and accommodations that are generally available for students for whom there is documentation on an Individualized Education Program (IEP) or Section 504 Plan. For English learners (ELs), Spanish language versions of the WVGSA mathematics and science are available.

Scores achieved by students using designated supports are included for federal accountability purposes. All educators making these decisions were trained on the process and understand the range of designated supports available.

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech) are provided digitally through instructional or assessment technology, and non-embedded accommodations (e.g., scribe) are non-digital. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. Such accommodations help students with a documented need generate valid outcomes of the assessments so that they can fully demonstrate what students know and are able to do. From the psychometric point of view, the purpose of providing accommodations is to "increase the validity of inferences about students with special needs by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

This potential for an alteration of the construct of interest with the use of an accommodation is a primary concern whenever they are considered for use. CAI has completed two studies to evaluate the use of dictionaries and glossaries as accommodations. The results of these studies are presented in Appendix K, Investigating the Effects of Dictionary Availability on Item Performance and Appendix L, Effectiveness of Computer-Based, Pop-Up Glossaries, respectively.

West Virginia TAs and STCs are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students include braille, American Sign Language (ASL), closed captioning, streamline, abacus, assistive technology (e.g., adaptive keyboards, touch screens, switches), calculation device, print-on-demand, multiplication table, and scribe. Descriptions for each of these accommodations can be found in Volume 5, Test Administration.

Table 7 through Table 9 list the number of testing sessions in which a student was provided with each accommodation during the spring 2022 test administration.

Table 7. Number of Testing Sessions with Allowed Embedded and Non-Embedded Accommodations, ELA

Accommodations	G3	G4	G5	G6	G7	G8
Embedded Accommodations						
American Sign Language	4	5	1	2	3	5
Braille	-	-	-	1	-	-
Color Choices	-	-	2	3	2	2
Closed Captioning	8	15	18	10	20	19
Dictation (Speech-to-Text)	2	1	1	4	4	3
Emboss: Stimuli and Items	-	-	1	2	2	-
Line Tracker	1	2	-	-	-	-
Masking	30	23	17	24	12	9
Mouse Pointer	1	-	-	-	-	-
Permissive Mode	38	41	53	54	55	64
Print-on-Demand: Stimuli & Items	-	1	2	1	2	--

Accommodations	G3	G4	G5	G6	G7	G8
Streamlined Interface Mode	16	28	25	25	23	14
Text-to-Speech: Instructions, Passages, & Items	2,533	2,868	3,014	2,834	2,867	2,746
Text-to-Speech: Instructions & Items	1,004	1,077	1,179	1,250	1,278	1,276
Zoom	6	3	8	-	2	-
Non-Embedded Accommodations						
Magnification	-	-	-	-	-	-
Print-on-Demand: Stimuli & Items	-	1	2	1	2	-
Scribe	-	-	-	-	-	-

Table 8. Number of Testing Sessions with Allowed Embedded and Non-Embedded Accommodations, Mathematics

Accommodations	G3	G4	G5	G6	G7	G8
Embedded Accommodations						
American Sign Language	-	-	-	-	-	-
Braille	1	-	1	2	2	-
Closed Captioning	-	-	-	-	-	-
Color Choices	-	-	2	3	1	21
Dictation (Speech-to-Text)	44	65	45	29	19	15
Emboss: Stimuli and Items	1	-	1	2	2	-
Language Format (Spanish)	3	10	2	9	16	24
Line Tracker	27	20	12	21	6	8
Masking	26	25	19	24	10	6
Mouse Pointer	1	-	-	-	-	-
Permissive Mode	38	39	57	55	58	62
Print-on-Demand: Stimuli & Items	-	1	2	1	2	-
Streamlined Interface Mode	17	28	26	26	25	13
Text-to-Speech: Instructions, Passages, & Items	2,436	2,770	2,950	2,773	2,805	2,659
Translations – Stacked	-	-	-	-	-	-
Zoom	6	3	5	-	3	-
Non-Embedded Accommodations						
Magnification	-	-	-	-	-	-
Print-on-Demand: Stimuli & Items	-	1	2	1	2	-
Scribe	-	-	-	-	-	-

Table 9. Number of Testing Sessions with Allowed Accommodations, Science

Accommodations	G5	G8
Color Choices	2	19
Dictation	30	6
Emboss	1	-
Language- Spanish	2	24
Line Reader	12	8
Masking	12	6
Permissive Mode	57	65
Print on Demand: Stimuli & Items	2	-
Streamlined Mode	25	14
Text-to-Speech: Stimuli & Items	2,949	2,644
Zoom/Print Size	6	-

3. Item Bank and Test Design

Content specialists and psychometricians reviewed all items in the Independent College and Career Readiness (ICCR) item banks with respect to item statistics, bias, and sensitivity for West Virginia. Selected items after these reviews were used for the West Virginia operational item pool. In this section, we describe the characteristics of the spring 2022 operational item pool for the computer-adaptive tests (English language arts [ELA] and mathematics) and the online tests administered adaptively (science). The characteristics include both content (e.g., item types) and statistical summaries. Test design and methodology of field testing new items are also discussed.

3.1 ELA and Mathematics Item Bank

For ELA and mathematics, all operational items used on the WVGSA test forms are drawn from Cambium Assessment, Inc.’s (CAI) West Virginia Item Authoring Tool (IAT) item bank. Volume 2, Test Development, is a separate, stand-alone report containing complete details on the ICCR bank. Here, we note that it is a pre-equated item bank with item parameters estimated under a multi-group item response theory framework, described in a later section of this volume.

The operational item pool included an array of item types. Each of the item types is described in Table 10 and Table 11. Table 12 and Table 13 show the number of items by item type that were available in the item pool. Examples are available in Volume 2, Part 1, Appendix C, Example Item Types.

Table 10. Item Types and Descriptions, ELA

Response Type	Description
Editing Task Choice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
Multiple Choice/Select + Editing Task Choice (Two-part ETC)	Student selects the correct answer from Part A and Part B. Part A is multiple choice or multiple select, and Part B is editing task choice.
Evidence-Based, Selected-Response (EBSR)	Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A.
Extended Response (ER)	Student is directed to provide a longer, written response in the form of an essay.
External Copy [block/line]	Student is directed to select text to support an analysis or make an inference.
Grid (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot Text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference.
Multiple Choice/Select + Hot Text (Two-part HT)	Student selects the correct answer from Part A and Part B. Part A is multiple choice or multiple select, and Part B is hot text.
Multiple Choice (MC)	Student selects one correct answer from a number of options.
Matching (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Natural Language (NL)	Student is directed to provide a short, written response.
Text Entry (TE)	Student is directed to type their response in a text box.

Table 11. Item Types and Descriptions, Mathematics

Response Type	Description
Editing Task Choice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
Multiple Choice/Select + Editing Task Choice (Two-part ETC)	Student selects the correct answer from Part A and Part B. Part A is multiple choice or multiple select, and Part B is editing task choice.
Equation (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. Responses can include numbers, fractions, expressions, inequalities, functions, and equations.
Multiple Choice/Select + Equation (Two-part EQ)	Student selects the correct answer from Part A and Part B. Part A is multiple choice or multiple select, and Part B is equation.
Grid (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot Text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference.

Response Type	Description
Multiple Choice (MC)	Student selects one correct answer from four options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Table Input (TI)	Student types numeric values into a given table.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.

Table 12. Operational Item Pool by Item Type, ELA

Item Type	Grade					
	3	4	5	6	7	8
MC	338	308	281	378	290	278
MS	23	34	38	47	52	23
MI	16	7	16	10	2	2
GI	-	-	1	-	-	-
ETC	49	58	45	43	43	39
Two-part ETC	-	-	-	-	1	-
HT	33	35	54	33	39	41
Two-part HT	2	4	4	6	4	1
EBSR	2	1	5	3	10	1
TE	2	2	2	2	2	2
NL	-	2	-	-	3	-

Table 13. Operational Item Pool by Item Type, Mathematics

Item Type	Grade					
	3	4	5	6	7	8
MC	138	113	115	180	123	188
MS	55	99	48	54	24	48
GI	84	56	31	43	37	53
ETC	1	6	5	3	1	4
Two-part ETC	-	1	-	-	2	1
TI	15	15	10	29	3	8
MI	13	34	16	14	10	9
EQ	347	366	333	352	298	253
Two-part EQ	-	1	3	-	-	2

3.1.1 Field Test

The adaptive 2021–2022 ELA and mathematics tests contained new field-test items in the non-scored embedded field-test (EFT) slots. The EFT slots are embedded into position in the middle of tests such that item location and motivation effects, if they exist, do not propagate into the estimates of the item parameters. To obtain high-quality responses to the EFT items, students were unaware of which items were operational and which were EFT.

For ELA reading, six to eight EFT items per test were administered; for mathematics, eight EFT items were administered in grades 3 – 5 and 7 – 8, four EFT items were administered in grade 6.

The spring 2022 ELA and mathematics EFT items were put onto the West Virginia reporting scale by using a fixed anchor item calibration method. The field-test items were administered in multiple states, such as Arizona, Wyoming, New Hampshire, North Dakota, and West Virginia. All operational (treated as fixed anchor) and field-test items were merged into a single incomplete data matrix for a multiple-group item response theory (MGIRT) calibration. Operational item parameters were fixed to their bank values, while field-test item parameters were estimated in a single run. If a calibration run did not converge, the reason was investigated. Usually, one or two items with negative item-total correlations were the cause. Those items were removed from the calibration and sent to the CAI content team for further action, such as a revision or rejection. The state group means, provided in Appendix J, Calibration Group Means and Standard Deviation for Spring 2022 Field-Test Items, were obtained during free estimations.

3.1.2 Operational Test Design

ELA and mathematics tests were assembled using CAI’s adaptive testing algorithm. The adaptive item-selection algorithm selects items based on their content value and information value. The algorithm ensures that each student receives a unique test that adheres to the content requirements described in the WVGSA test specifications, ensuring comparable and sufficient coverage of the content of the West Virginia College- and Career-Readiness (WVCCR) Standards. In addition, each student’s unique test assembled by the algorithm contains the items that best match students’ achievement levels, as defined by the blueprint. The details of the adaptive item selection algorithm for ELA and mathematics are presented in Volume 2, Test Development, of this technical report.

3.1.3 Operational Item Pool Statistics

As reported in Section 2.2, Simulations, a simulation approach to configure the adaptive algorithm was conducted prior to the operational testing window to maximize test score precision while meeting blueprint specifications based on the available pool of test items. The blueprint match was monitored for both simulation and operational administration. The summary of the simulation versus operational blueprint match for spring 2022 ELA and mathematics is provided in Appendix B, Simulation vs. Operational Blueprint Match. The summary shows that, across all grades and subjects, most tests met the blueprint specifications with a 100% match at the reporting category level in both simulation and operational administrations. There were a few exceptions in grade 7 and grade 8 ELA operational administrations, as a small number of students took the test for the same grade in both 2021 and 2022. The Test Delivery System (TDS) prevents administration of any item more than once to the same student, resulting in a smaller item pool available for students retaking the same test.

The item response theory statistical properties of the operational item pool used for the 2022 WVGSA are summarized in Table 14 through Table 19 for reading and mathematics. 3PL and 2PL refer to the three-parameter logistic model and the two-parameter logistic model, respectively, while GPCM is the generalized partial credit model. Minimum, maximum, and five-point percentiles are summarized for discrimination (*a*), difficulty (*b*), and guessing (*c*) parameters for 3PL items, and *a* and *b* parameters for 2 PL items. For GPCM, step parameters (*b1* and *b2*) are summarized.

Table 14. 3PL Operational Item Parameters Five-Point Summary and Range, ELA

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	306	0.2983	0.5748	0.8893	1.1571	1.4878	2.1034	11.8346
	b	306	-2.3642	-1.8515	-1.3163	-0.8862	-0.5021	0.3144	1.8919
	c	306	0.0281	0.0769	0.1426	0.1912	0.2483	0.3288	0.5934
4	*a*	273	0.1935	0.4309	0.7394	1.0017	1.3241	1.8111	2.4456
	b	273	-2.5463	-1.8547	-1.3222	-0.77	-0.1304	0.7009	2.3762
	c	273	0.0092	0.0507	0.1023	0.168	0.2228	0.3087	0.3707
5	*a*	254	0.2343	0.5045	0.8025	1.0577	1.335	1.7657	2.49
	b	254	-2.1055	-1.4472	-0.749	-0.3061	0.212	0.9553	2.5438
	c	254	0.0345	0.0651	0.1361	0.1856	0.2408	0.3119	0.425
6	*a*	309	0.1774	0.3976	0.7226	0.9827	1.259	1.6554	3.203
	b	309	-2.3266	-1.057	-0.4082	0.0388	0.6256	1.5053	5.6711
	c	309	0.005	0.0583	0.1271	0.1827	0.247	0.3254	0.4157
7	*a*	253	0.1115	0.446	0.7025	0.9049	1.1745	1.5904	2.7642
	b	253	-1.9767	-0.9815	-0.2399	0.3689	0.8184	1.7485	7.4043
	c	253	0.0076	0.0288	0.1101	0.1671	0.2388	0.3163	0.4147
8	*a*	239	0.0539	0.44	0.6701	0.9139	1.1082	1.437	2.0425
	b	239	-2.2389	-0.8547	-0.1415	0.3622	1.1166	2.0295	3.7823
	c	239	0.0039	0.0459	0.1153	0.1752	0.2547	0.3265	0.4308

Table 15. 2PL Operational Item Parameters Five-Point Summary and Range, ELA

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	134	0.0325	0.4073	0.6887	0.8636	1.0606	1.3677	1.9227
	b	134	-4.9942	-2.6384	-1.3653	-0.6894	-0.1649	0.7754	2.2434
4	*a*	153	0.0435	0.33	0.4847	0.6826	0.885	1.2581	1.5932
	b	153	-2.9592	-2.007	-1.1441	-0.4182	0.1849	1.895	5.3057
5	*a*	165	0.188	0.3632	0.5746	0.7576	0.946	1.2292	1.4684
	b	165	-2.1508	-1.6067	-0.7888	-0.1251	0.8003	1.7489	4.9807
6	*a*	182	0.1233	0.3089	0.5735	0.7256	0.901	1.1228	1.5377
	b	182	-2.1346	-1.5654	-0.1711	0.325	1.0523	2.8179	6.7526
7	*a*	166	0.1864	0.3125	0.478	0.6869	0.8559	1.2855	1.5975
	b	166	-2.3078	-1.3298	-0.278	0.4807	1.247	2.6569	4.9101
8	*a*	120	0.0568	0.2888	0.4597	0.6352	0.8365	1.0366	1.2206
	b	120	-4.599	-1.3045	-0.1246	0.5423	1.158	2.4216	5.3089

Table 16. GPCM Operational Item Parameters Five-Point Summary and Range, ELA

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	29	0.2509	0.2871	0.4989	0.788	1.1118	1.6213	1.7954
	b1	29	-3.698	-3.2067	-2.5037	-2.2593	-2.0032	-0.2972	1.0753
	b2	29	-4.3081	-1.8205	-1.4779	-0.9004	-0.3466	1.091	1.6529
	b3	29	-1.1574	-1.0974	-0.7743	0.5754	0.7203	0.789	0.8016
4	*a*	29	0.1682	0.3288	0.443	0.6356	0.8646	1.1802	1.2362
	b1	29	-3.4514	-3.264	-2.4722	-2.308	-1.5986	-0.5413	-0.2401
	b2	29	-2.1662	-1.7269	-1.0132	-0.5447	1.1484	2.143	3.9502

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b3	29	1.7205	1.7214	1.7251	1.8051	2.0451	2.4327	2.5296
5	*a*	32	0.2352	0.2571	0.4586	0.4968	0.7308	1.5161	1.5892
	b1	32	-3.1989	-2.8277	-2.1986	-1.8992	-1.39	-0.6471	0.4765
	b2	32	-1.4931	-1.1405	-0.5474	-0.1226	0.5558	1.4906	4.556
	b3	32	-1.9509	-1.825	-1.184	-0.4849	1.5015	2.4551	2.7386
6	*a*	35	0.2606	0.2962	0.3973	0.4919	0.633	1.4869	1.6881
	b1	35	-4.7118	-2.9879	-2.0591	-1.702	-0.5103	1.1278	2.9023
	b2	35	-2.7477	-1.548	-0.6144	0.1992	1.1584	1.739	1.9061
	b3	35	-1.4101	-1.2152	-0.4706	2.3372	2.6207	2.8477	2.8883
7	*a*	31	0.2057	0.2882	0.3825	0.514	0.767	1.5609	1.598
	b1	31	-3.1413	-2.214	-1.9718	-1.2935	-0.6047	1.9601	4.2022
	b2	31	-1.2501	-0.78	-0.1752	0.241	1.3056	2.6791	3.367
	b3	31	-0.3411	0.294	2.8345	3.0505	3.0905	3.0954	3.0967
8	*a*	32	0.2458	0.3168	0.3863	0.574	0.7306	1.3245	1.407
	b1	32	-3.0658	-2.6381	-1.7255	-1.3143	-0.9137	1.3579	2.0549
	b2	32	-1.0479	-0.851	-0.3704	-0.0793	0.8623	1.8441	2.4453
	b3	32	-0.4342	-0.4179	0.2528	2.1561	2.1981	2.2139	2.2187

Table 17. 3PL Operational Item Parameters Five-Point Summary and Range, Mathematics

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	138	0.4188	0.8655	1.2061	1.5115	1.8952	2.6063	4.1722
	b	138	-4.6052	-3.7391	-2.7339	-2.3416	-1.8784	-1.3518	-0.4886
	c	138	0.0124	0.066	0.1255	0.1887	0.2489	0.371	0.5925
4	*a*	113	0.2404	0.6536	0.9682	1.2161	1.528	1.8371	2.9673

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b	113	-3.8682	-3.2116	-2.399	-1.7444	-1.2504	-0.6068	0.3446
	c	113	0.033	0.0722	0.1332	0.1766	0.2606	0.3954	0.5991
5	*a*	114	0.2222	0.4639	0.769	0.9452	1.3401	1.9268	2.8706
	b	114	-5.696	-2.4702	-1.6962	-1.1375	-0.4529	0.1162	1.1471
	c	114	0.0391	0.0712	0.1423	0.1817	0.2379	0.3367	0.5628
6	*a*	180	0.1108	0.4317	0.7243	0.9336	1.1489	1.6891	4.7853
	b	180	-3.1622	-2.3607	-1.3385	-0.3575	0.2488	1.1189	2.9383
	c	180	0.0096	0.0561	0.116	0.1786	0.2321	0.3326	0.4021
7	*a*	123	0.1043	0.419	0.6118	0.8251	1.0184	1.4367	7.6175
	b	123	-4.0936	-1.6544	-0.4808	0.6475	1.5402	2.1451	2.9122
	c	123	0.0283	0.0652	0.1127	0.1903	0.2559	0.3522	0.4781
8	*a*	188	0.0791	0.3567	0.5242	0.7412	0.952	1.2812	2.7566
	b	188	-2.1497	-1.342	-0.028	1.1108	1.9923	3.1266	5.899
	c	188	0.0198	0.0496	0.1244	0.2029	0.2551	0.3801	0.5093

Table 18. 2PL Operational Item Parameters Five-Point Summary and Range, Mathematics

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	502	0.269	0.757	1.2352	1.5302	1.7767	2.1336	2.5996
	b	502	-5.6062	-3.2778	-2.7142	-2.3051	-1.8649	-1.2307	1.2483
4	*a*	552	0.354	0.6919	0.9954	1.2182	1.4684	1.7693	2.294
	b	552	-3.4152	-2.766	-2.0479	-1.5388	-1.0413	-0.3217	0.84
5	*a*	424	0.2	0.5489	0.8008	1.0281	1.2529	1.5592	2.0584
	b	424	-4.1089	-2.4179	-1.4596	-0.9269	-0.3906	0.4472	3.2196

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
6	*a*	471	0.0996	0.5183	0.7549	0.9548	1.1481	1.4396	1.9178
	b	471	-4.0426	-2.1598	-0.8864	-0.1161	0.553	1.5087	6.9734
7	*a*	355	0.1625	0.4358	0.6585	0.884	1.1094	1.4324	2.4711
	b	355	-1.7461	-0.9333	-0.0062	0.7583	1.5927	2.635	3.8524
8	*a*	356	0.1023	0.3662	0.5911	0.75	0.8906	1.1645	1.7156
	b	356	-5.5051	-0.1559	1.1327	1.9455	2.5844	3.8639	6.691

Table 19. GPCM Operational Item Parameters Five-Point Summary and Range, Mathematics

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	*a*	13	0.7322	0.7725	0.9615	1.1407	1.4309	1.6419	1.6621
	b1	13	-3.4077	-3.0349	-2.366	-2.04	-1.7875	-1.0562	-0.6762
	b2	13	-2.8526	-2.7842	-2.6719	-1.8579	-1.3667	-0.4869	-0.1856
4	*a*	26	0.4633	0.5317	0.657	0.9029	1.0509	1.2149	1.3409
	b1	26	-4.0185	-3.1121	-2.0584	-1.8572	-1.3267	-0.0076	0.4006
	b2	26	-3.2142	-3.1195	-2.1539	-1.6929	-1.0246	-0.3368	0.5371
	b3	26	-2.0179	-2.0179	-2.0179	-2.0179	-2.0179	-2.0179	-2.0179
5	*a*	23	0.4303	0.483	0.5333	0.7227	0.7944	1.1259	1.1949
	b1	23	-2.7811	-2.2461	-1.877	-1.1477	-0.4483	0.166	0.4743
	b2	23	-2.9968	-2.6656	-1.4904	-0.3708	-0.0499	0.436	0.8479
6	*a*	24	0.4922	0.5127	0.6931	0.7879	0.8421	0.9145	1.0632
	b1	24	-2.0784	-1.7929	-0.9813	-0.5586	0.2216	2.0603	2.3836
	b2	24	-2.0082	-1.2594	-0.4526	0.0659	0.6158	2.1943	3.3842
7	*a*	20	0.4375	0.4924	0.5432	0.6453	0.7173	1.1219	1.2202

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b1	20	-1.1742	-0.7311	0.1071	0.5283	1.1574	1.786	3.6687
	b2	20	-0.3315	-0.1628	0.4511	1.0649	1.5255	2.5005	2.8556
8	*a*	22	0.2187	0.297	0.4099	0.5946	0.6704	0.7741	0.7868
	b1	22	-1.4992	-1.3619	-0.6011	0.7469	2.2607	2.8974	4.7471
	b2	22	-3.1536	-0.7341	1.5114	2.2263	2.7758	4.1569	6.9449
	b3	22	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842

3.2 SCIENCE ITEM BANK AND TEST DESIGN

CAI works with a group of states and one US territory to develop science assessments to assess the Next Generation Science Standards (NGSS) and other standards influenced by the same science framework. Many of these states have signed a Memorandum of Understanding (MOU) to share item specifications and items. CAI has coordinated this group of states and holds contracts to develop and deliver the items for most of them.

CAI also built the ICCR science item bank in partnership with these states and one US territory. These CAI-owned items make up a substantial part of the item bank and are shared with partner states and one U.S. territory. West Virginia has signed the MOU, and therefore, the item pool available for WVGSA includes items from three sources:

1. Items owned by West Virginia
2. Items shared by other states and in the MOU collaboration
3. Items shared from the ICCR item bank

A detailed description of the Shared Science Assessment Item Bank development process is included in Volume 2, Test Development. All these items follow the same specifications, test development processes, and review processes. In 2018, CAI field tested 394 item clusters and stand-alone items in elementary and middle school, of which 338 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 244 item clusters and stand-alone items in elementary and middle school were field tested, of which 185 were accepted and made available for operational use in future years. In 2021, 373 item clusters and stand-alone items in elementary and middle school were field tested, of which 317 were accepted and made available for operational use in future years. In 2022, 360 item clusters and stand-alone items in elementary and middle school were field tested, of which 313 were accepted and made available for operational use in the future years.

The Shared Science Assessment Item Bank was used for operational accountability tests in 13 states and one U.S. territory in 2022, including West Virginia.

CAI's process for developing and field testing science items is detailed in Volume 2, Test Development. Here, note that best practices have been implemented at every turn.

- The goals, uses, and claims that the test would be designed to support were identified in a collaborative meeting on August 22–23, 2016, as an attempt to facilitate the transition from NGSS content standards to statewide summative assessments for science. CAI invited content and assessment leaders from 10 states (most of them MOU participants), as well as four nationally recognized experts who helped co-author the NGSS standards. Two nationally recognized psychometricians also participated.

- CAI staff and participating states and collaborated to develop items and item specifications, which are documents designed to guide the work of item writers as they craft test questions and the reviews of those items by stakeholders. The item specifications were generally accompanied by sample items meeting those specifications. All specifications and sample items were reviewed by state content experts and committees of educators in at least one state.

- Items were reviewed by science experts in at least one state.
- Every item was reviewed by a content advisory committee (composed of state educators) in at least one state, or in a cross-state educator review process.
- Every item was reviewed by a committee of educators charged with evaluating language accessibility and bias and sensitivity in at least one state or a cross-state educator review.
- Every item was field tested, and items with questionable data were re-reviewed by committees of educators.
- All scoring protocols (i.e., rubrics) were validated.
- In 2017, cognitive lab studies were carried out to evaluate and refine the process of developing item clusters aligned to three-dimensional science standards. Results of the cognitive lab studies confirmed the feasibility of the approach (see Volume 4, Section 6.2, Cognitive Laboratory Studies for Science, of this technical report).
- A second set of cognitive lab studies was carried out in 2018 and 2019 to determine if students using braille can understand the task demands of selected accommodated three-dimensional science aligned item clusters and navigate the interactive features of these item clusters in a manner that allows them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille and/or Job Access With Speech (JAWS) and those who had some vision and were able to read the screen with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time (see Volume 4, Section 6.2, Cognitive Laboratory Studies for Science, of this technical report).

3.2.1 Field Testing

All items that are part of the operational pool were field tested in 2018, 2019, 2021, and 2022 as described in Section 3.2.1.1, 2018 Field Test, Section 3.2.1.2, 2019 Field Test, Section 3.2.1.3, 2021 Field Test, and Section 3.2.1.4, 2022 Field Test.

3.2.1.1 2018 Field Test

In 2018, a large pool of items was field tested in nine states. For three states (Hawaii, Oregon, and Wyoming), unscored field-test items were added as an additional segment to the operational (scored) legacy science test. Two other states conducted an independent field test in which all students participated and were administered a full set of items, but no scores were reported (Connecticut and Rhode Island). In the remaining four states (New Hampshire, Utah, Vermont, and West Virginia), an operational field test was administered, meaning tests consisted of field-test items, but items became operational and were scored after the test administration if they were not rejected during rubric validation or item data review. In total, 257 item clusters and 137 stand-alone items were administered in the elementary and middle school grade bands. Table 20 presents the number of item clusters and stand-alone items administered in each grade for each state.

Table 20. Number of Field-Test Items Administered in Spring 2018

Grade Band and Item Type	CT	HI	MSSA (RI, VT)	NH	OR	UT	WV	WY	Whole Bank
Elementary School	135	24	69	58	26	-	91	14	153
Cluster	78	13	40	34	20	-	56	6	86
Stand-Alone	57	11	29	24	6	-	35	8	67
Middle School	174	27	56	55	28	98	123	17	241
Cluster	115	13	26	30	22	98	90	5	171
Stand-Alone	59	14	30	25	6	-	33	12	70
Total	**309**	**51**	**125**	**113**	**54**	**98**	**214**	**31**	**394**

For the states with a separate field-test segment (states with a legacy science test) and one of the states with an operational field test (Utah), field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent and operational field tests (except for Utah), including West Virginia, items were administered using a linear-on-the-fly (LOFT) test design. The difference between the test design for the independent field tests and operational field tests depended on the test blueprint. For the independent field tests, the only blueprint constraint imposed was that students received four stand-alone items and two cluster items for each of the three science disciplines, whereas a full blueprint was implemented for the states with an operational field test. The blueprint for the WVGSA science test is discussed in Section 3.2.2, Operational Test Design.

For any given state, a minimum sample size of 1,500 per item was targeted. Most items were administered in two or more states so that the item pools for all individual states were linked through common items. Table 21 and Table 22 present the number of cluster and stand-alone items that were in common between the item pools of any two states. The numbers below the diagonal represent the numbers for all the field-test items, and the numbers above the diagonal represent the number of common items at the time of 2018 calibration. The shaded diagonal elements represent the number of items that were administered only in the given state (in parentheses, the number of unique items at the time of calibration). Table 21 presents the results for elementary school, and Table 22 presents the results for middle school. The numbers at field testing are slightly different from the numbers at calibration for a variety of reasons, such as items being rejected during rubric validation and versioning issues for some items in some states.

Table 21. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2018, Science

	State	Connecticut	Hawaii	MSSA (RI, VT)	New Hampshire	Oregon	Utah	West Virginia	Wyoming
Cluster	**CT**	**3 (3)**	9	36	28	16	0	49	6
	HI	10	**0 (0)**	7	8	5	0	12	1
	MSSA (RI, VT)	36	8	**0 (2)**	15	12	0	26	2

	State	Connecticut	Hawaii	MSSA (RI, VT)	New Hampshire	Oregon	Utah	West Virginia	Wyoming
	NH	30	8	17	**1 (3)**	5	0	22	2
	OR	17	5	13	5	**1 (1)**	0	5	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	49	12	27	25	5	0	**0 (4)**	2
	WY	6	1	2	2	1	0	2	**0 (0)**
Stand-Alone	**CT**	**1 (3)**	5	25	22	2	0	33	7
	HI	5	**6 (6)**	0	0	0	0	4	0
	MSSA (RI, VT)	26	0	**0 (1)**	10	4	0	13	3
	NH	24	0	11	**0 (2)**	0	0	15	2
	OR	2	0	4	0	**1 (1)**	0	0	0
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	35	4	14	17	0	0	**0 (2)**	1
	WY	8	0	3	3	0	0	2	**0 (1)**
Grade-Band Total	**CT**	**4 (6)**	14	61	50	18	0	82	13
	HI	15	**6 (6)**	7	8	5	0	16	1
	MSSA (RI, VT)	62	8	**0 (3)**	25	16	0	39	5
	NH	54	8	28	**1 (5)**	5	0	37	4
	OR	19	5	17	5	**2 (2)**	0	5	1
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	84	16	41	42	5	0	**0 (6)**	3
	WY	14	1	5	5	1	0	4	**0 (1)**

Table 22. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2018, Science

	State	Connecticut	Hawaii	MSSA (RI, VT)	New Hampshire	Oregon	Utah	West Virginia	Wyoming
Cluster	**CT**	**2 (6)**	12	22	26	19	44	77	5
	HI	11	**1 (0)**	3	6	6	0	9	1
	MSSA (RI, VT)	23	3	**0 (1)**	9	1	7	22	2
	NH	26	6	10	**1 (2)**	7	0	17	3
	OR	19	6	1	7	**2 (2)**	0	5	1
	UT	48	0	7	0	0	**48 (52)**	43	0
	WV	83	10	21	18	6	48	**1 (9)**	2
	WY	5	1	2	3	1	0	2	**0 (0)**
Stand-Alone	**CT**	**2 (3)**	6	27	25	3	0	33	12
	HI	6	**8 (8)**	2	0	0	0	2	0
	MSSA (RI, VT)	27	2	**0 (0)**	18	3	0	20	2

	State	Connecticut	Hawaii	MSSA (RI, VT)	New Hampshire	Oregon	Utah	West Virginia	Wyoming
	NH	25	0	18	**0 (0)**	0	0	21	3
	OR	3	0	3	0	**0 (0)**	0	0	0
	UT	0	0	0	0	0	**0 (0)**	0	0
	WV	33	2	20	21	0	0	**0 (0)**	2
	WY	12	0	2	3	0	0	2	**0 (0)**
Grade-Band Total	CT	**4 (9)**	18	49	51	22	44	110	17
	HI	17	**9 (8)**	5	6	6	0	11	1
	MSSA (RI, VT)	50	5	**0 (1)**	27	4	7	42	4
	NH	51	6	28	**1 (2)**	7	0	38	6
	OR	22	6	4	7	**2 (2)**	0	5	1
	UT	48	0	7	0	0	**48 (52)**	43	0
	WV	116	12	41	39	6	48	**1 (9)**	4
	WY	17	1	4	6	1	0	4	**0 (0)**

The common item design was used to calibrate all the items on a common NGSS scale. The calibration model is explained in detail in Section 5, Item Calibration and Equating, of this volume.

Following the (operational) field test, items went through a substantial validation process. The process begins with rubric validation. Rubric validation is a process in which a committee of state educators reviews student responses and the proposed scoring. The responses reviewed are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents: (a) low-scored responses from otherwise high-scoring students, and (b) high-scored responses from otherwise low-scoring students.

During rubric validation, educators recommend revisions to rubrics where necessary. CAI staff revise the rubrics and rescore the entire sample to ensure that the rubric changes have all and only the intended effects.

Following rubric validation, classical item statistics were computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning statistics. The states establish standards for the statistics. Any items violating these standards are flagged for a second educator review. Even though the scoring assertions were the basic units of analysis to compute classical item statistics, the business rules to flag items for another educator review were established at the item level because assertions cannot be reviewed in isolation. A common set of business rules was defined for all the states participating in the (operational) field test, although some states decided to include additional items for data review. The item statistics were computed on the student data of the students testing in the state that owned the item. For Rhode Island and Vermont, which share their item development, the statistics were computed on the combined data. For ICCR items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (states that used ICCR items and with either an independent or operational field test) were combined. For each state, a data review committee consisting of educators (science teachers) and supported by CAI content experts reviewed the items that were owned by the state and flagged for data review according to the established business rules. For ICCR, cross-state review committees were established. Table 23 presents the number of items field tested, the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review.

Table 23. Overview of Field-Test Item Science Administration, Rubric Validation, and Item Data Review in Spring 2018

Grade Band and Owner	Number of Items Field Tested			Number of Items Rejected Before/During Rubric Validation			Number of Items Sent to Data Review			Number of Items Rejected at Data Review[b]			Number of Items Remaining		
	Clusters	Stand-Alone	Total	Cluster	Stand-Alone	Total	Cluster	Stand-Alone	Total	Cluster	Stand-Alone	Total	Cluster	Stand-Alone	Total
Elementary School	**86**	**67**	**153**	**3**	**0**	**3**	**23**	**41**	**64**	**5**	**8**	**13**	**78**	**59**	**137**
ICCR	34	31	65	0	0	0	7	19	26	1	2	3	33	29	62
Administered in WV	26	17	43	0	0	0	7	10	17	1	0	1	25	17	42
Other MOU States[a]	43	36	79	1	0	1	15	22	37	4	6	10	38	30	68
Administered in WV	21	18	39	1	0	1	5	9	14	2	4	6	18	14	32
West Virginia	9	0	9	2	0	2	1	0	1	0	0	0	7	0	7
Middle School	**171**	**70**	**241**	**12**	**4**	**16**	**67**	**36**	**103**	**15**	**9**	**24**	**144**	**57**	**201**
ICCR	31	28	59	0	0	0	11	15	26	1	2	3	30	26	56
Administered in WV	18	21	39	0	0	0	8	11	19	0	1	1	18	20	38
Other MOU States[a]	136	42	178	12	4	16	54	21	75	13	7	20	111	31	142
Administered in WV	68	12	80	6	1	7	15	3	18	1	2	3	61	9	70
West Virginia	4	0	4	0	0	0	2	0	2	1	0	1	3	0	3
Total	**257**	**137**	**394**	**15**	**4**	**19**	**90**	**77**	**167**	**20**	**17**	**37**	**222**	**116**	**338**

Note. [a]Other MOU states include Connecticut, Hawaii, MSSA (Rhode Island and Vermont), Oregon, Utah, and Wyoming. [b]Including three clusters rejected after item data review.

Table 24 summarizes the item pool that was used for West Virginia for each of three science disciplines.

Table 24. Overview of Shared Science Assessment Item Bank in Spring 2018, Science

Grade Band and Item Type	Items Field Tested in Spring 2018				Scored Operational Items			
	Total	Earth and Space Science	Life Science	Physical Science	Total	Earth and Space Science	Life Science	Physical Science
Elementary School	**91**	**25**	**32**	**34**	**81**	**22**	**29**	**30**
Cluster	56	15	20	21	50	14	17	19
Stand-Alone	35	10	12	13	31	8	12	11
Middle School	**123**	**34**	**45**	**44**	**109**	**30**	**40**	**39**
Cluster	90	26	31	33	80	23	28	29
Stand-Alone	33	8	14	11	29	7	12	10
Total	**214**	**59**	**77**	**78**	**190**	**52**	**69**	**69**

Note. [a]Excluding three Utah-owned middle school clusters that do not align to the NGSS.

3.2.1.2 2019 Field Test

In 2019, a second wave of items was field tested in nine states. For three states (Hawaii, Idaho elementary school, and Wyoming), unscored field-test items were added as a separate segment to the operational (scored) legacy science test. An independent field test in which students were administered a full set of items was conducted for a sample of Idaho middle schools. In the remaining six states (Connecticut, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia), field-test items were administered as unscored items embedded in the operational items. In total, 88 item clusters and 156 stand-alone items were administered as field-test items in the elementary and middle school grade bands. Table 25 presents the number of field-tested item clusters and stand-alone items administered in each grade for each state. The numbers in parentheses in the column representing West Virginia present the number of items owned by West Virginia.

Table 25. Number of Field-Test Items Administered in Spring 2019, Science

Grade Band and Item Type	CT	HI	ID	MSSA (RI, VT)	NH	OR	WV	WY	Whole Bank
Elementary School	**47**	**31**	**53**	**42**	**18**	**27**	**18 (6)**	**16**	**117**
Cluster	18	19	20	17	0	16	10 (3)	5	50
Stand-Alone	29	12	33	25	18	11	8 (3)	11	67
Middle School	**56**	**23**	**53**	**46**	**28**	**26**	**26 (7)**	**15**	**127**
Cluster	14	9	17	10	4	9	8 (1)	5	38
Stand-Alone	42	14	36	36	24	17	18 (6)	10	89
Total	**103**	**54**	**106**	**88**	**46**	**53**	**44 (13)**	**31**	**244**

For the three states with a separate field-test segment (states with a legacy science test), field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent field test, items were administered under a LOFT design, where the only blueprint constraint imposed was that students received four stand-alone items and two clusters for each of the three science disciplines. For the states with an operational test, field-test items were embedded within the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Other states opted for a test design in which the items were not grouped by discipline (Connecticut, Oregon, West Virginia). In these three states, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of five field-test stand-alone items. The test design for the WVGSA science test is discussed in Section 3.2.2, Operational Test Design.

A minimum sample size of 1,500 students per field-test item was targeted for any given state. Most items were administered in two or more states. Table 26 and Table 27 present the number of clusters and stand-alone items that were shared between the field-test pools of any two states. The numbers below the diagonal represent the numbers for all the field-test items, and the numbers above the diagonal represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state (in parentheses, the number of unique field-test items at the time of calibration). Table 26 presents the results for elementary schools and Table 27, the results for middle schools. The numbers at field testing are slightly different from the numbers at calibration because some items were rejected during rubric validation.

Table 26. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2019, Science

	State	Connecticut	Hawaii	Idaho	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia	Wyoming
Cluster	**CT**	**2 (2)**	2	10	3	0	2	1	4
	HI	2	**0 (0)**	3	8	0	14	2	0
	ID	10	3	**4 (4)**	0	0	1	3	3
	MSSA	3	8	0	**3 (3)**	0	9	4	1
	NH	0	0	0	0	**0 (0)**	0	0	0
	OR	2	14	1	9	0	**1 (1)**	0	0
	WV	1	2	3	4	0	0	**1 (0)**	1
	WY	4	0	3	1	0	0	1	**0 (0)**
Stand-Alone	**CT**	**5 (5)**	1	13	1	9	0	0	2
	HI	1	**0 (0)**	10	6	0	6	0	0
	ID	13	11	**1 (1)**	12	1	9	2	4
	MSSA	1	7	13	**3 (3)**	5	8	5	6
	NH	9	0	1	5	**2 (3)**	0	0	6
	OR	0	7	10	9	0	**1 (1)**	0	0
	WV	0	0	2	5	0	0	**1 (1)**	0
	WY	2	0	4	6	7	0	0	**0 (0)**
Grade Band	**CT**	**7 (7)**	3	23	4	9	2	1	6
	HI	3	**0 (0)**	13	14	0	20	2	0
	ID	23	14	**5 (5)**	12	1	10	5	7

	State	Connecticut	Hawaii	Idaho	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia	Wyoming
	MSSA	4	15	13	**6 (6)**	5	17	9	7
	NH	9	0	1	5	**2 (3)**	0	0	6
	OR	2	21	11	18	0	**2 (2)**	0	0
	WV	1	2	5	9	0	0	**2 (1)**	1
	WY	6	0	7	7	7	0	1	**0 (0)**

Table 27. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2019, Science

	State	Connecticut	Hawaii	Idaho	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia	Wyoming
Cluster	**CT**	**5 (5)**	3	4	2	0	2	1	0
	HI	3	**0 (0)**	4	4	0	5	1	0
	ID	4	4	**2 (2)**	4	0	4	3	3
	MSSA	2	4	4	**1 (1)**	0	2	3	1
	NH	0	0	1	0	**3 (0)**	0	0	0
	OR	2	5	4	2	0	**1 (1)**	1	2
	WV	1	1	3	3	0	1	**0 (0)**	2
	WY	0	0	3	1	0	2	2	**0 (0)**
Stand-Alone	**CT**	**10 (9)**	2	13	9	10	3	6	0
	HI	2	**0 (0)**	9	9	0	6	3	0
	ID	13	9	**2 (2)**	11	1	12	6	5
	MSSA	9	9	11	**1 (1)**	6	11	9	7
	NH	10	0	2	6	**3 (1)**	0	0	2
	OR	3	6	12	11	0	**0 (0)**	2	7
	WV	6	3	6	9	1	2	**0 (0)**	0
	WY	0	0	5	7	2	7	0	**0 (0)**
Grade-Band Total	**CT**	**15 (14)**	5	17	11	10	5	7	0
	HI	5	**0 (0)**	13	13	0	11	4	0
	ID	17	13	**4 (4)**	15	1	16	9	8
	MSSA	11	13	15	**2 (2)**	6	13	12	8
	NH	10	0	3	6	**6 (1)**	0	0	2
	OR	5	11	16	13	0	**1 (1)**	3	9
	WV	7	4	9	12	1	3	**0 (0)**	2
	WY	0	0	8	8	2	9	2	**0 (0)**

The calibration and linking of the items field tested in 2019 is explained in detail in Section 5.2, Item Calibration and Linking for Science, of this volume.

Following essentially the same process as explained in Section 3.2.1.1, 2018 Field Test, items went through a substantial validation process. The modifications to the process followed in 2018 were minor, including the following:

- In 2018, all the item statistics were computed on the student data of the students testing in the state that owned the item. In 2019, all the item statistics were computed on the student data of the students testing in the state that owned the item *except for the statistics related to differential item functioning* (DIF). Following recommendations of several technical advisory committees, the data of states were combined in the calculation of DIF statistics whenever possible (i.e., for states with an independent field test or an operational test for which the relevant demographic variable was available).
- In 2018, for ICCR items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (states that used ICCR items and with either an independent or operational field test) were combined. In 2019, these states were Connecticut, Idaho (only for middle school), New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia.
- The business rule to flag an item cluster for DIF was slightly modified (i.e., made more liberal) following recommendations of several technical advisory committees. The modification is discussed in Section 4.5, Differential Item Functioning Analysis.

Table 28 presents the number of items field tested in West Virginia, or another state, the number of items rejected before or during rubric validation, the number of items sent out for data review, and the number of items rejected during data review. The numbers in parentheses present the number of items owned by West Virginia.

Table 28. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2019

Grade Band and Item Type	Number of Items Field Tested	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining[a]
Elementary School	**117 (6)**	**2 (1)**	**72 (3)**	**24 (2)**	**91 (3)**
Cluster	50 (3)	1 (1)	16 (0)	10 (0)	39 (2)
Stand-Alone	67 (3)	1 (0)	56 (3)	14 (2)	52 (1)
Middle School	**127 (7)**	**6 (0)**	**66 (4)**	**21 (4)**	**97 (3)**
Cluster	38 (1)	1 (0)	12 (1)	5 (1)	29 (0)
Stand-Alone	89 (6)	5 (0)	54 (3)	16 (3)	68 (3)
Total	**244 (13)**	**8 (1)**	**138 (7)**	**45 (6)**	**188 (6)**

Note. West Virginia-owned items are indicated in the parentheses.

[a] Number of items remaining excludes three artificial intelligence (AI) scoring items field tested in spring 2019 that were not brought to item data review.

Table 29 summarizes the science item pool after adding the items that were field tested in 2019 and survived rubric validation and item data review. The numbers in parentheses present the number of items owned by West Virginia.

Table 29. Overview of Shared Science Assessment Item Bank in Spring 2019, Science

Grade Band and Item Type	Total	Combined Science Item Pool			
		Earth and Space Sciences	Engineering and Technology	Life Sciences	Physical Sciences
Elementary School	**225 (9)**	**67 (3)**	**0 (0)**	**77 (4)**	**81 (2)**
Cluster	115 (8)	34 (3)	0 (0)	40 (3)	41 (2)
Stand-Alone	110 (1)	33 (0)	0 (0)	37 (1)	40 (0)
Middle School	**287 (6)**	**81 (2)**	**1 (0)**	**109 (1)**	**96 (3)**
Cluster	165 (3)	44 (1)	1 (0)	63 (1)	57 (1)
Stand-Alone	122 (3)	37 (1)	0 (0)	46 (0)	39 (2)
Total	**512 (15)**	**148 (5)**	**1 (0)**	**186 (5)**	**177 (5)**

3.2.1.3 2021 Field Test

In 2021, a third wave of items was field tested in 12 states. For one state (Wyoming), unscored field-test items were added as a separate segment to the operational scored legacy science test. An independent field test, in which students were administered a full set of items, was conducted in Idaho and Montana. In the remaining nine states (Connecticut, Hawaii, New Hampshire, North Dakota, Rhode Island, South Dakota, Vermont, Utah, and West Virginia), field-test items were administered as unscored items embedded among the operational items. In total, 166 item clusters and 207 stand-alone items were administered as field-test items in the elementary and middle school grade bands. Table 30 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing West Virginia presents the number of field-test items owned by West Virginia.

Table 30. Number of Field-Test Items Administered in Spring 2021

Grade Band and Item Type	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY	Total Field-Test Items Administered
Elementary School	**36**	**22**	**140**	**55**	**21**	**11**	**19**	**8**	**54**	**19 (5)**	**17**	**214**
Cluster	16	6	58	18	7	3	3	3	54	7 (1)	5	106
Stand-Alone	20	16	82	37	14	8	16	5	0	12 (4)	12	108
Middle School	**33**	**19**	**129**	**54**	**20**	**11**	**18**	**11**	**45**	**19 (7)**	**20**	**159**
Cluster	17	6	44	18	7	3	2	2	45	7 (3)	4	60
Stand-Alone	16	13	85	36	13	8	16	9	0	12 (4)	16	99
Total	**69**	**41**	**269**	**109**	**41**	**22**	**37**	**19**	**99**	**38 (12)**	**37**	**373**

Note. West Virginia-owned items are indicated in the parentheses.

[a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment

For Wyoming, the state with a separate field-test segment, field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent field test, items were administered under a LOFT design, where the only blueprint constraint imposed was that students receive four stand-alone items and two item clusters for each of the three science disciplines.

For the states with an operational test, field-test items were embedded in the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, and Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Six other states (Connecticut, Hawaii, North Dakota, South Dakota, Utah, and West Virginia) opted for a test design in which the items were not grouped by discipline. In these states, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of four field-test stand-alone items. The test design for the WVGSA is discussed in Section 3.2.2, Operational Test Design.

For any given state, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two or more states. Table 31 and Table 32 present the number of item clusters and stand-alone items that were shared between the field-test pools of any two states. The numbers below the shaded diagonal elements represent the numbers for all administered field-test items, and the numbers above the shaded diagonal elements represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state (with the number of unique field-test items at the time of calibration in parentheses). Table 31 presents the results for elementary schools, and Table 32 presents the results for middle schools. The numbers of field-test items administered are slightly different from the numbers of field-test items at calibration because some items were rejected during rubric validation.

Table 31. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Cluster	CT	3 (3)	0	13	0	0	0	0	0	0	0	0
	HI	0	1 (1)	3	0	0	0	0	0	0	1	0
	ID	13	4	3 (2)	5	5	2	0	2	20	1	4
	MSSA	0	0	6	2 (2)	2	0	0	0	7	0	0
	MT	0	0	5	2	0 (0)	0	0	0	0	0	0
	ND	0	0	2	0	0	0 (0)	0	1	0	1	0
	NH	0	0	0	0	0	0	0 (0)	0	0	3	0
	SD	0	0	2	0	0	1	0	0 (0)	0	2	0
	UT	0	0	20	8	0	0	0	0	25 (24)	0	2
	WV	0	1	1	0	0	1	3	2	0	1 (1)	0
	WY	0	0	4	0	0	0	0	0	2	0	0 (0)
Stand-Alone	CT	3 (3)	0	14	2	0	0	0	0	0	0	1
	HI	0	0 (0)	12	1	0	0	2	3	0	1	0
	ID	14	12	3 (3)	30	13	4	3	3	0	4	9
	MSSA	2	1	30	0 (0)	12	0	3	1	0	0	0
	MT	0	0	13	12	0 (0)	0	0	0	0	0	0
	ND	0	0	4	0	0	0 (0)	2	0	0	0	1
	NH	0	2	4	3	0	2	0 (0)	2	0	3	1
	SD	0	3	3	1	0	0	2	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	1	4	0	0	1	3	0	0	3 (3)	0
	WY	1	0	9	0	0	1	1	0	0	0	0 (0)
Grade Band	CT	6 (6)	0	27	2	0	0	0	0	0	0	1
	HI	0	1 (1)	15	1	0	0	2	3	0	2	0

State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
ID	27	16	6 (5)	35	18	6	3	5	20	5	13
MSSA	2	1	36	2 (2)	14	0	3	1	7	0	0
MT	0	0	18	14	0 (0)	0	0	0	0	0	0
ND	0	0	6	0	0	0 (0)	2	1	0	1	1
NH	0	2	4	3	0	2	0 (0)	2	0	6	1
SD	0	3	5	1	0	1	2	0 (0)	0	2	0
UT	0	0	20	8	0	0	0	0	25 (24)	0	2
WV	0	2	5	0	0	2	6	2	0	4 (4)	0
WY	1	0	13	0	0	1	1	0	2	0	0 (0)

Note. [a]MSSA = Rhode Island and Vermont’s Multi-State Science Assessment.

Table 32. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2021

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
Cluster	**CT**	0 (0)	0	9	2	0	0	0	0	10	0	0
	HI	0	0 (0)	2	3	0	0	0	0	3	1	0
	ID	11	2	1 (1)	10	6	2	1	1	31	0	4
	MSSA	4	3	11	0 (0)	0	2	0	0	9	1	1
	MT	0	0	6	0	1 (1)	0	1	1	4	0	0
	ND	0	0	3	2	0	0 (0)	0	0	2	0	0
	NH	0	0	1	0	1	0	0 (0)	1	0	1	0
	SD	0	0	1	0	1	0	1	0 (0)	0	0	0
	UT	14	3	36	11	4	3	0	1	0 (0)	2	2
	WV	0	1	1	1	0	0	1	1	5	0 (0)	0
	WY	0	0	4	1	0	0	0	0	2	0	0 (0)
Stand-Alone	**CT**	2 (2)	0	12	2	0	0	0	3	0	0	2
	HI	0	0 (0)	10	1	0	0	0	0	0	2	0
	ID	13	10	2 (2)	29	10	6	12	7	0	5	15
	MSSA	2	1	29	0 (0)	10	2	1	1	0	2	4
	MT	0	0	12	10	0 (0)	0	0	0	0	0	0
	ND	0	0	7	2	0	0 (0)	1	0	0	0	0
	NH	0	0	12	1	0	1	0 (0)	2	0	1	3
	SD	3	0	7	1	0	0	2	0 (0)	0	3	4
	UT	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	0	2	6	3	0	1	1	3	0	0 (0)	0
	WY	2	0	15	4	0	0	3	4	0	0	0 (0)
Grade Band	**CT**	2 (2)	0	21	4	0	0	0	3	10	0	2
	HI	0	0 (0)	12	4	0	0	0	0	3	3	0

State	CT	HI	ID	MSSA[a]	MT	ND	NH	SD	UT	WV	WY
ID	24	12	3 (3)	39	16	8	13	8	31	5	19
MSSA	6	4	40	0 (0)	10	4	1	1	9	3	5
MT	0	0	18	10	1 (1)	0	1	1	4	0	0
ND	0	0	10	4	0	0 (0)	1	0	2	0	0
NH	0	0	13	1	1	1	0 (0)	3	0	2	3
SD	3	0	8	1	1	0	3	0 (0)	0	3	4
UT	14	3	36	11	4	3	0	1	0 (0)	2	2
WV	0	3	7	4	0	1	2	4	5	0 (0)	0
WY	2	0	19	5	0	0	3	4	2	0	0 (0)

Note. [a]MSSA = Rhode Island and Vermont’s Multi-State Science Assessment.

The calibration and linking of the field-test items in 2021 are explained in detail in Section 5.2.2, Item Calibration.

Table 33 presents the number of field-test items administered in West Virginia, or another state, the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of field-test items owned by West Virginia.

Table 33. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2021

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
Elementary School	**214 (5)**	**7 (0)**	**100 (3)**	**19 (1)**	**188 (4)**
Cluster	106 (1)	5 (0)	24 (0)	7 (0)	94 (1)
Stand-Alone	108 (4)	2 (0)	76 (3)	12 (1)	94 (3)
Middle School	**159 (7)**	**15 (2)**	**87 (4)**	**13 (2)**	**129 (3)**
Cluster	60 (3)	10 (1)	22 (1)	5 (1)	43 (1)
Stand-Alone	99 (4)	5 (1)	65 (3)	8 (1)	86 (2)
Total	**373 (12)**	**22 (2)**	**187 (7)**	**32 (3)**	**317 (7)**

Note. West Virginia-owned items are indicated in the parentheses.

Table 34 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2021 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by West Virginia.

Table 34. Overview of Shared Science Assessment Item Bank in Spring 2021

Grade Band and Item Type	**Science Discipline**			**Total**
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**136 (4)**	**128 (7)**	**149 (3)**	**413 (14)**
Cluster	65 (3)	66 (4)	76 (2)	207 (9)
Stand-Alone	71 (1)	62 (3)	73 (1)	206 (5)
Middle School	**114 (2)**	**156 (2)**	**137 (6)**	**407 (10)**
Cluster	55 (1)	76 (2)	67 (1)	198 (4)
Stand-Alone	59 (1)	80 (0)	70 (5)	209 (6)
Total	**250 (6)**	**284 (9)**	**286 (9)**	**820 (24)**

Note. West Virginia-owned items are indicated in the parentheses.

3.2.1.4 2022 Field Test

In 2022, a fourth wave of items was field tested in 13 states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Vermont, Utah, West Virginia, Wyoming, and U.S. Virgin Islands,). Field-test items were administered as unscored items embedded among the operational items. In total, 176 item clusters and 184 stand-alone items were administered as field-test items in the elementary and middle school grade bands. Table 35 presents the number of field-test item clusters and stand-alone items administered in each grade band for each state. The numbers in parentheses in the column representing West Virginia presents the number of field-test items owned by West Virginia.

Table 35. Number of Field-Test Items Administered in Spring 2022

Grade Band and Item Type	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY	Total Field-Test Items Administered
Elementary School	**34**	**28**	**22**	**66**	**12**	**12**	**17**	**41**	**10**	**1**	**62**	**19 (6)**	**10**	**170**
Cluster	22	8	11	22	4	4	5	15	4	1	62	11 (1)	2	88
Stand-Alone	12	20	11	44	8	8	12	26	6	0	-	8 (5)	8	82
Middle School	**40**	**30**	**35**	**64**	**12**	**12**	**17**	**39**	**10**	**1**	**76**	**33 (8)**	**10**	**190**
Cluster	20	10	7	21	4	4	5	16	4	1	76	5 (4)	2	88
Stand-Alone	20	20	28	43	8	8	12	23	6	0	-	28 (4)	8	102
Total	**74**	**58**	**57**	**130**	**24**	**24**	**34**	**80**	**20**	**2**	**138**	**52 (14)**	**20**	**360**

Note. West Virginia-owned items are indicated in the parentheses.

[a] MSSA = Rhode Island and Vermont’s Multi-State Science Assessment.

For the states with an operational test, field-test items were embedded in the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, and Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Ten other states and one U.S. territory (Connecticut, Hawaii, Idaho, Montana, North Dakota, Oregon, South Dakota, Utah, West Virginia, Wyoming, and U.S. Virgin Islands,) opted for a test design in which the items were not grouped by discipline. In these ten states and one US territory, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of four field-test stand-alone items. The test design for the WVGSA is discussed in Section 3.2.2, Operational Test Design.

For any given state or territory, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two states (or territory). Table 36 and Table 37 present the number of item clusters and stand-alone items that were shared between the field-test pools of any two states (or territory). The numbers below the shaded diagonal elements represent the numbers for all administered field-test items, and the numbers above the shaded diagonal elements represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state or territory (with the number of unique field-test items at the time of calibration in parentheses). Table 36 presents the results for elementary schools, and Table 37 presents the results for middle schools. The numbers of field-test items administered are slightly different from the numbers of field-test items at calibration because some items were rejected during rubric validation.

Table 36. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY
Item Clusters	**CT**	0 (0)	0	3	1	0	0	0	3	0	0	15	0	0
	HI	0	0(0)	0	0	0	0	0	5	0	0	2	0	0
	ID	3	0	0(0)	3	0	0	0	0	0	0	5	0	0
	MSSA	1	0	3	0(0)	0	0	0	5	1	0	12	0	0
	MT	0	0	0	0	0(0)	0	0	0	0	0	4	0	0
	ND	0	0	0	0	0	0(0)	4	0	0	0	0	0	0
	NH	0	0	0	0	0	4	0(0)	0	0	1	1	0	0
	OR	3	6	0	5	0	0	0	0(0)	0	0	1	0	0
	SD	0	0	0	1	0	0	0	0	0 (0)	0	3	0	0
	USVI	0	0	0	0	0	0	1	0	0	0(0)	1	0	0
	UT	15	2	5	12	4	0	1	1	3	1	6 (6)	11	2
	WV	0	0	0	0	0	0	0	0	0	0	11	0(0)	0
	WY	0	0	0	0	0	0	0	0	0	0	2	0	0 (0)
Stand-Alone Items	**CT**	0 (0)	2	0	4	4	0	0	0	2	0	0	0	0
	HI	2	0(0)	3	7	0	0	0	8	0	0	0	0	0
	ID	0	3	0(0)	1	1	4	0	2	0	0	0	0	0
	MSSA	4	7	1	0(0)	3	0	1	7	4	0	0	8	8
	MT	4	0	1	3	0(0)	0	0	0	0	0	0	0	0
	ND	0	0	4	0	0	0(0)	3	1	0	0	0	0	0
	NH	0	0	0	1	0	3	1 (0)	7	0	0	0	0	0
	OR	0	8	2	8	0	1	7	0(0)	0	0	0	0	0
	SD	2	0	0	4	0	0	0	0	0 (0)	0	0	0	0
	USVI	0	0	0	0	0	0	0	0	0	0(0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0(0)	0	0
	WV	0	0	0	8	0	0	0	0	0	0	0	0(0)	0

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY
	WY	0	0	0	8	0	0	0	0	0	0	0	0	0 (0)
Grade Band Total	**CT**	0(0)	2	3	5	4	0	0	3	2	0	15	0	0
	HI	2	0(0)	3	7	0	0	0	13	0	0	2	0	0
	ID	3	3	0(0)	4	1	4	0	2	0	0	5	0	0
	MSSA	5	7	4	0(0)	3	0	1	12	5	0	12	8	8
	MT	4	0	1	3	0(0)	0	0	0	0	0	4	0	0
	ND	0	0	4	0	0	0(0)	7	1	0	0	0	0	0
	NH	0	0	0	1	0	7	1 (0)	7	0	1	1	0	0
	OR	3	14	2	13	0	1	7	0(0)	0	0	1	0	0
	SD	2	0	0	5	0	0	0	0	0 (0)	0	3	0	0
	USVI	0	0	0	0	0	0	1	0	0	0(0)	1	0	0
	UT	15	2	5	12	4	0	1	1	3	1	6 (6)	11	2
	WV	0	0	0	8	0	0	0	0	0	0	11	0(0)	0
	WY	0	0	0	8	0	0	0	0	0	0	2	0	0 (0)

Note. [a]MSSA = Rhode Island and Vermont’s Multi-State Science Assessment.

Table 37. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2022

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY
Item Clusters	CT	0 (0)	1	1	0	0	0	0	1	0	0	17	0	0
	HI	1	0 (0)	0	1	0	0	0	1	0	0	5	0	0
	ID	1	0	0 (0)	0	0	0	0	0	1	0	5	0	0
	MSSA	0	1	0	0 (0)	0	0	0	2	0	0	18	0	0
	MT	0	0	0	0	0 (0)	0	0	0	0	0	4	0	0
	ND	0	0	0	0	0	0 (0)	0	0	0	0	4	0	0
	NH	0	0	0	0	0	0	0 (0)	3	0	0	2	0	0
	OR	1	2	0	2	0	0	3	0 (0)	0	0	8	0	0
	SD	0	0	1	0	0	0	0	0	0 (0)	1	2	0	0
	USVI	0	0	0	0	0	0	0	0	1	0 (0)	1	0	0
	UT	17	6	5	18	4	4	2	8	3	1	2 (2)	5	2
	WV	0	0	0	0	0	0	0	0	0	0	5	0 (0)	0
	WY	0	0	0	0	0	0	0	0	0	0	2	0	0 (0)
Stand-Alone Items	CT	0 (0)	0	0	12	0	0	0	4	1	0	0	3	0
	HI	0	0 (0)	8	5	0	0	0	6	0	0	0	1	0
	ID	0	8	0 (0)	5	8	0	0	3	4	0	0	0	0
	MSSA	12	5	5	0 (0)	0	0	0	4	0	0	0	9	8
	MT	0	0	8	0	0 (0)	0	0	0	0	0	0	0	0
	ND	0	0	0	0	0	0 (0)	0	0	0	0	0	8	0
	NH	0	0	0	0	0	0	0 (0)	6	0	0	0	5	0
	OR	4	6	3	4	0	0	6	0 (0)	0	0	0	0	0
	SD	1	0	4	0	0	0	0	0	0 (0)	0	0	1	0
	USVI	0	0	0	0	0	0	0	0	0	0 (0)	0	0	0
	UT	0	0	0	0	0	0	0	0	0	0	0 (0)	0	0
	WV	3	1	0	9	0	8	6	0	1	0	0	0 (0)	0

	State	CT	HI	ID	MSSA[a]	MT	ND	NH	OR	SD	USVI	UT	WV	WY
	WY	0	0	0	8	0	0	0	0	0	0	0	0	0 (0)
Grade Band Total	**CT**	0 (0)	1	1	12	0	0	0	5	1	0	17	3	0
	HI	1	0 (0)	8	6	0	0	0	7	0	0	5	1	0
	ID	1	8	0 (0)	5	8	0	0	3	5	0	5	0	0
	MSSA	12	6	5	0 (0)	0	0	0	6	0	0	18	9	8
	MT	0	0	8	0	0 (0)	0	0	0	0	0	4	0	0
	ND	0	0	0	0	0	0 (0)	0	0	0	0	4	8	0
	NH	0	0	0	0	0	0	0 (0)	9	0	0	2	5	0
	OR	5	8	3	6	0	0	9	0 (0)	0	0	8	0	0
	SD	1	0	5	0	0	0	0	0	0 (0)	1	2	1	0
	USVI	0	0	0	0	0	0	0	0	1	0 (0)	1	0	0
	UT	17	6	5	18	4	4	2	8	3	1	2 (2)	5	2
	WV	3	1	0	9	0	8	6	0	1	0	5	0 (0)	0
	WY	0	0	0	8	0	0	0	0	0	0	2	0	0 (0)

Note. [a]MSSA = Rhode Island and Vermont’s Multi-State Science Assessment.

The calibration and linking of the field-test items in 2022 are explained in detail in Section 5.2.2, Item Calibration.

Table 38 presents the number of field-test items administered in West Virginia, or another state (or territory), the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of field-test items owned by West Virginia.

Table 38. Overview of Field-Test Item Administration, Rubric Validation, and Item Data Review in Spring 2022

Grade Band and Item Type	Number of Field-Test Items Administered	Number of Items Rejected Before/During Rubric Validation	Number of Items Sent to Data Review	Number of Items Rejected at Data Review	Number of Items Remaining
Elementary School	**170 (6)**	**3 (0)**	**82 (5)**	**14 (2)**	**153 (4)**
Cluster	88 (1)	1 (0)	18 (1)	4 (0)	83 (1)
Stand-Alone	82 (5)	2 (0)	64 (4)	10 (2)	70 (3)
Middle School	**190 (8)**	**4 (0)**	**94 (5)**	**26 (2)**	**160 (6)**
Cluster	88 (4)	3 (0)	26 (3)	13 (2)	72 (2)
Stand-Alone	102 (4)	1 (0)	68 (2)	13 (0)	88 (4)
Total	**360 (14)**	**7 (0)**	**176 (10)**	**40 (4)**	**313 (10)**

Note. West Virginia-owned items are indicated in parentheses.

Table 39 summarizes the Shared Science Assessment Item Bank after adding the field-test items that were administered in 2022 and passed rubric validation and item data review. The numbers in parentheses present the number of items owned by West Virginia.

Table 39. Overview of Shared Science Assessment Item Bank in Spring 2022

Grade Band and Item Type	Science Discipline			Total[a]
	Earth and Space Sciences	*Life Sciences*	*Physical Sciences*	
Elementary School	**180 (5)**	**162 (7)**	**214 (6)**	**556 (18)**
Cluster	96 (4)	82 (4)	111 (2)	289 (10)
Stand-Alone	84 (1)	80 (3)	103 (4)	267 (8)
Middle School	**150 (3)**	**220 (6)**	**187 (7)**	**557 (16)**
Cluster	70 (2)	110 (3)	90 (1)	270 (6)
Stand-Alone	80 (1)	110 (3)	97 (6)	287 (10)
Total	**330 (8)**	**382 (13)**	**401 (13)**	**1113 (34)**

Note. West Virginia-owned items are indicated in parentheses.

[a] Count excludes eight MOU items that do not align to the NGSS.

3.2.2 Operational Test Design

The science assessments were assembled under an adaptive test design, using CAI's adaptive testing algorithm. The adaptive item selection algorithm selects items based on their content value and information value. At any given point during the test, an item's content value is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered. During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Vice versa, the content value decreases for items with content features for which the minimum has been met. The information value of an item is based on the item information function evaluated at the estimated proficiency. The proficiency estimate is updated throughout the test. Under an adaptive test design, operational items are selected on the fly based on the performance of a student on past items while ensuring the test blueprint is followed for each individual student. The science blueprint is given in Table 40 and Table 41. Details of CAI's item selection algorithm are described in Volume 2: Test Development, Appendix I, Adaptive Algorithm Design.

Table 40. Test Blueprint, Grade 5 Science

Grade 5	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Min Stand-Alone Items	Max Clusters + Max Stand-Alone Items
Discipline – Physical Science, PE Total = 17	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
3-PS2-1: Forces, balanced and unbalanced forces	0	1	0	1	0	1
3-PS2-2: Forces, pattern predicts future motion	0	1	0	1	0	1
3-PS2-3: Forces, between objects not in contact	0	1	0	1	0	1
3-PS2-4: Forces, magnets*	0	1	0	1	0	1
5-PS2-1: Space systems	0	1	0	1	0	1
DCI – Energy	**0**	**1**	**0**	**2**	**0**	**3**
4-PS3-1: Energy, relationship between speed and energy of object	0	1	0	1	0	1
4-PS3-2: Energy, transfer of energy	0	1	0	1	0	1
4-PS3-3: Energy, changes in energy when objects collide	0	1	0	1	0	1
4-PS3-4: Energy, converting energy from one form to another*	0	1	0	1	0	1
5-PS3-1: Matter & Energy	0	1	0	1	0	1
DCI – Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
4-PS4-1: Waves, waves can cause objects to move	0	1	0	1	0	1
4-PS4-2: Structure, function, information processing	0	1	0	1	0	1
4-PS4-3: Waves, using patterns to transfer information*	0	1	0	1	0	1
DCI – Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
5-PS1-1: Structure & Properties of Matter	0	1	0	1	0	1
5-PS1-2: Structure & Properties of Matter	0	1	0	1	0	1
5-PS1-3: Structure & Properties of Matter	0	1	0	1	0	1
5-PS1-4: Structure & Properties of Matter	0	1	0	1	0	1

Grade 5	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Min Stand-Alone Items	Max Clusters + Max Stand-Alone Items
Discipline – Life Science, PE Total = 12	**2**	**2**	**4**	**4**	**6**	**6**
DCI – From Molecules to Organisms: Structure and Function	**0**	**1**	**0**	**2**	**0**	**3**
3-LS1-1: Inheritance	0	1	0	1	0	1
4-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
4-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
5-LS1-1: Matter & Energy	0	1	0	1	0	1
DCI – Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
3-LS2-1: Ecosystems	0	1	0	1	0	1
5-LS2-1: Matter & Energy	0	1	0	1	0	1
DCI – Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
3-LS3-1: Inheritance	0	1	0	1	0	1
3-LS3-2: Inheritance	0	1	0	1	0	1
DCI – Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
3-LS4-1: Ecosystems	0	1	0	1	0	1
3-LS4-2: Inheritance	0	1	0	1	0	1
3-LS4-3: Ecosystems	0	1	0	1	0	1
3-LS4-4: Ecosystems*	0	1	0	1	0	1
Discipline – Earth and Space Science, PE Total = 13	**2**	**2**	**4**	**4**	**6**	**6**
DCI – Earth's Systems	**0**	**1**	**0**	**3**[a]	**0**	**3**
3-ESS2-1: Weather & Climate	0	1	0	1	0	1
3-ESS2-2: Weather & Climate	0	1	0	1	0	1
4-ESS2-1: Earth's Systems & Processes	0	1	0	1	0	1
4-ESS2-2: Earth's Systems & Processes	0	1	0	1	0	1
5-ESS2-1: Earth's Systems	0	1	0	1	0	1

Grade 5	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Min Stand-Alone Items	Max Clusters + Max Stand-Alone Items
5-ESS2-2: Earth's Systems	0	1	0	1	0	1
DCI – Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS3-1: Weather & Climate*	0	1	0	1	0	1
4-ESS3-2: Earth's Systems & Processes*	0	1	0	1	0	1
4-ESS3-1: Energy	0	1	0	1	0	1
5-ESS3-1: Earth's Systems	0	1	0	1	0	1
DCI – Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
4-ESS1-1: Earth's Systems & Processes	0	1	0	1	0	1
5-ESS1-1: Space Systems	0	1	0	1	0	1
5-ESS1-2: Space Systems	**0**	**1**	**0**	**1**	**0**	**1**
PE Total = 42	**6**	**6**	**12**	**12**	**18**	**18**

Note. Constraints on sampling across grades per discipline (except grade 3 LS): at most one cluster per grade, at most three SAs per grade, at most four clusters plus SAs per grade; for grade 3 LS, at most two clusters per grade, at most three SAs per grade, at most four clusters plus SAs per grade.

[a]Because of the limitation of the item pool in the ESS discipline, the maximum number of stand-alone items allowed was changed from two to three while keeping the maximum number of items (clusters plus SAs) allowed at three in ESS2.

*These PEs have an engineering component.

Table 41. Test Blueprint, Grade 8 Science

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
Discipline - Physical Science, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI - Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
8-MS-PS1-1: Structure & Properties of Matter	0	1	0	1	0	1
8-MS-PS1-2: Chemical Reactions	0	1	0	1	0	1
8-MS-PS1-3: Structure & Properties of Matter	0	1	0	1	0	1
8-MS-PS1-4: Structure & Properties of Matter	0	1	0	1	0	1
8-MS-PS1-5: Chemical Reactions	0	1	0	1	0	1
8-MS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
DCI - Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
7-MS-PS2-1: Forces & Interactions*	0	1	0	1	0	1
7-MS-PS2-2: Forces & Interactions	0	1	0	1	0	1
7-MS-PS2-3: Forces & Interactions	0	1	0	1	0	1
7-MS-PS2-4: Forces & Interactions	0	1	0	1	0	1
7-MS-PS2-5: Forces & Interactions	0	1	0	1	0	1
DCI - Energy	**0**	**1**	**0**	**2**	**0**	**3**
7-MS-PS3-1: Energy	0	1	0	1	0	1
7-MS-PS3-2: Energy	0	1	0	1	0	1
7-MS-PS3-3: Energy*	0	1	0	1	0	1
7-MS-PS3-4: Energy	0	1	0	1	0	1
7-MS-PS3-5: Energy	0	1	0	1	0	1
DCI - Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
6-MS-PS4-1: Waves & Electromagnetic Radiation	0	1	0	1	0	1
6-MS-PS4-2: Waves & Electromagnetic Radiation	0	1	0	1	0	1

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
6-MS-PS4-3: Waves & Electromagnetic Radiation	0	1	0	1	0	1
Discipline - Life Science, PE Total = 21	**2**	**2**	**4**	**4**	**6**	**6**
DCI - From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
7-MS-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
7-MS-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
7-MS-LS1-3: Structure, Function, Information Processing	0	1	0	1	0	1
8-MS-LS1-4: Growth, Development, Reproduction	0	1	0	1	0	1
8-MS-LS1-5: Growth, Development, Reproduction	0	1	0	1	0	1
6-MS-LS1-6: Matter & Energy	0	1	0	1	0	1
6-MS-LS1-7: Matter & Energy	0	1	0	1	0	1
7-MS-LS1-8: Structure, Function, Information Processing	0	1	0	1	0	1
DCI - Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
6-MS-LS2-1: Matter & Energy	0	1	0	1	0	1
6-MS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
6-MS-LS2-3: Matter & Energy	0	1	0	1	0	1
6-MS-LS2-4: Matter & Energy	0	1	0	1	0	1
6-MS-LS2-5: Interdependent Relationships in Ecosystems*	0	1	0	1	0	1
DCI - Hereditary: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
8-MS-LS3-1: Growth, Development, Reproduction	0	1	0	1	0	1
8-MS-LS3-2: Growth, Development, Reproduction	0	1	0	1	0	1
DCI - Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
8-MS-LS4-1: Natural Selection & Adaptation	0	1	0	1	0	1
8-MS-LS4-2: Natural Selection & Adaptation	0	1	0	1	0	1
8-MS-LS4-3: Natural Selection & Adaptation	0	1	0	1	0	1

Grade 8	Min Clusters	Max Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Clusters + Stand-Alone Items	Max Clusters + Stand-Alone Items
8-MS-LS4-4: Natural Selection & Adaptation	0	1	0	1	0	1
8-MS-LS4-5: Growth, Development, Reproduction	0	1	0	1	0	1
8-MS-LS4-6: Natural Selection & Adaptation	0	1	0	1	0	1
Discipline - Earth and Space Science, PE Total = 15	**2**	**2**	**4**	**4**	**6**	**6**
DCI - Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
6-MS-ESS1-1: Space Systems	0	1	0	1	0	1
6-MS-ESS1-2: Space Systems	0	1	0	1	0	1
6-MS-ESS1-3: Space Systems	0	1	0	1	0	1
7-MS-ESS1-4: History of Earth	0	1	0	1	0	1
DCI - Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
7-MS-ESS2-1: Earth's Systems	0	1	0	1	0	1
7-MS-ESS2-2: History of Earth	0	1	0	1	0	1
7-MS-ESS2-3: History of Earth	0	1	0	1	0	1
7-MS-ESS2-4: Earth's Systems	0	1	0	1	0	1
6-MS-ESS2-5: Weather & Climate	0	1	0	1	0	1
6-MS-ESS2-6: Weather & Climate	0	1	0	1	0	1
DCI - Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
7-MS-ESS3-1: Earth's Systems	0	1	0	1	0	1
6-MS-ESS3-2: Human Impacts	0	1	0	1	0	1
7-MS-ESS3-3: Human Impacts*	0	1	0	1	0	1
8-MS-ESS3-4: Human Impacts	0	1	0	1	0	1
6-MS-ESS3-5: Weather & Climate	0	1	0	1	0	1
Total PE= 55	**6**	**6**	**12**	**12**	**18**	**18**

Note. Constraints on sampling across grades per discipline: at most one cluster per grade, at most three SAs per grade, at most four clusters plus SAs per grade.

*These PEs have an engineering component.

The main characteristics of the blueprint were that any performance expectation could be tested only once (indicated by the values of 0 and 1 for the Min and Max values of the individual Performance Expectations [PEs] in Table 40 and Table 41); no more than one cluster or two stand-alone items could be sampled from the same domain core idea, and no more than three total items could be sampled from the same domain core idea (as indicated by the Min and Max values in the rows representing domain core ideas). Furthermore, one cluster and three stand-alone items, at most, could be sampled from a given grade in the grade band for each discipline.

In 2018, a segmented test design was used; items were administered grouped by science discipline. In 2019, a non-segmented test design was used; items from different disciplines were no longer grouped by science discipline. Instead, students received items from different disciplines in random order. The change of design was partially motivated by a possible move to a fully adaptive test in future years. In an adaptive test, the use of a non-segmented test design gives more freedom when selecting items targeting a current best estimate of proficiency. Embedded field-test items were randomly positioned in the test and randomly distributed across students. Every student received either one cluster or five stand-alone items as field-test items throughout the test. In 2021 and 2022, a non-segmented test design was used. Students received items from different disciplines in random order. Embedded field-test items were randomly positioned in the test and randomly distributed across students. Every student received either one item cluster or four stand-alone items as field-test items throughout the test.

4. Field Test Classical Analysis

Following test administration, all field-test items are evaluated for discrimination, difficulty, and differential item functioning (DIF). In addition, distractor analysis is conducted on multiple-choice (MC) items in English language arts (ELA) and mathematics, and response time analysis is performed for science items. Any items flagged for out-of-range statistics are reviewed by the Cambium Assessment, Inc. (CAI) content and psychometric staff; poorly performing items are then rejected from the item bank. The criteria for flagging and reviewing ELA and mathematics items is provided in Table 42.

Table 42. Thresholds for Flagging in Classical Item Analysis, ELA and Mathematics

Analysis Type	Flagging Criteria
Item Discrimination	Point biserial correlation for the correct response is < 0.20.
Distractor Analysis	Point biserial correlation for any distractor response is > 0.
Item Difficulty (MC items)	The proportion of students (*p*-value) is < 0.15 or > 0.90.
Item Difficulty (non-MC items)	Relative mean is < 0.10 or > 0.95.
Differential Item Functioning	Item DIF categorization of “C” in either direction.

As explained in Section 3.2, Science Item Bank and Test Design, of this volume, science items administered as field-test items underwent rubric validation and data review. Items were flagged for data review based on business rules defined on classical item statistics. Except for response times, the classical item statistics are computed for individual assertions, whereas the business rules for flagging are defined at the item level. In general, item statistics used to flag items for data review were computed using the student responses of the state that owned the item. However, for Independent College and Career Readiness (ICCR) item bank items, the flagging rules were defined on the item statistics computed from the combined data of states or territory that used ICCR items and that administered either an independent or operational field test (in 2022, those states were Connecticut, Idaho, New Hampshire, North Dakota, Oregon, Rhode Island, South Dakota, Utah, Vermont, and West Virginia). Furthermore, for the computation of DIF statistics, the data of all eligible states with an operational or independent field test were combined to obtain a sufficient number of students for each demographic group. The criteria for flagging and reviewing items is provided in Table 43, and a description of the statistics is provided below. Items that were flagged for data review were reviewed by a committee, as explained in Section 3.2.

Table 43. Thresholds for Flagging in Classical Item Analysis, Science

Analysis Type	Flagging Criteria
Item Discrimination	Average biserial correlation < 0.25 (across the assertions within an item)
	One or more assertions with a biserial correlation < 0.05
Item Difficulty (Clusters)	Average *p*-value < .30 or > 0.85 (across the assertions within a cluster)
Item Difficulty (Stand-Alones)	Average *p*-value < .15 or > 0.95 (across the assertions within a stand-alone)
Timing (Clusters)	Percentile 80* > 15 minutes
Timing (Stand-Alones)	Percentile 80* > 3 minutes
Timing	Assertions per minute < 0.5
DIF (Clusters)	Two or more assertions show ‘C’ DIF in the same direction
DIF (Stand-Alones)	One or more assertions show ‘C’ DIF in the same direction

*Note. *Percentile 80 of x minutes: 80% of the students spent x minutes or less on the item.*

4.1 ITEM DISCRIMINATION

The item discrimination index indicates the extent to which each item differentiates between those test takers who possessed the skills being measured and those who did not. Generally, the higher the value, the better the item is able to differentiate between high- and low-achieving students.

In science, for each assertion within an item, the discrimination index was calculated as the biserial correlation between the assertion score and the ability estimate for students. The average biserial correlation was then calculated across the assertions within an item.

4.2 ITEM DIFFICULTY

Items that are either very difficult or very easy are flagged for review but are not necessarily removed if they are grade-level appropriate and aligned with the test specifications. For multiple-choice items, the proportion of students in the sample selecting the correct answer (the *p*-value) is computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty is calculated using the item’s relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item’s mean score divided by the maximum possible score points). For science, both the *p*-value for individual assertions and the average across all assertions of an item are calculated. Acceptable item *p*-values are summarized in Table 42 for ELA and mathematics and Table 43 for science.

4.3 ELA AND MATHEMATICS DISTRACTOR ANALYSIS

Distractor analysis for multiple-choice items is used to identify items that may have marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracts high-scoring students. For multiple-choice items, the correct response should be the most frequently selected option by high-scoring students. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative.

4.4 SCIENCE RESPONSE TIME

Given that the science clusters consist of multiple student interactions, they require more time for students to complete. To ensure a good balance between the amount of information an item provides and the time students spend on the item, item response time was recorded and analyzed. Specifically, the statistic "percentile 80" was computed for each item. A percentile 80 of *x* minutes means that 80% of the students spent *x* minutes or fewer on the item. An item was flagged for review when the

- percentile 80 > 15 minutes, if the item is a cluster;
- percentile 80 > 3 minutes, if the item is a stand-alone; or
- assertions per (percentile 80) minute < 0.5.

4.5 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important, because it provides a statistical indicator that an item may contain cultural or other bias. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) provides a guideline for when sample sizes permitting subgroup differences in performance should be examined and appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors.

CAI uses a generalized Mantel-Haenszel (MH) procedure to calculate DIF. The generalizations include (a) adaptation to polytomous items and (b) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's estimated theta score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the generalized MH chi-square ($GMH\chi^2$) DIF statistic for balancing the stability and sensitivity of the DIF scoring category selection. The standardized mean difference (SMD [Dorans & Schmitt, 1991]) was also computed.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as:

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})},$$

where $k = \{1, 2, \dots K\}$ for the strata, $n_{R1k}$is the number of students with correct responses for the reference group in stratum k, and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k} n_{R+k}}{n_{++k}},$$

where n_{+1k} is the number of students with correct responses, n_{R+k} is the number of students in the reference group, n_{++k} is the number of students in stratum k, and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k} n_{F+k} n_{+1k} n_{+0k}}{n_{++k}^2 (n_{++k} - 1)},$$

n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses in stratum k.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k} n_{F0k} / n_{++k}}{\sum_k n_{R0k} n_{F1k} / n_{++k}}.$$

The MH-delta (Δ_{MH}, Holland & Thayer, 1988) is then defined as

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

The GMH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = (\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k))' (\sum_k var(\boldsymbol{a}_k))^{-1} (\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)),$$

where $\boldsymbol{a}_k$ is a $(T-1) \times 1$ vector of item response scores and $E(\boldsymbol{a}_k)$ is a $(T-1) \times 1$ mean vector, both corresponding to the T response categories of a polytomous item (excluding one response); $var(\boldsymbol{a}_k)$ is a $(T-1) \times (T-1)$ covariance matrix calculated analogously to the corresponding elements in $MH\chi^2$ in stratum k.

The standardized mean difference (SMD, Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Fk} m_{Rk},$$

where

$$p_{Fk} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum k,

$$m_{Fk} = \frac{1}{n_{F+k}} \left(\sum_t a_t n_{Ftk} \right)$$

is the mean item score for the focal group in stratum k, and

$$m_{Rk} = \frac{1}{n_{R+k}}\left(\sum_{t} a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum k.

DIF was evaluated for the embedded field-test items for spring 2022 ELA and mathematics. Appendix H, DIF Statistics for Spring 2022 Field-Test Items, presents the DIF analysis results using the generalized Mantel-Haenszel (MH) procedure. The generalized MH classified items into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF (refer to Table 46 for classification rules). Furthermore, items were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American, Hispanic, female), or negatively (i.e., –A, –B, or–C), signifying that an item favored the reference group (e.g., White, male). Items were flagged if their DIF statistics indicated the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF analysis was conducted for all field-test items with at least 200 responses per item in each subgroup (Zwick, 2012) to detect potential item bias for major demographic groups. Due to the limited number of students in some groups, DIF analyses were performed for the following groups in ELA and mathematics:

- Male/Female
- White/African American
- White/Hispanic
- White/Asian or Pacific Islander
- White/American Indian or Alaskan Native
- White/Multi-Racial
- Special Education (SPED) vs. Non-SPED

Table 44 and Table 45 illustrate the minimum to maximum number of field test item responses for each group.

Table 44. Range of Field-Test Item Responses by DIF Group, ELA

Grade	Group	Non-SPED/SPED	Male/Female	White/Asian	White/American Indian	White/African American	White/Hispanic	White/Multiracial
3	Reference	579-1056	895-1217	1390-1918	1390-1918	1390-1918	1390-1918	1390-1918
3	Focal	102-231	837-1155	13-37	52-96	57-98	49-144	55-126
4	Reference	573-1099	330-1457	498-2274	498-2274	498-2274	498-2274	498-2274
4	Focal	84-205	340-1266	3-29	48-83	25-100	31-148	30-135
5	Reference	596-1239	893-2065	1462-3410	1462-3410	1462-3410	1462-3410	1462-3410
5	Focal	95-206	894-1975	15-37	55-92	65-138	49-179	84-190
6	Reference	528-1080	620-1511	882-2391	882-2391	882-2391	882-2391	882-2391
6	Focal	73-192	568-1404	13-27	50-89	28-102	55-168	46-123
7	Reference	516-1131	294-1791	440-3009	440-3009	440-3009	440-3009	440-3009
7	Focal	62-182	299-1824	9-27	48-85	26-146	27-174	20-148
8	Reference	492-1152	312-1764	444-2879	444-2879	444-2879	444-2879	444-2879
8	Focal	64-170	280-1682	6-35	39-84	26-105	24-168	25-127

Table 45. Range of Field-Test Item Responses by DIF Group, Mathematics

Grade	Group	Non-SPED/SPED	Males/Females	White/Asian	White/American Indian	White/African American	White/Hispanic	White/Multiracial
3	Reference	945-2020	1250-2127	1832-3320	1832-3320	1832-3320	1832-3320	1832-3320
3	Focal	159-428	1127-2062	20-58	80-167	67-162	86-262	113-229
4	Reference	607-3116	373-3127	514-4905	514-4905	514-4905	514-4905	514-4905
4	Focal	105-555	348-2925	13-70	84-248	41-226	34-364	35-304
5	Reference	891-1036	1062-1232	1691-1939	1691-1939	1691-1939	1691-1939	1691-1939
5	Focal	138-196	990-1160	14-33	52-86	52-94	95-148	75-127
6	Reference	1182-4293	1551-3286	2320-5131	2320-5131	2320-5131	2320-5131	2320-5131
6	Focal	171-663	1535-3248	29-74	117-308	82-218	116-488	135-317
7	Reference	671-785	1035-1239	1748-1978	1748-1978	1748-1978	1748-1978	1748-1978
7	Focal	83-138	1014-1177	12-35	59-97	76-129	53-95	74-118
8	Reference	582-687	950-1076	1545-1747	1545-1747	1545-1747	1545-1747	1545-1747
8	Focal	64-107	891-1033	11-33	43-90	58-100	48-87	64-103

Table 46. DIF Classification Rules, ELA and Mathematics

Dichotomous Items	
Category	*Rule*
C	MH_{X^2} is significant and $\lvert\hat{\Delta}_{MH}\rvert$ ≥1.5.
B	MH_{X^2} is significant and 1 ≤ $\lvert\hat{\Delta}_{MH}\rvert$<1.5.
A	MH_{X^2} is not significant or $\lvert\hat{\Delta}_{MH}\rvert$<1.
Polytomous Items	
Category	*Rule*
C	MH_{X^2} is significant and $\lvert SMD\rvert / \lvert SD\rvert \ > .25$.
B	MH_{X^2} is significant and $.17 < \ \lvert SMD\rvert / \lvert SD\rvert \ \leq .25$.
A	MH_{X^2} is not significant or $\lvert SMD\rvert / \lvert SD\rvert \ \leq \ .17$.

In science, a similar DIF categorization rule was applied at the assertion level. Items were flagged for review according to additional item-level criteria sets based on the results of the assertion level categorizations. The item-level criteria also considered the item type (i.e., cluster or stand-alone). All science DIF statistics were computed after the testing windows closed. Student responses from multiple states were combined to minimize the number of items with insufficient sample sizes for one or more demographic groups. DIF statistics were calculated at the assertion level and DIF analyses were performed for the following groups in science (some items had insufficient sample sizes for DIF analyses in some groups):

- Male vs. Female
- American Indian/Alaskan Natives vs. White
- Hawaiian/Pacific Islander vs. White
- Asian vs. White
- African American vs. White
- Hispanic vs. White
- Multi-Racial vs. White
- English Learner (EL) vs. Non-EL
- Special Education (SPED) vs. Non-SPED
- Economically Disadvantaged vs. Non-Economically Disadvantaged

Similar to how the general MH statistic is used to classify items of traditional tests, assertions were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. The classification rules shown in Table 47 were applied to the assertions in each item (cluster or stand-alone). Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American, Hispanic, female), or negatively (i.e., -A, -B, or –C), signifying that an item favored the reference group (e.g., White, male).

An item is flagged for data review according to the following criterion:

- **Item Clusters.** Two or more assertions showed ‘C’ DIF in the same direction.
- **Stand-Alone Items.** One or more assertions showed ‘C’ DIF in the same direction.

Table 47. DIF Classification Rules, Science

Assertions	
Category	***Rule***
C	MH_{X^2} is significant and $\lvert SMD\rvert/\lvert SD\rvert \geq 0.25$.
B	MH_{X^2} is significant and $\frac{\lvert SMD\rvert}{\lvert SD\rvert} < 0.25$.
A	MH_{X^2} is not significant.

Note that, for the 2018 field test, a slightly less strict criterion was used for item clusters with 10 or more assertions (i.e., three or more assertions with “C” DIF in the same direction). The change was made taking into consideration the feedback received from several Technical Advisory Committees (TACs) and modified such that the rate of flagging items for DIF was similar for item clusters and stand-alone items (based on the flagging rates computed on items field tested in 2018).

Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal or reference group. DIF-flagged items are further examined by content experts who are asked to re-examine each flagged item to decide if the item should be excluded from the item pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in certain areas are less likely to offer rigorous mathematics classes, students at those schools might perform more poorly on mathematics items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but rather the instruction. However, DIF can indicate bias, so all items are evaluated for DIF. The spring 2022 field-test items that were flagged with a C rating were reviewed by CAI content team. In their evaluation, they were unable to determine any reason that these items might be functioning differently between the groups and made the determination that the items should be retained.

In addition to the DIF analyses on the field-tested items, a special study was conducted on the operational items in the ICCR ELA and mathematics item bank in 2020 to examine them for the presence of DIF for accommodated versus non-accommodated students. The results of these analyses are presented in Volume 7 of this technical report.

4.6 CLASSICAL ANALYSIS RESULTS

This section presents a summary of results from classical item analysis of the 2022 field-test items in ELA, mathematics, and science. ELA and mathematics include both West Virginia-owned field-test items and shared ICCR items. Science includes West Virginia-owned field-test items, shared Memorandum of Understanding (MOU) items, and shared ICCR items.

Table 48 through Table 53 provide the summary of the *p*-values and biserial correlations for the field-tested items in West Virginia for ELA, mathematics, and science, respectively. The statistics were computed using West Virginia data only, as well as data across the ICCR states (New Hampshire, North Dakota, West Virginia, and Wyoming) for shared ELA and mathematics field-tested items, and MOU states and one U.S. territory for science. For science, the average values across the assertions within an item were used to compute percentiles and ranges.

*Table 48. Distribution of p-Values for Field-Test Items, ELA**

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	118	0.16	0.25	0.36	0.44	0.54	0.72	0.77
4	118	0.10	0.18	0.36	0.50	0.61	0.79	0.84
5	115	0.09	0.22	0.41	0.54	0.65	0.79	0.91
6	117	0.15	0.26	0.43	0.53	0.63	0.76	0.84
7	123	0.08	0.19	0.35	0.49	0.61	0.76	0.83
8	121	0.11	0.22	0.38	0.51	0.64	0.76	0.87

**Results presented excluded flagged items.*

*Table 49. Distribution of Item Point-Biserial Correlations for Field-Test Items, ELA**

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	118	0.07	0.13	0.27	0.34	0.43	0.51	0.77
4	118	-0.12	0.09	0.26	0.37	0.45	0.51	0.84
5	115	0.01	0.14	0.33	0.42	0.46	0.53	0.91
6	117	0.03	0.16	0.30	0.36	0.44	0.50	0.84
7	123	-0.01	0.16	0.29	0.36	0.44	0.51	0.83
8	121	-0.02	0.16	0.30	0.37	0.45	0.52	0.87

**Results presented excluded flagged items.*

*Table 50. Distribution of p-Values for Field-Test Items, Mathematics**

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	84	0.03	0.11	0.27	0.43	0.61	0.82	0.90
4	61	0.04	0.12	0.28	0.37	0.52	0.78	0.86
5	131	0.01	0.11	0.22	0.38	0.54	0.68	0.80
6	40	0.01	0.04	0.11	0.38	0.50	0.69	0.77
7	133	0.02	0.03	0.12	0.28	0.45	0.73	0.88
8	149	0.00	0.05	0.16	0.31	0.48	0.71	0.83

**Results presented excluded flagged items.*

Table 51. Distribution of Item Point-Biserial Correlations for Field-Test Items, Mathematics*

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	84	0.11	0.23	0.32	0.41	0.49	0.55	0.60
4	61	-0.05	0.15	0.39	0.45	0.50	0.58	0.63
5	131	0.05	0.17	0.31	0.43	0.50	0.55	0.59
6	40	0.00	0.12	0.21	0.34	0.41	0.57	0.58
7	133	0.00	0.09	0.25	0.35	0.47	0.56	0.63
8	149	-0.05	0.11	0.28	0.39	0.46	0.54	0.62

**Results presented excluded rejected flagged items.*

Table 52. Distribution of p-Values for Field-Test Items, Science

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	19	0.20	0.21	0.30	0.38	0.51	0.69	0.71
8	32	0.00	0.09	0.22	0.31	0.41	0.63	0.78

Table 53. Distribution of Item Biserial Correlations for Field-Test Items, Science

Grade	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	19	0.09	0.19	0.32	0.38	0.49	0.62	0.72
8	32	0.13	0.16	0.28	0.34	0.48	0.67	0.69

Table 54 presents the summary of the percentile 80 response times by item type (item cluster or stand-alone item) for science field-test items administered in 2022.

Table 54. Summary of Percentile 80 Response Times for Field-Test Items Administered in Spring 2022

Grade	Item Type	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	Cluster	11	5.20	6.30	7.85	8.30	9.90	10.85	11.20
	Stand-Alone	8	1.50	1.61	1.88	2.75	3.75	4.36	4.60

Grade	Item Type	Total FT Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
8	Cluster	5	4.70	4.72	4.80	4.90	5.60	7.44	7.90
	Stand-Alone	27	1.30	1.40	1.85	2.10	2.70	3.27	4.90

Table 55 presents, for each item type, the number of field-test items flagged for DIF for each demographic group included in the DIF analyses in 2022.

Table 55. Differential Item Functioning Classifications for Field-Test Items Administered in Spring 2022

DIF Flag	Item Type	Female/ Male	American Indian[a]/ White	Asian/ White	African American / White	Hawaiian[b] / White	Hispanic/ White	Multi-Racial/ White	EL/ Non-EL	SPED/ Non-SPED	Low Income[c]/ Non-Low Income
						Grade 5					
Items Evaluated	Cluster	11	0	0	0	0	11	0	2	7	11
	Stand-Alone	8	0	0	0	0	8	0	0	0	8
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	0	0	0	0	0	0	0	0	0	0
% Items Flagged C	Cluster	0	-	-	-	-	0	-	0	0	0
	Stand-Alone	0	-	-	-	-	0	-	-	-	0
						Grade 8					
Items Evaluated	Cluster	5	0	0	0	0	5	0	0	1	5
	Stand-Alone	27	1	0	3	0	12	0	0	5	21
Items Flagged C	Cluster	0	0	0	0	0	0	0	0	0	0
	Stand-Alone	0	0	0	0	0	0	0	0	0	0
% Items Flagged C	Cluster	0	-	-	-	-	0	-	-	0	0
	Stand-Alone	0	0	-	0	-	0	-	-	0	0

Note. Full DIF Group names: [a]American Indian/Alaskan Native; [b]Hawaiian/Pacific Islander; [c]Economically Disadvantaged vs. Non-Economically Disadvantaged

Among the 51 science items that were field tested in West Virginia and passed the rubric validation in 2022, 12 items were flagged for item discrimination, 7 items were flagged for p-value, 17 items were flagged for response time, and no item were flagged for DIF according to the criteria outlined in the earlier sections. Some items were flagged due to multiple reasons. Flagged items were reviewed by educators during the process of data review. The total number of field-test items flagged and the total number of field-test items that passed item data review in 2022 were summarized in Table 38.

5. Item Calibration and Equating

5.1 ELA and Mathematics Item Calibration and Equating

Item response theory (IRT; van der Linden & Hambleton, 1997) is used to calibrate all items and derive scores for all Independent College and Career Readiness (ICCR) item bank items. IRT is a general framework that models test responses resulting from an interaction between students and test items.

IRT encompasses many related measurement models that allow for varied assumptions about the nature of the data. Simple unidimensional models are the most common models used in K–12 operational testing programs, and items are often calibrated using a sample of students from within a state population. ICCR items are administered across samples of students in different states. This grouping structure leads to a natural extension of the basic IRT models to data collected from multiple populations, hence the multiple group IRT (MGIRT) model (Bock & Zimowski, 1997) is used to calibrate all ICCR items.

5.1.1 Item Response Theory Methods

All individuals in the calibration sample are considered to have the observed responses z_{ijk_j}, corresponding to test taker j in group k to the ith item. The MGIRT assumes local (conditional) independence of item responses and further assumes that the jth individual is a member of the k_jth population with density function $f(\theta; \mu_{k_j}, \sigma^2_{k_j})$.

The generalized approach to item calibration begins with familiar probability models, including the three-parameter logistic model (3PL; Lord & Novick, 1968) for binary items and the generalized partial credit model (GPCM; Muraki, 1992) for items scored in multiple categories.

The probability model for binary items is denoted as

$$\mathrm{P_{ij}}(z_{ijk_j} = 1|\theta_{jk_j}) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i\left(\theta_{jk_j} - b_i\right)\right]},$$

where $\mathrm{p_{ij}}\left(z_{ijk_j} = 1|\theta_{jk_j}\right)$ is the probability of test taker j answering item i correct, c_i is the lower asymptote of the item response curve (the pseudo-guessing parameter), b_i is the location parameter, a_i is the slope parameter (the discrimination parameter), and D is a constant fixed at 1.7, bringing the logistic into coincidence with the probit model. Student ability is represented by θ_{jk_j}.

The GPCM is typically expressed as the probability for individual j of scoring in the $(z_{ijk_j} + 1)^{\text{th}}$ category to the i^{th} item as

$$\mathrm{P_{ij}}\left(z_{ijk_j}\middle|\theta_{\mathrm{jk_j}}\right) = \frac{\exp\sum_{l=1}^{z_{ijk_j}} Da_i\left(\theta_{jk_j} - b_{il}\right)}{1 + \sum_{h=1}^{m_i} \exp\sum_{l=1}^{h} Da_i\left(\theta_{jk_j} - b_{il}\right)},$$

where b_{ki} is the kth step value, $z_{ijk_j} = \{0,1,..,m_i\}$, and m_i is the maximum possible score of the item.

The conditional independence assumption then provides for the likelihood of the individual response pattern to be expressed as

$$\Pr\left(\mathbf{z}_{jk_j}\middle|\theta_{jk_j},\boldsymbol{\gamma}\right) = \prod_{i=1}^{I} Pr\left(z_{ijk_j}|\theta_{jk_j},\boldsymbol{\gamma}\right)$$

where $\boldsymbol{\gamma}$ is a vector of item parameters, leading to the marginal likelihood of the responses within group k as

$$L_j(\boldsymbol{\gamma}) = \int \prod_{i=1}^{I} Pr\left(z_{ij\boldsymbol{\gamma}}|\theta_{jk_j},\boldsymbol{\gamma}\right) f\left(\theta_{jk_j}|\mu_{k_j},\sigma_{k_j}^2\right) d\theta_{jk_j}.$$

Then, assuming independence between different groups, the overall likelihood to be maximized with respect to the item parameters is

$$\arg\max L(\boldsymbol{\gamma}) = \prod_{j=1}^{N} L_{jk_j}(\boldsymbol{\gamma}).$$

All item parameter estimates were obtained with IRTPRO version 4.1 (Cai, Thissen, & du Toit, 2011). IRTPRO uses marginal maximum likelihood estimation (MLE). The identification of the model requires fixing the population parameters for one group to $N(0,1)$, and then the means of all other groups are freely estimated relative to the reference group. Each group's means and standard deviations are reported in Appendix C, Calibration Group Means and Standard Deviations for ICCR Bank Items of this volume.

5.1.2 Equating to the Scale for ELA and Mathematics

Equating to the established reporting scale is conducted using the Stocking-Lord procedure (Stocking & Lord, 1983). The methods are implemented by calibrating the item response data using the same MGIRT model described in this report and then using the methods described in this section to equate them to the ICCR item bank. Without loss of generality, the subscript notation is simplified here as the grouping structure for the MGIRT is not used to establish linkages between tests.

First, the probability of response for the class of binary IRT models is defined on the *bank* scale, which is the scale we are linking items to, and the subscripts I and J denote the item parameters for the bank and items to be rescaled, respectively:

$$p\left(z_{i,I} = 1\middle|\theta\right) = c_{i,I} + \frac{1 - c_{i,I}}{1 + \exp\left[-Da_{i,I}\left(\theta - b_{i,I}\right)\right]}$$

and for the polytomous IRT models

$$p\left(z_{i,I}\middle|\theta\right) = \frac{\exp\left(\sum_{l=1}^{z_i} Da_i\left(\theta - b_{il,I}\right)\right)}{1 + \sum_{h=1}^{m_i} \exp \sum_{l=1}^{h} Da_{i,I}\left(\theta - b_{il,I}\right)}$$

where z_i denotes score point $\mathrm{z}_i = \{1, \dots, \mathrm{m}_i\}$ to item i. The expected score for the polytomous models is

$$E\left(z_{i,I}|\theta\right) = \sum_{z_{i,I}=1}^{m_i} z_{i,I}\,p\left(z_{i,I}|\theta\right).$$

The form of the IRT models for the new items that are to be linked onto the bank scale, or the *rescaled* items, have a similar form, but the transformation coefficients A and B are introduced as

$$p\left(z_{i,I}^{*} = 1\middle|\theta\right) = c_{i,J} + \frac{1 - c_{i,J}}{1 + \exp\left[-D\,\frac{a_{i,J}}{A}\left(\theta - \left(b_{i,J} * A + B\right)\right)\right]}$$

and

$$p\left(z_{i,I}^{*}\middle|\theta\right) = \frac{\exp\left(\sum_{k=0}^{z_i} D\,\frac{a_{i,J}}{A}\left(\theta - \left(b_{ki,J} * A + B\right)\right)\right)}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} D\,\frac{a_{i,J}}{A}\left(\theta - \left(b_{ki,J} * A + B\right)\right)}.$$

The “*” is used when transformation coefficients appear in the IRT model. The notation $p\left(z_{i,J}\middle|\theta\right)$ denotes the same IRT model, but without the transformation of coefficients A and B.

The symmetric approach uses the reverse transform for the bank items,

$$p\left(z_{i,I}^{*} = 1\middle|\theta\right) = c_{i,I} + \frac{1 - c_{i,I}}{1 + \exp\left[-DAa_{i,I}\left(\theta - \frac{\left(b_{i,I} - B\right)}{A}\right)\right]},$$

and for the polytomous IRT models,

$$p\left(z_{i,I}^{*}\middle|\theta\right) = \frac{\exp\left(\sum_{k=0}^{z_i} DAa_i\left(\theta - \frac{\left(b_{ki,I} - B\right)}{A}\right)\right)}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} DAa_{i,I}\left(\theta - \frac{\left(b_{ki,I} - B\right)}{A}\right)}.$$

And then the objective function to be minimized with respect to the transformation coefficients, A and B, is

$$\arg\min SL = \int \left[\sum_{i=1}^{K} E\left(z_{i,I}|\theta\right) - \sum_{i=1}^{K} E\left(z_{i,J}^{*}|\theta\right) \right]^2 f(\theta|\mu_1, \sigma_1^2)\, d\theta$$
$$+ \int \left[\sum_{i=1}^{K} E\left(z_{i,I}^{*}|\theta\right) - \sum_{i=1}^{K} E\left(z_{i,J}|\theta\right) \right]^2 f(\theta|\mu_2, \sigma_2^2)\, d\theta$$

where $f(\theta|\mu_1, \sigma_1^2)$ is the normal population density associated with putting operational items onto the bank scale, and $f(\theta|\mu_2, \sigma_2^2)$ is the density associated with putting bank items onto the operational scale. Implementation is performed using Gauss-Hermite quadrature and the integral is replaced with summation over q quadrature points, so

$$\arg\min SL = \sum_{q_1=1}^{Q_1} \left[\sum_{i=1}^{K} E\left(z_{i,I}\middle|\theta_{q_1}\right) - \sum_{i=1}^{K} E\left(z_{i,J}^{*}\middle|\theta_{q_1}\right) \right]^2 w_{q_1}$$
$$+ \sum_{q_2=1}^{Q_2} \left[\sum_{i=1}^{K} E\left(z_{i,I}^{*}\middle|\theta_{q_2}\right) - \sum_{i=1}^{K} E\left(z_{i,I}\middle|\theta_{q_2,}\right) \right]^2 w_{q_2}$$

where θ_{q_1} is node q_1 associated with $f(\theta|\mu_1, \sigma_1^2)$ w_{q_1}, is the weight at node q_1, θ_{q_2} is node q_2 associated with $f(\theta|\mu_2, \sigma_2^2)$, and w_{q_2} is the weight at node q_2.

5.1.3 Establishing the Initial ICCR Bank

Establishing the initial set of item parameters and equating the items over the years in which they were used are described in this section. The ICCR initial item bank spanned three different years (i.e., 2015–2017) of field testing with multiple states. Every grade and subject was calibrated separately within a given year using MGIRT. For example, grade 5 mathematics items in 2015 were calibrated and then, separately, the grade 5 mathematics items in 2016 were calibrated. These year-over-year separate item calibrations were then equated using the Stocking-Lord method to place all ICCR items from the separate calibrations onto a single scale.

This equating chain was established using a common item non-equivalent groups design in which a set of common items was administered in the pools each year. All common items in the pool were used unless the item's *A* parameter was less than .1 or greater than 3, and the absolute *B* parameter larger than 6, were not included. Table 56 displays year-to-year equating constants.

Table 56. Linking Across Years Results

Subject	Grade	2015 to 2016			2016 to 2017		
		Number of Anchors	Slope	Intercept	Number of Anchors	Slope	Intercept
ELA	3	113	0.9413	0.0085	138	0.9749	0.1082
	4	128	0.8711	0.0091	185	0.9531	0.1451
	5	125	1.0497	-0.0374	172	1.0340	0.0708
	6	173	1.0635	0.0953	184	0.9756	0.0750

Subject	Grade	2015 to 2016			2016 to 2017		
		Number of Anchors	Slope	Intercept	Number of Anchors	Slope	Intercept
	7	163	1.1462	-0.0069	178	1.0259	0.1838
	8	135	0.9785	-0.1097	155	1.0279	-0.1285
Mathematics	3	101	0.9765	0.0563	255	0.9444	0.0570
	4	96	1.0017	0.0011	229	1.0287	0.0394
	5	218	1.0586	0.0284	271	1.0392	0.0682
	6	194	1.0266	0.0949	228	1.0530	0.0961
	7	178	1.0682	-0.0574	259	1.0901	-0.0606
	8	194	1.1290	-0.1380	269	1.0763	-0.0296

5.1.4 Linking the Initial ICCR Bank to SAGE Bank

These methods are used to calibrate and equate the initial ICCR bank. Once that bank was established, these items were linked to the Utah Student Assessment of Growth and Excellence (SAGE) item bank, which provides a vertical reporting scale. Linking the ICCR and SAGE bank also used the Stocking-Lord method (Stocking & Lord, 1983) using the same common-item non-equivalent groups design. Table 57 shows linking constants for each grade and subject between the initial ICCR bank and SAGE. These linking constants were used to put the initial ICCR bank onto the SAGE on-grade-level scale. Appendix D, Vertical Scaling in SAGE, documents the design and results of the vertical linking study that was implemented to develop the SAGE ELA and mathematics item bank.

Table 57. Linking to SAGE Results

Subject	Grade	Number of Anchors	Slope	Intercept
ELA	3	177	1.0026	0.0729
	4	227	1.0267	-0.0131
	5	182	0.9873	0.0860
	6	244	1.0085	0.0228
	7	159	1.0189	-0.0243
	8	160	0.9983	0.1773
Mathematics	3	295	1.1081	0.1386
	4	276	1.0609	0.0979
	5	247	1.0406	0.1034
	6	211	1.0056	0.0525
	7	217	1.0125	0.1035
	8	252	0.9671	0.2525

Table 58 and Table 59 display the number of students in each participating state contributing to the ICCR multi-group IRT model.

Table 58. Number of Students Used in ICCR MGIRT Calibration, ELA

Grade	Year	Utah	Florida	Arizona	Oregon (2015)/Ohio (2016)
3	2015	39,279	-	33,687	9,323
	2016	46,901	-	62,242	85,972
	2017	47,317	-	72,754	-
4	2015	39,753	-	33,091	11,858
	2016	43,190	207,867	61,065	95,211
	2017	45,537	206,341	73,195	-
5	2015	38,976	35,780	32,398	8,398
	2016	36,196	199,326	60,210	97,451
	2017	43,825	209,984	72,289	-
6	2015	38,340	42,565	33,114	8,234
	2016	38,106	196,409	57,635	101,799
	2017	39,662	200,039	69,837	-
7	2015	36,082	56,752	30,911	10,688
	2016	45,469	193,186	58,050	105,249
	2017	45,484	197,752	69,754	-
8	2015	36,445	82,159	32,277	13,590
	2016	42,530	195,125	57,349	104,360
	2017	42,018	197,269	69,481	-

Table 59. Number of Students Used in ICCR MGIRT Calibration, Mathematics

Grade	Year	Utah	Florida	Arizona	Oregon (2015)/Ohio (2016)
3	2015	48,473	-	43,543	27,642
	2016	49,762	-	62,586	94,869
	2017	49,688	185,609	72,857	-
4	2015	47,088	-	43,464	27,102
	2016	48,367	-	61,384	95,765
	2017	49,727	173,825	73,438	-
5	2015	47,098	87,436	42,419	26,957
	2016	46,702	201,278	60,448	97,308
	2017	48,021	212,008	72,428	-

Grade	Year	Utah	Florida	Arizona	Oregon (2015)/Ohio (2016)
6	2015	46,160	87,831	40,512	27,550
	2016	46,380	193,158	57,868	101,015
	2017	46,263	195,425	70,034	-
7	2015	43,517	79,949	39,887	26,753
	2016	43,718	170,453	57,467	102,933
	2017	43,623	171,940	68,366	-
8	2015	43,745	60,958	39,997	26,969
	2016	43,377	125,120	49,781	78,629
	2017	44,035	120,321	59,171	-

5.2 Item Calibration and Linking for Science

5.2.1 Model Description

In discussing item response theory (IRT) models for the West Virginia science assessment, we distinguish between the underlying latent structure of a model and the parameterization of the item response function conditional on that assumed latent structure. Subsequently, we discuss how group effects are considered.

5.2.1.1 Latent Structure

Most operational assessment programs rely on a unidimensional IRT model for item calibration and computing scores for students. These models assume a single underlying trait, and that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This assumption of conditional independence implies that the conditional probability of a pattern of *I* item responses takes the relatively simple form of a product over items for a single student:

$$P(\boldsymbol{z_j}|\theta_j) = \prod_{i=1}^{I} P(z_{ij}|\theta_j),$$

where z_{ij} represents the scored response of student *j* (*j* = 1, …, *N*) to item *i* (*i* = 1, …, *I*), $\boldsymbol{z_j}$ represents the pattern of scored item responses for student *j*, and θ_j represents student *j*'s proficiency. Unidimensional IRT models differ with respect to the functional relation between the proficiency θ_j and the probability of obtaining a score z_{ij} on item *i*.

Items of the West Virginia science assessment are more complex than traditional item types. A single item may contain multiple parts, and each part may contain multiple student interactions. For example, a student may be asked to select a term from a set of terms at several places in a single item. Instead of receiving a single score for each item, multiple inferences are made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses to the item. These scoring units are called assertions and are the basic unit of analysis in our IRT analysis. That is, they fulfill the role of items in traditional assessments; however, for the West Virginia science items, multiple assertions are typically developed around a single item so that assertions are clustered within items.

One approach is to apply one of the traditional IRT models to the scored assertions; however, a substantial complexity that arises from the use of this new item types is that local dependencies exist between assertions pertaining to the same stimulus (item or item cluster). The local dependencies between the assertions pertaining to the same stimulus constitute a violation of the assumption that a single latent trait can explain all dependencies between assertions. Fitting a unidimensional model in the presence of local dependencies may result in biased item parameters and standard errors of measurement (SEM). In particular, it is well documented that ignoring local item dependencies leads to an overestimation of the amount of information conveyed by a set of responses and an underestimation of the SEM (e.g., Sireci, Thissen, & Wainer, 1991; Yen, 1993).

Many current ELA assessments also contain groups of items that pertain to the same stimulus. For example, often, several items share the same reading passage. Currently, item clustering effects and the resulting conditional dependencies are typically ignored, an approach that seems to work reasonably well in practice. This may be because in ELA assessments, the individual items within a group of items pertaining to the same passage are often written so that the effects of sharing the same stimulus material are kept to a minimum, for example, by relating items to different parts of the reading passage. However, for the West Virginia science items, the conditional dependencies between the assertions of an item (and item cluster) are too substantial to be ignored because those assertions are more intrinsically related to one another. For example, the assertions within an item are organized around a single performance expectation.

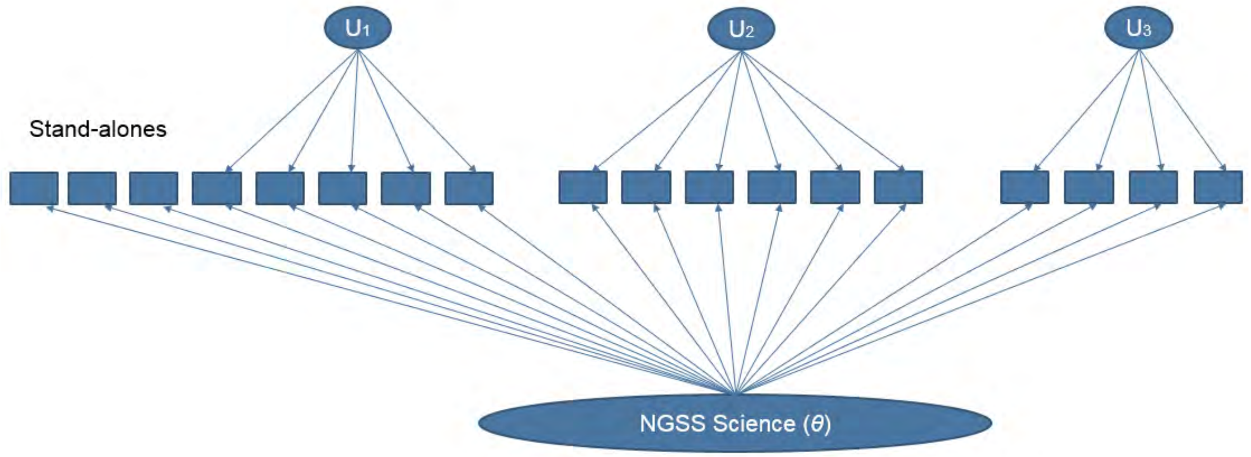
The effects of groups of assertions developed around a common stimulus can be accounted for by including additional dimensions corresponding to those groupings in the IRT model. These dimensions are considered to be nuisance dimensions. Whereas traditional unidimensional IRT models assume that all assertions (the basic units of analysis) are independent given a single underlying trait θ, we now assume the conditional independence of assertions, given the underlying latent trait θ and all nuisance dimensions

$$P(\boldsymbol{z_j}|\theta_j, \mathbf{u}_j) = \prod_{i\in \mathrm{SA}} P(z_{ij}|\theta_j) \prod_{g=1}^{G} \prod_{i\in g} P(z_{ij}|\theta_j, u_{jg}),$$

where SA indicates stand-alone assertions, u_{g} indicates the nuisance dimension for assertion group g (with the position of student j on that dimension denoted as u_{jg}), and $\mathbf{u}$ is the vector of all G nuisance dimensions. It can be seen that the conditional probability $P(z_{ij}|\theta_j, u_{jg})$ now becomes a function of two latent variables: the latent trait θ, representing a student's proficiency in science (the underlying trait of interest), and the nuisance dimension $u_{g,}$ accounting for the conditional dependencies between assertions of the same group. Furthermore, we assume that the nuisance dimensions are all uncorrelated with one another and with the general dimension. It is important to point out that even though every group of assertions introduces an additional dimension, models with this latent structure do not suffer from the curse of dimensionality like other multidimensional IRT models because one can take advantage of this special structure during model calibration (Gibbons & Hedeker, 1992). In this regard, Rijmen (Rijmen, 2010) showed that it is unnecessary to assume that all nuisance dimensions are uncorrelated; it is sufficient that they are independent, given the general dimension θ.

The model structure of the IRT model for science is illustrated in Figure 1. Note that stand-alone items can be scored with more than one assertion. The assertions of stand-alone items with more than one assertion, but fewer than four assertions, were also modeled as stand-alone assertions. Even though these assertions are likely to exhibit conditional dependencies, the variance of the nuisance dimension cannot be reliably estimated if it is based on a very small number of assertions. The few stand-alone items with four or more assertions were treated as item clusters to consider the conditional dependencies.

Figure 1. Directed Graph of the Science IRT Model



5.2.1.2 Item Response Function

The item response functions of the stand-alone assertions are modeled with a unidimensional model. For the grouped assertions, like in unidimensional models, different parametric forms can be assumed for the conditional probability of obtaining a score of z_{ij}. The Rasch testlet model (Wang & Wilson, 2005) is adopted as the IRT model for the science assessment. For binary data, it is defined as

$$P\left(z_{ij}|\theta_j, u_{jg}; b_i\right) = \frac{\exp\left(\theta_j + u_{jg} - b_i\right)}{1+\exp\left(\theta_j + u_{jg} - b_i\right)}.$$

The item response function of the Rasch testlet model is the probability of a correct answer (i.e., a true assertion), as a function of the overall proficiency θ, the nuisance dimension u_g, and the item (i.e., assertion) difficulty b_i. The IRT model for science does not include the item discrimination parameter; however, the same model structure as presented in Figure 1 could be employed with discrimination parameters included. Furthermore, only models for binary data are considered. Assertions are always binary because they are either true or false. Nevertheless, the model could easily accommodate polytomous responses by using the same response function that is incorporated in unidimensional models for polytomous data.

5.2.1.3 Multigroup Model

The science item bank was calibrated concurrently using all the items administered in any of the states that collaborate with CAI on their new science assessments. In the calibration, each state was treated as a population of students or a group. Overall group differences were considered by allowing a group-specific distribution of the overall proficiency variable θ. Specifically, for every student *j* belonging to group *k*, *k* = 1, …, *K*, a normal distribution was assumed,

$$\theta_j \sim N\left(\mu_k, \sigma_k^2\right),$$

where μ_k and σ_k^2 are the mean and variance of a normal distribution. The mean of the reference distribution (*k* = 1) was set to 0 to identify the model. For each of the nuisance variables u_g, a common variance parameter across groups was assumed, and the means were set to 0 in order to identify the model,

$$u_{jg} \sim N\left(0, \sigma_{u_g}^2\right).$$

5.2.2 Item Calibration

5.2.2.1 Estimation

A separate IRT model was fit for each grade band. The parameters of the IRT model were estimated using the marginal maximum likelihood (MML) method. In the MML method, the latent proficiency variable θ_j and the vector of nuisance parameters uj for each student *j* are treated as random effects and integrated out to obtain the marginal log likelihood corresponding to the observed response pattern $\mathbf{z}_j$ for student *j*,

$$\ell_j = \log \int \int P\left(\boldsymbol{z}_j | \theta_j, \boldsymbol{u}_j\right) N\left(\theta_j | \mu_k, \sigma_k^2\right) N\left(\boldsymbol{u}_j | \mathbf{0}, \boldsymbol{\Sigma}\right) d\boldsymbol{u}_j d\theta_j,$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements $\sigma_{u_k}^2$, denoting nuisance variance for group k. Across all students and groups, the overall log likelihood to be maximized with respect to the vector $\boldsymbol{\gamma}$ of all model parameters (item difficulty parameters, and the mean and variance parameters of the latent variables) is

$$\ell(\boldsymbol{\gamma}) = \sum_k \sum_{j \in k} \ell_j.$$

Even though the number of latent variables in the equation above is very high, the curse of dimensionality can be avoided because the integration over the high-dimensional latent $(\theta, \boldsymbol{u})$ space can be carried out as a sequence of computations in two-dimensional space $(\theta,\ \boldsymbol{u_g})$ (Gibbons & Hedeker, 1992; Rijmen, 2010).

The item bank was calibrated in 2018 after the 2018 science test administrations concluded and recalibrated in 2019 following the 2019 test administrations. The scores reported in 2019 were computed using the 2018 parameters since West Virginia reports scores before the testing window closes (immediate score reporting). The 2019 parameters were used for the 2021 and future test administrations. Because the calibration sequence was somewhat different between 2018 and 2019, the calibration sequence for both years is presented in detail below for both years. In addition, a summary of the 2021 and 2022 field-test items calibration and an overview of the 2022 operational item bank are also provided.

In 2018 and 2019, the IRT models were fitted using the BNL (Bayesian networks with logistic regression) suite of Matlab functions (Rijmen, 2006) and flexMIRT® (Cai, 2017). The resulting parameters from BNL were used as starting values for flexMIRT®, in order to speed up the estimation time for flexMIRT®. The flexMIRT® estimates were taken to be the operational parameters, except for the middle school items calibrated in 2018 during the core calibration (see Section 5.2.2.2, 2018 Calibration Sequence). For the 2018 core calibration of middle school items, flexMIRT® did not converge after several weeks, and the estimates obtained from BNL were used as operational parameters. Note that the parameters estimates were very similar across software packages.

Starting in 2021, the field-test items were calibrated with one multi-group calibration per grade band. In each calibration, the parameters of the operational items were fixed to their bank values (anchor items), and the item parameters of the field-test items as well as the mean and variance of each group were estimated using the MML method. Because the estimation time in flexMIRT® became prohibitive, CAIRT (Cambium Assessment IRT) was used. CAIRT was specifically developed by CAI to calibrate the multigroup Rasch model on very large data sets. It relies on the same estimation methods as BNL. CAI has cross-validated parameter estimates from CAIRT with BNL and flexMIRT under a variety of scenarios (Rijmen, Liao, & Lin, 2021). In 2022, field-test items were calibrated in CAIRT using the same procedure as 2021.

5.2.2.2 2018 Calibration Sequence

Table 60 provides an overview of the groups per grade for the 2018 calibration.

Table 60. Groups Per Grade Band for the Spring 2018 Core Calibration

GROUP	ELEMENTARY SCHOOL	MIDDLE SCHOOL	HIGH SCHOOL
CONNECTICUT	X	X	X
HAWAII	X	X	X
NEW HAMPSHIRE	X	X	X
RHODE ISLAND	X	X	X
UTAH GRADE 6		X	
UTAH GRADE 7		X	
UTAH GRADE 8		X	
VERMONT	X	X	X
WEST VIRGINIA	X	X	X

Items were calibrated in three steps for two reasons. First, the rubric validations for some states took place at a later date, and the student responses for the items owned by those states could not be included in the first round of calibrations without jeopardizing the reporting schedule of the two states with operational field tests (those two states did not have any of the items with late rubric validation in their item pool). Second, in order to divide the very large set of items and assertions into more manageable pieces, a separate calibration was carried out for two states with a large number of items administered only in those states. Specifically, the following sequence of calibrations was carried out:

1. Core calibration. The core calibration was performed on the following:

a. All the item responses of New Hampshire and West Virginia. These states administered items from (see bank sharing matrix in Table 61). A more detailed overlap of the common items at the time of the 2018 calibration was given in Section 3.2.1.1, 2018 Field Test (see Table 21 through Table 22).

   i. ICCR

   ii. Connecticut

   iii. Hawaii

   iv. Rhode Island

   v. Vermont

   vi. Utah

   vii. West Virginia

b. All the item responses of Connecticut, Rhode Island, and Vermont, except for the responses to Wyoming and Oregon items. These states administered items from the following sources:

   i. ICCR

   ii. Connecticut

   iii. Hawaii

   iv. Rhode Island

   v. Vermont

   vi. Utah

   vii. West Virginia

   viii. Wyoming (items were treated as not administered; responses were replaced by missing code)

   ix. Oregon (items were treated as not administered; responses were replaced by missing code)

c. Item responses from Hawaii to items also administered in another state (Hawaii items were used in Hawaii, Connecticut, Rhode Island, Vermont, and West Virginia)

d. Item responses from Utah to items also administered in another state (Utah items were used in Connecticut, Rhode Island, Utah, Vermont, and West Virginia). Utah tested middle school students only but included every grade in middle school. One third of students was selected at random to balance the large population size for Utah.

Table 61. Spring 2018 State-Sharing Matrix

Source Bank and State-Owned	CT	HI	MSSA	NH (from ITS Sandbox)	OR	UT	WV	WY
ICCR	X	X	X	X	X		X	X
Connecticut	X		X				X	
Hawaii	X	X	X				X	
MSSA	X		X				X	
Oregon	X		X		X			
Utah	X		X			X	X	
West Virginia	X		X				X	
Wyoming	X		X					X

Note. The core calibration provided parameters for all items used in New Hampshire and West Virginia.

2. Calibration of state-specific items.

   Both Utah and Hawaii had a substantial proportion of items that were administered only in Utah and Hawaii, respectively. Hawaii has both Hawaii and ICCR items in common with the states of the core calibration (Hawaii administered only Hawaii and ICCR items); Utah has only Utah items in common (Utah administered Utah items only). The parameters for the unique Hawaii items depend only on responses from Hawaii students, and the parameters for the unique Utah items depend only on responses from Utah students. For both states, the state-specific items were calibrated through a separate calibration based on the state data only, with the items in common with the core states mentioned in step 1 anchored to the estimates from step 1. These calibrations were conducted separately for each group, under a single-group IRT model. The mean and variance of the groups were fixed to the estimated mean and variance from core calibration 1.

3. Calibration of states with late rubric validation.

   Oregon and Wyoming items were administered in some of the states from the core calibration (Connecticut, Rhode Island, and Vermont) but could not be calibrated in step 1 because of their late rubric validation dates. In a later stage, items from Oregon and Wyoming were calibrated by the following methods:

   a. Adding Oregon and Wyoming student responses to the core calibration

   b. Keeping the responses from Connecticut, Rhode Island, and Vermont to Wyoming and Oregon items (as opposed to treating them as missing in step 1)

   c. Removing the responses from the states that did not administer Oregon or Wyoming items (as the item parameters for the Oregon and Wyoming items did not depend on the students from these states). The removed states were Hawaii, New Hampshire, Utah, and West Virginia.

d. Fixing the parameters of all other items to the values obtained in step 1, as well as the group means and standard deviations that were estimated in step 1.

5.2.2.3 2019 Calibration Sequence

The calibration was executed in two steps. CAI calibrated all items in operational use in 2019, for which 1,000 or more student responses were available (among these, there were 1,500 or more student responses for all but three items). In this step, only the data of states with an operational test were included. Table 62 provides an overview of the groups per grade for this first calibration. All students who attempted the test were included in the calibration. The assertions of skipped items were scored as incorrect. Note that only Rhode Island allowed students to skip items. There were nine items administered as operational items in 2019 for which the sample size was smaller than 1,000, out of a total of 438 items.

Table 63 and Table 64 present the number of operational clusters and stand-alone items that were shared between the item pools of any two states. The numbers below the diagonal elements represent the numbers for all the operational items, and the numbers above the diagonal elements represent the number of common operational items at the time of the 2019 calibration. The shaded diagonal elements represent the number of operational items that were administered only in the given state (in parentheses, the number of unique operational items at the time of calibration). Since the items that were administered but not calibrated were administered only in one state, the numbers above the diagonal are the same as the numbers below the diagonal. Table 63 presents the results for elementary schools, and Table 64 presents the results for middle schools. The numbers at the operational administration are slightly different from the numbers at the calibration because items with a sample size smaller than 1,000 were excluded from the calibration.

Table 62. Groups per Grade Band for the Spring 2019 Calibration of Operational Items

GROUP	ELEMENTARY SCHOOL	MIDDLE SCHOOL	HIGH SCHOOL
CONNECTICUT	X	X	X
NEW HAMPSHIRE	X	X	X
OREGON	X	X	X
RHODE ISLAND	X	X	X
VERMONT	X	X	X
WEST VIRGINIA	X	X	

Table 63. Number of Common Elementary School Operational Items Administered and Calibrated in Spring 2019, Science

	State	Connecticut	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia
Cluster	**CT**	**1 (1)**	44	24	42	55
	MSSA	44	**0 (0)**	17	37	41
	NH	24	17	**0 (0)**	14	27

	State	Connecticut	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia
	OR	42	37	14	**0 (0)**	41
	WV	55	41	27	41	**1 (1)**
Stand-Alone	**CT**	**3 (3)**	34	26	30	47
	MSSA	34	**0 (0)**	20	23	32
	NH	26	20	**0 (0)**	14	25
	OR	30	23	14	**0 (0)**	25
	WV	47	32	25	25	**1 (1)**
Grade-Band Total	**CT**	**4 (4)**	78	50	72	102
	MSSA	78	**0 (0)**	37	60	73
	NH	50	37	**0 (0)**	28	52
	OR	72	60	28	**0 (0)**	66
	WV	102	73	52	66	**2 (2)**

Table 64. Number of Common Middle School Operational Items Administered and Calibrated in Spring 2019, Science

	State	Connecticut	MSSA (RI, VT)	New Hampshire	Oregon	West Virginia
Cluster	**CT**	**3 (3)**	26	24	54	92
	MSSA	26	**0 (0)**	11	14	21
	NH	24	11	**1 (1)**	9	18
	OR	54	14	9	**2 (2)**	56
	WV	92	21	18	56	**12 (4)**
Stand-Alone	**CT**	**0 (0)**	42	26	34	50
	MSSA	42	**0 (0)**	25	30	37
	NH	26	25	**0 (0)**	16	21
	OR	34	30	16	**1 (0)**	29
	WV	50	37	21	29	**0 (0)**
Grade-Band Total	**CT**	**3 (3)**	68	50	88	142
	MSSA	68	**0 (0)**	36	44	58
	NH	50	36	**1 (1)**	25	39
	OR	88	44	25	**3 (2)**	85
	WV	142	58	39	85	**12 (4)**

In a second step, the field-test items were calibrated. The calibration included the operational items that were calibrated in step 1, and the field-test items across all states that administered field-test items. All students who attempted at least one field-test item were included in the calibration. Table 65 provides an overview of the groups per grade for calibration of the field-test items.

Table 65. Groups Per Grade Band for the Calibration of Field-Test Items

GROUP	ELEMENTARY SCHOOL	MIDDLE SCHOOL	HIGH SCHOOL
CONNECTICUT	X	X	X
HAWAII	X	X	X
IDAHO	X	X	
NEW HAMPSHIRE	X	X	X
OREGON	X	X	X
RHODE ISLAND	X	X	X
VERMONT	X	X	X
WEST VIRGINIA	X	X	
WYOMING	X	X	X

5.2.2.4 Linking the 2018 Scale to the 2019 Scale

The item parameter estimates obtained from the 2018 student responses were highly correlated with the item parameters obtained from the 2019 student responses. For the item difficulties, the correlation between the 2018 and 2019 estimates was 0.993 for elementary school and 0.986 for middle school. For the standard deviations of the clusters, these correlations were 0.971 and 0.972, respectively. These high correlations indicate that items functioned similarly in 2018 and 2019. Nevertheless, item parameters from separate calibrations cannot be directly compared because the scale of an IRT model is not determined. In the multigroup Rasch testlet model, the only scale indeterminacy is the origin of the scale. The models can be identified by setting the mean of the overall proficiency variable θ to 0 for the reference distribution. As a result, the 2018 and 2019 variable θ and item parameters are on the same scale except for an overall shift parameter B. Specifically, the 2018 scale can be linked to the 2019 scale as follows:

$$P\left(z_{ij}|\theta_{j\ 2018}, u_{jg}; b_{i\ 2018}\right) = \frac{\exp\left(\theta_{j\ 2018} + u_{jg} - b_{i\ 2018}\right)}{1 + \exp\left(\theta_{j\ 2018} + u_{jg} - b_{i\ 2018}\right)}$$

$$= \frac{exp\left(\theta_{j\ 2018} + B + u_{jg} - b_{i\ 2018} - B\right)}{1 + exp\left(\theta_{j\ 2018} + B + u_{jg} - b_{i\ 2018} - B\right)}$$

$$= \frac{exp\left(\theta_{j\ 2019} + u_{jg} - b_{i\ 2019}\right)}{1 + exp\left(\theta_{j\ 2019} + u_{jg} - b_{i\ 2019}\right)}.$$

Because $\theta_{j\ 2019} = \theta_{j\ 2018} + B$, the population means of θ must be transformed accordingly,

$$\theta_{j\ 2019} \sim N\ (\mu_{k\ 2018} + B, \sigma_k^2) \text{ and}$$

$$\theta_{j\ 2018} \sim N\ (\mu_{k\ 2018}, \sigma_k^2).$$

Item parameters based on 2018 student responses can be expressed on the 2019 scale by adding the constant *B* to the 2018 item parameter. The 2018 parameters were expressed on the 2019 scale for items that were part of the pool in both 2018 and 2019 but not administered in any states in 2019 (13 items) and for items that were administered in 2019. The number of student responses from the 2019 assessments was lower than 1,000 (9 items). Therefore, the linking process was performed for 22 items only.

All items that were operational in 2019 were also administered in 2018. Therefore, the shift parameter *B* can be estimated from a separate calibration of the items operational in 2019 using the 2019 student responses (of the six operational states), but with the item parameters fixed to the estimates obtained from the 2018 calibrations. By fixing a subset of the item parameters, the model is identified so that the means and variances of θ can be estimated for all groups. *B* can be obtained by equating the overall mean of θ across all groups for the 2019 student response data from the free calibration (2019 overall mean expressed on the 2019 scale) to the overall mean of θ across all groups for the 2019 student response data from the calibration with items anchored to their 2018 parameters values (2019 overall mean expressed on the 2018 scale):

$$\frac{1}{K}\sum_{k=1}^{K}\mu_{k\ 2019} = \frac{1}{K}\sum_{k=1}^{K}(\mu_{k\ 2018} + B).$$

Therefore, an estimate of *B* can be obtained as

$$\hat{B} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_{k\ 2019} - \hat{\mu}_{k\ 2018}).$$

The estimated means of θ under both the free and anchored calibrations as well as the number of students per state are presented in Table 66. The table also presents the overall means and estimated shift parameter *B*. Note that the parameters for three items were not anchored but freely estimated together with the means and variances in the anchored calibration. The reason for not treating these items as common items across the 2018 and 2019 administrations was that they had an omit rate of 4% or higher for the last item interaction in the 2018 administration in at least one state; in 2019, these interactions could no longer be omitted because all interactions of an item needed to be responded to in states where skipping was not allowed (all states except for Rhode Island). So, out of an abundance of caution, these three items were not anchored to their 2018 parameter values.

Table 66. Estimated Latent Means and Number of Students Per State

GROUP	ELEMENTARY SCHOOL			MIDDLE SCHOOL		
	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	*N*	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	*N*
CONNECTICUT	0.0000	0.0518	38549	0.0000	0.0234	39347
NEW HAMPSHIRE	0.0631	0.1083	13187	0.0940	0.1108	12060
OREGON	-0.0101	0.0096	44989	0.0028	0.0156	42043
RHODE ISLAND	-0.0312	0.0142	10751	-0.1044	-0.0692	10306
VERMONT	0.1069	0.1504	6017	0.0781	0.1133	5894
WEST VIRGINIA	-0.1970	-0.1529	19540	-0.3012	-0.2783	19043

GROUP	ELEMENTARY SCHOOL			MIDDLE SCHOOL		
	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	N	$\hat{\mu}_{k\ 2019}$	$\hat{\mu}_{k\ 2018}$	N
	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$	$\hat{B}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$	$\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$	$\hat{B}$
OVERALL	-0.0114	0.0303	-0.0416	-0.0385	-0.0141	-0.0244

5.2.2.5 Calibration of Field-Test Items in 2021 and Beyond

Starting in 2021, field-test items were calibrated with one multigroup calibration per grade band. In each calibration, the parameters of the operational items were fixed to their bank values (anchor items), and the item parameters of the field-test items as well as the mean and variance of each group were estimated using the MML method. The calibration included the field-test items across all states in which they were administered. All students who attempted at least one field-test item were included in the calibration. Table 67 and Table 68 provide an overview of the groups per grade band for calibration of the field-test items in 2021 and 2022, respectively.

Table 67. Groups Per Grade Band for the Spring 2021 Calibration of Field-Test Items

GROUP	ELEMENTARY SCHOOL	MIDDLE SCHOOL	HIGH SCHOOL
CONNECTICUT	X	X	X
HAWAII	X	X	X
IDAHO	X	X	
MONTANA	X	X	
NORTH DAKOTA	X	X	X
NEW HAMPSHIRE	X	X	X
RHODE ISLAND	X	X	X
SOUTH DAKOTA	X	X	X
UTAH	X	X	
VERMONT	X	X	X
WEST VIRGINIA	X	X	
WYOMING	X	X	X

Table 68. Groups Per Grade Band for the Spring 2022 Calibration of Field-Test Items

Group	Elementary School	Middle School	High School
Connecticut	X	X	X
Hawaii	X	X	X
Idaho	X	X	

Group	Elementary School	Middle School	High School
Montana	X	X	
North Dakota	X	X	X
New Hampshire	X	X	X
Oregon	X	X	X
Rhode Island	X	X	X
South Dakota	X	X	X
Utah	X	X	
Vermont	X	X	X
West Virginia	X	X	
Wyoming	X	X	X

5.2.2.7 Overview of the Operational Bank

Figure 2 and Figure 3 display a histogram of the difficulty parameters for grade 5 and grade 8 respectively, for all items that are part of the WVGSA operational pool. The figures also display the student proficiency distributions. The distribution of the difficulty parameter overlaps well with the proficiency distribution in grade 5. The grade 8 items are slightly more difficult compared to the student proficiency level.

Figure 2. WVGSA Item Difficulty and Student Proficiency Distributions, Grade 5

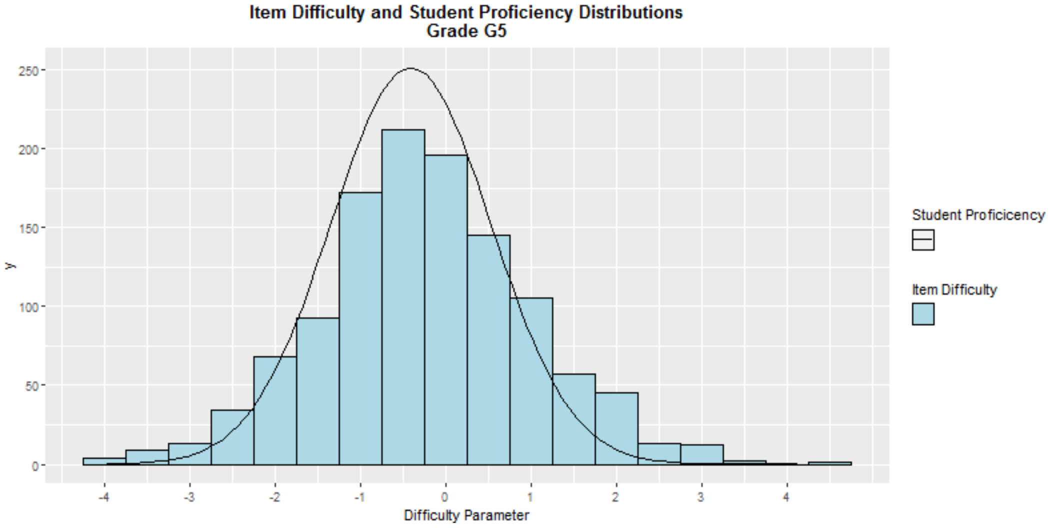
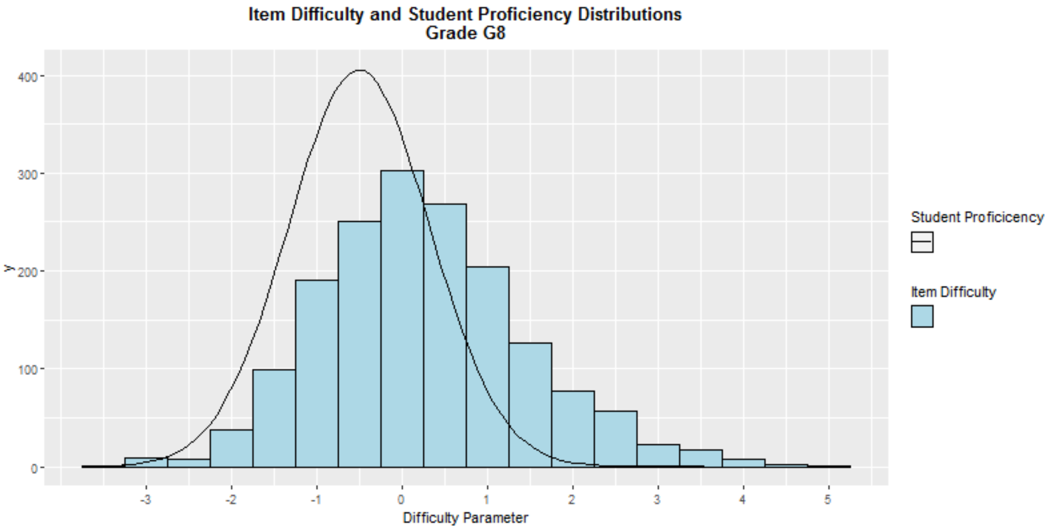


Figure 3. WVGSA Item Difficulty and Student Proficiency Distributions, Grade 8



6. Scoring

6.1 Maximum Likelihood Estimation for ELA and Mathematics

Ability estimates were generated using *pattern scoring*, a method that scores students depending on how they answer individual items. Scoring details for all ELA and mathematics online and paper forms are provided in the following paragraphs.

6.1.1 Likelihood Function

The likelihood function for generating the maximum likelihood estimates (MLEs) is based on a mixture of item types and can therefore be expressed as

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N_{MC}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{exp \sum_{l=1}^{z_i} Da_i(\theta - b_{il})}{1 + \sum_{h=1}^{m_i} exp \sum_{l=1}^{h} Da_i(\theta - b_{il})}$$

$$p_i = c_i + \frac{1 - c_i}{1 + exp\,[-Da_i(\theta - b_i)]}$$

$$q_i = 1 - p_i$$

where c_i is the lower asymptote of the item response curve (i.e., the pseudo-guessing parameter), a_i is the slope of the item response curve (i.e., the discrimination parameter), b_i is the location parameter, z_i is the observed response to the item, *I* indices the item, *h* indices step of the item, m_i is the maximum possible score point, b_{il} is the *l*th step for item *i* with *m* total categories, and $D = 1.7$.

A student's theta (i.e., MLE) is defined as $\arg\max_{\theta} og\left(L(\theta)\right)$, given the set of items administered to the student.

6.1.2 Derivatives

Finding the maximum of the likelihood requires an iterative method, such as Newton-Raphson iterations. The estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} \Big/ \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t}$$

where

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} = \frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} = \sum_{i=1}^{N_{3PL}} Da_i \frac{(P_i - c_i)Q_i}{1 - c_i} \left(\frac{z_i}{P_i} - \frac{1 - z_i}{Q_i} \right)$$

$$\frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} = -\sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1 - c_i)^2} \left(1 - \frac{z_i c_i}{P_i^2} \right)$$

$$\frac{\partial \ln L(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N_{CR}} Da_i \left(exp\left(\sum_{k=1}^{z_i} Da_i(\theta - \delta_{ki}) \right) \right) \left(\frac{z_i}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} - \frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{\left(1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)\right)^2} \right)$$

$$\frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} \right)$$

and where $_t$ denotes the estimated θ at iteration *t*. N$_{CR}$ is the number of items that are scored using the generalized partial credit model (GPCM), and N$_{3PL}$ is the number of items scored using a three-parameter logistic model (3PL) or a two-parameter logistic model (2PL).

6.1.3 Standard Errors of Estimate

When the MLE is available, the standard error (SE) of the MLE is estimated by

$$se(\hat{\theta}) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta}\right)}},$$

where

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} j Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1 - c_i)^2} \left(1 - \frac{z_i c_i}{P_i^2}\right)$$

where N$_{CR}$ is the number of items that are scored using the GPCM model, and N$_{3PL}$ is the number of items scored using the 3PL or 2PL model.

6.1.4 Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded, and an MLE cannot be generated. Additionally, when a student's raw score is lower than the expected raw score due to guessing, the likelihood is not identified. For WVGSA scoring, the extreme cases were handled as follows:

i. Assign the Lowest Obtainable Theta (LOT) value of –4 to a raw score of 0.
ii. Assign the Highest Obtainable Theta (HOT) value of 4 to a perfect score.
iii. Generate MLE for every other case and apply the following rule:
    a. If MLE is lower than –4, assign theta to –4.
    b. If MLE is higher than 4, assign theta to 4.

As WVGSA used a vertical score for scoring, the truncated LOT and HOT were converted to the vertical scale before being applied. These truncated LOT and HOT in a vertical scale and the associated scale scores for each grade and subject are provided in Table 69 and Table 70.

Table 69. Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates, ELA

Grade	Lowest Obtainable Theta (LOT)	Highest Obtainable Theta (HOT)	Lowest Obtainable Scale Score (LOSS)	Highest Obtainable Scale Score (HOSS)
3	−4.61	2.03	420	750
4	−4.39	2.73	430	790
5	−4.01	3.11	450	810
6	−3.72	3.48	460	830
7	−3.75	3.77	470	850
8	−3.84	4.24	480	870

Table 70. Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates, Mathematics

Grade	Lowest Obtainable Theta (LOT)	Highest Obtainable Theta (HOT)	Lowest Obtainable Scale Score (LOSS)	Highest Obtainable Scale Score (HOSS)
3	−4.85	-0.05	300	550
4	−4.77	1.15	310	610
5	−4.63	2.17	320	660
6	−4.52	3.40	330	720
7	−4.05	4.03	340	750
8	−4.28	5.64	350	830

6.1.5 Standard Error of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the SE is estimated based on Fisher information.

When the MLE is not available (such as for extreme score cases), or the MLE is censored to the LOT or HOT, the SE for student *s* with ability θ_s is estimated by

$$se(\theta_s) = \frac{1}{\sqrt{I(\theta_s)}},$$

where $I(\theta_s)$ is the test information for students. The WVGSA included items that were scored using the 3PL, 2PL, and GPCM from the item response theory (IRT). The 2PL can be seen as either 3PL items with no pseudo-guessing parameter or dichotomously scored GPCM items. The test information was calculated as

$$I(\theta_s) = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} D a_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} D a_i(\theta_s - b_{ik})\right)} - \left(\frac{\sum_{j=1}^{m_i} j Exp\left(\sum_{k=1}^{j} D a_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} D a_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{Q_i}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2 \right),$$

where N_{CR} is the number of items that are scored using the GPCM model, and N_{3PL} is the number of items scored using the 3PL or 2 PL model.

For SE of LOT/HOT scores, theta in the formula above is replaced with the LOT/HOT values. The upper bound of the SE was set to 1.5 and converted to the vertical scale. Any value larger than 1.5 was truncated at 1.5. Truncated standard error of measurement (SEM) values on the vertical scale are provided in Table 71.

Table 71. SEM Truncation Values for Each Grade and Subject

Subject	Grade	SEM Truncation Values on Theta Metric	SEM Truncation Values on Vertical Scale
ELA	3	1.5	1.25
	4	1.5	1.34
	5	1.5	1.34
	6	1.5	1.35
	7	1.5	1.41
	8	1.5	1.52
Mathematics	3	1.5	0.90
	4	1.5	1.11
	5	1.5	1.28
	6	1.5	1.49
	7	1.5	1.52
	8	1.5	1.86
Science	5	1	1.4
	8	1	1.4

6.1.6 Transforming Vertical Scores to Reporting Scale Scores, ELA and Mathematics

For spring 2022, WVGSA scale scores were reported for each student who took the English language arts (ELA) and mathematics assessments. The scale scores are based on the operational items presented to each student and do not include any field-test or linking items. Independent College and Career Readiness (ICCR) item parameters are converted to a vertical scale in the item bank, and a single scale across all grades is used within ELA and mathematics. The reporting scale scores were calculated as

$$SS = slope * \theta_{vertical} + intercept,$$

where $slope$ and $intercept$ are the reporting scaling constants, and $\theta_{vertical}$ is the post-vertically-scaled IRT ability estimate. For ELA, the slope and intercept were fixed at 50 and 650, and for mathematics, the slope and intercept were fixed at 50 and 550. In this transformation, the following rules were applied:

1. The same linear transformation was used for all students within a grade.
2. Scale scores were rounded to the nearest integer (e.g., 302.4 to 302; 302.5 to 303).
3. An SE was provided for each score, using the same set of items used to derive the score. The SE of the scaled score is calculated as

$$se(SS) = se(\theta_{vertical}) * slope.$$

4. Truncated scale scores use actual SEs from the vertical scale theta estimates.

The summary of WVGSA scale scores for each test is provided in Appendix E, Distribution of Scale Scores and Achievement Levels by Subgroup, and the summary of scale scores for each reporting category is provided in Appendix F, Distribution of Reporting Category Scores. All scores are based on the operational items presented to the student.

6.1.7 Overall Performance Classification

Each student is assigned an overall achievement category according to his or her overall scale score. Table 72 and Table 73 provide the scale score ranges for achievement standards for ELA and mathematics. The lower bound of level 3, *Meets Standard*, marks the minimum cut score for proficiency. Appendix I, Achievement Level Distribution Comparison Between 2018, 2019, 2021 and 2022, provides a comparison of percentages of students classified into each level across spring 2018, spring 2019, spring 2021, and spring 2022 for ELA and mathematics. Across the four school years, the proportions of students within each level vary, given the circumstances revolving around Covid-19.

Table 72. Achievement Levels by Grade, ELA

Grade	Level 1 Does Not Meet Standard	Level 2 Partially Meets Standard	Level 3 Meets Standard	Level 4 Exceeds Standard
3	420–549	550–585	586–615	616–750
4	430–562	563–598	599–628	629–790
5	450–587	588–621	622–654	655–810
6	460–596	597–638	639–679	680–830
7	470–601	602–643	644–684	685–850
8	480–612	613–655	656–697	698–870

Table 73. Achievement Levels by Grade, Mathematics

Grade	Level 1 Does Not Meet Standard	Level 2 Partially Meets Standard	Level 3 Meets Standard	Level 4 Exceeds Standard
3	300–400	401–425	426–447	448–550
4	310–421	422–455	456–477	478–610
5	320–448	449–486	487–512	513–660
6	330–473	474–517	518–549	550–720
7	340–502	503–547	548–582	583–750
8	350–528	529–586	587–616	617–830

6.1.8 Reporting Category Performance Classification

In addition to overall performance classification, subscale-level classification is computed to determine student performance levels for each of the content standard subscales. For each subscale, classification into one of three performance levels is determined by following the rules:

- If $(\theta_{tt} < \theta_{Proficient} - 1.5 \times SE_{RC})$, then performance is classified as *Below Standard.*
- If $(\theta_{Proficient} - 1.5 \times SE_{RC} \leq \theta_{tt} < \theta_{Proficient} + 1.5 \times SE_{RC})$, then performance is classified as *At or Near Standard.*
- If $(\theta_{tt} \geq \theta_{Proficient} + 1.5 * SE_{RC})$, then performance is classified as *Above Standard.*

Where $\theta_{Proficient}$ is the minimum proficiency cut score based on the overall test, θ_{tt} is the student's score on a given subscale, and SE_{RC} is the standard error of the given subscale. Zero and perfect scores were assigned *Below Standard* and *Above Standard*, respectively.

6.1.9 Strength and Weakness Scores

For individual students, strengths and weaknesses within reporting categories are computed relative to the student's estimated ability.

For each item *i*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij})$$

where $E(z_{ij})$ is the expected score on item *i* for student *j* with estimated ability $\hat{\theta}_j$.

Residuals are summed up for items within a reporting category. The sum of residuals is divided by the total number of points possible for items within the reporting category, *T*,

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score for the reporting category is computed by averaging the target scores of individual students with different abilities who receive different items that measure the same reporting category at different levels of difficulty,

$$\bar{\delta}_{Tg} = \frac{1}{n_g}\sum_{j\in g}\delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)}\sum_{j\in g}\left(\delta_{jT} - \bar{\delta}_{Tg}\right)^2},$$

where n_g is the number of students who responded to any of the items that belong to the reporting category T for an aggregate unit g. If a student did not happen to see any items on a particular reporting category, the student is not included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (negative $\bar{\delta}_{Tg}$) at teaching a given target.

For reporting category-level strengths/weakness, the following is reported:

- If $\bar{\delta}_{Tg} \leq -1 * \; se\left(\bar{\delta}_{Tg}\right)$, then performance is worse than on the overall test.
- If $\bar{\delta}_{Tg} \geq +1 * \; se\left(\bar{\delta}_{Tg}\right)$, then performance is better than on the overall test.
- Otherwise, performance is similar to performance on the overall test.
- If $se\left(\bar{\delta}_{Tg}\right)$ > 0.2, data are insufficient.

6.1.10Lexile® and Quantile® Scores

WVGSA reports Lexile and Quantile measures with ELA and mathematics. MetaMetrics provides conversion tables between ELA scale scores and Lexile measures, and between mathematics scale scores and Quantile measures for each grade. A linking study for ELA and mathematics took place in June 2018 to determine final conversions. (The linking study report can be found in the 2017–2018 technical report, Volume 7, for special studies.) Lexile and Quantile measures are reported for all tests including online and paper tests.

6.2 MARGINAL MAXIMUM LIKELIHOOD ESTIMATION FOR SCIENCE

6.2.1 Marginal Maximum Likelihood Function

Student scores are obtained by marginalizing out the nuisance dimensions **u**$_j$ from the likelihood of the observed response pattern **z**$_j$ for student *j*,

$$\ell_i\left(\theta_j\right) = \log\int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \boldsymbol{\Sigma}\right) d\mathbf{u}_j,$$

and maximizing this marginalized likelihood function for θ_j. The marginal maximum likelihood estimation (MMLE) is a hybrid of the expected a posteriori (EAP) estimator (by marginalizing out the nuisance dimensions) and the MLE estimator (by maximizing the resulting marginal likelihood for θ). The marginal likelihood is maximized with respect to θ using the Newton Raphson method.

The proposed model reduces to the unidimensional Rasch model when the nuisance variances are zero for all g. Likewise, the proposed MMLE is equivalent to the MLE of the unidimensional Rasch model when all the nuisance variances are zero. This can be shown by using the variable transformation $\mathbf{v} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{u}$. Then we have

$$\int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \boldsymbol{\Sigma}\right) d\mathbf{u}_j = \int_{\mathbf{v}_j} P\left(\mathbf{z}_j \middle| \theta_j, \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{v}_j\right) N\left(\mathbf{v}_j \middle| \mathbf{0}, \mathbf{I}\right) d\mathbf{v}_j .$$

If $\sigma^2_{u_g} = 0$ for all g, then

$$\int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \boldsymbol{\Sigma}\right) d\mathbf{u}_j = P\left(\mathbf{z}_j \middle| \theta_j\right),$$

which is the likelihood under the unidimensional Rasch model.

6.2.2 Derivatives

The marginal log likelihood function based on the item response theory (IRT) model with one overall dimension and one nuisance dimension for each grouping of assertions can be written as

$$l(\theta) = \sum_{i \in \mathrm{SA}} \log\big(P(z_i|\theta)\big) + \sum_{g=1}^{G} \log\left\{\int \mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma^2_{u_g}\right) du_g\right\}.$$

The first derivative of the marginal log likelihood function with respect to θ is

$$\frac{dl(\theta)}{d\theta} = \sum_{i \in \mathrm{SA}} \frac{\frac{dP(z_i|\theta)}{d\theta}}{P(z_i|\theta)} + \sum_{g=1}^{G} \frac{\int \left\{\mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i \in g} \frac{\frac{dP\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)}\right) N\left(u_g \middle| 0, \sigma^2_{u_g}\right)\right\} du_g}{\int \left\{\mathrm{Exp}\left[\sum_{i \in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma^2_{u_g}\right)\right\} du_g},$$

and the second derivative of the marginal log likelihood function with respect to θ is

$$
\begin{aligned}
&\frac{d^2 l(\theta)}{d\theta^2} \\
&= \sum_{i \in \mathrm{SA}} \left[\frac{\frac{d^2\, P(z_i|\theta)}{d\theta^2}}{P(z_i|\theta)} - \left(\frac{\frac{d\, P(z_i|\theta)}{d\theta}}{P(z_i|\theta)} \right)^2 \right] \\
&+ \sum_{g=1}^{G} \frac{\int \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i\in g} \frac{\frac{d\, P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right)^2 N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g} \\
&+ \sum_{g=1}^{G} \frac{\int \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i\in g} \left[\frac{\frac{d^2\, P\left(z_{ig}|\theta, u_g\right)}{d\theta^2}}{P\left(z_{ig}|\theta, u_g\right)} - \left(\frac{\frac{d\, P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right)^2 \right] \right) N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g} \\
&- \sum_{g=1}^{G} \left\{ \frac{\int \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] \left(\sum_{i\in g} \frac{\frac{d\, P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)} \right) N\left(u_g \middle| 0, \sigma_{u_g}^2\right) du_g}{\int \left\{ \mathrm{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right] N\left(u_g \middle| 0, \sigma_{u_g}^2\right) \right\} du_g} \right\}^2
\end{aligned}
$$

Based on these equations, we need only to define the ratios of the first and second derivatives of the item response probabilities with respect to θ to the response probabilities. For the Rasch testlet model, these are obtained as

$$p_i = P(z_i = 1|\theta) = \frac{E\mathrm{xp}(\theta - b_i)}{1 + E\mathrm{xp}(\theta - b_i)},\ q_i = P(z_i = 0|\theta) = 1 - p_i,$$

and

$$p_{ig} = P\left(z_{ig} = 1|\theta, u_g\right) = \frac{E\mathrm{xp}\left(\theta + u_g - b_i\right)}{1 + E\mathrm{xp}\left(\theta + u_g - b_i\right)},\ q_{ig} = P\left(z_{ig} = 0|\theta, u_g\right) = 1 - p_{ig}.$$

Therefore, we have,

$$\frac{\frac{dp_i}{d\theta}}{p_i} = q_i\ ,\ \frac{\frac{dq_i}{d\theta}}{q_i} = -p_i,$$

$$\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}} = q_{ig}\ ,\ \frac{\frac{dq_{ig}}{d\theta}}{q_{ig}} = -p_{ig},$$

$$\frac{\frac{d^2\, p_i}{d\theta^2}}{p_i} - \left(\frac{\frac{d\, p_i}{d\theta}}{p_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2\, q_i}{d\theta^2}}{q_i} - \left(\frac{\frac{d\, q_i}{d\theta}}{q_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2\, p_{ig}}{d\theta^2}}{p_{ig}} - \left(\frac{\frac{d\, p_{ig}}{d\theta}}{p_{ig}}\right)^2 = -p_{ig} q_{ig}, \text{ and}$$

$$\frac{\frac{d^2\, q_{ig}}{d\theta^2}}{q_{ig}} - \left(\frac{\frac{d\, q_{ig}}{d\theta}}{q_{ig}}\right)^2 = -p_{ig} q_{ig}.$$

6.2.3 Extreme Case Handling

Just like the MLE, the MMLE is not defined for zero and perfect scores. These cases are handled by assigning the lowest (LOT) and highest (HOT) obtainable theta scores, respectively. Table 71 contains the LOT and HOT values for each grade.

6.2.4 Standard Errors of Estimate

The SEM of the MMLE score estimate is:

$$SEM(\hat{\theta}_{MMLE}) = \frac{1}{\sqrt{I(\hat{\theta}_{MMLE})}}$$

where $I(\hat{\theta}_{MMLE})$ is the observed information evaluated at $\hat{\theta}_{MMLE}$. The observed information is calculated as $I(\theta^2) = -\frac{d^2 l(\theta)}{d\theta^2}$, where $\frac{d^2 l(\theta)}{d\theta^2}$ is defined in the previous section on derivatives. Note that the calculation of the SEM depends on the unique set of items that each student answers and their estimate of θ. Different students have different SEM values, even if they have the same raw score and/or theta estimate. Standard errors are truncated at 1 for the overall science scores and truncated at 1.4 for the discipline scores.

Standard errors for MMLE estimates truncated at the LOT and HOT are computed by evaluating the observed information at the MMLE before truncation. For all incorrect or all correct answers, the reported standard are set at the truncation value for the standard error.

6.2.5 Scoring Incomplete Tests

The science assessments are assembled on the fly using an adaptive testing design. Tests are considered complete if students respond to all the operational items. Otherwise, the tests are "incomplete." Tests that are incomplete but attempted (Attempt = Y) are scored. In order to receive a discipline score (i.e., Life Sciences, Physical Sciences, and Earth and Space Sciences), a student must have attempted the corresponding discipline of the test. The MMLE is used to score the attempted incomplete tests, counting unanswered items as incorrect. If the unanswered items are unknown due to the test being assembled on the fly, the item parameters for a typical item are used. If a missing item is an item cluster, the simulated item parameters of the missing item are the item parameters of item cluster 21910 for grade 5 and 22081 for grade 8, which are operational item clusters that are typical for the WVGSA item pool used in West Virginia in terms of the number of assertions and estimated parameters. Likewise, if a missing item is a stand-alone item, the simulated item parameters of the missing item are the item parameters of stand-alone item 22068 for grade 5 and 21830 for grade 8, which are operational stand-alone items that are typical for the WVGSA item pool used in West Virginia.

If the identities of items that have not been answered to are known because they have already been lined up through the pre-fetch process, the item parameters of the lined-up items will be used. Similarly, for the accommodated forms that are fixed forms, the item parameters of the unanswered items on the form will be used.

6.2.6 Student-Level Scale Scores

At the student level, scale scores are computed for

1. Overall Science
2. Life Sciences
3. Physical Sciences
4. Earth and Space Sciences

Scores are computed using the MMLE method outlined, with all items from overall science or only items within the given discipline. Scores are truncated on the “theta” scale at the LOT and HOT values specified in Table 74, which correspond to values of the estimated mean plus or minus four times of the estimated standard deviation of θ.

The reporting scales are linear transformations of the theta scales

$$\text{SS} = \text{a}*\hat{\theta}_{MMLE} + b,$$

where a and b are the slope and intercept of the linear transformation that transforms $\hat{\theta}_{MMLE}$ to the reporting scale (refer to Table 74). The standard error of estimate for the estimated scale score is obtained as

$$SEM_{SS} = a * SEM_{\hat{\theta}_{MMLE}}.$$

In 2018, the slope a and intercept b were chosen so that the center of the reporting scale of each grade (550 and 850 respectively) is centered at the grade mean of the 2018 base year and has a standard deviation of 12.5. Furthermore, for each grade, the reporting scale ranges from the base-year mean minus four times the standard deviation to the base-year mean plus four times the standard deviation. Specifically, for grade 5, the slope and intercept were obtained as

$$SS = 12.5\theta^* + 550$$

$$= 12.5\frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta} + 550$$
$$= \frac{12.5}{\hat{\sigma}_\theta}\theta + \left(550 - \frac{12.5\hat{\mu}_\theta}{\hat{\sigma}_\theta}\right),$$

where the second line stems from standardizing theta, $\theta^* = \frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta}$. For grade 8, the slope and intercept can also be derived in a similar fashion.

Per grade, Table 74 presents the intercept, slope, LOT, HOT, LOSS, and HOSS values that were used for the 2018 and 2019 reporting scale.

As explained in Section 5.2.2, Item Calibration, the item bank was recalibrated in 2019 and the 2019 item parameter and θ scale will be the underlying scale going forward. Because $\theta_{j\ 2019} = \theta_{j\ 2018} + B$, the reporting scale is linear transformation of the 2019 scale, with the slope and intercept updated as follows:

$$SS = a * \hat{\theta}_{MMLE,2018} + b_{2018}$$
$$= a * \left(\hat{\theta}_{MMLE,2019} - B\right) + b_{2018} = a * \hat{\theta}_{MMLE,2019} + b_{2019},$$

with b_{2019}=$b_{2018} - a * B$. Table 75 represents the updated slope and intercept for the linear transformation of the 2019 θ scale. Because the LOT and HOT are specified to correspond to values of the estimated mean minus/plus four times the estimated standard deviation of θ, they are updated, as well. The updated linear transformation ensures that the scales remain comparable across years.

Table 74. Science Reporting Scale Linear Transformation Constants & Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2018 θ scale)

Grade	Slope (*a*)	Intercept (*b*)	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
5	16.089	551.740	-3.21	2.99	500	600
8	17.180	852.600	-3.06	2.75	800	900

Table 75. Science Reporting Scale Linear Transformation Constants & Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2019 θ scale)

Grade	Slope (*a*)	Intercept (*b*)	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
5	16.089	552.410	-3.25	2.95	500	600
8	17.180	853.019	-3.08	2.73	800	900

6.2.7 Rules for Calculating Achievement Levels

Achievement levels and corresponding cut scores were set during standard setting in the summer of 2018. Students are classified into one of four achievement levels, based on their total score.

Table 76 contains the cut scores on the reporting scale metrics for each of the grades.

Table 76. Achievement-Level Cut Scores

Grade	Cut 1	Cut 2	Cut 3
5	537	555	568
8	837	855	867

6.2.7.1 Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score

Discipline-level classifications are computed to classify student achievement levels for each of the science disciplines. The classification rules are:

- If ($\hat{\theta}_{discipline} < \theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{discipline}\right)$), then achievement is classified as *Below Mastery*.
- If ($\theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{discipline}\right) \leq \hat{\theta}_{discipline} < \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{discipline}\right)$), then achievement is classified as *At/Near Mastery*.
- If ($\hat{\theta}_{discipline} \geq \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{discipline}\right)$), then achievement is classified as *Above Mastery*.

Where $\theta_{Proficient}$ is the proficiency cut score of the overall test. Standard errors are truncated at 1.4. The LOT is always classified as *Below Mastery*, and the HOT is always classified as *Above Mastery*.

6.2.8 Disciplinary Core Ideas and Performance Expectation-Level Reporting

6.2.8.1 Relative to Overall Achievement

For aggregated units (classrooms, schools, and districts), there is reporting at levels below the science discipline level. In 2017–2018, reports were provided at the level of Disciplinary Core Ideas (DCI). In 2018–2019, reports were provided at the level of both DCI and Performance Expectations (PE). The method for reporting at levels below the science discipline level is based on the use of residuals. The equations are presented first for DCIs. The underlying method for PEs is the same, but the residuals for an individual student were aggregated within a PE rather than within a DCI. Also, the final reported measure for the PE was an estimated percentage correct, which is discussed later in this section.

For each assertion *i*, the residual between observed and expected score for each student *j* is defined as

$$\delta_{ij} = z_{ij} - E\left(z_{ij}\right).$$

The expected score is computed for a student's estimated overall ability. For the assertions clustered within an item, the expected score is marginalized over the nuisance dimensions for the assertions clustered within an item,

$$E(z_{ijg} = 1; \theta_{j,overall}, \boldsymbol{\tau}_i) = \int P(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \boldsymbol{\tau}_i) N(u_{jg}) du_{jg},$$

where $\boldsymbol{\tau}_i$ is the vector of parameters for assertion *i* (e.g., for the Rasch testlet model, $\boldsymbol{\tau}_i = b_i$), and $P(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \boldsymbol{\tau}_i)$ is defined in Section 6.2.2, Derivatives. Next, residuals are aggregated over assertions within each student,

$$\delta_{jDCI} = \frac{\sum_{i \in DCI} \delta_{ij}}{n_{jDCI}},$$

and over students of the group on which is reported,

$$\bar{\delta}_{DCIm} = \frac{1}{n_m} \sum_{j \in m} \delta_{jDCI},$$

where n_{jDCI} is the number of assertions related to the DCI for student j, and n_m is the number of students in a group assessed on the DCI. If a student did not see any items on a DCI, the student is not included in the n_m count for the aggregate. The standard error of the average residual is computed as

$$SEM(\bar{\delta}_{DCIm}) = \sqrt{\frac{1}{n_m(n_m - 1)} \sum_{j \in m} (\delta_{jDCI} - \bar{\delta}_{DCIm})^2}.$$

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{DCIm}$is positive) or less effective (negative $\bar{\delta}_{DCIm}$) in teaching a given DCI or PE.

We do not suggest the direct reporting of the statistic $\bar{\delta}_{DCIm}$; instead, we recommend reporting whether, in the aggregate, a group of students perform better, worse, or as expected on this DCI or PE. In some cases, sufficient information is not available and that is indicated, as well.

For DCI-level strengths/weakness, the following is reported:

- If $\bar{\delta}_{DCIm} \leq -1.5 * SEM(\bar{\delta}_{DCIm})$, then achievement is worse than on the overall test.
- If $\bar{\delta}_{DCIm} \geq 1.5 * SEM(\bar{\delta}_{DCIm})$, then achievement is better than on the overall test.
- Otherwise, achievement is similar to the overall test.
- If $SEM(\bar{\delta}_{DCIm})$ > 0.2, data are insufficient.

6.2.8.2 Relative to Proficiency Cut Score

DCI- and PE-level scores for aggregated units can be computed using the same method as outlined in the previous section, but with the expected score computed at the theta value corresponding to the proficiency cut score:

$$E\left(z_{ijg} = 1; \theta_{proficiency}, \boldsymbol{\tau}_i\right) = \int P\left(z_{ijg} = 1 | u_{jg};\ \theta_{proficiency}, \boldsymbol{\tau}_i\right) N\left(u_{jg}\right) du_{jg}.$$

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIm} \leq -\ 1.5 * SEM\left(\bar{\delta}_{DCIm}\right)$, then achievement is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIm} \geq 1.5 *\ SEM\left(\bar{\delta}_{DCIm}\right)$, then achievement is *above* the proficiency cut score.
- Otherwise, achievement is *near* the proficiency cut score.
- If $SEM\left(\bar{\delta}_{DCIm}\right)$ > 0.2, data are insufficient.

6.2.8.3 Performance Expectations: Estimated Percentage Correct

There are two differences between how scores are reported for PEs and DCIs. First, for PEs, only weaknesses (achievement is worse than on the overall test) are flagged rather than both weaknesses and strengths. Second, rather than directly reporting the relative performance on a PE, the performance is expressed on a scale of 0 to 100, like an estimated percentage correct. The measure is computed as described in the next paragraph.

In the first step, average residuals $\bar{\delta}_{PEm}$ are computed for PEs using the same residual-based method that is used for DCIs. The same rules as for DCIs are used to indicate whether, in the aggregate, a group of students performs better, worse, or as expected on a PE:

- If $\bar{\delta}_{PEm} \leq -1.5 * SEM\left(\bar{\delta}_{PEm}\right)$, then achievement is worse than on the overall test.
- If $\bar{\delta}_{PEm} \geq 1.5 * SEM(\bar{\delta}_{PEm})$, then achievement is better than on the overall test.
- Otherwise, achievement is similar to the overall test.
- If $SEM(\bar{\delta}_{PEm})$ > 0.2, data are insufficient.

In the second step, a reference cluster is selected for each grade. The reference clusters are operational clusters that are typical for the item bank used in West Virginia in terms of the number of assertions and estimated parameters. The reference clusters are item cluster 21910 for grade 5 and item cluster 22081 for grade 8. The expected score is computed for a student's estimated overall ability for each of the assertions of the reference cluster, and then averaged across the assertions,

$$E_{jref} = \frac{\sum_{i \in ref} E\left(z_{ijref} = 1; \theta_{j,overall}, \boldsymbol{\tau}_i\right)}{n_{AS_{ref}}},$$

with $E\left(z_{ijref} = 1; \theta_{j,overall}, \boldsymbol{\tau}_i\right) = \int P\left(z_{ijref} = 1 | u_{jref};\ \theta_{j,overall}, \boldsymbol{\tau}_i\right) N\left(u_{jref}\right) du_{jref}$,

and $n_{AS_{ref}}$ is the number of assertions in the reference cluster.

Based on students in the group where $\bar{\delta}_{PEm}$ is calculated, one can define the mean expected correct value as

$$\bar{E}_{PEm} = \frac{1}{N_m} \sum_{j \in m} E_{jref},$$

where N_m is all the students in the group.

The reported score is defined as:

$$\bar{P}_{PEm} = \begin{cases} 1 \text{ if } 100\left(\bar{E}_{PEm} + \bar{\delta}_{PEm}\right) < 1 \\ 99 \text{ if } 100\left(\bar{E}_{PEm} + \bar{\delta}_{PEm}\right) > 99 \\ 100\left(\bar{E}_{PEm} + \bar{\delta}_{PEm}\right) \quad \text{O.W.} \end{cases}$$

When reporting, numbers are rounded to whole numbers. For a given group, if the PE score for a given standard is significantly lower than the expected PE score based on the overall scores of the students in the group, this $\bar{P}_{PEm}$ has a flag indicating a statistically significant weakness. Otherwise, only the value of $\bar{P}_{PEm}$ is reported. If the data are insufficient or no data are available for a given PE, “NA” is reported.

7. QUALITY CONTROL PROCEDURES

Cambium Assessment, Inc’s (CAI) quality assurance (QA) procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

Although the quality of any test is monitored as an ongoing activity, several sources of CAI’s quality control system are described here. First, QA reports are routinely generated and evaluated throughout the testing window to ensure that each test is performing as anticipated. Second, the quality of scores is ensured by employing a second independent scoring verification system.

7.1 QUALITY ASSURANCE REPORTS

Test monitoring occurs while tests are administered in a live environment to ensure that item behavior is consistent with expectations. This is accomplished using CAI’s quality monitoring system that yields item statistics, blueprint match rates, and item exposure rate reports. Table 77 provides the summary of indicators generated from each QA report.

Table 77. Overview of Quality Assurance Reports

QA Report	Purpose	Rationale
*Item Statistics**	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology-enhanced items)
Blueprint Match Rates	To monitor unexpected low blueprint match rates	Early detection of unexpected blueprint match issues

QA Report	Purpose	Rationale
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

**No item statistics report for science*

7.1.1 Item Analysis

The item analysis report is a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine the performance of test items, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as item fit statistics based on the item response theory (IRT). The report is configurable and can be produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. The criteria for flagging and reviewing English language arts (ELA) and mathematics items are provided in Table 78, and a description of the statistics is provided later in this section. For science, statistics reports at the assertion level (which are the units of analysis for science) are not yet available; however, as a routine and continuing practice, our psychometricians compute and monitor classical item statistics at the end of each testing window.

Table 78. Thresholds for Flagging Items in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Point biserial correlation for the correct response is < 0.10.
Distractor Analysis	Point biserial correlation for any distractor response is > 0.
Item Difficulty	The proportion of students (*p*-value) is 0 or 1.

7.1.1.1 Item Discrimination

As described in Section 4.1, Item Discrimination, the item discrimination index indicates the extent to which each item differentiated between those test takers who possessed the skills being measured and those who did not. Most of the operational items had a higher-point biserial correlation than the flagging criteria. Fewer than 2% of the operational items were flagged by low-point biserial for both ELA and mathematics. Items with low-point biserial correlations were reviewed by CAI content experts, and all items behaved as expected.

7.1.1.2 Item Difficulty

Items that are either extremely difficult or extremely easy are flagged for review but are not necessarily removed if they are grade-level appropriate and align with test specifications. For further detail, refer back to Section 4.2, Item Difficulty. Most of the operational items had *p*-values within the expected range, but one item across all test grades and subjects was flagged for a *p*-value of 0. CAI content experts and psychometricians verified that this item behaved as expected and was scored correctly.

7.1.1.3 Distractor Analysis

Discussed in Section 4.3, ELA and Mathematics Distractor Analysis, distractor analysis for multiple-choice items is used to identify items that may have had marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracted high-scoring students. Most operational items had a negative distractor. CAI content experts reviewed items with positive distractor correlations and did not find any issue.

7.1.2 Blueprint Match

The QA system generates blueprint match reports at the content standards level and for other content requirements such as strand or Depth of Knowledge (DOK) level for ELA and mathematics, or strand and affinity group for science. For each blueprint element, the report indicates the minimum and maximum number of items specified in the blueprint, the number of test administrations in which those specifications were met, the number of administrations in which the blueprint requirements were not met, and, for administrations in which specifications were not met, the number of items by which the requirement was not met.

While simulation results described in Appendix A, Simulation Summary Report, indicated that the configuration resulted in test administrations meeting all blueprint match requirements, it is important to evaluate the blueprint match rate for actual test administrations. Appendix B, Simulation vs. Operational Blueprint Match, shows the detailed comparison for simulation and operational blueprint match for ELA and mathematics.

Across grades in ELA and mathematics, all tests, except ELA grade 7 and grade 8, met the blueprint specifications with a 100% match at the reporting category level. ELA grade 7 and grade 8 included a few exceptions, as a small number of students took the test for the same grade in both 2018 and 2019. The Test Delivery System (TDS) prevents administration of any item more than once to the same student, resulting in slight blueprint mismatch for certain reporting categories.

For science, blueprint match is discussed in detail in Volume 2, Test Development, Part 2, for both simulated and operational test administrations.

7.1.3 Item Exposure Rates

The QA system also generates Item Exposure reports that allow test items to be monitored for unexpectedly large exposure rates or unusually low item-pool usage throughout the testing window. As with other reports, it is possible to examine the exposure rate for all items or flag items with exposure rates that exceed an acceptable range. Often, item overexposure indicates a blueprint element or combination of blueprint elements that are underrepresented in the item pool, and which should be targeted for future item development. Such item overexposure is also usually anticipated in the simulation studies used to configure the adaptive algorithm.

As is consistent with the simulation results described in Appendix A, Simulation Summary Report, most test items were administered to 20% or fewer test takers across all grades and subjects. Appendix G, Operational Item Exposure, shows the item exposure rates for the operational test administrations for ELA and mathematics. Similarly, for science, most of the test items were administered to 20% or fewer test takers in both grades. More details are discussed in Volume 2, Test Development.

7.1.4 Cheating Detection Analysis

As part of the QA procedures, a forensics report can also be provided to identify possible irregularities in test administration for further investigation. Unusual patterns of responding at the student level can be aggregated to the test session, test administrator, and school levels to identify possible group-level testing anomalies. CAI psychometricians can monitor testing anomalies throughout the testing window. Evidence can be evaluated with respect to item response times and irregular item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. The analyses used to detect the testing anomalies can be run anytime within the testing window.

7.2 SCORE QUALITY CHECK

All student test scores are produced using CAI’s scoring engine. Prior to releasing any scores, a second score verification system is used to verify that all test scores match with 100% agreement in all tested grades. This second system is independently constructed and maintained from the main scoring engine and separately estimates MLEs for ELA and mathematics and MMLEs for science, using the procedures described within this report.

8. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: Author.

Bock, R. D., & Zimowski, M. F. (1997). Multiple Group IRT. In: van der Linden W. J., Hambleton R.K. (Eds.) Handbook of Modern Item Response Theory. Springer, New York, NY.

Cai, L. (2017). flexMIRT® version 3.51: Flexible multilevel multidimensional item analysis and test scoring (computer software). Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows (Computer software). Lincolnwood, IL: Scientific Software International.

Chen, W., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22(3*), 265–289. Retrieved from http://www.jstor.org/stable/1165285.

Council of Chief State School Officers (2020). Restart and Recovery: Accountability interrupted: Guidance for collecting, evaluating, and reporting data in 2020–2021. Washington, DC: Author.

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

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

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

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Rijmen, F. (2006). BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes (Technical Report). Amsterdam, the Netherlands: VU University Medical Center.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the BiFactor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*(3), 361–372.

Rijmen, F., Liao, D., & Lin, Z. (2021). *The Rasch testlet model for the calibration of three-dimensional science assessments: A software comparison* [White paper]. Washington, DC Cambium Assessment, Inc.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory*. New York: Springer-Verlag.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement* (ETS Research Report No. 12-08). Princeton, NJ: Educational Testing Service.

Appendix A

Simulation Summary Report

INTRODUCTION

This document describes the results of simulated test administrations used to configure and evaluate the adequacy of the item selection algorithm used to administer the WVGSA 2021-2022 assessments. The purpose of the simulations is to configure the adaptive algorithm to optimize item selection to both meet blueprint specifications while targeting test information to student ability. When the adaptive algorithm is optimized, the observed score is measured more precisely than would otherwise be possible in a fixed-form environment, especially for high- and low-performing students. Consequently, the test administrations (forms) generated by the adaptive algorithm will not and should not be statistically parallel. Nevertheless, scores from the assessment should be comparable, and each test form should measure the same content, albeit with a different set of test items.

Test administrations were simulated separately for the following tests:

1. ELA grades 3-8
2. Mathematics Grades 3-8

TESTING PLAN

This report summarizes the results of the test item selection algorithm properties and resulting test simulations.

The testing plan begins by generating a sample of examinees from a Normal (μ, σ) distribution for each grade and subject. The parameters for the normal distribution were based on operational test scores obtained from administration in the previous years.

STATISTICAL SUMMARIES

Some of the tables in this document provide statistical summaries of the data by grade and by subject. The statistics computed include the statistical bias of the estimated theta parameter; mean squared error (MSE); significance of the bias; average standard error of the estimated theta; the standard error at the 5th, 25th, 75th, and 95th percentiles; and the percentage of students falling inside the 95% and 99% confidence intervals.

Statistical bias refers to whether test scores systematically underestimate or overestimate the student’s true ability and is distinguished from differential item functioning analyses which are used to detect “bias” or unfairness in the performance of test items across subgroups.

Computational details of each statistic are provided below.

$$bias = N^{-1}\sum_{i=1}^{N}(\theta - \hat{\theta}) \quad ,$$
$$MSE = N^{-1}\sum_{i=1}^{N}(\theta - \hat{\theta})^2 \quad , \qquad (1)$$

where θ is the true ability and $\hat{\theta}$ is its estimate. For the variance of the bias, we use a first-order Taylor series of Equation (1) as:

$$var(bias) = \sigma^2 * g'(\hat{\theta})^2$$
$$= \frac{1}{N(N-1)}\sum_{i=1}^{N}(\theta_i - \bar{\hat{\theta}})^2 \quad . \qquad (2)$$

Significance of the bias is then tested as:

$$z = \frac{bias}{\sqrt{var(bias)}} \quad . \qquad (3)$$

A *p*-value for the significance of the bias is reported from this *z* test. The average standard error is computed as:

$$mean(se) = \sqrt{N^{-1}\sum_{i=1}^{N} se_i^2} \quad , \qquad (4)$$

where se_i^2 is the standard error of the estimate, $\hat{\theta}$ for individual *i*.

To determine the number of students falling outside the 95% and 99% confidence interval coverage, a t-test is performed:

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} \quad . \qquad (5)$$

where $\hat{\theta}_i$ is the ability estimate for individual *i* and θ_i is the true ability for individual *i*. The percentage of students falling outside the coverage is determined by comparing the absolute value of the *t*-statistic to a critical value of 1.96 for the 95% coverage and to 2.58 for the 99% coverage.

TEST BLUEPRINTS

The adaptive item selection algorithm must administer each student a unique test that adheres to the content requirements described in the WVGSA test specifications, ensuring a comparable and sufficient coverage of the content of the West Virginia content Standards.

The tables in Appendix A provide a detailed summary of the blueprint configuration used in the simulations. The tables include the minimum and maximum items to be delivered for a given content area, as well as whether a strict maximum was imposed, indicating that the constraint is required to be met exactly (TRUE = imposition of a strict maximum).

FACTORS AFFECTING SIMULATION RESULTS

There are a number of factors that may influence simulation results for an adaptive test administration. These include:

1. *The proportional relationship between the pool and the constraints to be met.* Proportionally distributed pools tend to make better use of the pool (i.e., more uniform item exposure) and make it easier to meet blueprint and other constraints. For example, if the specifications call for 50% of the items to be technology enhanced (TE) items, but the pool only contains 6% TE items, it may be difficult to meet this constraint.
2. *The correlational structure between constraints*. It is easier to satisfy a constraint if there are instances of the constraint at all levels of another constraint. For example, if DOK3 items are only associated with a specific content area, it may be difficult to meet both the desired distribution of content and the desired distribution of DOK.
3. *Whether or not there is a "strict maximum" on a given constraint*. This means that the requirement must be met exactly in each test administration.

RESULTS OF SIMULATED TEST ADMINISTRATIONS

Simulations were evaluated for all content areas using 1,000 simulated cases.

SUMMARY OF STATISTICAL ANALYSES

Each simulated record includes a true score and an ability estimate based on the adaptive test administration. Table 1 shows the correlations between the true score and estimated ability for each of the WVGSA assessments. As Table 1 shows, correlations between true and estimated ability are nearly one, indicating that the adaptive test administrations reliably estimate student ability.

Table 1. Correlations between True and Estimated Ability by Subject and Grade

Subject	Grade	Correlation
English Language Arts	3	0.96
	4	0.95
	5	0.96
	6	0.95
	7	0.96
	8	0.96
Mathematics	3	0.97
	4	0.97
	5	0.96
	6	0.94
	7	0.95
	8	0.93

Table 2 presents the mean of the biases, which is the average of the biases of the estimated abilities across all students, the *p*-value for the significance of the estimated bias reported from the z-test and mean square error (MSE) of the estimated theta by subject and grade. In most cases, the mean bias of the estimated abilities is very small and statistically insignificant, providing further evidence that the true score is adequately recovered in the estimated score. There are instances, however, where the bias is statistically significant, especially, in the upper grade math assessments, where the distribution of item difficulties is substantially greater than the distribution of student abilities. On average, when the distribution of item difficulties is greater than the distribution of student abilities, the student abilities are somewhat underestimated, especially for low ability students; when the distribution of item difficulties is lower than the student abilities, the student abilities are somewhat overestimated, especially for high ability students.

Table 2. Statistical Summaries of Ability Estimation – Bias of the Estimated Abilities by Subject and Grade

Subject	Grade	Mean of the biases	*P*-value for the T-Test	MSE
ELA	3	0.026	0.009	0.100
	4	0.027	0.011	0.114
	5	0.015	0.136	0.101
	6	0.023	0.034	0.118
	7	0.030	0.003	0.102
	8	-0.002	0.843	0.102
Mathematics	3	0.001	0.892	0.054
	4	0.026	0.001	0.061
	5	0.021	0.031	0.095
	6	0.021	0.069	0.133
	7	0.034	0.003	0.129
	8	0.036	0.006	0.170

Table 3 shows the mean standard errors of the ability estimate across the 1,000 simulated test administrations, as well as the standard error across the ability distribution. As the table indicates, in most of the tests, the standard error is highest at the very low end of the ability spectrum, and relatively lower in the middle range of the ability distribution.

Table 3. Statistical Summaries of Ability Estimation – Standard Errors of the Estimated Abilities by Subject and Grade

Grade	Average SE	5th PR of SE	25th PR of SE	50th PR of SE	75th PR of SE	95th PR of SE
				ELA		
3	0.357	0.202	0.213	0.229	0.288	0.685
4	0.337	0.226	0.247	0.265	0.306	0.557
5	0.311	0.208	0.229	0.251	0.297	0.485
6	0.335	0.233	0.248	0.269	0.329	0.551
7	0.331	0.242	0.257	0.274	0.312	0.503
8	0.313	0.243	0.262	0.281	0.313	0.433

Grade	Average SE	5th PR of SE	25th PR of SE	50th PR of SE	75th PR of SE	95th PR of SE
				Mathematics		
3	0.273	0.14	0.148	0.164	0.215	0.595
4	0.281	0.167	0.178	0.194	0.230	0.523
5	0.321	0.205	0.220	0.236	0.281	0.553
6	0.364	0.225	0.255	0.287	0.339	0.608
7	0.397	0.244	0.261	0.282	0.351	0.640
8	0.428	0.296	0.330	0.374	0.434	0.627

The summary statistics of the estimated abilities show that the item selection algorithm is generally choosing items that are conditional on each examinee’s ability, where available. This is limited in the case of ELA by selection of item groups for passages and other stimulus based items, and by relatively difficulty of the upper grade mathematics item banks relative to student ability. Given that we know the true ability for each examinee is known in a simulation, these data show that the true ability is almost always recovered—an indication that the algorithm is working as expected for a computer-adaptive test.

GLOBAL ITEM EXPOSURE

The simulator output also reports the degree to which the constraints set forth in the blueprints may yield greater exposure of items to students. This is reported by examining the percentage of test administrations in which an item appears. For instance, in a fixed paper form, 100% of the items appear on 100% of the test administrations because every examinee sees the same items. In an adaptive test with a sufficiently large item pool, we would expect that most of the items would appear on only a relatively small percentage of the test administrations.

When this condition holds, it suggests that test administrations between students are more or less unique. Therefore, we calculated the item exposure rate for each item across by dividing the total number of test administrations in which an item appears by the total number of tests administered. Then we report the distribution of the item exposure rate (r) in six bins. The bins are r=0% (unused), 0%<r<=1%, 1%<r<=5%, 5%<r<=20%, 20%<r<=40%, 40%<r<=60%, 60%<r<=80%, and 80%<r<=100%. If global item exposure is minimal, we would expect the largest proportion of items to appear in the bins of 0%<r<=20%, an indication that most of the items appear on a very small percentage of the test forms.

Table 4 presents the percentage of items that fall into each exposure bin for all grades. As expected, most test items are administered and they are administered in 20% or fewer test administrations.

Table 4. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Test Administrations

Grade	Total Items	[0,0]%	(1,20]%	(20,40]%	(40,60]%	(60,80]%	(80,100]%
				ELA			
3	430	7.53	79.78	9.68	3.01	0.00	0.00
4	398	11.75	72.73	13.75	1.77	0.00	0.00
5	434	2.91	85.01	8.95	3.13	0.00	0.00
6	488	6.51	83.33	7.47	2.68	0.00	0.00
7	409	8.30	80.49	9.19	1.35	0.67	0.00
8	354	8.53	70.28	17.05	2.58	1.55	0.00
				Mathematics			
3	619	5.21	92.65	2.14	0.00	0.00	0.00
4	647	6.37	91.90	1.74	0.00	0.00	0.00
5	548	2.32	92.87	4.81	0.00	0.00	0.00
6	666	1.33	94.37	4.30	0.00	0.00	0.00
7	474	4.82	83.53	9.44	2.21	0.00	0.00
8	551	2.65	89.40	7.77	0.18	0.00	0.00

APPENDIX A – SIMULATION TEST BLUEPRINT FOR WVGSA

Test Blueprint for Grade 3 ELA

ContentLevelID	MinItems	MaxItems	isStrictMax
IT|3.CS	4	6	FALSE
IT|3.CS|ELA.3.10	0	3	FALSE
IT|3.CS|ELA.3.11	0	3	FALSE
IT|3.CS|ELA.3.12	0	3	FALSE
IT|3.IKI	1	3	FALSE
IT|3.IKI|ELA.3.15	0	2	FALSE
IT|3.IKI|ELA.3.16	0	2	FALSE
IT|3.IKI|ELA.3.17	1	2	FALSE
IT|3.KID	5	7	FALSE
IT|3.KID|ELA.3.4	0	3	FALSE
IT|3.KID|ELA.3.5	0	3	FALSE
IT|3.KID|ELA.3.6	0	3	FALSE
L|3.CSE	6	8	FALSE
L|3.CSE|ELA.3.36a	0	5	FALSE
L|3.CSE|ELA.3.36b	0	5	FALSE
L|3.CSE|ELA.3.36d	0	5	FALSE
L|3.CSE|ELA.3.36e	0	5	FALSE
L|3.CSE|ELA.3.36f	0	5	FALSE
L|3.CSE|ELA.3.36g	0	5	FALSE
L|3.CSE|ELA.3.36h	0	5	FALSE
L|3.CSE|ELA.3.36i	0	5	FALSE
L|3.CSE|ELA.3.37a	0	5	FALSE
L|3.CSE|ELA.3.37c	0	5	FALSE
L|3.CSE|ELA.3.37d	0	5	FALSE
L|3.CSE|ELA.3.37e	0	5	FALSE
L|3.CSE|ELA.3.37f	0	5	FALSE
L|3.KL	0	0	FALSE
L|3.KL|ELA.3.38a	0	0	FALSE
L|3.VAU	1	2	TRUE
L|3.VAU|ELA.3.39a	0	2	FALSE
L|3.VAU|ELA.3.39b	0	2	FALSE
L|3.VAU|ELA.3.40a	0	2	FALSE
L|3.VAU|ELA.3.40b	0	2	FALSE
L|3.VAU|ELA.3.40c	0	2	FALSE
LT|3.CS	6	8	FALSE

Test Blueprint for Grade 3 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
LT|3.CS|ELA.3.7	0	3	FALSE
LT|3.CS|ELA.3.8	0	3	FALSE
LT|3.CS|ELA.3.9	0	3	FALSE
LT|3.IKI	1	3	FALSE
LT|3.IKI|ELA.3.13	0	2	FALSE
LT|3.IKI|ELA.3.14	1	2	FALSE
LT|3.KID	6	8	FALSE
LT|3.KID|ELA.3.1	0	3	FALSE
LT|3.KID|ELA.3.2	0	3	FALSE
LT|3.KID|ELA.3.3	0	3	FALSE
SL|3.CaC	0	3	FALSE
SL|3.CaC|ELA.3.31	0	2	FALSE
SL|3.CaC|ELA.3.32	0	2	FALSE
W|3.TTP	1	1	FALSE
W|3.TTP|ELA.3.20a	0	1	FALSE
W|3.TTP|ELA.3.21a	0	1	FALSE

Test Blueprint for Grade 4 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
IT|4.CS	4	6	FALSE
IT|4.CS|ELA.4.10	0	3	FALSE
IT|4.CS|ELA.4.11	0	3	FALSE
IT|4.CS|ELA.4.12	0	3	FALSE
IT|4.IKI	1	3	FALSE
IT|4.IKI|ELA.4.15	0	2	FALSE
IT|4.IKI|ELA.4.16	0	2	FALSE
IT|4.IKI|ELA.4.17	1	2	FALSE
IT|4.KID	5	7	FALSE
IT|4.KID|ELA.4.4	0	3	FALSE
IT|4.KID|ELA.4.5	0	3	FALSE
IT|4.KID|ELA.4.6	0	3	FALSE
L|4.CSE	6	8	FALSE
L|4.CSE|ELA.4.36a	0	5	FALSE
L|4.CSE|ELA.4.36b	0	5	FALSE
L|4.CSE|ELA.4.36c	0	5	FALSE
L|4.CSE|ELA.4.36d	0	5	FALSE
L|4.CSE|ELA.4.36e	0	5	FALSE

Test Blueprint for Grade 4 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
L|4.CSE|ELA.4.36f	0	5	FALSE
L|4.CSE|ELA.4.36g	0	5	FALSE
L|4.CSE|ELA.4.37a	0	5	FALSE
L|4.CSE|ELA.4.37b	0	5	FALSE
L|4.CSE|ELA.4.37c	0	5	FALSE
L|4.CSE|ELA.4.37d	0	5	FALSE
L|4.VAU	1	2	TRUE
L|4.VAU|ELA.4.39a	0	2	FALSE
L|4.VAU|ELA.4.39b	0	2	FALSE
L|4.VAU|ELA.4.40a	0	2	FALSE
L|4.VAU|ELA.4.40b	0	2	FALSE
L|4.VAU|ELA.4.40c	0	2	FALSE
LT|4.CS	6	8	FALSE
LT|4.CS|ELA.4.7	0	3	FALSE
LT|4.CS|ELA.4.8	0	3	FALSE
LT|4.CS|ELA.4.9	0	3	FALSE
LT|4.IKI	1	3	FALSE
LT|4.IKI|ELA.4.13	0	2	FALSE
LT|4.IKI|ELA.4.14	1	2	FALSE
LT|4.KID	6	8	FALSE
LT|4.KID|ELA.4.1	0	3	FALSE
LT|4.KID|ELA.4.2	0	3	FALSE
LT|4.KID|ELA.4.3	0	3	FALSE
SL|4.CaC	0	3	FALSE
SL|4.CaC|ELA.4.31	0	2	FALSE
SL|4.CaC|ELA.4.32	0	2	FALSE
W|4.TTP	1	1	FALSE
W|4.TTP|ELA.4.20a	0	1	FALSE
W|4.TTP|ELA.4.21a	0	1	FALSE

Test Blueprint for Grade 5 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
IT|5.CS	4	6	FALSE
IT|5.CS|ELA.5.10	0	3	FALSE
IT|5.CS|ELA.5.11	0	3	FALSE
IT|5.CS|ELA.5.12	0	3	FALSE
IT|5.IKI	1	3	FALSE

Test Blueprint for Grade 5 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
IT|5.IKI|ELA.5.15	0	2	FALSE
IT|5.IKI|ELA.5.16	0	2	FALSE
IT|5.IKI|ELA.5.17	1	2	FALSE
IT|5.KID	5	7	FALSE
IT|5.KID|ELA.5.4	0	3	FALSE
IT|5.KID|ELA.5.5	0	3	FALSE
IT|5.KID|ELA.5.6	0	3	FALSE
L|5.CSE	6	8	TRUE
L|5.CSE|ELA.5.36a	0	5	FALSE
L|5.CSE|ELA.5.36b	0	5	FALSE
L|5.CSE|ELA.5.36c	0	5	FALSE
L|5.CSE|ELA.5.36d	0	5	FALSE
L|5.CSE|ELA.5.36e	0	5	FALSE
L|5.CSE|ELA.5.37a	0	5	FALSE
L|5.CSE|ELA.5.37b	0	5	FALSE
L|5.CSE|ELA.5.37c	0	5	FALSE
L|5.CSE|ELA.5.37e	0	5	FALSE
L|5.VAU	1	2	TRUE
L|5.VAU|ELA.5.39a	0	2	FALSE
L|5.VAU|ELA.5.39b	0	2	FALSE
L|5.VAU|ELA.5.39c	0	2	FALSE
L|5.VAU|ELA.5.40a	0	2	FALSE
L|5.VAU|ELA.5.40b	0	2	FALSE
L|5.VAU|ELA.5.40c	0	2	FALSE
LT|5.CS	6	8	FALSE
LT|5.CS|ELA.5.7	0	3	FALSE
LT|5.CS|ELA.5.8	0	3	FALSE
LT|5.CS|ELA.5.9	0	3	FALSE
LT|5.IKI	1	3	FALSE
LT|5.IKI|ELA.5.13	0	2	FALSE
LT|5.IKI|ELA.5.14	1	2	FALSE
LT|5.KID	6	8	FALSE
LT|5.KID|ELA.5.1	0	3	FALSE
LT|5.KID|ELA.5.2	0	3	FALSE
LT|5.KID|ELA.5.3	0	3	FALSE
SL|5.CaC	0	3	FALSE
SL|5.CaC|ELA.5.31	0	2	FALSE
SL|5.CaC|ELA.5.32	0	2	FALSE

Test Blueprint for Grade 5 ELA

ContentLevelID	MinItems	MaxItems	isStrictMax
W|5.TTP	1	1	FALSE
W|5.TTP|ELA.5.20a	0	1	FALSE
W|5.TTP|ELA.5.21a	0	1	FALSE

Test Blueprint for Grade 6 ELA

ContentLevelID	MinItems	MaxItems	isStrictMax
IT|6.CS	6	8	FALSE
IT|6.CS|ELA.6.10	0	3	FALSE
IT|6.CS|ELA.6.11	0	3	FALSE
IT|6.CS|ELA.6.12	0	3	FALSE
IT|6.IKI	1	3	FALSE
IT|6.IKI|ELA.6.15	0	2	FALSE
IT|6.IKI|ELA.6.16	0	2	FALSE
IT|6.IKI|ELA.6.17	1	2	FALSE
IT|6.KID	6	8	FALSE
IT|6.KID|ELA.6.4	0	3	FALSE
IT|6.KID|ELA.6.5	0	3	FALSE
IT|6.KID|ELA.6.6	0	3	FALSE
L|6.CSE	0	8	FALSE
L|6.CSE|ELA.6.36a	0	5	FALSE
L|6.CSE|ELA.6.36b	0	5	FALSE
L|6.CSE|ELA.6.36c	0	5	FALSE
L|6.CSE|ELA.6.36d	0	5	FALSE
L|6.CSE|ELA.6.36e	0	5	FALSE
L|6.CSE|ELA.6.37a	0	5	FALSE
L|6.CSE|ELA.6.37b	0	5	FALSE
L|6.VAU	1	2	TRUE
L|6.VAU|ELA.6.39a	0	2	FALSE
L|6.VAU|ELA.6.39b	0	2	FALSE
L|6.VAU|ELA.6.40a	0	2	FALSE
L|6.VAU|ELA.6.40c	0	2	FALSE
LT|6.CS	4	6	FALSE
LT|6.CS|ELA.6.7	0	3	FALSE
LT|6.CS|ELA.6.8	0	3	FALSE
LT|6.CS|ELA.6.9	0	3	FALSE
LT|6.IKI	1	3	FALSE
LT|6.IKI|ELA.6.14	1	2	FALSE

Test Blueprint for Grade 6 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
LT|6.KID	5	7	FALSE
LT|6.KID|ELA.6.1	0	3	FALSE
LT|6.KID|ELA.6.2	0	3	FALSE
LT|6.KID|ELA.6.3	0	3	FALSE
SL|6.CaC	0	3	FALSE
SL|6.CaC|ELA.6.31	0	2	FALSE
SL|6.CaC|ELA.6.32	0	2	FALSE
W|6.TTP	1	1	FALSE
W|6.TTP|ELA.6.20a	0	1	FALSE
W|6.TTP|ELA.6.21a	0	1	FALSE

Test Blueprint for Grade 7 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
IT|7.CS	6	8	FALSE
IT|7.CS|ELA.7.10	0	3	FALSE
IT|7.CS|ELA.7.11	0	3	FALSE
IT|7.CS|ELA.7.12	0	3	FALSE
IT|7.IKI	1	3	FALSE
IT|7.IKI|ELA.7.15	0	2	FALSE
IT|7.IKI|ELA.7.16	0	2	FALSE
IT|7.IKI|ELA.7.17	1	2	FALSE
IT|7.KID	6	8	FALSE
IT|7.KID|ELA.7.4	0	3	FALSE
IT|7.KID|ELA.7.5	0	3	FALSE
IT|7.KID|ELA.7.6	0	3	FALSE
L|7.CSE	0	8	FALSE
L|7.CSE|ELA.7.36a	0	5	FALSE
L|7.CSE|ELA.7.36b	0	5	FALSE
L|7.CSE|ELA.7.36c	0	5	FALSE
L|7.CSE|ELA.7.37a	0	5	FALSE
L|7.CSE|ELA.7.37b	0	5	FALSE
L|7.VAU	1	2	TRUE
L|7.VAU|ELA.7.39a	0	2	FALSE
L|7.VAU|ELA.7.39b	0	2	FALSE
L|7.VAU|ELA.7.40a	0	2	FALSE
L|7.VAU|ELA.7.40b	0	2	FALSE
L|7.VAU|ELA.7.40c	0	2	FALSE

Test Blueprint for Grade 7 ELA

ContentLevelID	MinItems	MaxItems	isStrictMax
LT|7.CS	4	6	FALSE
LT|7.CS|ELA.7.7	0	3	FALSE
LT|7.CS|ELA.7.8	0	3	FALSE
LT|7.CS|ELA.7.9	0	3	FALSE
LT|7.IKI	1	3	FALSE
LT|7.IKI|ELA.7.13	0	2	FALSE
LT|7.IKI|ELA.7.14	1	2	FALSE
LT|7.KID	5	7	FALSE
LT|7.KID|ELA.7.1	0	3	FALSE
LT|7.KID|ELA.7.2	0	3	FALSE
LT|7.KID|ELA.7.3	0	3	FALSE
SL|7.CaC	0	3	FALSE
SL|7.CaC|ELA.7.31	0	2	FALSE
W|7.TTP	1	1	FALSE
W|7.TTP|ELA.7.20a	0	1	FALSE
W|7.TTP|ELA.7.21a	0	1	FALSE

Test Blueprint for Grade 8 ELA

ContentLevelID	MinItems	MaxItems	isStrictMax
IT|8.CS	6	8	FALSE
IT|8.CS|ELA.8.10	0	3	FALSE
IT|8.CS|ELA.8.11	0	3	FALSE
IT|8.CS|ELA.8.12	0	3	FALSE
IT|8.IKI	1	3	FALSE
IT|8.IKI|ELA.8.15	0	2	FALSE
IT|8.IKI|ELA.8.16	0	2	FALSE
IT|8.IKI|ELA.8.17	1	2	FALSE
IT|8.KID	6	8	FALSE
IT|8.KID|ELA.8.4	0	3	FALSE
IT|8.KID|ELA.8.5	0	3	FALSE
IT|8.KID|ELA.8.6	0	3	FALSE
L|8.CSE	0	8	FALSE
L|8.CSE|ELA.8.36a	0	5	FALSE
L|8.CSE|ELA.8.37a	0	5	FALSE
L|8.CSE|ELA.8.37c	0	5	FALSE
L|8.KL	0	8	FALSE
L|8.KL|ELA.8.38a	0	5	FALSE

Test Blueprint for Grade 8 ELA			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
L|8.KL|ELA.8.38b	0	5	FALSE
L|8.KL|ELA.8.38c	0	5	FALSE
L|8.VAU	1	2	TRUE
L|8.VAU|ELA.8.39a	0	2	FALSE
L|8.VAU|ELA.8.39b	0	2	FALSE
L|8.VAU|ELA.8.39c	0	2	FALSE
L|8.VAU|ELA.8.40a	0	2	FALSE
L|8.VAU|ELA.8.40b	0	2	FALSE
L|8.VAU|ELA.8.40c	0	2	FALSE
LT|8.CS	4	6	FALSE
LT|8.CS|ELA.8.7	0	3	FALSE
LT|8.CS|ELA.8.8	0	3	FALSE
LT|8.CS|ELA.8.9	0	3	FALSE
LT|8.IKI	1	3	FALSE
LT|8.IKI|ELA.8.14	1	2	FALSE
LT|8.KID	5	7	FALSE
LT|8.KID|ELA.8.1	0	3	FALSE
LT|8.KID|ELA.8.2	0	3	FALSE
LT|8.KID|ELA.8.3	0	3	FALSE
SL|8.CaC	0	3	FALSE
SL|8.CaC|ELA.8.31	0	2	FALSE
SL|8.CaC|ELA.8.32	0	2	FALSE
W|8.TTP	1	1	FALSE
W|8.TTP|ELA.8.20a	0	1	FALSE
W|8.TTP|ELA.8.21a	0	1	FALSE

Test Blueprint for Grade 3 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
MDG|3.G.c1	0	3	FALSE
MDG|3.G.c1|M.3.24	0	2	FALSE
MDG|3.G.c1|M.3.25	0	2	FALSE
MDG|3.MD.c1	0	3	FALSE
MDG|3.MD.c1|M.3.16	0	2	FALSE
MDG|3.MD.c1|M.3.17	0	2	FALSE
MDG|3.MD.c2	0	3	FALSE
MDG|3.MD.c2|M.3.18	0	2	FALSE
MDG|3.MD.c2|M.3.19	0	2	FALSE
MDG|3.MD.c3	0	3	FALSE
MDG|3.MD.c3|M.3.20-21	0	2	FALSE
MDG|3.MD.c3|M.3.20-21|M.3.20	0	2	FALSE
MDG|3.MD.c3|M.3.20-21|M.3.20|M.3.20a	0	2	FALSE
MDG|3.MD.c3|M.3.20-21|M.3.20|M.3.20b	0	2	FALSE
MDG|3.MD.c3|M.3.20-21|M.3.21	0	2	FALSE
MDG|3.MD.c3|M.3.22	0	2	FALSE
MDG|3.MD.c3|M.3.22|M.3.22a	0	2	FALSE
MDG|3.MD.c3|M.3.22|M.3.22b	0	2	FALSE
MDG|3.MD.c3|M.3.22|M.3.22c	0	2	FALSE
MDG|3.MD.c3|M.3.22|M.3.22d	0	2	FALSE
MDG|3.MD.c4	0	2	FALSE
MDG|3.MD.c4|M.3.23	0	2	FALSE
NBTF|3.NBT.c1	6	9	FALSE
NBTF|3.NBT.c1|M.3.10	1	3	FALSE
NBTF|3.NBT.c1|M.3.11	1	3	FALSE
NBTF|3.NBT.c1|M.3.12	1	3	FALSE
NBTF|3.NF.c1	6	9	FALSE
NBTF|3.NF.c1|M.3.13	1	3	FALSE
NBTF|3.NF.c1|M.3.14	1	3	FALSE
NBTF|3.NF.c1|M.3.14|M.3.14a	0	3	FALSE
NBTF|3.NF.c1|M.3.14|M.3.14b	0	3	FALSE
NBTF|3.NF.c1|M.3.15	1	3	FALSE
NBTF|3.NF.c1|M.3.15|M.3.15a	0	3	FALSE
NBTF|3.NF.c1|M.3.15|M.3.15b	0	3	FALSE
NBTF|3.NF.c1|M.3.15|M.3.15c	0	3	FALSE
NBTF|3.NF.c1|M.3.15|M.3.15d	0	3	FALSE
OAT|3.OAT.c1	0	5	FALSE
OAT|3.OAT.c1|M.3.1	0	2	FALSE

Test Blueprint for Grade 3 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
OAT|3.OAT.c1|M.3.2	0	2	FALSE
OAT|3.OAT.c1|M.3.3	0	2	FALSE
OAT|3.OAT.c1|M.3.4	0	2	FALSE
OAT|3.OAT.c2	0	4	FALSE
OAT|3.OAT.c2|M.3.5	0	2	FALSE
OAT|3.OAT.c2|M.3.6	0	2	FALSE
OAT|3.OAT.c3	0	2	FALSE
OAT|3.OAT.c3|M.3.7	0	2	FALSE
OAT|3.OAT.c4	0	4	FALSE
OAT|3.OAT.c4|M.3.8	0	2	FALSE
OAT|3.OAT.c4|M.3.9	0	2	FALSE

Test Blueprint for Grade 4 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
MDG|4.G.c1	0	5	FALSE
MDG|4.G.c1|M.4.26	0	3	FALSE
MDG|4.G.c1|M.4.27	0	3	FALSE
MDG|4.G.c1|M.4.28	0	3	FALSE
MDG|4.MD.c1	0	4	FALSE
MDG|4.MD.c1|M.4.19	0	2	FALSE
MDG|4.MD.c1|M.4.20	0	2	FALSE
MDG|4.MD.c1|M.4.21	0	2	FALSE
MDG|4.MD.c2	0	2	FALSE
MDG|4.MD.c2|M.4.22	0	2	FALSE
MDG|4.MD.c3	0	5	FALSE
MDG|4.MD.c3|M.4.23	0	2	FALSE
MDG|4.MD.c3|M.4.23|M.4.23a	0	2	FALSE
MDG|4.MD.c3|M.4.23|M.4.23b	0	2	FALSE
MDG|4.MD.c3|M.4.24	0	2	FALSE
MDG|4.MD.c3|M.4.25	0	2	FALSE
NBTF|4.NBT.c1	0	5	FALSE
NBTF|4.NBT.c1|M.4.6	0	3	FALSE
NBTF|4.NBT.c1|M.4.7	0	3	FALSE
NBTF|4.NBT.c1|M.4.8	0	3	FALSE
NBTF|4.NBT.c2	0	5	FALSE
NBTF|4.NBT.c2|M.4.10	0	3	FALSE
NBTF|4.NBT.c2|M.4.11	0	3	FALSE

Test Blueprint for Grade 4 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
NBTF|4.NBT.c2|M.4.9	0	3	FALSE
NBTF|4.NF.c1	0	3	FALSE
NBTF|4.NF.c1|M.4.12	0	2	FALSE
NBTF|4.NF.c1|M.4.13	0	2	FALSE
NBTF|4.NF.c2	0	3	FALSE
NBTF|4.NF.c2|M.4.14	0	2	FALSE
NBTF|4.NF.c2|M.4.14|M.4.14a	0	2	FALSE
NBTF|4.NF.c2|M.4.14|M.4.14b	0	2	FALSE
NBTF|4.NF.c2|M.4.14|M.4.14c	0	2	FALSE
NBTF|4.NF.c2|M.4.14|M.4.14d	0	2	FALSE
NBTF|4.NF.c2|M.4.15	0	2	FALSE
NBTF|4.NF.c2|M.4.15|M.4.15a	0	2	FALSE
NBTF|4.NF.c2|M.4.15|M.4.15b	0	2	FALSE
NBTF|4.NF.c2|M.4.15|M.4.15c	0	2	FALSE
NBTF|4.NF.c3	0	5	FALSE
NBTF|4.NF.c3|M.4.16	0	3	FALSE
NBTF|4.NF.c3|M.4.17	0	3	FALSE
NBTF|4.NF.c3|M.4.18	0	3	FALSE
OAT|4.OAT.c1	2	6	FALSE
OAT|4.OAT.c1|M.4.1	0	3	FALSE
OAT|4.OAT.c1|M.4.2	0	3	FALSE
OAT|4.OAT.c1|M.4.3	0	3	FALSE
OAT|4.OAT.c2	0	3	FALSE
OAT|4.OAT.c2|M.4.4	0	3	FALSE
OAT|4.OAT.c3	0	3	FALSE
OAT|4.OAT.c3|M.4.5	0	3	FALSE

Test Blueprint for Grade 5 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
MDG|5.G.c1	0	2	TRUE
MDG|5.G.c1|M.5.23	0	2	FALSE
MDG|5.G.c1|M.5.24	0	2	FALSE
MDG|5.G.c2	0	2	TRUE
MDG|5.G.c2|M.5.25	0	2	FALSE
MDG|5.G.c2|M.5.26	0	2	FALSE
MDG|5.MD.c1	0	2	FALSE
MDG|5.MD.c1|M.5.18	0	2	FALSE

Test Blueprint for Grade 5 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
MDG|5.MD.c2	0	2	FALSE
MDG|5.MD.c2|M.5.19	0	2	FALSE
MDG|5.MD.c3	0	5	FALSE
MDG|5.MD.c3|M.5.20	0	2	FALSE
MDG|5.MD.c3|M.5.20|M.5.20a	0	2	FALSE
MDG|5.MD.c3|M.5.20|M.5.20b	0	2	FALSE
MDG|5.MD.c3|M.5.21	0	2	TRUE
MDG|5.MD.c3|M.5.22	0	2	FALSE
MDG|5.MD.c3|M.5.22|M.5.22a	0	2	FALSE
MDG|5.MD.c3|M.5.22|M.5.22b	0	2	FALSE
MDG|5.MD.c3|M.5.22|M.5.22c	0	2	FALSE
NBTF|5.NBT.c1	0	5	FALSE
NBTF|5.NBT.c1|M.5.4-5	0	2	FALSE
NBTF|5.NBT.c1|M.5.4-5|M.5.4	0	2	FALSE
NBTF|5.NBT.c1|M.5.4-5|M.5.5	0	2	FALSE
NBTF|5.NBT.c1|M.5.6	0	2	FALSE
NBTF|5.NBT.c1|M.5.6|M.5.6a	0	2	FALSE
NBTF|5.NBT.c1|M.5.6|M.5.6b	0	2	FALSE
NBTF|5.NBT.c1|M.5.7	0	2	FALSE
NBTF|5.NBT.c2	0	5	FALSE
NBTF|5.NBT.c2|M.5.10	0	2	FALSE
NBTF|5.NBT.c2|M.5.8	0	2	FALSE
NBTF|5.NBT.c2|M.5.9	0	2	FALSE
NBTF|5.NF.c1	0	4	FALSE
NBTF|5.NF.c1|M.5.11	0	2	FALSE
NBTF|5.NF.c1|M.5.12	0	2	FALSE
NBTF|5.NF.c2	0	6	FALSE
NBTF|5.NF.c2|M.5.13	0	2	FALSE
NBTF|5.NF.c2|M.5.14	0	2	FALSE
NBTF|5.NF.c2|M.5.14|M.5.14a	0	2	FALSE
NBTF|5.NF.c2|M.5.14|M.5.14b	0	2	FALSE
NBTF|5.NF.c2|M.5.15	0	2	FALSE
NBTF|5.NF.c2|M.5.15|M.5.15a	0	2	FALSE
NBTF|5.NF.c2|M.5.15|M.5.15b	0	2	FALSE
NBTF|5.NF.c2|M.5.16	0	2	FALSE
NBTF|5.NF.c2|M.5.17	0	2	FALSE
NBTF|5.NF.c2|M.5.17|M.5.17a	0	2	FALSE
NBTF|5.NF.c2|M.5.17|M.5.17b	0	2	FALSE

Test Blueprint for Grade 5 Mathematics

ContentLevelID	MinItems	MaxItems	isStrictMax
NBTF|5.NF.c2|M.5.17|M.5.17c	0	2	FALSE
OAT|5.OAT.c1	0	8	FALSE
OAT|5.OAT.c1|M.5.1	0	5	FALSE
OAT|5.OAT.c1|M.5.2	0	5	FALSE
OAT|5.OAT.c2	0	4	FALSE
OAT|5.OAT.c2|M.5.3	0	4	FALSE

Test Blueprint for Grade 6 Mathematics

ContentLevelID	MinItems	MaxItems	isStrictMax
GSP|6.G.c1	0	6	FALSE
GSP|6.G.c1|M.6.21	0	2	FALSE
GSP|6.G.c1|M.6.22	0	2	FALSE
GSP|6.G.c1|M.6.23	0	2	FALSE
GSP|6.G.c1|M.6.24	0	2	FALSE
GSP|6.SP.c1	0	4	FALSE
GSP|6.SP.c1|M.6.25	0	2	FALSE
GSP|6.SP.c1|M.6.26	0	2	FALSE
GSP|6.SP.c1|M.6.27	0	2	FALSE
GSP|6.SP.c2	0	3	FALSE
GSP|6.SP.c2|M.6.28	0	2	FALSE
GSP|6.SP.c2|M.6.29	0	2	FALSE
GSP|6.SP.c2|M.6.29|M.6.29a	0	2	FALSE
GSP|6.SP.c2|M.6.29|M.6.29c	0	2	FALSE
EE|6.EE.c1	1	7	FALSE
EE|6.EE.c1|M.6.12	0	2	FALSE
EE|6.EE.c1|M.6.13	0	2	FALSE
EE|6.EE.c1|M.6.13|M.6.13a	0	2	FALSE
EE|6.EE.c1|M.6.13|M.6.13b	0	2	FALSE
EE|6.EE.c1|M.6.13|M.6.13c	0	2	FALSE
EE|6.EE.c1|M.6.14	0	2	TRUE
EE|6.EE.c1|M.6.15	0	2	FALSE
EE|6.EE.c2	1	7	FALSE
EE|6.EE.c2|M.6.16	0	1	TRUE
EE|6.EE.c2|M.6.17	0	2	FALSE
EE|6.EE.c2|M.6.18	0	3	FALSE
EE|6.EE.c2|M.6.19	0	3	FALSE
EE|6.EE.c3	0	2	FALSE

Test Blueprint for Grade 6 Mathematics

ContentLevelID	MinItems	MaxItems	isStrictMax
EE|6.EE.c3|M.6.20	0	2	FALSE
RPNS|6.NS.c1	0	2	FALSE
RPNS|6.NS.c1|M.6.4	0	2	FALSE
RPNS|6.NS.c2	0	4	FALSE
RPNS|6.NS.c2|M.6.5	0	2	FALSE
RPNS|6.NS.c2|M.6.6	0	2	FALSE
RPNS|6.NS.c2|M.6.7	0	2	FALSE
RPNS|6.NS.c3	0	4	FALSE
RPNS|6.NS.c3|M.6.10	0	2	FALSE
RPNS|6.NS.c3|M.6.10|M.6.10a	0	2	FALSE
RPNS|6.NS.c3|M.6.10|M.6.10b	0	2	FALSE
RPNS|6.NS.c3|M.6.10|M.6.10c	0	2	FALSE
RPNS|6.NS.c3|M.6.10|M.6.10d	0	2	FALSE
RPNS|6.NS.c3|M.6.11	0	2	FALSE
RPNS|6.NS.c3|M.6.8	0	2	FALSE
RPNS|6.NS.c3|M.6.9	0	2	FALSE
RPNS|6.NS.c3|M.6.9|M.6.9a	0	2	FALSE
RPNS|6.NS.c3|M.6.9|M.6.9b	0	2	FALSE
RPNS|6.NS.c3|M.6.9|M.6.9c	0	2	FALSE
RPNS|6.RP.c1	3	8	FALSE
RPNS|6.RP.c1|M.6.1	0	4	FALSE
RPNS|6.RP.c1|M.6.2	0	4	FALSE
RPNS|6.RP.c1|M.6.3	0	4	FALSE
RPNS|6.RP.c1|M.6.3|M.6.3a	0	4	FALSE
RPNS|6.RP.c1|M.6.3|M.6.3b	0	4	FALSE
RPNS|6.RP.c1|M.6.3|M.6.3c	0	4	FALSE
RPNS|6.RP.c1|M.6.3|M.6.3d	0	4	FALSE

Test Blueprint for Grade 7 Mathematics

ContentLevelID	MinItems	MaxItems	isStrictMax
EE|7.EE.c1	2	6	FALSE
EE|7.EE.c1|M.7.7	0	3	FALSE
EE|7.EE.c1|M.7.8	0	3	FALSE
EE|7.EE.c2	2	6	FALSE
EE|7.EE.c2|M.7.10	0	3	FALSE
EE|7.EE.c2|M.7.10|M.7.10a	0	3	FALSE

Test Blueprint for Grade 7 Mathematics

ContentLevelID	MinItems	MaxItems	isStrictMax
EE|7.EE.c2|M.7.10|M.7.10b	0	3	FALSE
EE|7.EE.c2|M.7.9	0	3	FALSE
G|7.G.c1	2	6	FALSE
G|7.G.c1|M.7.11	0	2	FALSE
G|7.G.c1|M.7.12	0	2	FALSE
G|7.G.c1|M.7.13	0	2	FALSE
G|7.G.c2	2	6	FALSE
G|7.G.c2|M.7.14	0	2	FALSE
G|7.G.c2|M.7.15	0	2	FALSE
G|7.G.c2|M.7.16	0	2	FALSE
RPNS|7.NS.c1	2	6	FALSE
RPNS|7.NS.c1|M.7.4	0	3	FALSE
RPNS|7.NS.c1|M.7.4|M.7.4a	0	3	FALSE
RPNS|7.NS.c1|M.7.4|M.7.4b	0	3	FALSE
RPNS|7.NS.c1|M.7.4|M.7.4c	0	3	FALSE
RPNS|7.NS.c1|M.7.4|M.7.4d	0	3	FALSE
RPNS|7.NS.c1|M.7.5	0	3	FALSE
RPNS|7.NS.c1|M.7.5|M.7.5a	0	3	FALSE
RPNS|7.NS.c1|M.7.5|M.7.5b	0	3	FALSE
RPNS|7.NS.c1|M.7.5|M.7.5c	0	3	FALSE
RPNS|7.NS.c1|M.7.5|M.7.5d	0	3	FALSE
RPNS|7.NS.c1|M.7.6	0	3	FALSE
RPNS|7.RP.c1	2	6	FALSE
RPNS|7.RP.c1|M.7.1	0	3	TRUE
RPNS|7.RP.c1|M.7.2	0	3	FALSE
RPNS|7.RP.c1|M.7.2|M.7.2a	0	3	FALSE
RPNS|7.RP.c1|M.7.2|M.7.2b	0	3	FALSE
RPNS|7.RP.c1|M.7.2|M.7.2c	0	3	FALSE
RPNS|7.RP.c1|M.7.2|M.7.2d	0	3	FALSE
RPNS|7.RP.c1|M.7.3	0	3	FALSE
SP|7.SP.c1	0	3	FALSE
SP|7.SP.c1|M.7.17	0	2	FALSE
SP|7.SP.c1|M.7.18	0	2	FALSE
SP|7.SP.c2	0	3	FALSE
SP|7.SP.c2|M.7.19	0	2	FALSE
SP|7.SP.c2|M.7.20	0	2	FALSE
SP|7.SP.c2|M.7.20|M.7.20a	0	2	FALSE
SP|7.SP.c2|M.7.20|M.7.20c	0	2	FALSE

Test Blueprint for Grade 7 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
SP|7.SP.c2|M.7.21	0	2	FALSE
SP|7.SP.c2|M.7.22	0	2	FALSE
SP|7.SP.c3	2	6	FALSE
SP|7.SP.c3|M.7.23	0	2	FALSE
SP|7.SP.c3|M.7.24	0	2	FALSE
SP|7.SP.c3|M.7.25	0	2	FALSE
SP|7.SP.c3|M.7.25|M.7.25a	0	2	FALSE
SP|7.SP.c3|M.7.25|M.7.25b	0	2	FALSE
SP|7.SP.c3|M.7.26	0	2	FALSE
SP|7.SP.c3|M.7.26|M.7.26a	0	2	FALSE
SP|7.SP.c3|M.7.26|M.7.26b	0	2	FALSE
SP|7.SP.c3|M.7.26|M.7.26c	0	2	FALSE

Test Blueprint for Grade 8 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
EENS|8.EE.c1	0	4	TRUE
EENS|8.EE.c1|M.8.3	0	2	FALSE
EENS|8.EE.c1|M.8.4	0	2	FALSE
EENS|8.EE.c1|M.8.5	0	2	FALSE
EENS|8.EE.c1|M.8.6	0	2	FALSE
EENS|8.EE.c2	0	4	FALSE
EENS|8.EE.c2|M.8.7	0	2	FALSE
EENS|8.EE.c2|M.8.8	0	2	FALSE
EENS|8.EE.c3	0	4	FALSE
EENS|8.EE.c3|M.8.10	0	2	FALSE
EENS|8.EE.c3|M.8.10|M.8.10a	0	2	FALSE
EENS|8.EE.c3|M.8.10|M.8.10b	0	2	FALSE
EENS|8.EE.c3|M.8.10|M.8.10c	0	2	FALSE
EENS|8.EE.c3|M.8.9	0	2	FALSE
EENS|8.EE.c3|M.8.9|M.8.9a	0	2	FALSE
EENS|8.EE.c3|M.8.9|M.8.9b	0	2	FALSE
EENS|8.NS.c1	0	2	TRUE
EENS|8.NS.c1|M.8.1	0	2	FALSE
EENS|8.NS.c1|M.8.2	0	2	FALSE
F|8.F.c1	2	6	FALSE
F|8.F.c1|M.8.11	0	3	FALSE
F|8.F.c1|M.8.12	0	3	FALSE

Test Blueprint for Grade 8 Mathematics			
ContentLevelID	**MinItems**	**MaxItems**	**isStrictMax**
F|8.F.c1|M.8.13	0	3	FALSE
F|8.F.c2	2	6	FALSE
F|8.F.c2|M.8.14	0	3	FALSE
F|8.F.c2|M.8.15	0	3	FALSE
GSP|8.G.c1	1	6	FALSE
GSP|8.G.c1|M.8.16	0	2	FALSE
GSP|8.G.c1|M.8.16|M.8.16a	0	2	FALSE
GSP|8.G.c1|M.8.16|M.8.16b	0	2	FALSE
GSP|8.G.c1|M.8.16|M.8.16c	0	2	FALSE
GSP|8.G.c1|M.8.17	0	2	FALSE
GSP|8.G.c1|M.8.18	0	2	FALSE
GSP|8.G.c1|M.8.19	0	2	FALSE
GSP|8.G.c1|M.8.20	0	2	FALSE
GSP|8.G.c2	0	4	FALSE
GSP|8.G.c2|M.8.21	0	2	FALSE
GSP|8.G.c2|M.8.22	0	2	FALSE
GSP|8.G.c2|M.8.23	0	2	FALSE
GSP|8.G.c3	0	2	FALSE
GSP|8.G.c3|M.8.24	0	2	FALSE
GSP|8.SP.c1	1	6	FALSE
GSP|8.SP.c1|M.8.25	0	2	FALSE
GSP|8.SP.c1|M.8.26	0	2	FALSE
GSP|8.SP.c1|M.8.27	0	2	FALSE
GSP|8.SP.c1|M.8.28	0	2	FALSE

Appendix B

Simulation vs. Operational Blueprint Match

Table 1: Grade 3 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	94.70	5.20	0.10									94.52	5.43	0.05								
DOK2	100											100										
DOK3	100											100										

Table 2: Grade 4 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	99						1					99.02						0.98				
DOK2	100											100										
DOK3	100											100										

Table 3: Grade 5 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	97.50						2.50					97.62						2.38				
DOK2	100											100										
DOK3	100											100										

Table 4: Grade 6 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	86.10	13.30	0.10	0.40	0.10							87.36	12.22	0.22	0.15	0.05						
DOK2	100											100										
DOK3	100											100										

Table 5: Grade 7 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	99.70	0.30										99.71	0.28					0.01				
DOK2	100											99.98	0.02									
DOK3	100											100										

Table 6: Grade 8 ELA

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
RI	100											100										
RL	100											100										
L	100											100										
W	100											100										
DOK1	97.50	2.40					0.10					97.83	2.16					0.01				
DOK2	100											100										
DOK3	100											100										

Table 7: Grade 3 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
MDG	100											100										
NBTF	100											100										
OA	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Table 8: Grade 4 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
MDG	100											100										
NBTF	100											100										
OA	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Table 9: Grade 5 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
MDG	100											100										
NBTF	100											100										
OA	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Table 10: Grade 6 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
EE	100											100										
GSP	100											100										
RPNS	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Table 11: Grade 7 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
EE	100											100										
G	100											100										
RPNS	100											100										
SP	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Table 12: Grade 8 Mathematics

Content Level	% of Simulated Cases Violating Blueprint											% of Actual Cases Violating Blueprint										
	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5	% of Cases Meeting BP	1	2	3	4	≥5	-1	-2	-3	-4	≤-5
EENS	100											100										
F	100											100										
GSP	100											100										
DOK1	100											100										
DOK2	100											100										
DOK3	100											100										

Appendix C

Calibration Group Means and Standard Deviations

for ICCR Bank Items

Table 1: Calibration Group Means and Standard Deviations, Grade 3 ELA

Test	Year	Group	Mean	SD
Grade 3 ELA	2015	Utah	0.00	1.00
Grade 3 ELA	2016	Utah	0.00	1.00
Grade 3 ELA	2017	Utah	0.00	1.00
Grade 3 ELA	2018	Utah	0.00	1.00
Grade 3 ELA	2015	Arizona	-0.24	0.92
Grade 3 ELA	2016	Arizona	-0.22	0.95
Grade 3 ELA	2017	Arizona	-0.09	0.88
Grade 3 ELA	2018	Arizona	-0.25	0.93
Grade 3 ELA	2015	Oregon	-0.23	1.06
Grade 3 ELA	2016	Ohio	0.08	0.97

Table 2: Calibration Group Means and Standard Deviations, Grade 4 ELA

Test	Year	Group	Mean	SD
Grade 4 ELA	2015	Utah	0.00	1.00
Grade 4 ELA	2016	Utah	0.00	1.00
Grade 4 ELA	2017	Utah	0.00	1.00
Grade 4 ELA	2018	Utah	0.00	1.00
Grade 4 ELA	2016	Florida	-0.06	0.84
Grade 4 ELA	2017	Florida	-0.04	0.89
Grade 4 ELA	2018	Florida	0.64	0.97
Grade 4 ELA	2015	Arizona	-0.46	0.97
Grade 4 ELA	2016	Arizona	-0.44	1.07
Grade 4 ELA	2017	Arizona	-0.31	1.01
Grade 4 ELA	2018	Arizona	-0.44	0.90
Grade 4 ELA	2015	Oregon	-0.24	0.91
Grade 4 ELA	2016	Ohio	-0.12	0.95
Grade 4 ELA	2015	Utah	0.00	1.00

Table 3: Calibration Group Means and Standard Deviations, Grade 5 ELA

Test	Year	Group	Mean	SD
Grade 5 ELA	2015	Utah	0.00	1.00
Grade 5 ELA	2016	Utah	0.00	1.00
Grade 5 ELA	2017	Utah	0.00	1.00
Grade 5 ELA	2018	Utah	0.00	1.00
Grade 5 ELA	2015	Florida	0.20	0.93
Grade 5 ELA	2016	Florida	0.09	0.85
Grade 5 ELA	2017	Florida	0.05	0.84
Grade 5 ELA	2018	Florida	0.05	0.93
Grade 5 ELA	2015	Arizona	-0.50	1.01
Grade 5 ELA	2016	Arizona	-0.24	0.92
Grade 5 ELA	2017	Arizona	-0.28	0.96
Grade 5 ELA	2018	Arizona	-0.31	1.00
Grade 5 ELA	2015	Oregon	-0.12	0.94
Grade 5 ELA	2016	Ohio	0.09	0.79

Table 4: Calibration Group Means and Standard Deviations, Grade 6 ELA

Test	Year	Group	Mean	SD
Grade 6 ELA	2015	Utah	0.00	1.00
Grade 6 ELA	2016	Utah	0.00	1.00
Grade 6 ELA	2017	Utah	0.00	1.00
Grade 6 ELA	2018	Utah	0.00	1.00
Grade 6 ELA	2015	Florida	0.11	0.98
Grade 6 ELA	2016	Florida	-0.12	0.87
Grade 6 ELA	2017	Florida	0.02	1.02
Grade 6 ELA	2018	Florida	0.02	0.91
Grade 6 ELA	2015	Arizona	-0.27	1.01
Grade 6 ELA	2016	Arizona	-0.28	0.90
Grade 6 ELA	2017	Arizona	-0.08	0.99
Grade 6 ELA	2018	Arizona	-0.28	0.97
Grade 6 ELA	2015	Oregon	-0.21	0.83
Grade 6 ELA	2016	Ohio	-0.16	0.84

Table 5: Calibration Group Means and Standard Deviations, Grade 7 ELA

Test	Year	Group	Mean	SD
Grade 7 ELA	2015	Utah	0.00	1.00
Grade 7 ELA	2016	Utah	0.00	1.00
Grade 7 ELA	2017	Utah	0.00	1.00
Grade 7 ELA	2018	Utah	0.00	1.00
Grade 7 ELA	2015	Florida	0.12	0.96
Grade 7 ELA	2016	Florida	0.04	0.75
Grade 7 ELA	2017	Florida	-0.17	0.92
Grade 7 ELA	2018	Florida	0.19	0.94
Grade 7 ELA	2015	Arizona	-0.36	1.02
Grade 7 ELA	2016	Arizona	-0.07	0.80
Grade 7 ELA	2017	Arizona	-0.17	0.95
Grade 7 ELA	2018	Arizona	-0.33	0.95
Grade 7 ELA	2015	Oregon	-0.18	1.05
Grade 7 ELA	2016	Ohio	-0.03	0.84

Table 6: Calibration Group Means and Standard Deviations, Grade 8 ELA

Test	Year	Group	Mean	SD
Grade 8 ELA	2015	Utah	0.00	1.00
Grade 8 ELA	2016	Utah	0.00	1.00
Grade 8 ELA	2017	Utah	0.00	1.00
Grade 8 ELA	2018	Utah	0.00	1.00
Grade 8 ELA	2015	Florida	0.10	0.92
Grade 8 ELA	2016	Florida	0.06	0.93
Grade 8 ELA	2017	Florida	0.18	0.93
Grade 8 ELA	2018	Florida	0.24	0.88
Grade 8 ELA	2015	Arizona	-0.27	0.91
Grade 8 ELA	2016	Arizona	-0.12	0.89
Grade 8 ELA	2017	Arizona	-0.16	1.00
Grade 8 ELA	2018	Arizona	-0.11	0.96
Grade 8 ELA	2015	Oregon	-0.18	1.01
Grade 8 ELA	2016	Ohio	-0.03	0.87

Table 7: Calibration Group Means and Standard Deviations, Grade 3 Mathematics

Test	Year	Group	Mean	SD
Grade 3 Math	2015	Utah	0.00	1.00
Grade 3 Math	2016	Utah	0.00	1.00
Grade 3 Math	2017	Utah	0.00	1.00
Grade 3 Math	2018	Utah	0.00	1.00
Grade 3 Math	2017	Florida	0.29	0.97
Grade 3 Math	2018	Florida	0.39	1.08
Grade 3 Math	2015	Arizona	-0.27	0.93
Grade 3 Math	2016	Arizona	-0.21	1.02
Grade 3 Math	2017	Arizona	-0.15	1.00
Grade 3 Math	2018	Arizona	-0.24	1.03
Grade 3 Math	2015	Oregon	-0.21	0.97
Grade 3 Math	2016	Ohio	-0.06	0.96

Table 8: Calibration Group Means and Standard Deviations, Grade 4 Mathematics

Test	Year	Group	Mean	SD
Grade 4 Math	2015	Utah	0.00	1.00
Grade 4 Math	2016	Utah	0.00	1.00
Grade 4 Math	2017	Utah	0.00	1.00
Grade 4 Math	2018	Utah	0.00	1.00
Grade 4 Math	2017	Florida	0.40	0.87
Grade 4 Math	2018	Florida	0.33	0.95
Grade 4 Math	2015	Arizona	-0.29	0.97
Grade 4 Math	2016	Arizona	-0.24	1.02
Grade 4 Math	2017	Arizona	-0.18	0.93
Grade 4 Math	2018	Arizona	-0.16	1.01
Grade 4 Math	2015	Oregon	-0.34	1.00
Grade 4 Math	2016	Ohio	0.07	0.98

Table 9: Calibration Group Means and Standard Deviations, Grade 5 Mathematics

Test	Year	Group	Mean	SD
Grade 5 Math	2015	Utah	0.00	1.00
Grade 5 Math	2016	Utah	0.00	1.00
Grade 5 Math	2017	Utah	0.00	1.00
Grade 5 Math	2018	Utah	0.00	1.00
Grade 5 Math	2015	Florida	0.20	0.95
Grade 5 Math	2016	Florida	0.20	0.96
Grade 5 Math	2017	Florida	0.23	0.98
Grade 5 Math	2018	Florida	0.33	0.98
Grade 5 Math	2015	Arizona	-0.34	1.02
Grade 5 Math	2016	Arizona	-0.27	1.00
Grade 5 Math	2017	Arizona	-0.29	0.99
Grade 5 Math	2018	Arizona	-0.25	0.97
Grade 5 Math	2015	Oregon	-0.19	1.03
Grade 5 Math	2016	Ohio	0.02	0.92

Table 10: Calibration Group Means and Standard Deviations, Grade 6 Mathematics

Test	Year	Group	Mean	SD
Grade 6 Math	2015	Utah	0.00	1.00
Grade 6 Math	2016	Utah	0.00	1.00
Grade 6 Math	2017	Utah	0.00	1.00
Grade 6 Math	2018	Utah	0.00	1.00
Grade 6 Math	2015	Florida	-0.16	0.99
Grade 6 Math	2016	Florida	-0.30	1.03
Grade 6 Math	2017	Florida	-0.25	0.98
Grade 6 Math	2018	Florida	-0.22	1.00
Grade 6 Math	2015	Arizona	-0.36	1.05
Grade 6 Math	2016	Arizona	-0.40	1.06
Grade 6 Math	2017	Arizona	-0.28	1.00
Grade 6 Math	2018	Arizona	-0.31	1.02
Grade 6 Math	2015	Oregon	-0.38	1.06
Grade 6 Math	2016	Ohio	-0.11	0.97

Table 11: Calibration Group Means and Standard Deviations, Grade 7 Mathematics

Test	Year	Group	Mean	SD
Grade 7 Math	2015	Utah	0.00	1.00
Grade 7 Math	2016	Utah	0.00	1.00
Grade 7 Math	2017	Utah	0.00	1.00
Grade 7 Math	2018	Utah	0.00	1.00
Grade 7 Math	2015	Florida	-0.65	1.05
Grade 7 Math	2016	Florida	-0.33	1.04
Grade 7 Math	2017	Florida	-0.35	0.96
Grade 7 Math	2018	Florida	-0.32	0.97
Grade 7 Math	2015	Arizona	-0.70	1.20
Grade 7 Math	2016	Arizona	-0.49	1.16
Grade 7 Math	2017	Arizona	-0.42	1.13
Grade 7 Math	2018	Arizona	-0.28	0.98
Grade 7 Math	2015	Oregon	-0.37	1.18
Grade 7 Math	2016	Ohio	-0.09	1.04

Table 12: Calibration Group Means and Standard Deviations, Grade 8 Mathematics

Test	Year	Group	Mean	SD
Grade 8 Math	2015	Utah	0.00	1.00
Grade 8 Math	2016	Utah	0.00	1.00
Grade 8 Math	2017	Utah	0.00	1.00
Grade 8 Math	2018	Utah	0.00	1.00
Grade 8 Math	2015	Florida	-0.93	0.74
Grade 8 Math	2016	Florida	-0.56	0.78
Grade 8 Math	2017	Florida	-0.77	0.80
Grade 8 Math	2018	Florida	-0.74	0.78
Grade 8 Math	2015	Arizona	-0.55	1.10
Grade 8 Math	2016	Arizona	-0.46	0.93
Grade 8 Math	2017	Arizona	-0.53	1.06
Grade 8 Math	2018	Arizona	-0.40	1.03
Grade 8 Math	2015	Oregon	-0.52	1.34
Grade 8 Math	2016	Ohio	-0.27	0.87

Table 13: Calibration Group Means and Standard Deviations, Grade 5 Science

Test	Year	Group	Mean	SD
Grade 5 Science	2018	Connecticut	0	0.88
Grade 5 Science	2018	Oregon	0.12	0.79
Grade 5 Science	2018	Rhode Island	-0.14	0.95
Grade 5 Science	2018	Vermont	0.05	0.93
Grade 5 Science	2018	Wyoming	-0.37	0.76
Grade 5 Science	2018	New Hampshire	0.18	0.78
Grade 5 Science	2018	Hawaii	-0.57	0.78
Grade 5 Science	2018	West Virginia	-0.11	0.77

Table 14: Calibration Group Means and Standard Deviations, Grade 8 Science

Test	Year	Group	Mean	SD
Grade 8 Science	2018	Connecticut	0	0.81
Grade 8 Science	2018	Oregon	0.2	0.77
Grade 8 Science	2018	Rhode Island	-0.2	0.88
Grade 8 Science	2018	Vermont	0.09	0.87
Grade 8 Science	2018	Wyoming	0	0.66
Grade 8 Science	2018	New Hampshire	0.15	0.66
Grade 8 Science	2018	Hawaii	-0.02	0.82
Grade 8 Science	2018	Utah Grade 6	-0.05	0.84
Grade 8 Science	2018	Utah Grade 7	0.2	0.98
Grade 8 Science	2018	Utah Grade 8	0.29	1.04
Grade 8 Science	2018	West Virginia	-0.15	0.73

Appendix D

Vertical Scaling in SAGE

VERTICAL SCALING IN SAGE

Scoring and reporting student achievement on a vertical scale allows for the monitoring and evaluating of students' gains over time.

Because item parameters in the Independent College and Career Readiness (ICCR) item bank were equated to the SAGE vertical scale, this section documents the design and results of a vertical linking study that was implemented to develop the SAGE English language arts (ELA) and mathematics item bank.

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade-level assessment (g) were embedded in field-test slots of the assessment in the grade below (g - 1). This approach may take the risk of administering to students one or two items that measure contents that students may not yet have had the opportunity to learn. However, the resulting linkage represents student achievement of grade-level content for which they will receive instruction and thus can be interpreted as a pre-test score for measuring student acquisition of subsequent grade-level content.

1.1 SELECTING LINKING ITEMS

In order to adequately represent the content domain measured by each of the grade-level and subject-area assessments in the vertical linking design, approximately two forms' (test administrations) worth of items were identified for the vertical linking set at each grade. The vertical linking items were selected so that they met blueprints for test administrations both on grade-level assessments from which they were selected, as well as on the lower-grade assessments in which they were embedded.

Thus, a representative set of items from each grade-level assessment were identified for administration in the EFT (embedded field test) blocks in the below-grade level. All the linking items were fast-track items that had been run through rubric review but not data review. The performance of these vertical linking items was evaluated based on classical item analysis and calibration to ensure the high quality of the linking sets.

1.2 LINKING ANALYSIS

A chain linking approach was used to link the grade-level assessments within each subject area. An important advantage of chain linking approaches is that, because IRT calibrations proceed by establishing the within-grade scale, the achievement construct intended by the blueprint and enacted in the operational test form is preserved. The chain linking approach was also more practical given the very large number of items included in the SAGE adaptive item pools and the 3PL/GPC parameter estimation.

1.3 FINAL LINKING SET

To facilitate the development of a vertical scale that would be sensitive to student growth over time, we evaluated the performance of vertical linking items and removed items when the biserial/polyserial was less than 0.10, the proportion correct value was greater than .98 or less than .01, or the items were inactivated during administration. Additionally, items with poor fit due to underused categories were removed if they interfered with calibration. Table 1 shows the number of items removed, as well as the number of items remaining in the final vertical linking set. We note that the linking sets between grade 8 mathematics and SM I, and between SM I and SM II, had relatively higher proportions of items excluded from the final linking set. We also note that linking sets between grades 3 and 4 ELA and mathematics assessments had relatively higher proportions of items removed. Nevertheless, the number of items included in the final linking sets was large, and the content distribution approximated the blueprint distribution even after the removal of items from the original linking sets.

Table 1. Number of Items Dropped and Remaining in the Final Vertical Linking Set

	ELA			Mathematics	
Linkage	**Dropped Items**	**Final VL Set**	**Linkage**	**Dropped Items**	**Final VL Set**
G4 → G3	21	72	G4 → G3	16	82
G5 → G4	15	78	G5 → G4	5	94
G6 → G5	9	84	G6 → G5	7	92
G7 → G6	12	84	G7 → G6	7	92
G7 → G8	11	79	G7 → G8	3	95
G8 → G9	15	82	G8 → SMI	19	77
G9 → G10	15	80	SMI → SMII	35	65
G10 → G11	17	77	SMII → SMIII	10	87

1.4 Chain Linking

The chain linking approach proceeds from the within-grade item parameters identified in the initial calibrations of the operational and embedded field-test items. Because operational test items at each grade were administered in the EFT slots in the grade below, each item in the vertical linking set has two sets of item parameters: on-grade (g) and below-grade (g – 1). The chain linking proceeds by identifying the linking constants necessary to place the below-grade item parameters on the on-grade scale for the items in the final vertical linking set. The Stocking-Lord procedure (Stocking & Lord, 1983) was used to identify the linking constants to link each of the grade-level assessments.

For both SAGE ELA and mathematics, grade 7 served as the base grade, grades 6 and 8 were linked directly to grade 7, and the remaining assessments chained through intervening grades to be placed on the grade 7 scales. No additional items were dropped in the linking step. In this way, the vertical linking constants necessary to place the within-grade scales onto the vertical reporting scale were identified. The final vertical linking constants are shown in Table 2.

Table 2. Final Linking Constants for ELA and Mathematics

ELA			Mathematics		
Grade	**Slope**	**Intercept**	**Grade**	**Slope**	**Intercept**
3	0.83	–1.29	**3**	0.60	–2.45
4	0.89	–0.83	**4**	0.74	–1.81
5	0.89	–0.45	**5**	0.85	–1.23
6	0.90	–0.12	**6**	0.99	–0.56
7	0.94	0.01	**7**	1.01	–0.01
8	1.01	0.20	**8**	1.24	0.68
9	1.07	0.36	**SMI**	1.46	1.01
10	1.13	0.48	**SMII**	1.62	1.72
11	1.17	0.58	**SMIII**	1.69	2.33

To examine the properties of the vertical linking scale for mathematics and ELA, the mean ability (theta) and test characteristic curves were examined for each of the grade-level assessments on the vertical scale.

Table 3 shows descriptive statistics for ELA across grades on the vertical scale, with mean ability shown graphically in Figure 1. For ELA, achievement gains across grade levels are not as large as for mathematics, and results indicate a deceleration of reading gains as one moves from lower to higher grades. TCCs (test characteristic curve) for the reading item pools, shown in Figure 2, show less separation at the higher grade levels, indicating larger differences in item difficulty between elementary grade item pools than between upper-grade item pools.

Table 3. Descriptive Statistics for ELA Achievement on the Vertical Scale

Grade	N	Mean	Standard Deviation	Minimum	Maximum
3	46,762	–1.29	0.90	–5.46	2.89
4	46,613	–0.84	0.96	–5.29	2.91
5	44,348	–0.45	0.98	–4.92	4.03
6	38,092	–0.13	0.98	–4.62	3.83
7	36,304	0.00	1.02	–4.71	4.73
8	37,532	0.20	1.08	–4.86	4.57
9	31,746	0.35	1.16	–4.97	5.69
10	31,601	0.48	1.23	–5.15	6.12
11	32,341	0.57	1.27	–5.25	6.41

Figure 1. Mean ELA Achievement on the Vertical Scale

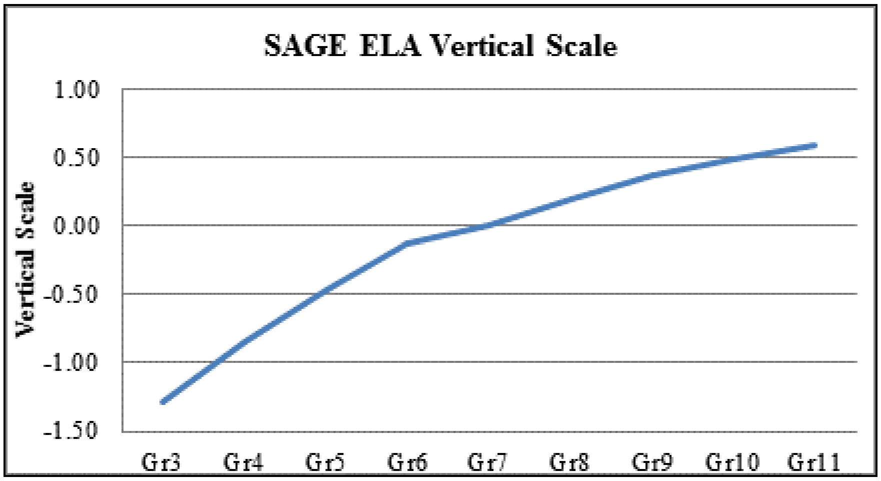


Figure 2. ELA Test Characteristic Curves

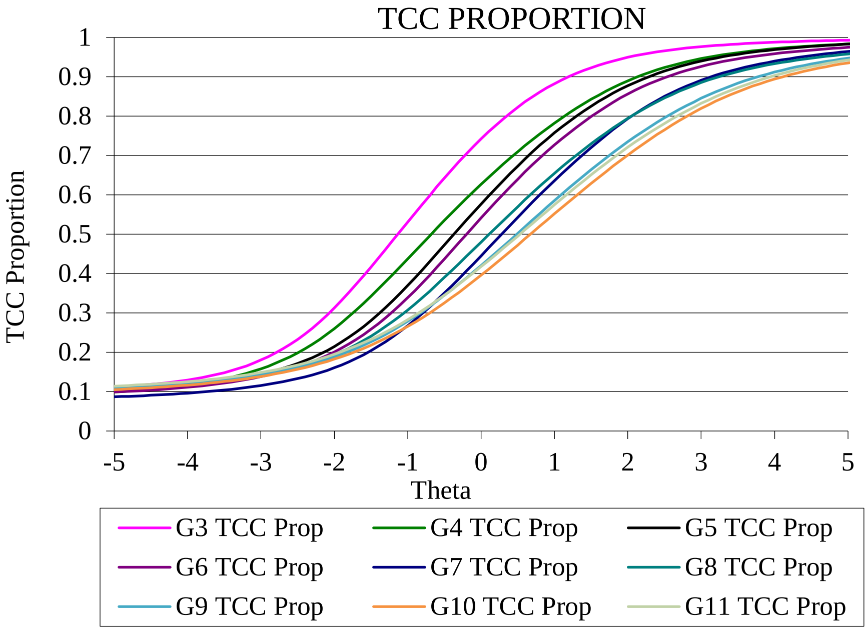


Table 4 shows descriptive statistics for mathematics across grades on the vertical scale, with mean ability shown graphically in Figure 3. For mathematics, results indicate relatively uniform and large achievement gains across most grades, with a somewhat smaller difference in means between grade 8 and SM I. Moreover, the mathematics TCCs shown in Figure 4 indicate uniform increases in the difficulty of the item pools across grades.

Table 4. Descriptive Statistics for Mathematics Achievement on the Vertical Scale

Grade	N	Mean	Standard Deviation	Minimum	Maximum
3	47,414	–2.46	0.66	–5.46	0.56
4	47,337	–1.83	0.84	–5.54	1.91
5	46,832	–1.26	1.00	–5.47	3.01
6	45,498	–0.58	1.12	–5.49	4.38
7	43,509	–0.05	1.15	–5.06	5.05
8	43,374	0.62	1.43	–5.51	6.88
SMI	44,527	0.87	1.85	–6.31	8.33
SMII	37,519	1.51	2.20	–6.40	8.57
SMIII	17,046	1.95	2.61	–6.13	10.60

Figure 3. Mean Mathematics Achievement on the Vertical Scale

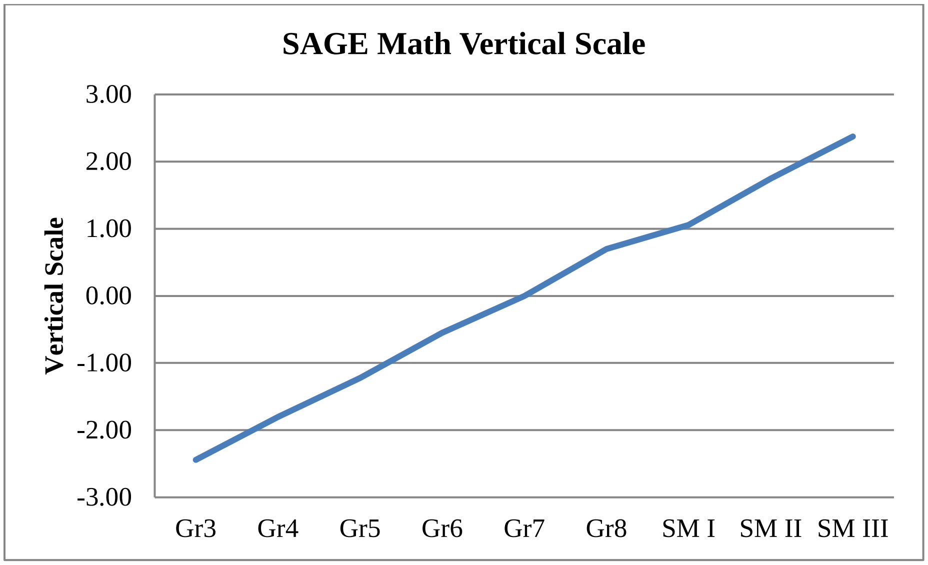
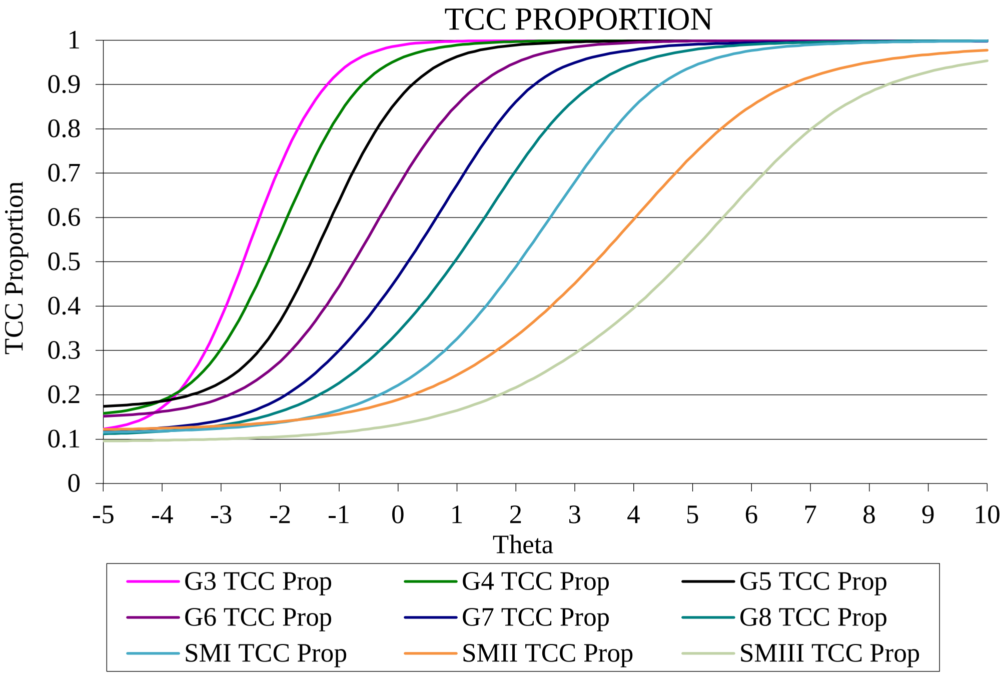


Figure 4. Mathematics Test Characteristic Curves



Appendix E

Distribution of Scale Scores and Achievement Levels by Subgroup

Table 1: Mean and Standard Deviation of Scale Scores by Subgroup, ELA

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,526	8,538	8,988	685	6	102	370	800	10	15,252	185
	Mean	567.53	570.81	564.42	549.01	-	584.01	563.21	562.82	-	568.42	544.06
	SD	43.53	42.39	44.37	41.62	-	48.56	39.86	42.13	-	43.52	37.25
4	*N*	17,323	8,454	8,869	630	8	98	370	794	8	15,129	131
	Mean	588.43	593.28	583.8	573.27	-	629.85	580.85	582.67	-	589.14	556.5
	SD	47.78	46.96	48.09	46.28	-	43.85	47.66	46.39	-	47.71	40.75
5	*N*	17,683	8,611	9,072	649	7	113	372	804	8	15,439	131
	Mean	608.94	613.69	604.43	591.06	-	643.35	600.41	603.78	-	609.75	574.45
	SD	45.87	44.73	46.49	43.76	-	43.57	45.39	44.87	-	45.82	37.41
6	*N*	17,697	8,678	9,019	683	13	100	410	723	9	15,282	100
	Mean	625.29	631.33	619.48	605.27	618.13	660.49	620.00	621.82	-	626.21	579.82
	SD	47.54	45.38	48.83	46.32	59.50	46.91	49.13	47.66	-	47.39	44.33
7	*N*	18,242	8,979	9,263	783	28	117	390	696	3	15,946	122
	Mean	628.37	636.84	620.15	611.14	600.33	660.36	624.72	620.54	-	629.45	581.54
	SD	50.25	47.72	51.27	47.3	56.01	53.36	51.73	48.95	-	50.24	42.10
8	*N*	18,698	9,012	9,686	717	20	120	440	729	9	16,371	133
	Mean	638.94	647.71	630.79	619.21	621.86	675.14	631.09	636.10	-	640.04	590.04
	SD	52.63	50.24	53.49	50.23	52.72	57.18	52.57	50.19	-	52.59	42.10

* The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.

Table 2: Mean and Standard Deviation of Scale Scores by Subgroup, Mathematics

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,548	8,551	8,997	686	6	102	370	802	10	15,270	184
	Mean	419.56	417.45	421.57	403.16	-	440.32	412.84	413.84	-	420.62	404.73
	SD	35.79	34.25	37.08	34.42	-	38.32	34.30	35.16	-	35.67	33.44
4	*N*	17,342	8,463	8,879	631	8	98	369	797	8	15,143	131
	Mean	444.53	443.08	445.92	426.75	-	490.56	435.57	436.98	-	445.50	422.03
	SD	43.63	41.16	45.83	43.71	-	40.96	42.63	41.56	-	43.44	45.97
5	*N*	17,727	8,637	9,090	651	7	113	372	806	8	15,483	131
	Mean	465.21	464.45	465.93	442.88	-	519.55	455.97	454.90	-	466.36	437.34
	SD	51.35	49.19	53.32	47.97	-	54.52	50.63	50.74	-	51.04	44.24
6	*N*	17,721	8,687	9,034	708	13	100	411	741	9	15,437	100
	Mean	481.89	482.36	481.45	455.97	465.38	537.97	473.61	474.93	-	483.40	443.01
	SD	56.82	54.03	35.79	54.31	73.45	58.46	55.83	57.98	-	56.47	57.10
7	*N*	18,302	9,006	9,296	788	29	118	392	698	3	15,995	124
	Mean	512.71	512.72	512.69	481.08	488.48	571.07	502.38	499.46	-	514.77	467.23
	SD	62.72	59.28	65.88	60.87	88.85	68.52	60.32	63.77	-	62.14	61.43
8	*N*	18,742	9,031	9,711	718	20	120	440	732	9	16,409	133
	Mean	537.28	539.57	535.15	501.09	503.96	605.22	526.77	525.85	-	539.51	494.87
	SD	77.07	72.90	80.69	72.93	75.89	88.67	75.02	77.53	-	76.55	66.61

* The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.

Table 3: Mean and Standard Deviation of Scale Scores by Subgroup, Science

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
5	*N*	17,698	8,621	9,077	647	7	112	371	800	8	15,305	448	131
	Mean	545.71	545.59	545.83	537.79	-	557.96	543.49	543.41	-	546.03	549.37	535.21
	SD	16.38	15.60	17.09	15.25	-	16.20	15.99	15.34	-	16.36	16.11	13.02
8	*N*	18,694	9,013	9,681	711	20	120	438	723	9	16,265	408	134
	Mean	844.08	844.46	843.72	837.66	839.41	856.34	842.24	842.35	-	844.42	843.35	834.99
	SD	15.67	14.67	16.54	13.67	13.86	18.02	15.62	15.24	-	15.66	15.36	12.35

*The descriptive statistics are not provided when the number of students for given group is 10 or less than 10.

Table 4: Percent of Students in Each Achievement Level by Subgroup, ELA

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,526	8,538	8,988	685	6	102	370	800	10	15,252	185
	L1	34.41	30.77	37.86	50.22	50	26.47	38.11	38.13	20	33.6	57.3
	L2	29.46	31.03	27.98	30.51	-	22.55	29.73	30.5	20	29.41	31.35
	L3	21.45	22.66	20.3	13.43	-	17.65	22.43	19.38	50	21.91	7.03
	L4	14.68	15.54	13.85	5.84	50	33.33	9.73	12	10	15.07	4.32
4	*N*	17,323	8,454	8,869	630	8	98	370	794	8	15,129	131
	L1	29.56	25.75	33.18	40.79	25	8.16	35.95	34.38	37.5	28.99	54.96
	L2	26.94	26.69	27.17	29.05	12.5	16.33	25.68	26.83	12.5	26.9	27.48
	L3	22.5	24.31	20.78	18.41	50	18.37	23.24	22.17	37.5	22.65	14.5
	L4	21.01	23.26	18.86	11.75	12.5	57.14	15.14	16.62	12.5	21.46	3.05
5	*N*	17,683	8,611	9,072	649	7	113	372	804	8	15,439	131
	L1	31.41	27.27	35.35	47.3	28.57	10.62	36.56	34.83	25	30.77	62.6
	L2	27.66	28.52	26.84	28.35	42.86	17.7	29.03	30.1	37.5	27.49	25.19
	L3	24.45	25.58	23.38	16.95	28.57	26.55	22.04	22.01	25	24.89	10.69
	L4	16.47	18.63	14.43	7.4	-	45.13	12.37	13.06	12.5	16.86	1.53
6	*N*	17,697	8,678	9,019	683	13	100	410	723	9	15,282	100
	L1	26.94	21.95	31.73	43.63	38.46	8	30.24	31.81	-	26.19	64
	L2	31.24	31.71	30.79	30.89	23.08	18	30.98	28.77	33.33	31.29	24
	L3	29.75	32.50	27.11	21.08	7.69	39	29.02	28.22	44.44	30.05	12
	L4	12.07	13.84	10.37	4.39	30.77	35	9.76	11.2	22.22	12.47	-

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
7	*N*	18,242	8,979	9,263	783	28	117	390	696	3	15,946	122
	L1	28.91	21.94	35.67	42.53	53.57	16.24	30.26	32.9	-	28.14	66.39
	L2	30.62	31.06	30.18	31.93	35.71	13.68	34.1	34.05	33.33	30.26	27.87
	L3	27.49	31.71	23.4	18.65	3.57	34.19	23.08	23.71	-	28.3	5.74
	L4	12.98	15.29	10.74	6.9	7.14	35.9	12.56	9.34	66.67	13.3	-
8	*N*	18,698	9,012	9,686	717	20	120	440	729	9	16,371	133
	L1	30.89	23.58	37.68	46.16	45	13.33	36.36	32.1	55.56	30.03	75.19
	L2	30	31.1	28.97	29.15	25	19.17	28.41	32.1	22.22	30.03	17.29
	L3	25.48	29.08	22.14	18.55	30	29.17	25.23	24.83	11.11	25.84	6.77
	L4	13.63	16.23	11.21	6.14	-	38.33	10	10.97	11.11	14.09	0.75

Table 5: Percent of Students in Each Achievement Level by Subgroup, Mathematics

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
3	*N*	17,548	8,551	8,997	686	6	102	370	802	10	15,270	184
	L1	0.27	0.29	0.25	0.46	0.17	0.13	0.32	0.33	0.10	0.26	0.41
	L2	0.27	0.29	0.26	0.27	0.33	0.17	0.29	0.29	0.30	0.27	0.32
	L3	0.25	0.25	0.25	0.19	0.17	0.29	0.25	0.22	0.40	0.25	0.20
	L4	0.21	0.18	0.24	0.08	0.33	0.41	0.14	0.16	0.20	0.22	0.08
4	*N*	17,342	8,463	8,879	631	8	98	369	797	8	15,143	131
	L1	0.26	0.26	0.26	0.44	0.12	0.07	0.33	0.33	0.25	0.25	0.48
	L2	0.33	0.35	0.31	0.32	0.38	0.08	0.32	0.35	0.25	0.33	0.29
	L3	0.20	0.21	0.20	0.13	0.25	0.21	0.20	0.16	0.25	0.21	0.15
	L4	0.21	0.18	0.24	0.11	0.25	0.63	0.15	0.16	0.25	0.22	0.08
5	*N*	17,727	8,637	9,090	651	7	113	372	806	8	15,483	131
	L1	0.34	0.35	0.34	0.54	0.43	0.10	0.42	0.44	0.12	0.33	0.59
	L2	0.31	0.32	0.31	0.29	0.43	0.20	0.32	0.31	0.38	0.32	0.28
	L3	0.18	0.19	0.17	0.12	0.14	0.19	0.15	0.14	0.38	0.18	0.09
	L4	0.17	0.15	0.18	0.06	-	0.50	0.11	0.12	0.12	0.17	0.04

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
6	*N*	17,721	8,687	9,034	708	13	100	411	741	9	15,437	100
	L1	0.42	0.42	0.42	0.64	0.62	0.09	0.49	0.48	0.11	0.41	0.75
	L2	0.31	0.31	0.30	0.24	0.08	0.26	0.29	0.29	0.56	0.31	0.17
	L3	0.16	0.17	0.15	0.09	0.15	0.19	0.14	0.14	0.22	0.17	0.04
	L4	0.11	0.10	0.12	0.03	0.15	0.46	0.08	0.09	0.11	0.11	0.04
7	*N*	18,302	9,006	9,296	788	29	118	392	698	3	15,995	124
	L1	0.41	0.41	0.41	0.64	0.48	0.14	0.51	0.52	0.33	0.39	0.75
	L2	0.30	0.31	0.29	0.23	0.31	0.21	0.25	0.25	0.67	0.31	0.15
	L3	0.17	0.17	0.17	0.08	0.14	0.19	0.17	0.14	-	0.18	0.06
	L4	0.12	0.11	0.13	0.05	0.07	0.46	0.08	0.09	-	0.13	0.04
8	*N*	18,742	9,031	9,711	718	20	120	440	732	9	16,409	133
	L1	0.44	0.42	0.45	0.66	0.60	0.17	0.50	0.50	0.56	0.42	0.73
	L2	0.30	0.31	0.28	0.22	0.30	0.20	0.30	0.31	0.33	0.30	0.17
	L3	0.12	0.13	0.12	0.06	0.05	0.16	0.10	0.07	0.11	0.13	0.06
	L4	0.15	0.14	0.15	0.06	0.05	0.48	0.09	0.12	-	0.15	0.05

Table 6: Percent of Students in Each Achievement Level by Subgroup, Science

Grade	Group	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
5	*N*	17,698	8,621	9,077	647	7	112	371	800	8	15,305	448	131
	L1	0.29	0.28	0.29	0.47	0.43	0.07	0.33	0.33	0.25	0.28	0.18	0.53
	L2	0.41	0.43	0.39	0.38	0.29	0.37	0.41	0.44	0.38	0.41	0.45	0.37
	L3	0.21	0.21	0.21	0.12	0.29	0.31	0.20	0.17	0.13	0.22	0.24	0.09
	L4	0.09	0.08	0.10	0.03	0.00	0.25	0.06	0.06	0.25	0.09	0.13	0.00
8	*N*	18,694	9,013	9,681	711	20	120	438	723	9	16,265	408	134
	L1	0.35	0.32	0.37	0.50	0.45	0.16	0.38	0.40	0.56	0.34	0.35	0.59
	L2	0.39	0.43	0.36	0.37	0.45	0.33	0.40	0.38	0.44	0.39	0.41	0.37
	L3	0.18	0.19	0.17	0.10	0.10	0.23	0.15	0.16	0.00	0.19	0.15	0.03
	L4	0.08	0.06	0.10	0.03	0.00	0.29	0.07	0.06	0.00	0.08	0.09	0.01

Appendix F

Distribution of Reporting Category Scores

Table 1: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 3 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	17,526	8,538	8,988	685	6	102	370	800	10	15,252	185
IT	*Mean*	557.64	558.6	556.72	535.11	601.15	575.89	553.46	550.21	569.45	558.84	530.11
	SD	69.19	67.66	70.59	66.59	114.81	70.26	64.4	68.6	35.83	69.26	62.87
LT	*Mean*	561.88	565.44	558.51	543.11	571.9	577.54	555.16	558.38	588.98	562.92	532.45
	SD	62.69	61.16	63.93	63.38	100.61	66.15	61.76	61.29	27.13	62.56	63.37
WL	*Mean*	565.03	569.81	560.48	545.3	594.54	586.57	561.55	559.74	600.35	565.85	542.25
	SD	49.48	48.84	49.65	46.81	61.05	54.5	45.81	47.89	47.12	49.47	43.53

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 2: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 4 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	17,323	8,454	8,869	630	8	98	370	794	8	15,129	131
IT	*Mean*	586.39	587.74	585.11	567.38	597.24	628.55	584.53	580.74	566.6	587.06	558.37
	SD	68.28	66.94	69.51	67.64	35.99	58.7	64.4	67.5	91.86	68.45	62.92
LT	*Mean*	585.38	589.99	580.99	568.76	600.59	634.53	576.85	577.49	571.47	586.22	548.41
	SD	60.71	59.89	61.17	60.61	23.24	52.19	62.72	61.97	84.99	60.47	61.77
WL	*Mean*	584.49	592.69	576.68	570.18	598.99	629.78	573.71	578.87	566.68	585.12	547.89
	SD	57.72	57.04	57.28	54.47	48.61	52.97	59.69	55.36	54.92	57.76	49.66

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 3: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 5 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	17,683	8,611	9,072	649	7	113	372	804	8	15,439	131
IT	*Mean*	607.99	610.29	605.8	589.22	586.1	644.59	591.97	604.17	627.75	608.99	561.31
	SD	66.55	65.68	67.3	67.88	76.11	55.53	72.46	66.18	60.01	66.24	68.75
LT	*Mean*	607.84	611.35	604.51	589.53	618.32	645.29	599.37	603.25	623.59	608.65	576.29
	SD	59.16	59.38	58.76	57.08	43.23	56.1	60.02	58.21	74.52	59.05	52.38
WL	*Mean*	603.68	611.19	596.54	583.6	574.58	641.74	596.59	597.5	602.25	604.48	565.35
	SD	54.86	53.21	55.45	53.4	56.02	49.92	53.41	52.49	66.02	54.89	46.47

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 4: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 6 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	17,697	8,678	9,019	683	13	100	410	723	9	15,282	100
IT	*Mean*	618.31	621.5	615.24	597.47	609	656.54	610.7	614.71	658.88	619.31	571.9
	SD	64.33	63.55	64.93	65.61	100.76	58.03	68.82	63.11	44.13	64.11	65.73
LT	*Mean*	612.52	619.13	606.17	590.41	610.41	652.98	610.19	606.4	649.54	613.72	571.09
	SD	69.8	68.06	70.85	70.15	69.6	65.14	68.41	71.15	75.49	69.55	62.46
WL	*Mean*	629.11	638.12	620.43	606.27	615.38	671.44	621.45	626.5	654.51	630.03	573.58
	SD	57.91	55.8	58.57	54.49	59.13	53.91	61.72	58.71	45.33	57.76	58.44

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 5: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 7 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	18,242	8,979	9,263	783	28	117	390	696	3	15,946	122
IT	*Mean*	624.43	631.12	617.94	608.2	588.54	660.31	619.13	614.72	689.26	625.54	577.59
	SD	63.23	60.37	65.23	60.19	71.26	64.61	66.61	63.83	46.05	63.14	59.5
LT	*Mean*	616.96	626.05	608.14	596.1	593.12	650.34	614.12	608.29	669.08	618.2	571.67
	SD	68.7	66.07	70.04	67.89	73.5	71.21	70.17	65.55	68.66	68.72	58.61
WL	*Mean*	631.45	643.14	620.13	612.28	600.55	666.77	628.73	624.06	671.64	632.59	577.2
	SD	59.71	56.67	60.39	57.51	68.51	61.57	61.98	58.7	68.92	59.63	50.94

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 6: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 8 ELA

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	LEP
Total	*N*	18,698	9,012	9,686	717	20	120	440	729	9	16,371	133
IT	*Mean*	632.6	639.62	626.06	612.53	633.36	672.45	623.67	628.84	625.84	633.71	594.71
	SD	69.94	67.5	71.52	67.41	51.08	65.7	69.29	66.14	80.95	70.12	55.88
LT	*Mean*	632.98	641.35	625.19	613.11	613.56	678.6	625.29	626.78	597.19	634.18	577.48
	SD	77.32	76.16	77.59	74.23	80.69	72.23	76.56	78.1	101.89	77.19	65.32
WL	*Mean*	638.67	649.87	628.24	617.11	613.93	677.14	630.42	637.31	609.07	639.8	584.01
	SD	58.96	56.97	58.88	56.9	62.58	68.71	59.63	56.53	68.56	58.84	52.14

IT = Informational Text; LT = Literary Text; WL = Writing and Language

Table 7: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 3 Mathematics

Rep Cat	Scale Score	All Students	Female	Male	Multi-Racial	American Indian/ Native Alaskan	Asian	Hispanic	African American	White	Pacific Islander	LEP
Total	*N*	17,548	8,551	8,997	802	6	102	370	686	15,270	10	184
MDG	*Mean*	415.55	412.83	418.14	409.72	433.82	440.11	407.35	396.54	416.81	423.06	398.18
	SD	43.00	42.27	43.53	40.67	45.86	46.10	42.60	43.16	42.78	48.60	40.93
NBTF	*Mean*	421.30	418.74	423.73	414.58	426.01	439.01	414.37	404.14	422.49	423.13	406.71
	SD	37.91	35.95	39.53	38.09	37.60	38.68	36.31	36.35	37.75	35.35	33.97
OAT	*Mean*	413.74	412.20	415.21	408.06	421.29	441.41	406.82	395.36	414.81	421.31	398.10
	SD	49.44	48.19	50.56	49.44	64.72	47.14	46.75	49.08	49.27	35.82	46.38

MDG = Measurement, Data, and Geometry; NBTF = Numbers and Operations in Base Ten & Fractions; OAT = Operations and Algebraic Thinking

Table 8: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 4 Mathematics

Rep Cat	Scale Score	All Students	Female	Male	Multi-Racial	American Indian/ Native Alaskan	Asian	Hispanic	African American	White	Pacific Islander	LEP
Total	*N*	17,342	8,463	8,879	797	8	98	369	631	15,143	8	131
MDG	*Mean*	438.22	436.71	439.66	429.41	445.21	486.04	429.41	412.23	439.57	456.18	410.61
	SD	55.43	53.02	57.60	54.86	38.83	51.60	54.37	59.14	54.92	57.00	55.42
NBTF	*Mean*	443.67	441.97	445.28	436.38	463.76	493.70	434.60	426.25	444.57	464.71	423.64
	SD	46.85	44.30	49.11	44.29	28.11	49.02	45.18	46.95	46.63	58.43	47.72
OAT	*Mean*	442.42	441.29	443.50	433.01	442.90	494.00	432.60	424.94	443.52	453.15	416.05
	SD	54.93	53.33	56.40	54.26	42.56	44.90	54.72	54.70	54.74	36.96	58.57

MDG = Measurement, Data, and Geometry; NBTF = Numbers and Operations in Base Ten & Fractions; OAT = Operations and Algebraic Thinking

Table 9: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 5 Mathematics

Rep Cat	Scale Score	All Students	Female	Male	Multi-Racial	American Indian/ Native Alaskan	Asian	Hispanic	African American	White	Pacific Islander	LEP
Total	*N*	17,727	8,637	9,090	806	7	113	372	651	15,483	8	131
MDG	*Mean*	460.32	457.05	463.42	449.96	468.61	517.68	448.44	432.13	461.78	482.40	423.89
	SD	63.53	61.55	65.21	62.23	24.74	68.78	63.46	63.00	63.02	81.14	62.42
NBTF	*Mean*	462.74	461.98	463.46	449.99	443.35	519.46	454.98	441.78	463.95	493.20	441.73
	SD	56.23	53.89	58.36	56.15	58.57	56.59	54.18	52.28	55.96	58.59	43.41
OAT	*Mean*	463.48	465.15	461.89	456.64	445.94	524.09	451.70	439.79	464.42	490.50	423.65
	SD	66.47	65.03	67.78	66.12	64.46	70.01	68.59	64.83	66.03	80.96	61.51

MDG = Measurement, Data, and Geometry; NBTF = Numbers and Operations in Base Ten & Fractions; OAT = Operations and Algebraic Thinking

Table 10: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 6 Mathematics

Rep Cat	Scale Score	All Students	Female	Male	Multi-Racial	American Indian/ Native Alaskan	Asian	Hispanic	African American	White	Pacific Islander	LEP
Total	*N*	17,721	8,687	9,034	741	13	100	411	708	15,437	9	100
EE	*Mean*	475.92	477.59	474.31	469.16	454.44	539.69	467.39	452.05	477.32	519.36	432.90
	SD	70.17	67.99	72.17	70.43	80.85	68.07	71.40	66.20	69.98	38.13	71.97
GSP	*Mean*	475.03	476.59	473.53	470.60	460.62	537.75	468.90	447.19	476.23	519.97	438.53
	SD	74.14	71.19	76.85	74.41	84.65	70.78	71.34	71.22	74.03	50.26	76.64
RPNS	*Mean*	478.75	478.13	479.36	470.64	457.11	535.80	468.64	450.63	480.48	510.41	437.14
	SD	62.19	59.59	64.59	64.15	90.07	62.46	63.78	60.48	61.68	43.02	64.23

EE = Expressions and Equations; GSP = Geometry & Statistics and Probability; RPNS = Ratios and Proportional Relationships & Number System

Table 11: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 7 Mathematics

Rep Cat	**Scale Score**	**All Students**	**Female**	**Male**	**Multi-Racial**	**American Indian/ Native Alaskan**	**Asian**	**Hispanic**	**African American**	**White**	**Pacific Islander**	**LEP**
Total	*N*	18,302	9,006	9,296	698	29	118	392	788	15,995	3	124
EE	*Mean*	498.76	499.61	497.94	480.92	481.42	564.16	488.04	464.06	501.20	505.87	448.52
	SD	83.68	81.42	85.81	86.85	96.97	77.68	83.12	81.70	83.06	27.36	79.96
G	*Mean*	501.07	503.29	498.92	485.63	486.43	572.26	488.81	472.01	503.08	476.84	457.37
	SD	76.53	74.09	78.76	77.10	103.67	80.62	75.72	73.40	76.13	42.79	73.72
RPNS	*Mean*	511.66	509.10	514.14	497.93	473.27	569.39	501.03	475.04	513.96	507.11	458.65
	SD	77.25	74.23	79.98	78.42	104.47	82.88	76.53	75.24	76.64	38.49	78.74
SP	*Mean*	504.90	505.42	504.39	489.14	485.43	568.45	492.68	468.78	507.36	512.92	457.07
	SD	82.55	78.94	85.90	84.67	107.02	88.79	80.66	80.78	81.91	85.84	73.48

EE = Expressions and Equations; G = Geometry; RPNS = Ratios and Proportional Relationships & Number System; SP = Statistics and Probability

Table 12: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 8 Mathematics

Rep Cat	Scale Score	All Students	Female	Male	Multi-Racial	American Indian/ Native Alaskan	Asian	Hispanic	African American	White	Pacific Islander	LEP
Total	*N*	18,742	9,031	9,711	732	20	120	440	718	16,409	9	133
EENS	*Mean*	535.89	537.15	534.72	522.23	509.39	609.09	523.25	501.67	538.14	521.19	492.16
	SD	88.69	85.45	91.59	90.13	85.72	100.20	86.17	83.01	88.31	56.82	78.23
F	*Mean*	532.26	533.01	531.56	519.86	495.04	610.98	518.84	494.36	534.53	540.81	489.57
	SD	92.22	88.52	95.54	92.97	84.80	108.50	90.24	86.68	91.69	62.46	80.39
GSP	*Mean*	530.19	534.03	526.61	519.82	477.59	597.12	522.26	490.49	532.51	517.49	484.01
	SD	86.69	82.57	90.22	85.79	101.69	92.23	84.75	82.88	86.30	75.97	80.31

EENS = Expressions and Equations & Number System; F = Functions; GSP = Geometry & Statistics and Probability

Table 13: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 5 Science

Rep Cat	Scale Score	All Students	Female	Male	African America n	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
Total	*N*	17,698	8,621	9,077	647	7	112	371	800	8	15,305	448	131
ESS	*Mean*	545.70	545.24	546.14	537.38	533.61	558.72	543.06	542.82	553.09	546.06	549.41	534.84
	SD	19.38	18.81	19.91	17.89	14.47	20.93	18.66	18.00	32.57	19.37	19.53	15.31
LS	*Mean*	545.34	545.58	545.12	536.72	545.01	557.73	543.39	543.38	543.91	545.65	549.24	533.75
	SD	19.02	18.40	19.59	18.23	17.87	18.41	18.54	17.80	23.28	19.03	17.92	14.60
PS	*Mean*	545.75	545.56	545.93	538.29	547.05	557.86	543.62	543.58	548.44	546.03	549.44	536.09
	SD	18.27	17.38	19.07	16.97	15.78	17.21	18.32	17.83	19.99	18.26	17.64	16.80

ESS = Earth Space Science; LS = Life Science; PS = Physical Science

Table 14: Mean and Standard Deviation of Reporting Category Scores by Demographic, Grade 8 Science

Rep Cat	Scale Score	All Students	Female	Male	African American	American Indian/ Native Alaskan	Asian	Hispanic	Multi-Racial	Pacific Islander	White	Declined to Report	LEP
Total	*N*	18,694	9,013	9,681	711	20	120	438	723	9	16,265	408	134
ESS	*Mean*	844.07	844.27	843.89	837.60	836.88	855.81	842.48	842.33	837.09	844.43	842.79	835.89
	SD	18.02	17.20	18.76	15.77	18.52	19.16	17.95	17.62	14.82	18.03	18.26	15.04
LS	*Mean*	843.42	843.99	842.88	836.57	839.71	857.40	841.81	841.43	838.27	843.77	842.77	833.71
	SD	18.82	17.91	19.62	17.25	16.81	22.42	18.49	18.00	14.07	18.81	18.79	15.18
PS	*Mean*	843.94	844.42	843.49	837.51	840.13	855.61	841.75	842.32	836.60	844.28	843.45	833.60
	SD	17.27	16.23	18.18	15.71	13.91	18.15	17.48	16.98	18.50	17.25	17.06	15.16

ESS = Earth Space Science; LS = Life Science; PS = Physical Science

Appendix G

Operational Item Exposure Rates

Table 1: Percentage of Items by Exposure Rate, ELA

Grade	Total N Items	Exposure Rate (Percentage of Students)					
		Unused	0 to 20	21 to 40	41 to 60	61 to 80	81 to 100
3	17,520	7.53	79.78	9.68	3.01	0.00	0.00
4	17,191	11.75	72.73	13.75	1.77	0.00	0.00
5	17,426	2.91	85.01	8.95	3.13	0.00	0.00
6	17,687	6.51	83.33	7.47	2.68	0.00	0.00
7	18,228	8.30	80.49	9.19	1.35	0.67	0.00
8	18,649	8.53	70.28	17.05	2.58	1.55	0.00

Table 2: Percentage of Item Exposure Rates by Grade, Mathematics

Grade	Total N Items	Exposure Rate (Percentage of Students)					
		Unused	0 to 20	21 to 40	41 to 60	61 to 80	81 to 100
3	17,540	5.21	92.65	2.14	0.00	0.00	0.00
4	17,339	6.37	91.90	1.74	0.00	0.00	0.00
5	17,715	2.32	92.87	4.81	0.00	0.00	0.00
6	17,711	1.33	94.37	4.30	0.00	0.00	0.00
7	18,286	4.82	83.53	9.44	2.21	0.00	0.00
8	18,741	2.65	89.40	7.77	0.18	0.00	0.00

Appendix H

DIF Statistics for Spring 2022 Field-Test Items

Table 1: ELA DIF Statistics

Grade	DIF Groups*	DIF Category		
		A	B	C
3	Female/Male	115	3	0
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White*	0	0	0
	Hispanic/White	82	3	0
	Multi-Racial/White	75	1	0
	SPED/Non-SPED	95	0	0
4	Female/Male	109	6	3
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	2	0	0
	Hispanic/White	81	6	0
	Multi-Racial/White	87	1	0
	SPED/Non-SPED	85	5	4
5	Female/Male	110	5	0
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	6	0	0
	Hispanic/White	109	4	2
	Multi-Racial/White	71	4	0
	SPED/Non-SPED	82	9	3
6	Female/Male	112	5	0
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	4	0	0
	Hispanic/White	85	7	1
	Multi-Racial/White	32	1	0
	SPED/Non-SPED	84	5	5
7	Female/Male	115	7	1
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	6	0	0
	Hispanic/White	85	4	1

Grade	DIF Groups*	DIF Category		
		A	B	C
	Multi-Racial/White	29	2	0
	SPED/Non-SPED	85	2	3
8	Female/Male	115	6	1
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	3	0	0
	Hispanic/White	71	6	2
	Multi-Racial/White	30	1	0
	SPED/Non-SPED	72	6	1

**DIF is not computed for items where N < 200 for either group*

Table 2: Mathematics DIF Statistics

Grade	DIF Groups*	DIF Category		
		A	B	C
3	Female/Male	81	3	0
	Asian or Pacific Islander/White*	0	0	0
	American Native/White	49	4	0
	African American/White	59	0	2
	Hispanic/White	52	1	1
	Multi-Racial/White	60	0	3
	SPED/Non-SPED	59	3	0
4	Female/Male	59	1	1
	Asian or Pacific Islander/White*	0	0	0
	American Native/White	33	3	1
	African American/White	36	0	1
	Hispanic/White	37	0	0
	Multi-Racial/White	56	0	0
	SPED/Non-SPED	34	3	1
5	Female/Male	124	5	2
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White*	0	0	0

Grade	DIF Groups*	DIF Category		
		A	B	C
	Hispanic/White	98	8	0
	Multi-Racial/White	50	0	0
	SPED/Non-SPED	105	1	3
6	Female/Male	35	3	2
	Asian or Pacific Islander/White*	0	0	0
	American Native/White	19	1	0
	African American/White	19	1	0
	Hispanic/White	20	0	0
	Multi-Racial/White	21	0	0
	SPED/Non-SPED	19	1	0
7	Female/Male	122	12	3
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	42	0	0
	Hispanic/White	31	1	0
	Multi-Racial/White	20	1	0
	SPED/Non-SPED	96	2	2
8	Female/Male	140	6	4
	Asian or Pacific Islander/White*	0	0	0
	American Native/White*	0	0	0
	African American/White	1	0	0
	Hispanic/White	32	0	0
	Multi-Racial/White	4	0	0
	SPED/Non-SPED	40	2	1

**DIF is not computed for items where N < 200 for either group*

Appendix I

Achievement Level Distribution Comparison

Between 2018, 2019, 2021, and 2022

Table 1: Spring 2018, Spring 2019, Spring 2021, and Spring 2022 Achievement Level Distribution, ELA

Grade	Year	Achievement Level			
		Does Not Meet Standard	Partially Meets Standard	Meets Standard	Exceeds Standard
3	2018	21%	32%	29%	18%
	2019	25%	32%	26%	17%
	2021	35%	31%	22%	12%
	2022	34%	29%	21%	15%
4	2018	26%	29%	24%	20%
	2019	24%	28%	25%	23%
	2021	33%	30%	21%	16%
	2022	30%	27%	23%	21%
5	2018	27%	29%	27%	17%
	2019	26%	27%	26%	21%
	2021	32%	28%	24%	16%
	2022	31%	28%	24%	16%
6	2018	27%	30%	29%	14%
	2019	25%	31%	31%	14%
	2021	27%	34%	28%	11%
	2022	27%	31%	30%	12%
7	2018	26%	31%	29%	15%
	2019	27%	31%	29%	13%
	2021	30%	32%	26%	12%
	2022	29%	31%	27%	13%
8	2018	27%	32%	28%	14%
	2019	25%	32%	29%	15%
	2021	27%	30%	28%	15%
	2022	31%	30%	25%	14%

Table 2: Spring 2018, Spring 2019, Spring 2021, and Spring 2022 Achievement Level Distribution, Mathematics

Grade	Year	Achievement Level			
		Does Not Meet Standard	Partially Meets Standard	Meets Standard	Exceeds Standard
3	2018	23%	29%	27%	21%
	2019	21%	29%	26%	25%
	2021	32%	29%	23%	16%
	2022	27%	27%	25%	21%
4	2018	22%	33%	22%	23%
	2019	20%	33%	22%	25%
	2021	31%	35%	18%	16%
	2022	26%	33%	20%	21%
5	2018	29%	31%	20%	20%
	2019	26%	34%	20%	20%
	2021	38%	33%	16%	13%
	2022	34%	31%	18%	17%
6	2018	34%	33%	19%	14%
	2019	34%	32%	19%	15%
	2021	46%	33%	14%	7%
	2022	42%	31%	16%	11%
7	2018	33%	31%	21%	14%
	2019	34%	30%	19%	17%
	2021	42%	32%	16%	10%
	2022	41%	30%	17%	12%
8	2018	34%	34%	14%	17%
	2019	33%	31%	15%	21%
	2021	45%	31%	12%	12%
	2022	43%	30%	12%	15%

Appendix J

Calibration Group Means and SD for Spring 2022 Field-Test Items

Table 1: Calibration Group Means and Standard Deviations, ELA Grade 3

Year	Group	Mean	SD
2022	New Hampshire	-0.15	0.98
2022	North Dakota	-0.30	0.87
2022	West Virginia	-0.42	0.98
2022	Wyoming	0.03	0.92

Table 2: Calibration Group Means and Standard Deviations, ELA Grade 4

Year	Group	Mean	SD
2022	New Hampshire	-0.22	0.96
2022	North Dakota	-0.40	0.87
2022	West Virginia	-0.45	0.99
2022	Wyoming	-0.03	0.90

Table 3: Calibration Group Means and Standard Deviations, ELA Grade 5

Year	Group	Mean	SD
2022	New Hampshire	-0.12	0.94
2022	North Dakota	-0.25	0.87
2022	West Virginia	-0.41	0.96
2022	Wyoming	0.05	0.88

Table 4: Calibration Group Means and Standard Deviations, ELA Grade 6

Year	Group	Mean	SD
2022	New Hampshire	-0.03	0.94
2022	North Dakota	-0.18	0.87
2022	West Virginia	-0.37	0.97
2022	Wyoming	0.05	0.88

Table 5: Calibration Group Means and Standard Deviations, ELA Grade 7

Year	Group	Mean	SD
2022	New Hampshire	-0.22	0.93
2022	North Dakota	-0.41	0.91
2022	West Virginia	-0.46	0.99
2022	Wyoming	-0.08	0.97

Table 6: Calibration Group Means and Standard Deviations, ELA Grade 8

Year	Group	Mean	SD
2022	New Hampshire	-0.14	0.95
2022	North Dakota	-0.36	0.85
2022	West Virginia	-0.41	0.97
2022	Wyoming	0.03	0.92

Table 7: Calibration Group Means and Standard Deviations, Mathematics Grade 3

Year	Group	Mean	SD
2022	New Hampshire	0.06	1.03
2022	North Dakota	-0.09	0.94
2022	West Virginia	-0.25	1.08
2022	Wyoming	0.31	1.04

Table 8: Calibration Group Means and Standard Deviations, Mathematics Grade 4

Year	Group	Mean	SD
2022	New Hampshire	-0.06	1.03
2022	North Dakota	-0.20	0.96
2022	West Virginia	-0.37	1.09
2022	Wyoming	0.19	1.02

Table 9: Calibration Group Means and Standard Deviations, Mathematics Grade 5

Year	Group	Mean	SD
2022	New Hampshire	-0.17	1.03
2022	North Dakota	-0.28	0.95
2022	West Virginia	-0.51	1.10
2022	Wyoming	0.14	1.05

Table 10: Calibration Group Means and Standard Deviations, Mathematics Grade 6

Year	Group	Mean	SD
2022	New Hampshire	-0.37	0.98
2022	North Dakota	-0.45	0.90
2022	West Virginia	-0.78	1.05
2022	Wyoming	-0.07	0.97

Table 11: Calibration Group Means and Standard Deviations, Mathematics Grade 7

Year	Group	Mean	SD
2022	New Hampshire	-0.33	1.09
2022	North Dakota	-0.40	1.01
2022	West Virginia	-0.68	1.13
2022	Wyoming	-0.09	1.10

Table 12: Calibration Group Means and Standard Deviations, Mathematics Grade 8

Year	Group	Mean	SD
2022	New Hampshire	-0.39	1.09
2022	North Dakota	-0.48	0.96
2022	West Virginia	-0.72	1.15
2022	Wyoming	-0.07	1.04

Appendix K

Investigating the Effects of Dictionary Availability on Item Performance

USOE would like to provide dictionary access to students during SAGE test administrations. The goal of providing a dictionary is to help improve access to test content for English language learners (ELLs). Providing students with a dictionary may reduce construct irrelevant barriers to accessing test content for ELL students, resulting in more valid estimates of student ability across subject area assessments. This memo describes the results of an initial investigation of the effects of providing students access to a dictionary on their performance on test items.

The principle concern with providing students access to a dictionary is that the assessed constructed may be altered. For example, if an item is designed to assess whether students can infer the meaning of complex terms from passage context, providing students a dictionary may changes the measured construct considerably so that the item measures instead dictionary usage. For ELA items in particular, it may be necessary to reevaluate the alignment of items in an assessment context in which students are provided with a dictionary. It is also worth noting that dictionary access may not simply alter the alignment of some items, but may render some standards unmeasurable, especially those related to acquisition of vocabulary and inferring meaning from context.

To identify whether an accommodation removes a construct irrelevant barrier to accessing test content or alters the construct being assessed can be evaluated by whether the effects of an accommodation are isolated to the group for whom the accommodation is intended or whether the accommodation impacts test performance across groups. When the impact of a test accommodation on student performance is localized to the population with the access limitation, then the accommodation can be said to mitigate construct irrelevant barriers to test content. However, when an accommodation impacts student performance across the general population, the accommodation is likely altering the construct assessed by the test.

To investigate whether providing students a dictionary reduces construct irrelevant barriers to accessing test content for English language learners, ELL and non-ELL students in participating schools were administered an abbreviated SAGE assessment, with students randomly assigned to a dictionary treatment condition.

Design

The study was conducted as a 2 (ELA vs. non-ELA) by 2 (dictionary vs. no dictionary) between subjects design. Students were randomly assigned to the dictionary vs. no dictionary treatment condition. Students assigned to the dictionary condition could use the online Merriam-Webster dictionary to look up the meaning of any word presented during the test administration. To control for wide variation in student achievement and increase the power of the design, student test scores from the spring 2014 administration of SAGE were included as covariates. Responses to math items were covaried using spring 2014 math scale scores, with responses to ELA and science items covaried using spring 2014 ELA and science scale scores, respectively.

Sample

Participation in the study was restricted to students eligible for the grade 6 SAGE assessments. USOE identified a sample of schools for participation in the study. Classification of students as English language learners (ELLs) was based on the demographic information provided in the test student enrollment files uploaded by districts.

The final sample included 1,341 students, including 323 (24%) ELL students, 962 (72%) non-ELL students, and 56 (4%) students with missing ELL information. Students were randomly assigned to treatment condition, with 688 (51%) students provided dictionary access, and 653 (49%) students assigned to the no dictionary condition. The distribution of ELL and non-ELL assigned to treatment and control groups are shown in the table 1.

Table 1. Assignment of Treatment Condition by ELL Status

ELL Status	Treatment Condition	
	Dictionary	No Dictionary
Non-ELL	493	469
ELL	171	152
Missing	24	32

Materials

A 24-item multi-subject test form was developed to investigate the effect of dictionary availability across subject area assessments. The assessment included an 8-item passage set to measure reading comprehension, as well as eight items each to measure math and science content. Passage and item selection were directed toward identification of items with subject specific and technical vocabulary for which students could use the dictionary to identify the meaning.

Test Delivery System

The assessment was administered using the same test delivery system used to administer the SAGE operationally. Item groups were selected randomly, so that the position of items varied across test administrations.

Analyses

For each item response, the likelihood providing a correct response was analyzed using a Probit random effects model. Since each student was administered multiple items, and the likelihood of correct responding across items within a student is not independent (e.g., high ability students have a higher likelihood of responding correctly across all items), item responses were grouped by student.

In the base model, the scored item response dependent variable was predicted by

1. students' previous year SAGE scale score in the appropriate subject area assessment (i.e., response to a science item was predicted by previous year science achievement), since likelihood of correct responding is determined in part by student ability;
2. the item on which the response is based, since likelihood of responding correctly is determined also by the characteristics of the item, including the item difficulty;
3. a main effect for student ELL status (ELL or non-ELL), to determine whether the ELL status affects likelihood of correct responding independent of other effects;
4. a main effect for treatment condition (dictionary or no dictionary), to determine whether the accommodation increases the likelihood of correct responding generally;
5. an interaction term between ELL status and treatment condition, to identify whether the treatment differentially affected ELL students.

In a second model, we also investigated whether there might be differential effects of dictionary access for ELL students across subject area assessments, so the second model also included:

6. an interaction term between subject area and treatment condition, to identify whether the treatment differentially affected student performance across subject areas;
7. three-way interaction terms between ELL status, subject area, and treatment condition, to determine whether the dictionary access differentially affected ELL performance across subject areas.

Results

The overall base model was statistically significant ($\chi^2_{(29)}$ = 3942.06; p < .0000). Table 2 shows the regression parameters and statistical tests for each of the modeled effects. As expected, students' ability estimates from the spring 2014 SAGE assessments significantly predicted their likelihood of responding correctly to test items, with previously high achieving students more likely to provide a correct response than lower achieving students. Also as anticipated, the items themselves influenced the likelihood of providing a correct response, with students more likely to respond correctly to easy than difficult items, for example. ELL status also contributed to the likelihood of responding correctly, indicating that ELL students were less likely to answer test items correctly even when accounting for previous achievement. The treatment main effect was not significant. Providing students access to a dictionary did not significantly increase their likelihood of responding correctly. The treatment by ELA status interaction, indicating differential effects of dictionary access for ELL students, also did not reach significance.

Table 2. Parameter Estimates for the Base Model

Parameter	Coefficient	Std. Error	*z*	P>|*z*|
Intercept	-8.952	0.5311	-16.85	0.0000
Math Scale Score	0.003	0.0004	7.38	0.0000
ELA Scale Score	0.002	0.0002	7.09	0.0000
Science Scale Score	0.009	0.0007	11.73	0.0000
Treatment	0.028	0.0242	1.14	0.2540
ELL Status	-0.113	0.0356	-3.19	0.0010
ELL*Treatment Interaction	0.029	0.0496	0.59	0.5540
Item_1	0.759	0.0550	13.81	0.0000
Item_2	-0.601	0.0570	-10.55	0.0000
Item_3	-0.068	0.0544	-1.25	0.2100
Item_4	0.336	0.0540	6.22	0.0000
Item_5	0.113	0.0540	2.10	0.0360
Item_6	-0.153	0.0543	-2.81	0.0050
Item_7	-0.385	0.0553	-6.95	0.0000
Item_8	0.550	0.0545	10.09	0.0000
Item_9	0.704	0.0546	12.88	0.0000
Item_10	0.139	0.0541	2.58	0.0100
Item_11	-0.233	0.0550	-4.24	0.0000
Item_12	-0.690	0.0584	-11.81	0.0000
Item_13	0.276	0.0539	5.11	0.0000
Item_14	0.256	0.0538	4.76	0.0000
Item_15	-0.579	0.0569	-10.18	0.0000
Item_16	-0.039	0.0546	-0.71	0.4800
Item_17	0.264	0.0539	4.90	0.0000
Item_18	0.086	0.0542	1.59	0.1110
Item_19	0.083	0.0541	1.52	0.1270

Parameter	Coefficient	Std. Error	*z*	P>|*z*|
Item_20	-0.287	0.0549	-5.23	0.0000
Item_21	-0.390	0.0560	-6.96	0.0000
Item_22	-1.295	0.0678	-19.11	0.0000
Item_23	0.013	0.0542	0.24	0.8110

The full model, which specified differential treatment by ELL interactions across subject area assessments was also statistically significant ($\chi^2_{(33)} = 3947.21$; p < .0000). However, the likelihood ratio between the base and full model was not significant ($\chi^2_{(4)} = 4.66$; n.s.), indicating that the full model did not account for significant variation beyond that of base model. Table 3 shows the parameter estimates and statistical tests for the modeled effects.

As in the base model, students' prior ability estimates significantly predicted the likelihood of responding correctly to the test items presented. Also as with the base model, the likelihood of providing a correct response was item dependent. ELL status continued to contribute to the likelihood of responding correctly. The treatment main effect was not significant. Providing students access to a dictionary did not significantly increase their likelihood of responding correctly. Moreover, there was no statistical support for subject area by treatment interactions, or differential effects of dictionary access for ELL students across subject area assessments.

Table 3. Parameter Estimates for the Full Model

Parameter	Coefficient	Std. Error	*z*	P>|*z*|
Intercept	-8.945	0.5313	-16.84	0.0000
Math Scale Score	0.003	0.0004	7.37	0.0000
ELA Scale Score	0.002	0.0002	7.11	0.0000
Science Scale Score	0.009	0.0007	11.73	0.0000
Treatment	0.028	0.0337	0.84	0.4030
ELL Status	-0.112	0.0356	-3.16	0.0020
Math*Treatment Interaction	-0.028	0.0428	-0.66	0.5090
ELA*Treatment Interaction	0.024	0.0414	0.57	0.5660
Science*ELL*Treatment Interaction	-0.025	0.0620	-0.40	0.6870
ELA*ELL*Treatment Interaction	0.070	0.0616	1.13	0.2590
Math*ELL*Treatment Interaction	0.040	0.0672	0.60	0.5480
Item_1	0.730	0.0583	12.53	0.0000
Item_2	-0.601	0.0569	-10.55	0.0000
Item_3	-0.097	0.0577	-1.67	0.0950
Item_4	0.330	0.0573	5.76	0.0000
Item_5	0.107	0.0573	1.87	0.0620
Item_6	-0.181	0.0576	-3.14	0.0020
Item_7	-0.384	0.0553	-6.95	0.0000
Item_8	0.521	0.0578	9.02	0.0000
Item_9	0.698	0.0579	12.05	0.0000
Item_10	0.111	0.0575	1.93	0.0540
Item_11	-0.240	0.0582	-4.12	0.0000
Item_12	-0.690	0.0584	-11.81	0.0000
Item_13	0.269	0.0572	4.71	0.0000

Parameter	Coefficient	Std. Error	*z*	P>|*z*|
Item_14	0.250	0.0571	4.38	0.0000
Item_15	-0.579	0.0569	-10.18	0.0000
Item_16	-0.046	0.0579	-0.79	0.4310
Item_17	0.236	0.0572	4.12	0.0000
Item_18	0.079	0.0574	1.38	0.1660
Item_19	0.054	0.0575	0.94	0.3480
Item_20	-0.287	0.0549	-5.23	0.0000
Item_21	-0.417	0.0592	-7.05	0.0000
Item_22	-1.295	0.0678	-19.11	0.0000
Item_23	0.013	0.0542	0.24	0.8120

Conclusion

The results of this investigation did not find evidence that providing students with access to a dictionary would differentially affect the performance of ELL students on the SAGE assessments. However, given the relatively low power of the study afforded by small sample size, there is a very real possibility that the study was not sufficiently sensitive to detect real effects, whether main effects of the treatment condition, differential effects of treatment by ELL status, or even differential effects of treatment across subjects by ELL status. Affirming that a dictionary accommodation removes construct irrelevant barriers to test content for ELL students without altering the construct being assessed may require very much larger samples of students. Moreover, effects of dictionary access could vary across grade level assessments as well, further complicating the situation.

Because the risk of a type II error (e.g., failing to reject a false null hypothesis) is substantial, care needs also to be taken to avoid over-interpretation of null results. One could, for example, be tempted to interpret the null results as indicating that, because there were no observed effects for dictionary access on student performance, students can safely be offered the dictionary accommodation without altering the measured construct. Such interpretations are always risky, and are only warranted when the risk of type II error is very low, which is not the case in this study.

Finally, providing students with a dictionary could alter the standards alignment for, and student performance on, only a subset of items, especially in ELA, and such effects would likely only be observed in a more focused investigation of item types. For example, the alignment of items measuring student ability to infer meaning of words from context or demonstrate understanding of grade level vocabulary would certainly be affected by providing students with a dictionary. Moreover, the difficulty of such items would also likely be affected by availability of a dictionary. But such effects would be difficult to detect except in study specifically targeting items measuring those impacted standards. Should USOE consider providing a dictionary during SAGE administrations, it would be necessary to ensure that the alignment of test items, especially in ELA, is still valid.

In the absence of evidence indicating that providing a dictionary impacts student performance, USOE's Technical Advisory Committee (TAC) recommended that USOE make the dictionary tool available to all students. The dictionary tool was available to all students for the spring 2015 SAGE administration.

Appendix L

Effectiveness of Computer-Based, Pop-Up Glossaries

Technical Report

On the Reliable Identification and Effectiveness of Computer-Based, Pop-Up Glossaries in Large Scale Assessments

Dale Cohen

University of North Carolina Wilmington

Jon Cohen

American Institutes for Research

Alesha Ballman and Frank Rijmen

Cambium Assessment, Inc.[1]

JUNE 2019

[1] Work was completed while they were employed with the American Institutes for Research.

Contents

Tables

Figures

On the Reliable Identification and Effectiveness of Computer-Based, Pop-Up Glossaries in Large Scale Assessments

Introduction

There are 4.6 million public school students with limited English proficiency (termed English Learners, or ELs) in the United States (U.S. Department of Education National Center for Education Statistics, 2017). Typically, ELs perform worse on required statewide standardized assessments than native English-speaking students (Avenia-Trapper & Llosa, 2015; Kieffer, Lesaux, Rivera, & Francis, 2009; Martiniello, 2008; Sato, Rabinowitz, Gallagher, & Huang, 2010). It has been hypothesized that the performance gap between ELs and non-ELs can, at least in part, be accounted for by the added difficulty of the linguistic structure and cultural bias of statewide assessment items, rather than the content being measured (Abedi, Courtney, & Leon, 2003; Abedi & Gándara, 2006; Abedi, Hofstetter, & Lord, 2004; Johnson & Monroe, 2004; Martiniello, 2009; Wolf & Leon, 2009). In an attempt to mitigate the influence of linguistic structure and cultural bias, the *Every Student Succeeds Act* of 2015 (ESSA) mandates that states provide ELs appropriate accommodations when administering statewide assessments (Council of Chief State School Officers, 2016). The form that the appropriate accommodation should take, however, is not specified. Here, we assess the effectiveness and impact on validity of a computerized pop-up glossary accommodation for EL students in a controlled, large-scale, randomized trial.

Assessment accommodations are adjustments of the assessment, environment, or procedure that permit a student to access the content of an item without changing the construct that the item is measuring. Assessment accommodations may range from adjusting the assessment items (e.g., adjustments to the language of the test, adding graphics, glossaries) to changing the conditions under which the assessment is taken (e.g., allowing additional time for completion) (Kieffer et al., 2009; Li & Suen, 2012; Pennock-Roman & Rivera, 2011). Accommodations that involve changes to the assessment items (i.e., Item Accommodations) must meet two criteria before they can be implemented: (1) the accommodation must not affect the validity of claims about the construct measured, and (2) the accommodation must demonstrate that it effectively mitigates the barriers to the students' access of the item content (termed *effectiveness* here) (Abedi, 2012; Keiffer et al., 2009; Sireci & Faulkner-Bond, 2015). Accommodations that mitigate the barriers to the students' access of the item content will, presumably, improve performance of EL students. However, a performance improvement alone may also indicate that the accommodation changes the construct that the item is measuring. To ensure that the accommodation did not change the construct the item measures, the accommodation should not influence the performance of non-EL students (Abedi & Ewers, 2013; Abedi, 2012; Sireci, Scarpati, & Li, 2005). Thus, an optimal pattern of results to validate any accommodation is both an improvement in EL performance (effectiveness) and no influence on non-EL performance (validity).

Computer-based, pop-up glossaries are perhaps the most promising accommodation aimed at mitigating the influence of linguistic structure and cultural bias on the performance of EL students on statewide assessments (Abedi, 2012). Glossary accommodations, "… contain simplifications of non-content related terms and phrases deemed complex for students. These simplifications include clarifications of complex sentences and synonyms or specific, context-related definitions of words" (p. 262, Cohen, Tracy, & Cohen, 2017). Both paper-based (e.g., Kiplinger, Haug, & Abedi, 2000; Pennock-Roman & Rivera, 2011) and computer-based (e.g., Abedi, 2009; Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007; Pennock-Roman & Rivera, 2011) glossaries have demonstrated some effectiveness in laboratory settings. Until recently, it remained unclear whether the results of these studies generalized to field-based assessments.

Cohen et al. (2017) conducted a large-scale, randomized controlled trial experiment to assess the effectiveness and impact on construct validity of a computer-based, pop-up glossary accommodation in a statewide assessment. The authors sampled all students taking the English Language Arts (ELA) and Mathematics online statewide accountability assessment in grades 3 and 7. About 12 percent of these students were ELs. Students were randomly assigned to either a Glossary or Control condition. All students were presented field test items. For those in the Glossary condition, a subset of the field test items contained a pop-up glossary accommodation. For those in the Control condition, the same field test items did not contain the accommodation. Counterintuitively, the results revealed that the computer-based, pop-up glossary accommodations inhibited EL student performance in Mathematics for both grades 3 and 7 and in ELA for grade 3. The computer-based, pop-up glossary accommodation only improved EL student performance in ELA for grade 7, while not influencing performance of non-EL grade 7 students.

The general inhibitory influence of the pop-up glossaries on EL students' performance was unexpected (Cohen et al., 2017). The authors suggest several possible sources of the inhibitory effect; these range from the general ineffectiveness of a pop-up glossary in a field-based assessment to a "cognitive load" hypothesis. The "cognitive load" hypothesis suggests that added cognitive resources, such as attention and memory, are required to effectively make use of pop-up glossaries. The extra cognitive load may have diverted cognitive resources that would typically be used to answer the item correctly. This effect may have been exasperated by the restrictions on the words glossed in the item. For example, in the Mathematics assessment, only construct-irrelevant words were glossed. This may have misdirected students' attention away from construct relevant words, thus inhibiting students' performance. For example, the word "hippopotamus" may be glossed in a math word problem because it is a low frequency/uncommon word, but its meaning may be irrelevant to the successful completion of the item. As such, a student who spent cognitive resources using the glossary in this instance may have put themselves at a disadvantage when completing the item. In contrast, the authors found that pop-up glossaries for grade 7 ELA were effective and otherwise did not appear to affect the construct measured. The words in these items were glossed often because they used figurative language. Because these glosses may have been directly relevant to understanding the item correctly, students who spent cognitive resources using the glossary in this instance may have benefited from the accommodation when completing the item.

Here, we replicated and extended Cohen et al. (2017). We recognized that the effectiveness of glossaries and their impact on the validity of claims about the construct depend critically on the specific rules for glossing words. We focused our efforts on reliably identifying words in the item for which glossing will aid EL students while not influencing construct the item measured. Because our focus was EL students, our primary targets for glossing were *culturally bound language*. We defined culturally bound language as, "Particular words and phrases that are well known to speakers for whom English is their primary language but would not be to a new arrival to the country or language minority." These words are neither literally interpretable, nor can their interpretation be immediately inferred, and therefore the linguistic structure and cultural bias for these words must be mitigated in order for ELs to be able to access the content of the test items.

Identifying non-literal language is notoriously difficult. For example, in a paper describing the performance of an algorithm that identifies figurative language, Neuman et al. (2013) trained raters to identify figurative language. The authors found that the trained raters missed over 60% of the metaphors in the text. This difficulty in achieving reliability in coding figurative language is relatively common (e.g., Loenneker-Rodman, & Narayanan, 2010; Pragglejaz Group, 2007; Roberts & Kreuz, 1994).

The source of the difficulty with reliably identifying figurative language is not known. However, published accounts often mention that judges are extensively trained on the various ways figurative language can manifest in text (e.g., Barlow, 1971; Neuman et al., 2013). This suggests that the difficulty arises, not from a misunderstanding of what constitutes figurative language, but from the automated interpretation of common non-literal phrases as literally meaningful. That is, when one reads the phrase, "pay attention," the reader may not recognize that the verb "pay" cannot be literally applied to the noun "attention." This lack of recognition may be a result of the frequency that the phrase is used. High frequency/common, non-literal phrases may be interpretively invisible to the reader. It is likely this obstacle that must be overcome when developing a reliable coding procedure for non-literal language. Therefore, as a precursor to studying the effect of glossaries, we established a coding method for non-literal language that was sufficiently reliable in order to investigate the effect of glossaries effectively.

In the Coding Procedure, we develop a reliable coding procedure for culturally bound language. The Coding Procedure three phases: each phase implements a change in the procedure intended to increase the reliability of the coding. In the current study, we implement a computerized, pop-up glossary accommodation of the words identified as requiring a gloss in the Coding Procedure in a randomized, controlled design embedded within an operational statewide assessment. The current study implements the same pop-up glossary accommodation as Cohen et al. (2017), with a more selective and reliable method for identifying the words to be glossed and assesses effectiveness and impact on validity across more grades (3-11) and items.

Coding Procedure: Phase I

In Phase I of the Coding Procedure, we attempted to develop a reliable protocol for identifying culturally bound language.

Method

Participants

"Raters" consisted of 10 temporary employees (M age = 29.10 years, SD age = 6.65 years) at the American Institutes for Research (AIR). Out of the 10 raters, 2 identified as female and 8 identified as male. In terms of raters' highest level of education, 5 reported that they held a bachelor's degree, and 5 reported they held a master's degree.

Materials

Training presentation. The training consisted of a PowerPoint presentation that covered the following topics.

Culturally bound language. The training presentation focused on the identification of culturally bound language. We defined culturally bound language as particular words and phrases that are well known to speakers for whom American English is their primary language but would not be to a new arrival to the country or language minority. For the purposes of this project, culturally bound language was operationalized as words and phrases which are not literally interpretable, and their non-literal interpretation cannot be immediately inferred from the context.

The raters were trained to determine whether the meaning can be inferred from the context by first finding the literal meaning of each word. They were then trained to ask the following questions:

> *Is the word or phrase used with its literal meaning?* Taking the sentence "the task was a piece of cake" for example. In this sentence, "piece of cake" is used with its non-literal meaning of "easy" as opposed to the literal meaning of "a piece that is cut from a cake." Raters used Google search to aid their judgment of non-literal language identification. Specifically, if the raters were unsure whether the word's meaning was literal, the raters entered a word or phrase into Google search and if the meaning in the text was not one of the first few definitions of

the word or phrase listed, the determination was that it is likely non-literal language.

If the word or phrase was identified as non-literal, then the following two questions were asked:

1. *How close is the relation between the non-literal meaning of each word and the literal meaning?* Taking the expression “spinning a tale,” for example, the literal meaning of “spinning” is not closely related to the non-literal meaning “creating.” Therefore, the meaning of “spinning a tale” also cannot be inferred from the phrase. However, taking the expression “soulful tale,” for example, the literal meaning of “soulful” is closely related to the non-literal meaning “deeply meaningful.” Therefore, the meaning of “soulful tale” can be inferred from the phrase.
2. *Is the word or phrase so commonly used that it has become an accepted part of the language?* For example, the phrase a “lead balloon” has the literal interpretation of a heavy balloon; but, the American cultural interpretation of this phrase is “failure will occur.” The cultural interpretation may be inferred from the individual words in the phase, but this phrase has become an accepted part of the language because of common usage. This phrase is identified as a Frozen Trope.

The word or phrase was “Tagged” as culturally bound if it was identified as (1) non-literal, and (2) the meaning could not be inferred from the context, or (3) it was a Frozen Trope. The word or phrase was tagged for the first two criteria because they required cultural knowledge to interpret correctly. Frozen Tropes were tagged because the native English speakers do not have to make the inference to understand Frozen Tropes, but the non-native English speakers do. As such, the non-native English speaker is at a disadvantage (relative to the native English speaker) when interpreting a Frozen Trope.

To aid raters in making their decisions about whether a word or phrase was non-literal, raters were also trained to identify the following non-literal categories. If a word or phrase belonged to one of these categories, raters were trained to tag them as they are culturally bound elements.

Idioms, which were defined as a group of words established by usage as having a meaning not deducible from those of the individual words. For example, the phrase “raining cats and dogs” is non-literal in meaning; the meaning of each word in the idiom is unrelated to the overall meaning of the idiom. The raters were provided with extensive lists of idioms to reference in support of idiomatic expression identification (Dixson, 2003; Holmes, n.d.; “Resource list. Idioms: Figurative language,” n.d.).

Euphemisms, which were defined as mild or pleasant language used instead of language that is unpleasant or offensive. This substitution requires culture specific knowledge that certain words and phrases carry negative meanings or connotations. For example, it is generally recommended when referring to a job from which one was terminated, to say that the company “downsized,” rather than saying that you were “fired.” The raters were trained to tag all words and phrases that they identified as euphemisms.

Acronyms, which were defined as letters that stand for words. For example, using the acronym “TP” to stand for toilet paper. The raters were instructed to tag all acronyms as needing glossing, as acronyms vary across cultures and therefore require culturally specific knowledge to understand.

Item/proper names, which were defined in terms of this project as the names of items or people that are intimately tied to the culture. Examples include John Adams (president), the House of Representatives, and the White House; these require cultural background knowledge to comprehend. Raters were trained to tag all instances of item/proper names when they were not the focus of the passage content.

Culturally specific terms, which were defined as the cultural use of terms to substitute for a descriptive word or phrase. Examples of culturally specific terms include the term "head count," as used in the phrase "let's take a head count to see how many showed up." This also extends to proper names, such as "Uncle Tom" or "Jim Crow." Raters were trained to tag instances of culturally specific terms such as the substitution of parts for wholes, proper names, and products or services.

Onomatopoeias, which were defined as the use of words whose sounds echo the sense. For example, using the word "zoom" to describe a vehicle driving past at a high rate of speed. Onomatopoeias are culturally bound because different cultures express sounds in different ways. Raters were trained to tag all instances of onomatopoeias, as onomatopoeic expressions vary across nationalities and cultures.

The training presentation also included a flow chart to aid raters in their determination of whether a word or phrase is culturally bound. This flow chart provided an illustrated overview of the questions they were to ask and the determinations that could be made as result of answering those questions (see Figure 1).

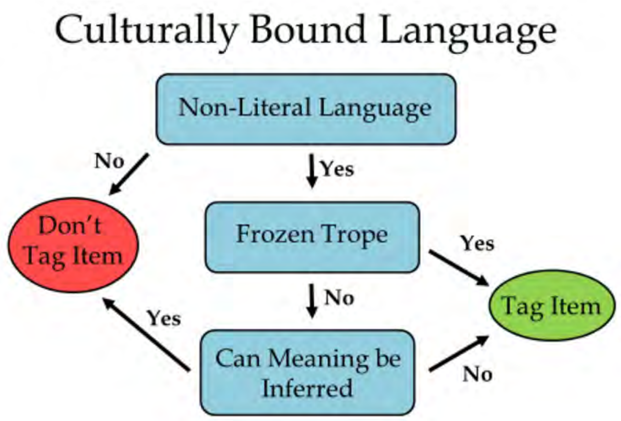


Figure 1. Flow chart for Phase I of Coding Procedure

Practice items. The raters were provided with paper versions of English Language Arts items from the AIR Common Core State Assessment. These items were chosen because they were released state assessment practice test items. There were 187 practice items selected from grades 3 to 11 for training, and the paper versions of the items were in the same format that students would see the items. Each item was printed on a separate page. The passages associated with the items were also provided to aid in the determination of whether meaning can be inferred from the context. Each passage contained a range of 15 to 2,500 words and the items contained a range of 25 to 2,000 words.

Certification test. A certification test was created to identity when each rater had achieved an acceptable level of tagging sensitivity and reliability. Forty items were selected from grades 4, 5, 6, 8, and 10 from the AIR Common Core State Assessment for the tagging certification test. The raters were provided with paper versions of the certification test items, which were in the same format as the practice items. The related passages were also included for rater reference. The items contained a total of 3,468 words; raters were required to identify a total of nine tagged words and phrases that had previously been identified and documented by project staff.

Procedure

Overview of training. Raters were assembled in a conference room and project staff explained the purpose and scope of the project. Staff described that raters would be tagging culturally bound expressions in English Language Arts (ELA) and Mathematics field test items to provide pop-up glossaries for those expressions. Raters were told that for the ELA tests, they would only be tagging the response items and not the passages themselves. The passages were not reviewed for tagging as the passage content had undergone fairness review processes; therefore, content related to the passage theme, even if it appears culturally bound, is assumed to be culturally fair.

Training took place over the course of one week, and the sequence of activities was as follows. Day one consisted of project staff going over the PowerPoint presentation with the raters and working through selected practice items with the raters as a group. Days two and three consisted of the raters working on practice items independently following the principles outlined from the presentation, and then discussing tagging determinations as a group. On days four and five, raters were tasked with individual completion of the certification test items to determine their tagging accuracy. The raters were provided paper copies of the certification test items, and they highlighted all culturally bound words and phrases following the training they had received. Raters submitted their completed work to staff one item page at a time, so that agreement with tags made by staff could be monitored without delay.

Rater process. The raters determined whether a word or phrase required tagging by going through the following steps. First, raters determined whether words were used with their literal meaning. If it was not clear whether words were used with their literal meaning, raters entered each word or phrase into a Google search. If the meaning of the word or phrase in the text was listed as one of the first few definitions in the search results, the word or phrase was identified as literal in meaning. The Google search results aid in this identification because common or literal meanings are often the most frequent definition; therefore, we assumed that the literal meanings will be one of the first definitions listed, as they are in other dictionaries ("There is no such thing as 'the dictionary,'" n.d.). If the word or phrase was determined to be used literally, the rater moved to the next word or phrase. If the meaning was identified as non-literal, the rater would proceed to make two additional judgments.

First, the rater would determine if the non-literal and literal meanings were similar. This determination helps establish whether the meaning of the word or phrase can be inferred. If one cannot infer a non-literal meaning from the text, then we identified the text as requiring culturally bound knowledge. If the meanings were closely related, then the non-literal meaning could be inferred from the text and the rater move on to the next question. If the meanings were not closely related, the rater tagged the word or phrase.

Second, the rater would determine whether the word or phrase was a Frozen Trope. Raters were to decide whether the word has become an accepted part of the language, so that native speakers immediately understand the meaning, but non-native speakers would not. If the rater determined that the word was a Frozen Trope, they would tag the word.

The raters would also follow the steps outlined for identifying additional non-literal categories. If the rater identified the word or phrase as an idiom, euphemism, acronym, item/proper name, culturally specific term, or an onomatopoeia, they were to tag it. To aid in determining whether a word or phrase was an idiom, they were to make the determination using one of the idiom lists that were provided.

Results and Discussion

After the raters completed the certification test, the number of tags and percent agreement with the tags made by the staff were computed. The number of tags made by raters ranged from two to 11, with an average of 6.9. The percent agreement with staff was computed as the number of unique words that were tagged both by the rater and by staff, divided by the number of unique words that were tagged by either the rater or staff. For example, Rater 9 tagged five words across the certification set of 40 items and 3,468 words; only one of those tagged words appeared in the set of nine words tagged by staff, and the four other tags were unique to the rater. So for Rater 9, the percent agreement was 1/(9+4) = .08. This measure of percent agreement differs from how percent agreement is usually computed in that the number of words that were tagged neither by staff nor the rater were ignored. Because most of the words were not tagged by anyone, including the number of words tagged by neither staff nor the rater would have resulted in very high percent agreements regardless of whether staff and rater were tagging the same set of words. Note that by excluding these counts, our proposed measure of percentage agreement is conservative. As a tentative criterion, an exact agreement of .50 was used for certification. An example of reaching this criterion would be a case were both the staff and rater tagged 12 words, with eight of those words (two-thirds) in common. Across raters, the percent agreement ranged from zero to .14, with an average of .05.

Because agreement with the staff certification tags was extremely poor, staff reviewed the principles outlined in the training. Raters were then put in pairs to re-evaluate the items in the certification set to facilitate dialogue of where misunderstandings remained. Agreement with the certification remained extremely poor, even after several iterations and group discussions. Having raters discuss the certification test items in pairs only marginally improved results; the percentage agreement ranged from .08 to .15 across the five pairs of raters, with an average of .11

Coding Procedure: Phase II

In Phase II of the Coding Procedure, we adjusted the tagging procedure with the expectation of increasing rater reliability.

Method

Participants

The raters in Phase II were same 10 temporary employees that were included in Phase I of the project.

Materials

The materials in this phase of the project were identical to Phase I, with the following exceptions. Merriam-Webster's Learner's Dictionary replaced the Google search for words. Merriam-Webster's Learner's Dictionary is designed for English language learners and provided raters with more consistent results than Google search. In addition, because raters had a relatively high miss rate for tagged words in Phase I, the raters were required to look up each word in Merriam-Webster's Learner's Dictionary. The materials described below reflect these and other, smaller, changes.

Training presentation. The PowerPoint presentation was similar to that of Phase I, with the following exceptions. First, the flow chart in the PowerPoint presentation from Phase I was expanded to reflect the required use of Merriam-Webster's Learner's Dictionary. Included was a note that every word was to be defined in isolation, along with the online dictionary website and instructions for defining the words (see Figure 2). Second, "phrasal verbs" were added to the list of culturally bound language to be tagged by the raters. Phrasal are idiomatic phrases consisting of a verb and another element, such as an adverb (such as "break down"), a preposition (such as "see to"), or a combination of both (such as "look down on"). This change was added into the flow chart; the Learner's dictionary lists phrasal verbs and idioms similarly at the bottom of the dictionary entry. Additionally, frozen tropes were removed from the flow chart, as it was determined that the use of the Learner's Dictionary would serve in this identification.

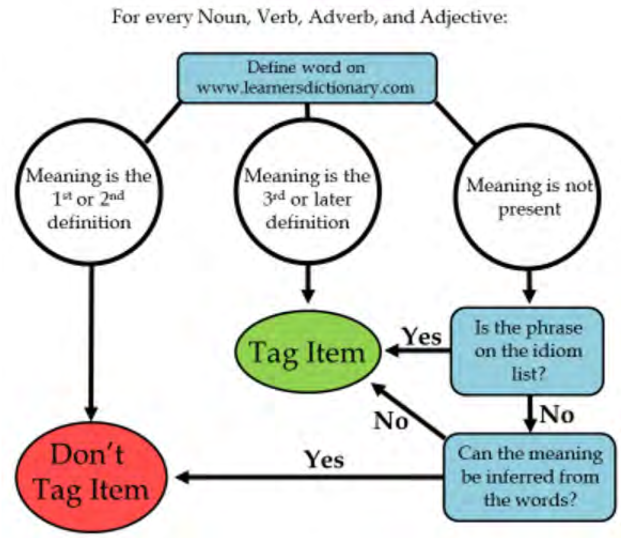

For every Noun, Verb, Adverb, and Adjective:
Define word on www.learnersdictionary.com
Meaning is the 1st or 2nd definition
Meaning is the 3rd or later definition
Meaning is not present
Tag Item
Yes
Is the phrase on the idiom list?
No
No
Don't Tag Item
Yes
Can the meaning be inferred from the words?

Figure 2. Flow chart for Phase II of Coding Procedure

Spreadsheets. Excel spreadsheets were developed to standardize the process of tagging. The raters were required to enter words and phrases they identified as needing to be glossed into individual rows in the spreadsheet. In addition to entering the tagged word or phrase, raters were also required to enter in their reason for tagging following the principles outlined in the Phase I training. If the word or phrase was identified as being non-literal due to the meaning not being listed as one of the first few definitions, the rater was to enter in the definition number that fit with the word or phrase into a specified column. In addition to entering the definition number, raters would check another column for words or phrases identified as an idiom or phrasal verb in the dictionary, or if it appeared on the idiom lists. Finally, if the word or phrase was tagged because the meaning could not be inferred from the context, the rater would enter this into another column.

Certification test. Twenty-two grade 6 ELA items were chosen for the certification test to identify when raters achieved an acceptable level of tagging sensitivity and reliability. The raters were provided with paper versions of the certification test items, which were in the same format as the practice items. Additionally, the related passages were included for rater reference. The items selected contained a range of 30 to 2,000 words; raters were required to identify a total of 39 tagged words and phrases that had previously been identified and documented by project staff.

Procedure

The procedure was the same as Phase I, with the following exceptions.

Overview of training. Because the participants were the same raters that were used in Phase I, a full training repetition was unnecessary. Rater training in this phase consisted of reviewing the main principles from the original Phase I presentation and explaining the changes. Staff provided raters with guidance for using Merriam-Webster's Learner's Dictionary, with examples. After the new task criteria was established, the raters completed practice items and a certification test as they did in Phase I. The items were the same items used in Phase I; all raters were aware of this fact, and that they were expected to have different tagging results from the procedural changes compared to Phase I.

The certification test items were re-tagged and re-documented by staff using Merriam-Webster's Learner's Dictionary and the new procedure. The raters were instructed to complete the certification test independently. Staff monitored rater progress and agreement directly with raters entering word and phrase tags into spreadsheets.

Rater process. The raters determined whether a word or phrase required tagging by completing the following steps. First, the rater would determine if each word was a noun, verb, adverb, or adjective. If the word was one of these types, the rater would enter it into Merriam-Webster's Learner's Dictionary search. If the meaning of the word in the text was listed as one of the first three definitions in the Learner's Dictionary entry, the rater would make the determination that the meaning of the word was literal or culturally common and move to the next word. However, if the meaning was not one of the first three definitions listed in the Learner's Dictionary, they would determine that the meaning of the word was non-literal or culturally uncommon and tag the word. Merriam-Webster's Learner's Dictionary definition order is positively correlated with frequency of usage in the English language ("There is no such thing as 'the dictionary,'" n.d.). If the rater could also not find the meaning in one of the listed definitions, he or she would check if the word was listed at the bottom of Merriam-Webster's Learner's Dictionary as an idiom or phrasal verb. If the meaning of the word was encompassed by the meaning of an idiom or phrasal verb, the rater would tag the word by entering it into the spreadsheet, marking the appropriate tagging reason column. They would also reference the idiom lists to aid in idiom identification.

Results and Discussion

After the raters completed the second certification test, the number of tags and percent agreement with the tags made by staff was computed. The number of tags ranged from seven to 59, with an average of 31.8. The percent agreement with the 39 tags generated by staff ratings was computed the same way as the first certification test, and ranged from .05 to .46, with an average of .28. To assess the effect of combining the judgment of individual raters, the two raters that showed the lowest agreement (of .05 and .16) were removed from the set of raters; the remaining six raters were randomly divided into two groups of three raters. Within each group, a word or phrase was identified as tagged if it was tagged by at least two of the three raters. Combining the judgments of three raters improved the reliability of the final tagging status: the percent agreement between the two groups of three raters was .48.

Although the Phase II tagging procedure had an increased reliability relative to Phase I, it did not meet our requirements for implementation on a large-scale assessment. As such, we adjusted the procedure again in Phase III.

Coding Procedure: Phase III

In Phase III of the Coding Procedure, we made major changes to the tagging procedure to increase reliability. First, we recruited new raters with more specialized educational training. We did so for two reasons: 1) novel trainers would allow us to remove the influence of the previous stages on the raters' performance, and 2) the more specialized educational training would presumably allow the trainers to make more knowledgeable decisions. Second, the raters coded the words and phrases, but did not make tagging decisions per se. The tagging decisions were made algorithmically based on the codes entered by the raters.

Method

Participants

Eight novice raters (M age = 32.88 years, SD age = 10.27 years) were selected for inclusion in the study. Out of the eight raters, six (75%) identified as female and two identified as male (25%). In terms of raters' highest level of education, four reported that they held a bachelor's degree, three reported they held a master's degree, and one reported that they held a Juris Doctor degree. These were individuals who had applied for temporary placement jobs through various employment agencies (i.e., Elite Personnel, the Midtown Group, and Randstad), and were deemed to have all the pre-requisite qualifications for inclusion in the project. These qualifications required they held at least a bachelor's degree in English or a related discipline (teaching background preferred), they were a native English speaker, and were available immediately for participation in the project. These qualifications were instituted to increase reliability of tagging, with the assumption that advanced knowledge in the discipline would aid in making some of the judgments. Two raters dropped out of the project prior to completing training for reasons unrelated to the study.

Materials

To provide Phase III raters with a more structured process and standardized task, the raters in Phase III no longer made tagging decisions. Rather, the raters input the information relevant to tagging into spreadsheets and the actual tagging decisions were made algorithmically from that information. The procedure and materials were altered to reflect this change.

Training presentation. An expanded PowerPoint presentation was developed to reflect the change in rater task and the standardization of the process. Raters were only making a series of judgments to aid in the identification of words eligible for tagging. The Phase I training presentation focused on raters tagging words, and therefore, a categorization of the tagging rationale was needed. The presentation in Phase III focused on training raters how to differentiate between common, uncommon, and non-literal language in the context of using Merriam-Webster's Learner's Dictionary. A more detailed flow chart was also developed to reflect the use of Merriam-Webster's Learner's Dictionary in this context (see Figure 3). In addition, the following training principles from Phase I were altered to reflect these changes.

The judgment process for determining whether meaning can be inferred was altered in the training presentation. Raters were to enter the definition number that best matched the meaning of the word in the item. If the meaning of the word was not present as a numbered definition, the rater was to determine if the meaning of the word or phrase was listed in Merriam-Webster's Learner's Dictionary as an idiom or a phrasal verb. In the case that the meaning of the word or phrase was not listed in Merriam-Webster's Learner's Dictionary, the rater judged whether meaning could be inferred.

A section for special cases was added to the presentation. Examples were given for acronyms and proper names as these were to be rated independent of how they were listed in Merriam-Webster's Learner's Dictionary. Compound nouns and compound adjectives were added to the presentation as they also required independent ratings regardless of how they were listed in Merriam-Webster's Learner's Dictionary. We defined compound nouns and compound adjectives as nouns and adjectives made up of two or more words. An example of a compound noun is the phrase "tall tale." This phrase is listed with its own entry in the Learner's Dictionary; but when the words of the phrase are searched independently, it is evident that the non-literal meaning "unbelievable" does not relate to the literal meaning of "tall." A rater would therefore incorrectly judge the compound by its dictionary entry; adding these to the presentation served to mitigate the task with use of the Learner's Dictionary.

Spreadsheet. The spreadsheets developed in Phase II of the project were also expanded upon to capture the full range of information the raters were identifying for words and phrases. This was accomplished by pre-populating all item words into individual rows. Pre-populating all item words also served to ensure that raters would not miss any words. The columns that were included in Phase II requiring raters to enter in the reasoning for tagging were extended and edited to reflect that the raters were no longer making tagging determinations. Columns were

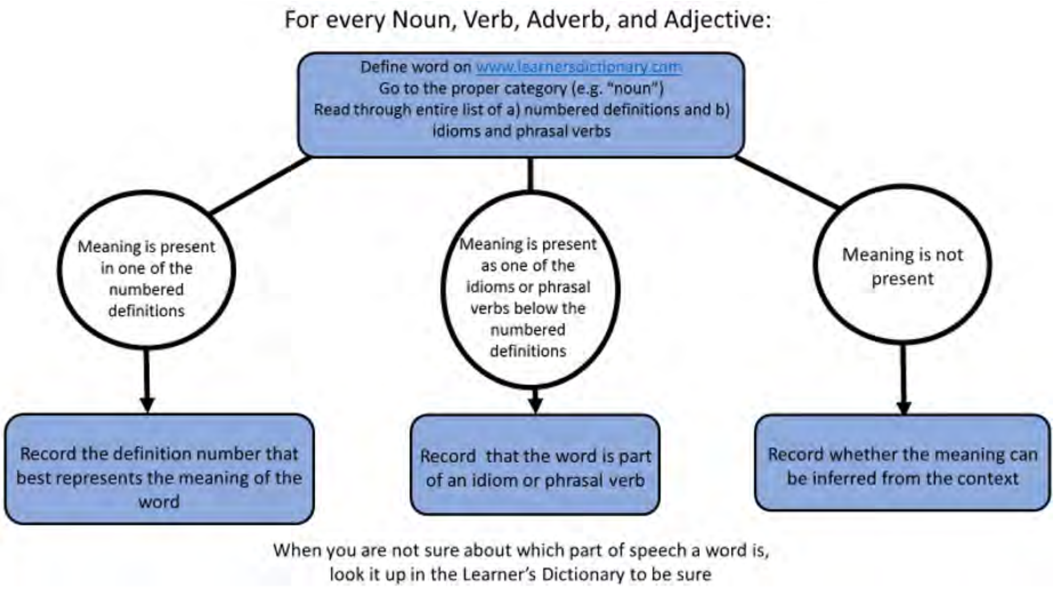

For every Noun, Verb, Adverb, and Adjective:
Define word on www.learnersdictionary.com
Go to the proper category (e.g. "noun")
Read through entire list of a) numbered definitions and b) idioms and phrasal verbs
Meaning is present in one of the numbered definitions
Meaning is present as one of the idioms or phrasal verbs below the numbered definitions
Meaning is not present
Record the definition number that best represents the meaning of the word
Record that the word is part of an idiom or phrasal verb
Record whether the meaning can be inferred from the context
When you are not sure about which part of speech a word is, look it up in the Learner's Dictionary to be sure

Figure 3. Flow chart for Phase III of Coding Procedure

also added to capture the special cases that were identified in the presentation.

Practice items. The raters were provided with the English Language Arts items from the AIR Common Core State Assessment established in Phase I of the project. However, the raters were no longer provided with paper versions. Rather, the raters were provided computerized versions and the spreadsheets with pre-populated practice item words.

Certification test. The raters were provided a condensed version of the certification test used in Phase II; ten grade 6 ELA items were selected for certification. There was a total of 785 words among these items. Raters were required to correctly enter the answers to the set of judgments to all words into pre-populated spreadsheets that were provided. Judgments for these words had previously been documented by staff in a master spreadsheet that was used for comparison.

Procedure

Overview of training. The training took place over the course of three days, and the sequence of activities was as follows. Day one began training after on-boarding procedures were completed, which consisted of AIR related employment paperwork. The raters were required to sign non-disclosure statements to discourage possible compromise of assessment items and were then equipped with laptops and access to necessary files for the project.

The training portion immediately followed the raters' completion of the new employee on-boarding process. As introduced to the raters in Phase I, Phase III raters were brought in and staff explained the purpose and scope of the project. Staff described that raters would be helping determine which words and phrases in ELA and Mathematics field test items need pop-up glossaries. The raters were told they would be completing a set of judgments for each word of an item. Staff introduced the material in the PowerPoint presentation, which took on average two hours to complete. During that time, raters were provided with specific examples to explain the set of judgments they would be making. Staff also provided guidance for using Merriam-Webster's Learner's Dictionary. After the task criteria was established, the raters completed practice items following the criteria outlined from the presentation. Staff guided raters through the practice items, going through each item word individually. Raters made the set of judgments for each word own their own, and then overall determination was discussed as a group. From this discussion, staff could determine which areas of the task raters were still confused about and provided additional training for those areas of confusion. Completion of the training presentation and practice items were completed on Day one.

Day two began with staff reviewing the training presentation and answering any additional questions. The raters were then tasked with individual completion of a certification test to determine their judgment accuracy. Completion of the certification test lasted the final two days of the training process. The certification test consisted of 10 grade six ELA items that staff had previously completed the series of judgments for and documented into a master spreadsheet. The raters were provided with computer versions of the items and spreadsheets with every item word pre-populated into individual rows. Raters were encouraged to save their work to the share folder often so that project staff could monitor their progress and accuracy.

Rater process. The raters made the following series of judgments for every word that was pre-populated into the spreadsheet. If the word was identified as a noun, verb, adverb, or adjective, the rater would enter the word into the Merriam-Webster's Learner's Dictionary search. If the word was not identified as a noun, verb, adverb, or adjective, the rater would move to the next word in the item. If the meaning of the word in the text was listed as a numbered definition, the rater would record the definition number that best represented the meaning of the word into the spreadsheet. The rater would then move to the next word in the item. If the meaning was not listed as a numbered definition, the rater would determine if the meaning of the word in the text was listed as an idiom or phrasal verb which are listed in the Merriam-Webster's Learner's Dictionary below the numbered definitions. If listed as an idiom or phrasal verb, the rater would check a column in the spreadsheet that identified it as such and then move to the next item word. If the word was not listed as a numbered definition or as an idiom or phrasal verb, the rater would then determine if the meaning of the word could be inferred from the text. If the meaning could not be inferred from the text, the rater would check a column in the spreadsheet to identify it as such. Finally, the rater would determine if the word was one of the special cases outlined in the presentation (proper name, compound noun, or compound adjective); if the word was one of these cases, the rater would specify this by checking an additional column in the spreadsheet.

Results and Discussion

In Phase III, raters did not directly determine whether a word or a phrase should be tagged. Instead, the raters answered a set of questions, and the authors established a decision rule for tagging that was applied to the raters' answers to those questions. Specifically, a word was tagged if one of the four following conditions applied:

1. The meaning of the word or phrase corresponded to a definition in Merriam-Webster's Learner's Dictionary that had a definition number larger than three (i.e., a less common meaning of the word)
2. The word was part of an idiom or phrasal verb listed below the numbered definitions in Merriam-Webster's Learner's Dictionary
3. The word was an acronym (independent of how it was listed in Merriam-Webster's Learner's Dictionary)
4. The meaning of the word was not listed in Merriam-Webster's Learner's Dictionary as one of the numbered definitions or as part of an idiom or phrasal verb, and the meaning of the word could not be inferred from the context

The number of tagged words across the six remaining raters ranged from 68 to 204, with an average of 115.8. The number of unique words that were tagged ranged from 41 to 148, with an average of 75.2. Similar to Phase II, the raters were grouped into two sets of three raters, and within each group a word was tagged if at least two out of the three raters had identified a tagging rule for it. The percent agreement between the two sets of three raters was .59.

Because .59 exceeded our reliability criterion for tagging, we implemented our Phase III protocol to tag the words and phrases for all the items in the current study.

Current Study

In the current study, we implemented a computerized, pop-up glossary accommodation of the words tagged in the Coding Procedure in a randomized, controlled design embedded within an operational statewide assessment. We were specifically interested in whether the new, reliable tagging of culturally bound words that are relevant to understanding the item contribute to both the effectiveness of pop-up glossary accommodations, and its impact on the validity of claims about the construct.

Method

Participants

Raters. An additional 31 raters (M age = 28.55 years, SD age = 6.36 years) were hired from the same agency and with the same qualifications to do the final coding in the current study. Out of the 31 raters, 17 (54.8%) identified as female and 14 identified as male (45.2%). In terms of raters' highest level of education, 21 (67.74%) reported that they held a bachelor's degree, eight (25.81%) reported they held a master's degree, and two (6.45%) reported that they held a Juris Doctor degree. These raters (plus the original six) were assigned to one of four cohorts based on hire date. Each cohort started with eight to twelve raters. To keep the ratio of project staff to raters low, all cohorts were trained separately.

Three of these raters were dismissed due to poor rater agreement that did not improve with one-on-one training iterations.

Students. Participants in this study consisted of students taking online English Language Arts (ELA) and Mathematics statewide accountability assessments during the spring 2017 administration. All students except 34,200 students who participated in a passage review study to determine the usability of reading passages for future assessments were eligible to participate in the current study. Approximately 60,000 students from each subject and grade were included, with EL students ranging from approximately 1,000 to 8,000 students per grade, with more in the lower grades.

Every student included in the study was randomly allocated with equal probabilities into one of the following three glossary conditions: English glossary only; English glossary with Spanish translation; and no glossary. Appendix A summarizes the number of students included in the study by grade, subject, EL status and glossary condition.

Materials

Items. Each operational test administration contains a subset of newly developed items which are included for data gathering purposes for potential operational use in future test administrations. These newly developed items are called field test items; field test items are not scored, and students do not know which questions on the test are the field test items.

For the current study, all items to be field-tested in the spring 2017 test administration were potentially glossed. The ELA assessment consisted of 1,014 field test items along with their associated reading passages (which were not glossed with reasoning explained in the Coding Procedure). The Mathematics assessment consisted of 581 field test items. Table 1 presents the number of field test items by subject and grade and the number of items that were glossed.

A subset of the field test items in each grade (46 on average) were also administered in the grade below to establish a vertical link across grades. The linking study is outside the scope of this paper, but as a result some of the glossed field test items were also presented in the grade below, without adapting the glossary entries. Because the grade-to-grade changes in glossary entries was expected to be minimal and to maximize the statistical power of the analyses, we decided to include this subset of items in the analyses.

Glossary. Glossaries were developed for the item words that were tagged from the procedure outlined in Phase III of the Coding Procedure; this included the pop-up English glossary

Table 1. Number of field test items and number of glossed items

	ELA			Mathematics	
Grade/Course	Passages	Items	Items with Glossaries	Items	Items with Glossaries
3	12	91	55	85	47
4	12	86	70	84	40
5	11	90	55	78	34
6	11	97	88	52	35
7	11	91	82	53	35
8	11	90	72	54	43
9/Algebra	15	154	147	59	49
10/Geometry	15	157	136	54	45
11/Algebra 2	15	158	142	62	37

entries with audio and Spanish translations of the tagged word with audio. The audio files that were developed only consisted of information within the glossary and did not contain any other test information. The glossaries were developed in a three-step process, with all the guidelines precisely documented for all responsible parties participating in the glossing process.

Students that were administered the glossed field test items were presented with instructions for how to use the glossary immediately before viewing the item. The instructions were brief and explained that help would be accessible to them on the next item; it specified that some of the words they would see on the next screen would have dotted lines above and below the word. If they moved their mouse over one of these words, the word would be highlighted in blue; and if they clicked on the word, a dialogue box would appear. For students in the English glossary only condition, the dialogue box would only contain the word and the English definition of the word. For students in the English glossary with Spanish translation condition, the dialogue box would contain the word, a Glossary tab with the English definition, and a Spanish Glossary tab with the Spanish translation of the glossed word. The instructions displayed a picture of the glossary dialogue box and pinpointed the speaker icon within the dialogue box that the student would click on to hear the words contained in the glossary.

Procedure

Glossing. For every field test item, the process outlined in Phase III of the Coding Procedure was completed by three independent raters for all the words in the items. A word was identified for potential tagging if two out of the three raters indicated one of the reasons for tagging was present; additionally, reliability of words identified by raters was confirmed through a 10% read behind. The same rules for tagging were applied for the read behind items. For the read behind items, the percentage agreement between the first and second group of three raters was .57 for ELA, and .59 for Mathematics. The final tagging decisions were made algorithmically from the information provided from the rater judgments. Once the final tagging decisions were completed, glossing began. Glossing was completed in three stages. The glossing of the tagged words was completed in the following process. First, the tag was removed if the word was included on the grade level vocabulary lists (“Construct relevant vocabulary for English language arts and literacy,” 2015, For Mathematics a vocabulary list was developed by AIR test developers for each grade) or if the word was part of the construct being assessed. For Mathematics items, the tag was also removed if the meaning of the tagged word was not necessary to solve the problem. Second, if the tag remained, the wording of the glossary was based on Merriam-Webster’s Learner’s Dictionary definition. As such, we first reviewed the definition or explanation of the idiom or phrasal verb from Merriam-Webster’s Learner’s Dictionary to ensure that it was (1) the correct definition for the context of the item, and (2) grade level appropriate. If the chosen dictionary definition was rejected, we based the gloss on the definition identified as correct. If the definition was not grade-level appropriate, we re-worded the definition (idiom or phrasal verb) so that it was written at the grade level of the student who would be reading it. Finally, AIR content staff reviewed the final glossary entries for simplicity. If necessary and possible, AIR staff simplified the glossary entry without losing meaning.

In addition to the words tagged using the method above, words were also tagged if they were identified as more challenging for the grade level than is typical for that grade. We made this determination using the Lexile word measure of difficulty for each word. A Lexile word measure is a validated estimate of the challenge a given word will present, on average, during independent reading (Elmore, Lattanzio, Stenner, & Sanford-Moore, 2016). Lexile word measures are calculated with a corpus-based, machine-learning model developed and validated by over 7 million student item responses. Lexile word measures and Lexile text measures contributed about equally to predicting student responses to words. Table 2 presents the cut scores that were used to identify tagged words for each grade level based on the Lexile word measure. If the Lexile word measure was higher than the ‘High Lexile’ for the grade, then the word was tagged for glossing.

The last step in the process involved producing audio files of the glossaries and Spanish translations. English audio files were produced for the glossary entries, and Spanish translations and audio files were produced for the glossed words.

Experiment. To examine the effectiveness of the computer-based pop-up glossary accommodation, the glossary was incorporated into the statewide accountability assessment; this followed the standardized state assessment guidelines, procedures, and rules. As such, the assessment is untimed, and all students requiring testing accommodations were provided their regular authorized accommodations.

The incorporation of the pop-up glossary accommodation into the statewide assessment allowed every eligible student taking the assessment to be randomly assigned into one of the three glossary conditions. In addition, each student was randomly assigned seven to nine of the field test items (using a pseudo-random number generator in the test delivery system) that were included in the study. In the Mathematics assessment tests, the field test items were individually assigned for administration. For ELA, the field test items were assigned in groups depending on the associated reading passage.

Field test items were randomly chosen by the test delivery system for students in all conditions. For students in the pop-up English glossary only condition and the pop-up English glossary with Spanish translation condition, instructions (as described above) were given prior to the student viewing the item. The item with the glossed words was presented after the student clicked the “next” button on the screen. The glossed words were identified as described above. After the student responded to the item, they would proceed to the next question presented.

Results and Discussion

The student responses to operational and field test items were analyzed with a mixed logistic regression model. The independent variables of the model were dummy coded for the item, EL status, glossary condition, and the interaction between glossary condition and EL status. Specifically, conditional on a person-specific random intercept u_i, the probability of a correct answer was modeled as

$$\Pr(Y_{ij} = 1|u_i) = \frac{\exp\ (1.7n_{ij})}{1+\exp\ (1.7n_{ij})}$$

$$n_{ij} = u_i + \beta_j + \alpha_1 ENG_{ij} + \alpha_2 ENG\_SP_{ij} + \alpha_3 EL_i ENG_{ij} + \alpha_4 EL_i ENG\_SP_{ij}$$

Table 2. Lexile word measure cut-scores

	ELA		Mathematics	
Grade/Course	**Low Lexile**	**High Lexile**	**Low Lexile**	**High Lexile**
3	520	820	520	820
4	740	940	740	940
5	830	1010	830	1010
6	925	1070	925	1070
7	970	1120	970	1120
8	1010	1185	1010	1185
9/Algebra	1050	1260	1010	1185
10/Geometry	1080	1335	1050	1260
11/Algebra 2	1185	1385	1080	1335

where,

$$u_i \sim \{ \begin{matrix} N(0, \sigma^2_{non\ EL}) \\ N(\mu_{EL}, \sigma^2_{EL}) \end{matrix}$$

β_j: effect of item j

ENG_{ij} = 1 if student i is in the *English glossary* condition and item j has glossaries, = 0 else

ENG_SP_{ij} = 1 if student i is in the *English glossary + Spanish translation* condition and item j has glossaries, = 0 else

EL_i = 1 if student i is an EL, = 0 else

To allow comparison to the results of Cohen et al. (2017), we added a scaling coefficient (i.e., 1.7) to the model that scaled the effects of the logistic regression to that of a probit regression model. The terms β_j, j = 1 ,..., J are fixed effects controlling for the differences in difficulty across the J items. The term u_i is a person-specific intercept capturing the dependencies of item responses from the same student. Specifically, u_i is a latent variable representing student achievement and modeled as a random effect. A separate mean and variance was specified for the distribution of the random effect for the two groups defined by EL status to allow for overall differences in performance across both groups. To identify the model, the mean of u_i was set to zero for the non-ELs, so that μ_{EL}, the mean of u_i for the ELs, represents the overall difference in performance between both groups. The coefficients α_1 and α_2 indicate the main effect of glossary conditions, and the coefficients α_3 and α_4 represent the interaction between glossary condition and EL status. For the non-EL group, the effect of English glossaries and English glossaries with Spanish translations are directly given by α_1 and α_2. For the group of ELs, the effect of English glossaries and English glossaries with Spanish translations are obtained as the sum of the main and interaction effects. Specifically, $\alpha_1 + \alpha_3$ represents the effect of English glossaries for ELs, and $\alpha_2 + \alpha_4$ represents the effect of English glossaries with Spanish translation.

A separate mixed logistic regression model was fit for every grade for both ELA and Mathematics. The models were estimated using the maximum likelihood method as implemented in the R package *flirt* (Jeon, Rijmen, & Rabe-Hesketh, 2014). Appendix B and C show the parameter estimates and standard errors for the overall group difference μ_{EL}, and for α_1 to α_4, representing the effects of glossaries and their interactions with EL status, for ELA and Mathematics respectively. To facilitate the interpretation of the results, the contrasts $\alpha_1 + \alpha_3$ and $\alpha_2 + \alpha_4$, and their standard errors are also shown. Positive coefficients are associated with a higher performance. The z statistic, defined as the parameter estimate divided by its standard error was used to test for statistical significance. Under a maximum likelihood framework, the z statistic is asymptotically distributed as the standard normal distribution under the null hypothesis. In Appendix B and C, significant results (at α = .05) are flagged.

For both ELA and Mathematics, there are substantial differences in overall performance, with the group of non-ELs significantly outperforming the group of ELs across grades. For the ELA assessments, the effects of providing the English glossary and the English glossary with Spanish translation were mostly positive for EL students in all grades. The estimated effects ranged from 0.01 to 0.09 for elementary schoolers and gradually increased for the middle school and high school EL students. The effect is significant for all grades and both types of glossaries except for the two glossary conditions in Grade 3. Providing a glossary on the ELA tests significantly improved the performance of EL students across all grades. For non-ELs, the effects are smaller in absolute value and tend to be negative. Significant effects were found for grades 3, 4, and 9 for the effect of English glossaries, and for grades 3 to 7 for the effect of English glossaries with Spanish translations. Providing a glossary had virtually no effect for non-EL students in grades 8 to 11, but a small negative effect for grades 3 to 7.

With respect to the Mathematics assessments, all significant effects for EL students indicate a positive effect of glossaries on performance. Some of the effects are quite substantial. There tends to be an increasing effect across grades, but the trend is less clear than the trend of increasing effects observed for ELA. For the group of non-ELs, the only significant effect is for Geometry, which has a positive effect that is small compared to the effects observed for ELs.

In sum, except for the lower grades, the pop-up glossaries were revealed to be both and effective (the glossaries improved EL student performance), and non-detrimental to the construct measured (the glossaries had little or no effect on non-EL student performance). As such, we conclude that the tagging protocol successfully identified words and phrases that are most likely to inhibit EL student performance.

General Discussion

We attempted to (1) develop a procedure to reliably identify words and phrases that are most likely to inhibit EL student performance on a large-scale assessment, and (2) determine whether using a pop-up glossary accommodation is an effective method of reducing the influence of these words on EL students' performance on a large-scale assessment, without damage to the construct being measured. In Phase III of the Coding Procedure, we implemented a protocol that resulted in reliable tagging of culturally bound words. In the current study, these tagged words were glossed to assess the effectiveness and impact on validity of claims regarding the construct of a computer-based, pop-up glossary accommodation for EL students. The current study was a large-scale, randomized controlled trial experiment in a statewide assessment across grades 3-11. The results demonstrated that generally the pop-up glossary accommodation was and effective for both the ELA and Mathematics assessments and did not harm the construct being measured.

Here, we demonstrate for the first time in a randomized, controlled, large scale experiment that a pop-up English glossary accommodation increased the performance of EL students in both ELA and Mathematics (demonstrating effectiveness), while minimally influencing the scores of non-EL students on those same items (demonstrating validity). First, our data replicated the general finding that EL students perform significantly worse than non-EL students in both ELA and Mathematics on large scale standardized assessments (Avenia-Tapper & Llosa, 2015; Kieffer et al., 2009; Martiniello, 2008; Sato et al., 2010). This finding demonstrates the performance gap between EL and non-EL students. The pop-up glossaries are an attempt to reduce this performance gap.

An accommodation must not affect the validity of claims about the construct being measured. Evidence for this is obtained by demonstrating that the accommodation does not influence the performance of the group that does not require the accommodation. Here, that group is the non-EL students. In the Mathematics assessment, neither of the glossary accommodations show any influence on non-EL students' performance. As such, one can conclude that the pop-up glossaries did not influence the construct of the items in the Mathematics assessment.

In the ELA assessment, the picture is slightly more complicated. For the ELA assessment, the English glossary inhibited the 3rd and 4th grade non-EL students' performance. This was consistent with the results of Cohen at al. (2017). The English glossary with Spanish translation, however, inhibited the 3rd through 7th grade non-EL students' performance. At a minimum, this suggests that we ought not provide Spanish glossaries to non-Spanish speakers. At the same time, we found no evidence that either variant of the glossaries significantly improved performance for the non-EL group. The negative effect observed for the non-EL students in the lower grades for ELA was quite small both in absolute terms and compared to the positive effects observed for the EL students.

For accommodations to be considered effective, the accommodation must demonstrate that it effectively mitigates the barriers to the students' access of the item. Evidence for effectiveness is obtained by demonstrating that the accommodation improves the performance of the group that requires the accommodation. Here, that group is the EL students. As stated above, the pop-up glossaries improved performance of the EL students in both the ELA and Mathematics assessments in most grades. The exceptions were generally the lowest grades whose effects were in the positive direction but did not reach significance. As such, we conclude that overall, pop-up glossaries are effective when implemented properly.

It is important to note that the effectiveness of the pop-up glossary accommodation and its impact on validity are a function of (1) the implementation of the glossary itself, and (2) the words identified and glossed by the accommodation. Whereas much of the discussion of accommodations focus on implementation, here we focused on the identification of words. The physical implementation of the pop-up glossaries in the present research are identical to those of Cohen et al. (2017). Despite that important similarity, the overall effect of the accommodation is quite different. The pop-up glossary accommodations of Cohen et al. (2017) were effective only in only one of the two grades assessed in ELA (grade 7) and neither of the two grades assessed in Mathematics. Here, the pop-up glossary accommodations were and effective across grades and subjects, and we found no evidence that it changed the intended construct. We attribute this difference to the choice of words that we identified to be glossed.

Cohen et al. (2017) hypothesized that the inhibitory effect of the pop-up glossaries that they observed was a function of the increased cognitive load that the glossaries created for the students. That is, the glossaries required focused attention to use effectively, and that focused attention on the glossed words and the glossary itself may have reduced the amount of attention that the students could commit to completing the item correctly. Cohen et al. (2017) termed this the *Cognitive Load Hypothesis*.

Here, we attempted to reduce the extraneous cognitive load of the glossaries by reducing the number of words that were glossed. Specifically, when deciding which words to gloss, we started with the assumption that every tagged word could potentially serve as a distractor to the students and thus inhibit performance. As such, our goal was to tag only those words that were most likely to aid EL students. In general, we chose to tag words (1) whose meaning was necessary to understand in order to respond to the item, and (2) whose meaning was only identifiable by having culturally relevant information or was a less common meaning of the word. These general rules identified words that EL students would most likely have difficulty interpreting and are necessary to know to respond to the item.

As discussed above, previous research revealed that the reliable identification of culturally relevant words is notoriously difficult (e.g., Loenneker-Rodman, & Narayanan, 2010; Pragglejaz Group, 2007; Roberts, & Kreuz, 1994). When we implemented procedures published in the literature, we were unable to reach an acceptable level of reliability. As such, we developed an algorithmic procedure that relied on expert rater's judgment of the meaning of every word in an item. Because even this subjective element was somewhat unreliable, we based our algorithm on the intersection of two of three expert raters. It was only with this level of care that we were able to reach an acceptable level of reliability.

Our results provide evidence that words we identified for tagging were words whose meaning inhibited EL students' performance on the assessments without influencing claims of the construct related validity of the item. Furthermore, our results support the Cognitive Load Hypothesis. Specifically, by reducing the number of tagged words to only those most likely to be misinterpreted, we reduced the interference of the pop-up glossary (relative to that of Cohen et al, 2017). Furthermore, similar to Cohen et al. (2017), the few grades in which we observed the inhibitory effect of the glossaries were the lowest (3rd and 4th).

These results provide an important empirical demonstration of the importance of the word selection on the effectiveness of the pop-up glossary accommodation, as well as its impact on the validity of claims about the construct measured. Indeed, the contrast between the present results and those of Cohen et al. (2017) suggest that the words identified for tagging are, at least, as important as the accommodation itself.

There are several limitations to the current research. First, although we provided pop-up glossaries, we did not track their use by the students. Such tracking was neither possible nor economical on the large scale of the current research. Because our findings confirm the effectiveness of pop-up glossaries, this limitation is not critical. Nevertheless, were we able to track the students' use of the pop-up glossaries, we may have found a larger effect size for those students that viewed the pop-up glossaries. Furthermore, we only assessed a Spanish translation version. Although the first language was Spanish for the majority of EL students in the state population we assessed, we did not vary language translation by EL language status. Again, such a manipulation for the current research was not economical. Nevertheless, the pop-up glossary with translation may produce a larger effect size (relative to the non-translation pop-up glossary) for those EL students whom the translation was their native language.

In summary, we have demonstrated that pop-up glossaries, when implemented according to the guidelines presented here, are effective accommodations for EL students, and do not change the mathematics or ELA constructs measured. This inference is supported by a randomized, controlled statewide, large scale assessments. The contrast between the present results and those of Cohen et al. (2017) demonstrate that these results are contingent on the reliable identification of words with culturally relevant or low frequency meanings. Here, we presented a procedure for reliably identifying such words.

References

Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment, 14*, 195-211. doi:10.1080/10627190903448851

Abedi, J. (2012). Validity issues in designing accommodations. In: Fulcher, G., Davidson, F. The Routledge Handbook of Language Testing in a Nutshell. Florence, KY: Routledge, Taylor & Francis Group.

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Technical Report No. 586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research based decision algorithm.* Retrieved from https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues & Practice, 25*(4), 36-46. doi:10.1111/j.1745-3992.2006.00077.x

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1-28. doi:10.3102/00346543074001001

Avenia-Tapper, B. & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment, 20*(2), 95-111. doi:10.1080/10627197.2015.1028622

Barlow, J. M. (1971). Training Manual for Identifying Figurative Language.

Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education, 30*(4), 259-272. doi:10.1080/08957347.2017.1353986

Construct relevant vocabulary for English Language Arts and literacy (2015). Smarter Balanced Assessment Consortium: The Regents of the University of California. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2015/08/ELA-Construct-Relevant-Vocabulary.pdf

Council of Chief State School Officers. (2016). *Major provisions of Every Student Succeeds Act (ESSA) related to the education of English learners.* Retrieved from https://ccsso.org/sites/default/files/2017-11/ccsso%20resource%20on%20els%20and%20essa.pdf

Dixson, R.J. (2003). Essential idioms in English. Englewood Cliffs, NJ: Prentice Hall Regents. Retrieved from https://abiiid.files.wordpress.com/2011/06/essential-idioms-in-english-phrasal-verbs-and-collocations.pdf

Elmore, J., Lattanzio, S., Stenner, A.J., & Sanford-Moore, E.E. (2016). Calculation of Lexile word measures using a corpus-based model and student performance data. Durham, NC: MetaMetrics, Inc. Retrieved from http://cdn.lexile.com/cms_page_media/147/Calculation%20of%20Lexile%20Word%20Measures.pdf

Holmes, D. (n.d.). Idioms and expressions. Retrieved from https://www.noblepath.info/idioms_and_expressions/idioms_and_expressions.pdf

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Measurement, 38*(5), 404-405. doi:10.1177/0146621614524982

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*(3), 35-45. doi:10.1177/073724770402900303

Kieffer, M.J., Lesaux, N.K., Rivera, M., & Francis, D.J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168-1201. doi:10.3102/0034654309332490

Kiplinger, V.L., Haug, C.A., & Abedi, J. (2000). *Measuring math—not reading—on a math assessment: A language accommodation study of English language learners and other special populations.* Marion, IN: Indiana Wesleyan Center for Educational Excellence. Retrieved from ERIC database. (ED441813)

Kopriva, R., Emick, J., Hipolito-Delgadp, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*, 11-20. doi:10.111/j.1745-3992.2007.00097.x

Li, H., & Suen, H.K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education, 25*, 327-346. doi :10.1080/08957347.2012.714690

Loenneker-Rodman, B., & Narayanan, S. (2010). Computational approaches to figurative language. *Cambridge Encyclopedia of Psycholinguistics*.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368. doi:10.17763/haer.78.2.70783570r1111t32

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*(3-4), 160-179. doi:10.1080/10627190903422906

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., & Frieder, O. (2013). Metaphor identification in large texts corpora. *PloS one, 8*(4), e62343. doi:10.1371/journal.pone.0062343

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*, 10-28. doi:10.111/empi.2011.30.issue-3

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol, 22*(1), 1-39. doi:10.1080/10926480709336752

Resource list. Idioms: Figurative language (n.d.). *Reading Manipulatives, Inc*. Retrieved from http://www.readskill.com/resources/SkillResourceLists/pdf/RM_Idioms.pdf

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science, 5*(3), 159-163. doi:10.1111/j.1467-9280.1994.tb00653.x

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.W. (2010). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets (NCEE 2009-4079)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sireci, S.G., Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education, 39*(1), 215-252. doi:10.3102/0091732X14557003

Sireci, S.G., Scarpati, S.E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. Review of Educational Research, 75(4), 457-490. doi:10.3102/00346543075004457

There is no such thing as "the dictionary." (n.d.). Retrieved from https://www.merriam-webster.com/words-at-play/dictionary-facts-and-trivia/there-is-no-such-thing-as-201Cthe-dictionary201D

U.S. Department of Education, National Center for Education Statistics. (2017). Fast facts: English language learners. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=96

Wolf, M.K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3-4), 139-159. doi:10.1080/10627190903425883

Appendix A. Number of students included in the study by subject, grade, EL status, and glossary condition

		Number of students					
Participants		ELA			Mathematics		
Grade/ Course	Group	Glossary	Glossary with Spanish translation	No Glossary	Glossary	Glossary with Spanish translation	No Glossary
3	All	21,920	22,229	22,148	22,011	22,355	22,241
	EL	2,535	2,449	2,532	2,569	2,481	2,563
	Non-EL	19,385	19,780	19,616	19,442	19,874	19,678
4	All	22,225	22,534	22,490	22,347	22,666	22,624
	EL	2,425	2,520	2,350	2,450	2,545	2,375
	Non-EL	19,800	20,014	20,140	19,897	20,121	20,249
5	All	21,726	22,110	21,952	21,833	22,176	22,053
	EL	1,924	1,928	1,906	1,935	1,941	1,920
	Non-EL	19,802	20,182	20,046	19,898	20,235	20,133
6	All	21,062	21,359	21,299	21,113	21,444	21,349
	EL	1,380	1,343	1,393	1,397	1,361	1,410
	Non-EL	19,682	20,016	19,906	19,716	20,083	19,939
7	All	21,082	21,399	21,270	20,723	21,018	20,958
	EL	1,241	1,307	1,316	1,251	1,306	1,323
	Non-EL	19,841	20,092	19,954	19,472	19,712	19,635
8	All	21,142	21,537	21,399	18,066	18,473	18,340
	EL	1,044	1,118	1,029	1,048	1,108	1,025
	Non-EL	20,098	20,419	20,370	17,018	17,365	17,315
9/Algebra	All	16,791	17,066	16,960	19,043	19,271	19,117
	EL	548	589	530	561	595	513
	Non-EL	16,243	16,477	16,430	18,482	18,676	18,604
10/Geometry	All	15,550	15,854	15,602	15,794	16,137	16,045
	EL	326	372	323	334	410	357
	Non-EL	15,224	15,482	15,279	15,460	15,727	15,688
11/Algebra 2	All	14,080	14,247	14,199	14,306	14,338	14,290
	EL	183	218	209	182	175	208
	Non-EL	13,897	14,029	13,990	14,124	14,163	14,082

Appendix B. Estimated Logistic Regression Coefficients and Standard Errors by Grade Level on Scores for the English Language Arts Assessment

	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade	9th grade	10th grade	11th grade
EL	-0.576*	-0.59*	-0.691*	-0.642*	-0.677*	-0.671*	-0.656*	-0.641*	-0.56*
	(0.005)	(0.005)	(0.006)	(0.008)	(0.008)	(0.009)	(0.009)	(0.011)	(0.017)
ENG	-0.022*	-0.024*	-0.011	-0.004	-0.01	-0.001	-0.013*	-0.002	-0.006
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
ENG_SP	-0.018*	-0.034*	-0.014*	-0.014*	-0.012*	0.003	0.001	0.003	0.001
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
EL by ENG	0.046*	0.067*	0.088*	0.104*	0.103*	0.111*	0.161*	0.104*	0.194*
	(0.014)	(0.014)	(0.016)	(0.017)	(0.018)	(0.02)	(0.027)	(0.034)	(0.045)
EL by ENG_SP	0.026*	0.092*	0.102*	0.082*	0.098*	0.115*	0.095*	0.102*	0.194*
	(0.014)	(0.014)	(0.016)	(0.018)	(0.017)	(0.02)	(0.027)	(0.032)	(0.042)
ENG + (EL by ENG)	0.024	0.043*	0.077*	0.1*	0.093*	0.11*	0.148*	0.102*	0.188*
	(0.014)	(0.014)	(0.016)	(0.017)	(0.018)	(0.02)	(0.027)	(0.034)	(0.045)
ENG_SP + (EL by ENG_SP)	0.008	0.058*	0.089*	0.067*	0.086*	0.118*	0.096*	0.106*	0.195*
	(0.014)	(0.013)	(0.016)	(0.018)	(0.017)	(0.02)	(0.026)	(0.032)	(0.042)

Note. EL: Overall group difference between the EL group (EL=1) and the non-EL group (EL=0); *ENG* = the main effect of English glossaries, also the effect of glossaries for the non-EL group; *ENG_SP* = the main effect of English glossaries with Spanish translations for the non-EL group, also the effect of glossaries with Spanish translations for the non-EL group; EL by ENG_SP= interaction between English glossary condition and EL status; EL by ENG_SP = interaction between English glossary with Spanish translation condition and EL status; ENG + (EL by ENG)= the effect of English glossaries for the EL group; ENG + (EL by ENG_SP)= the effect of the English glossaries with Spanish translation for the EL group. Significant results (p < .05) are indicated by *.

Appendix C. Estimated Logistic Regression Coefficients and Standard Errors by Grade Level on Scores for the Mathematics Assessment

	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade	Algebra	Geometry	Algebra 2
EL	-0.832*	-0.793*	-0.858*	-0.82*	-0.825*	-0.601*	-0.702*	-0.669*	-0.437*
	(0.01)	(0.009)	(0.01)	(0.012)	(0.012)	(0.011)	(0.011)	(0.018)	(0.021)
ENG	0.004	-0.009	-0.0002	-0.006	0.006	0.012	0.01	0.029*	-0.025
	(0.009)	(0.009)	(0.01)	(0.01)	(0.01)	(0.01)	(0.012)	(0.013)	(0.014)
ENG_SP	-0.007	-0.012	-0.006	0.003	0.006	-0.008	0.009	0.016	-0.016
	(0.009)	(0.009)	(0.01)	(0.01)	(0.01)	(0.01)	(0.012)	(0.013)	(0.014)
EL by ENG	0.114*	0.053*	0.011	0.09*	0.093*	0.178*	0.422*	0.216*	-0.043
	(0.02)	(0.023)	(0.027)	(0.032)	(0.032)	(0.034)	(0.051)	(0.071)	(0.103)
EL by ENG_SP	0.111*	0.135*	0.037	0.069	0.12*	0.172*	0.476*	0.074	0.129
	(0.021)	(0.022)	(0.026)	(0.032)	(0.031)	(0.034)	(0.049)	(0.071)	(0.091)
ENG + (EL by ENG)	0.118*	0.044	0.011	0.084*	0.099*	0.191*	0.433*	0.245*	-0.066
	(0.02)	(0.023)	(0.026)	(0.032)	(0.032)	(0.034)	(0.051)	(0.071)	(0.103)
ENG_SP + (EL by ENG_SP)	0.104*	0.123*	0.031	0.072*	0.126*	0.163*	0.484*	0.09	0.113
	(0.021)	(0.022)	(0.026)	(0.032)	(0.031)	(0.034)	(0.049)	(0.071)	(0.091)

Note. μ_{EL} = EL mean of random intercept; α_1 = the effect of English glossary only for non-EL group; α_2 = the effect of English glossaries with Spanish translations for non-EL group; α_3 = interaction between English glossary condition and EL status; α_4 = interaction between English glossary with Spanish translation condition and EL status; $\alpha_1 + \alpha_3$ = the effect of English glossary only for EL group; $\alpha_2 + \alpha_4$ = the effect of the English glossary with Spanish translation for EL group. Significant results (p < .05) are indicated by * .