NORTH DAKOTA DEPARTMENT OF
PUBLIC INSTRUCTION

North Dakota State Assessment for English Language Arts/Literacy and Mathematics

2020–2021

Volume 1
Annual Technical Report

ACKNOWLEDGMENTS

This technical report was produced on behalf of the North Dakota Department of Public Instruction (NDDPI). Requests for additional information concerning this technical report or the associated appendices should be directed to NDDPI at seschauer@nd.gov .

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Youngmi Cho, Kevin Clayton, Aida Diaz, Mattie MacDonald, Seyfullah Tingir, Pamela Trantham, and Ahmet Turhan. Major contributors from the NDDPI include Stanley Schauer and Bonnie Weisz.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

1. Introduction

The North Dakota State Assessment (NDSA) 2020–2021 technical report volumes are provided to document and make transparent all processes used in item development, test construction, psychometric analyses, standard setting, and score reporting. This includes summarizing student assessment results and documenting evidence for intended uses and interpretations of the test scores. The technical report is presented as seven separate, self-contained volumes that cover the following topics:

1. *Annual Technical Report*. This annually updated volume provides a general overview of the tests administered to students each year.
2. *Test Development*. This volume details the procedures used to construct tests and summarizes the Independent College and Career Readiness (ICCR) item bank and its item development process.
3. *Standard Setting*. This volume documents the methods and results of the 2018 NDSA standard setting process. This is a static volume until a new standard setting takes place.
4. *Evidence of Reliability and Validity*. This volume provides technical summaries of the test quality to support the intended uses and interpretations of the test scores.
5. *Summary of Test Administration Procedures*. This volume describes the methods used to administer all available test forms, security protocols, and modifications or accommodations.
6. *Score Interpretation Guide*. This volume describes the score types reported along with the appropriate inferences and intended uses of each score type.
7. *Special Studies*. This volume compiles any special studies conducted for the NDSA; it is updated annually to reflect studies relevant to the respective test administration.

The North Dakota Department of Public Instruction (NDDPI) communicates the quality of the NDSA by making this technical report accessible to the public.

1.1 Background and Historical Context of Test

The NDSA was constructed to measure student achievement in English language arts (ELA) and mathematics relative to the North Dakota Academic Content Standards. The NDSA ELA assessment consists of two segments: reading and writing. In this report, the term *ELA* refers to the combination of reading and writing subject areas, while *reading* refers only to the reading portion of the test.

The NDSA was first administered to students during spring 2018 and spring 2019 as a fixed-form test, replacing the English language arts (ELA) and mathematics assessments developed by the Smarter Balanced Assessment Consortium (SBAC). In an effort to reduce testing times and increase test security, NDDPI chose to adopt a computer-adaptive test (CAT) format for the spring 2021 NDSA ELA and mathematics assessments. The CAT algorithm developed by Cambium Assessment, Inc. (CAI) assembles a unique test form for each individual such that the items selected for each student best match their ability levels, while ensuring that each test form covers the required content standards defined by the state's blueprint (for details, refer to Volume 2).

The administration of the NDSA scheduled for spring 2020 was cancelled by NDDPI due to the statewide school closures that followed the onset of the COVID-19 pandemic. Although the use of this year's assessment data for accountability purposes has been suspended, the Council of Chief State School Officers has recognized that the data can serve as a valuable resource to better understand the overall impact of the pandemic on student learning (CCSSO, 2020).

1.2 Purpose and Intended Uses of the North Dakota State Assessments

The NDSA is a criterion-referenced test that applies principles of evidence-centered design to yield overall and reporting category-level test scores at the student level and at other levels of aggregation that reflect student achievement of North Dakota's Academic Content Standards. The NDSA supports instruction and student learning by providing immediate feedback to educators and parents, which can be used to inform instructional strategies that remediate or enrich instruction. An array of reporting metrics allows achievement to be monitored at both student and aggregate levels and growth to be measured at both student and group levels over time.

The NDSA draws all items from the ICCR item bank (refer to Volume 2, Test Development), which is a rigorously developed item bank aligned to nationally recognized career and college readiness standards. CAI and NDDPI worked together to ensure that the items in the test forms constructed for all grades uniquely measure student mastery of the North Dakota Content Standards in ELA and mathematics, which are aligned with knowledge and skills that are essential for college and career readiness.

Table 1 outlines the required uses and citations of the NDSA on the basis of the "Curriculum and Testing" chapter of the North Dakota legislative code (www.legis.nd.gov/cencode/t15-1c21.pdf) and the federal Every Student Succeeds Act (ESSA). The NDSA fulfills all the requirements described in Table 1.

Table 1: Required Uses and Citations of the NDSA

Required Use	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 in order to meet the requirements under Section 1111(b)(2)(B)(v)(I)(bb) of the Elementary and Secondary Education Act of 1965 (ESEA)	ESSA Plan Section 3 A
Test administration frequency and grade levels	North Dakota Statute 15.1-21-08.1
Compilation of test scores	North Dakota Statute 15.1-21-09
Publication of test scores	North Dakota Statute 15.1-21-10
Requirement for the alignment of the test to academic content standards	North Dakota Statute 15.1-21-11

1.3 Participants in the Test Development and Analysis of the North Dakota State Assessments

NDDPI manages the North Dakota state assessment program with the assistance of North Dakota educators, a Technical Advisory Committee (TAC), and several vendors (listed in the following paragraphs). NDDPI fulfills the diverse requirements of implementing the NDSA 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).

North Dakota Department of Public Instruction

The Office of Student Assessment oversees all aspects of the NDSA program, including coordination with other NDDPI offices, North Dakota public schools, and program vendors.

North Dakota Educators

North Dakota educators participate in most aspects of the conceptualization and test development of the NDSA. 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.

Technical Advisory Committee

Multiple times a year, NDDPI convenes an advisory committee panel to discuss psychometric, test development, administrative, and policy issues relevant to current and future North Dakota assessments. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple North Dakota school districts. A list of participating TAC members can be found in Appendix I.

Cambium Assessment, Inc

CAI is the vendor selected through the state-mandated competitive procurement process. In the winter of 2017, American Institutes for Research (now CAI) became the primary party responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the NDSA. Additionally, CAI is responsible for developing and maintaining the ICCR item bank, which is used for the NDSA test construction.

Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2021 test administration to ensure that any secure testing materials, such as items and prompts, were not leaked.

1.4 Available Test Formats and Special Versions

The spring 2021 NDSA ELA and mathematics assessments were administered online using an adaptive item selection algorithm (Volume 2, Appendix M) and making use of technology-enhanced item types. Students unable to participate in the online administration had the option to use print-on-demand, a feature that provides the same items administered to students online in a paper-pencil format. Spanish versions of mathematics tests (developed to meet the same content standards as the English versions) were available for all tested grades. Students participating in the computer-based NDSA could use standard online testing features in the Test Delivery System (TDS), which includes a selection of font colors and sizes, the ability to zoom in and out, and the ability to highlight text. In addition to the resources available to all students, options were available to accommodate students with an Individualized Education Program (IEP) or Section 504 Plan. These options included braille, American Sign Language (ASL), closed captioning, and large print. Students with disabilities could take either the NDSA, with or without accommodations, or an alternate assessment. English learners (ELs) could take the Spanish-language version of the NDSA mathematics. During test development, it was ensured that scores obtained on the Spanish-language version or other alternative modes of administrations were comparable to those received on the standard online test adhering to the same blueprints. The test summary comparison between the standard online form and the braille form, which matches the Spanish-language form, is provided in Volume 2.

1.5 Student Participation

All North Dakota public school students in grades 3–8 and 10 participated in the statewide assessments. The NDSA for ELA and mathematics assessments are administered in the spring.

Table 2 shows the number of students tested and the number of students reported for the spring 2021 NDSA by grade and subject. Table 3 and Table 4 present the distribution of students, in counts and percentages, by subgroups for mathematics and ELA, respectively. The subgroup categories reported here are gender, ethnicity, special education status (SPED), Title 1, and ELs.

Table 2: Number of Students Participating in the 2020–2021 NDSA

Mathematics			**ELA**		
Grade	Number Tested	Number Reported	Grade	Number Tested	Number Reported
3	8907	8904	3	8911	8852
4	8540	8539	4	8549	8504
5	8542	8540	5	8563	8523
6	8552	8545	6	8554	8496
7	8557	8548	7	8585	8518
8	8362	8358	8	8398	8319
10*	2964	2963	10	2953	2932

**There is a smaller number of participants for grade 10 because some high schools have an option for administering the ACT instead of the NDSA.*

Table 3: Distribution of Demographic Characteristics of Tested Population, Mathematics

Grade	Group	All Students	Female	Male	Multiracial	American Indian	Asian	Hispanic	African American	White	Pacific Islander	SPED	Title 1	ELs
3	*N*	8907	4344	4549	476	681	128	499	450	6634	25	1179	2662	408
	%	100	48.77	51.07	5.34	7.65	1.44	5.6	5.05	74.48	0.28	13.24	29.89	4.58
4	*N*	8540	4211	4317	389	705	123	431	416	6443	21	1173	2500	353
	%	100	49.31	50.55	4.56	8.26	1.44	5.05	4.87	75.44	0.25	13.74	29.27	4.13
5	*N*	8542	4159	4374	388	707	106	420	401	6487	24	1131	2567	300
	%	100	48.69	51.21	4.54	8.28	1.24	4.92	4.69	75.94	0.28	13.24	30.05	3.51
6	*N*	8552	4181	4337	362	757	98	419	399	6461	22	1104	2399	256
	%	100	48.89	50.71	4.23	8.85	1.15	4.9	4.67	75.55	0.26	12.91	28.05	2.99
7	*N*	8557	4143	4384	365	739	113	436	397	6457	20	1031	2348	280
	%	100	48.42	51.23	4.27	8.64	1.32	5.1	4.64	75.46	0.23	12.05	27.44	3.27
8	*N*	8362	4052	4279	329	693	108	413	379	6388	21	1025	2240	216
	%	100	48.46	51.17	3.93	8.29	1.29	4.94	4.53	76.39	0.25	12.26	26.79	2.58
10	N	2964	1410	1542	69	379	19	141	55	2286	3	334	896	46
	%	100	47.57	52.02	2.33	12.79	0.64	4.76	1.86	77.13	0.1	11.27	30.23	1.55

Table 4: Distribution of Demographic Characteristics of Tested Population, ELA

Grade	Group	All Students	Female	Male	Multiracial	American Indian	Asian	Hispanic	African American	White	Pacific Islander	SPED	Title 1	ELs
3	*N*	8911	4318	4515	471	665	126	496	448	6602	25	1166	2626	402
	%	100	48.46	50.67	5.29	7.46	1.41	5.57	5.03	74.08	0.28	13.08	29.47	4.51
4	*N*	8549	4194	4298	385	697	122	429	411	6427	21	1169	2476	346
	%	100	49.06	50.27	4.5	8.15	1.43	5.02	4.81	75.18	0.25	13.67	28.96	4.05
5	*N*	8563	4149	4363	389	704	105	417	396	6477	24	1130	2555	296
	%	100	48.45	50.95	4.54	8.22	1.23	4.87	4.62	75.64	0.28	13.2	29.84	3.46
6	*N*	8554	4169	4311	363	749	98	415	393	6440	22	1090	2369	248
	%	100	48.74	50.4	4.24	8.76	1.15	4.85	4.59	75.29	0.26	12.74	27.69	2.9
7	*N*	8585	4148	4351	365	717	113	432	392	6460	20	1032	2326	272
	%	100	48.32	50.68	4.25	8.35	1.32	5.03	4.57	75.25	0.23	12.02	27.09	3.17
8	*N*	8398	4039	4260	330	686	108	404	377	6374	20	1005	2223	209
	%	100	48.09	50.73	3.93	8.17	1.29	4.81	4.49	75.9	0.24	11.97	26.47	2.49
10	*N*	2953	1391	1535	71	371	19	139	54	2269	3	333	891	45
	%	100	47.1	51.98	2.4	12.56	0.64	4.71	1.83	76.84	0.1	11.28	30.17	1.52

2. Summary of Operational Procedures

2.1 Administration Procedures

Table 5 shows the testing window schedule for the 2020–2021 North Dakota State Assessment (NDSA) administration by subject.

Table 5: 2020–2021 NDSA Testing Windows

Assessment	Testing Window
ELA (Reading and Writing) 3–8 and 10	March 15–May 7, 2021
Mathematics 3–8 and 10	March 15–May 7, 2021

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

A secure browser developed by Cambium Assessment, Inc. (CAI) was required to access the online NDSA tests. The CAI Secure Browser provides 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 were able to pause a test, review previously answered questions, and modify their responses if the test had not been paused for more than 20 minutes.

2.2 Simulations

Prior to the operational testing window, CAI performs simulations for all NDSA assessments. Simulations are used to configure the adaptive algorithm (described further in Volume 2, Appendix M) and seek to maximize test score precision while meeting blueprint specifications based on the available pool of test items. Psychometricians review results of the ELA and mathematics simulations for the following key diagnostic factors:

- Match-to-test blueprint. Determines that the tests have the correct number of test items overall and the appropriate proportion by content standards, 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.
- 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 increased security resulting from a larger pool of items. The CAI simulation system 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 the 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 the blueprint, precision, and item exposure rate.

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;
- designated features that are available to students for whom the need has been identified by an informed educator or team of educators; and
- accommodations that are generally available for students for whom there is documentation on an Individualized Education Program (IEP), Section 504 Plan, or Individual Language Plan (ILP).

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 in an IEP, Section 504 Plan, or ILP to generate valid assessment outcomes so that students who require accommodations can fully demonstrate what they 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 disabilities by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

The 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. Two studies have been completed by CAI to evaluate the use of dictionaries and glossaries as accommodations. The results of these studies are presented in Appendices L and M, respectively.

The test administrators (TAs) and school test coordinators (SCs) in North Dakota 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 screen, switches), calculator, print-on-demand, multiplication table, and scribe. Detailed descriptions of each of these accommodations can be found in Appendix C of Volume 5.

Table 6 through Table 11 list the number of test sessions in which a student was provided with each accommodation during the spring 2021 test administration.

Table 6: ELA Total Sessions with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	10
Embedded Accommodations							
American Sign Language	2	2	4	7	2	5	–
Braille	–	1	1	–	–	–	–
Closed Captioning	1	1	4	6	1	5	–
Embedded Speech-to-Text	149	134	123	108	110	74	15
Permissive Mode	9	7	4	1	4	1	–
Streamlined Mode	3	4	3	5	4	15	2
Text-to-Speech: Items	60	43	48	77	52	52	1
Text-to-Speech: Passages	1	1	1	1	–	–	–
Text-to-Speech: Passages & Items	691	722	699	715	712	663	169
Non-Embedded Accommodations							
Alternate Response Options	3	2	2	1	4	2	1
Print-on-Demand: Stimuli and Items	9	4	5	8	7	7	4
Read-Aloud Stimuli	361	427	427	418	376	328	108
Scribe Items (Writing)	213	247	246	210	200	152	45
Speech-to-Text	139	163	166	181	193	143	37

Table 7: ELA Total Sessions with Allowed Embedded Designated Supports

Designated Supports	Grade						
	3	4	5	6	7	8	10
Color Contrast	6	4	23	6	3	7	2
Glossary (Spanish)	5	7	4	6	11	11	3
Line Reader (Enhanced)	–	–	–	2	–	–	–
Masking	22	18	39	40	43	46	5
Text-to-Speech: Items	60	43	48	77	52	52	1
Text-to-Speech: Passages	1	1	1	1	–	–	–
Text-to-Speech: Passages & Items	691	722	699	715	712	663	169

Table 8: ELA Total Sessions with Allowed Non-Embedded Designated Supports

Designated Supports	Grade						
	3	4	5	6	7	8	10
Color Contrast	4	5	5	5	4	7	–

Designated Supports	Grade						
	3	4	5	6	7	8	10
Color Overlay	7	12	12	12	9	7	1
Magnification	13	7	8	2	4	5	–
Noise Buffer	42	42	31	27	25	24	6
Read-Aloud: Items	473	500	515	475	426	380	102
Read-Aloud: Stimuli (Writing)	130	88	91	51	53	32	7
Separate Setting	863	930	934	877	837	790	243
Scribe Items (Non-Writing)	213	247	246	210	200	152	45
Simplified Test Directions	141	92	83	41	55	50	34
Translated Test Directions	9	13	16	22	20	17	12

Table 9: Mathematics Total Sessions with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	10
Embedded Accommodations							
American Sign Language	–	–	–	–	–	–	–
Braille	–	–	1	–	–	–	–
Closed Captioning	–	–	–	–	–	–	–
Embedded Speech-to-Text	142	132	118	88	112	76	19
Permissive Mode	1	2	5	1	4	2	–
Streamlined Mode	3	4	3	4	4	15	3
Text-to-Speech: Items	13	16	11	52	21	28	–
Text-to-Speech: Passages	1	–	–	–	–	–	–
Text-to-Speech: Passages & Items	688	713	699	630	620	624	141
Non-Embedded Accommodations							
Alternate Response Options	3	2	2	1	2	2	1
Print-on-Demand: Stimuli and Items	9	5	6	8	7	8	4
Read-Aloud Stimuli	–	101	94	46	54	36	7
Speech-to-Text	133	160	161	174	185	144	35
Calculator	64	96	128	215	302	412	174
100s Number Table	225	243	206	180	113	88	6
Multiplication Table	222	333	375	423	449	454	100
Abacus	4	2	2	1	–	–	–

Table 10: Mathematics Total Sessions with Allowed Embedded Designated Supports

Designated Supports	Grade						
	3	4	5	6	7	8	10
Color Contrast	6	3	18	3	3	7	2
Glossary (Spanish)	6	8	1	9	11	11	5
Line Reader (Enhanced)	–	1	–	–	–	–	–
Masking	24	21	38	31	46	49	5
Text-to-Speech: Items	13	16	11	52	21	28	–
Text-to-Speech: Passages	1	–	–	–	–	–	–
Text-to-Speech: Passages & Items	688	713	699	630	620	624	141
Translations Toggle (Spanish)	–	1	2	4	4	5	5

Table 11: Mathematics Total Sessions with Allowed Non-Embedded Designated Supports

Designated Supports	Grade						
	3	4	5	6	7	8	10
Color Contrast	4	6	3	5	4	7	–
Color Overlay	6	13	10	12	9	7	1
Magnification	13	7	6	2	4	3	–
Noise Buffer	39	43	30	27	25	24	6
Read-Aloud: Items	465	509	520	462	409	379	105
Read-Aloud: Stimuli	–	101	94	46	54	36	7
Read-Aloud: Stimuli/Items-Spanish	1	3	2	–	–	–	5
Separate Setting	858	929	937	882	832	801	247
Simplified Test Directions	139	93	80	39	60	53	33
Translated Test Directions	8	13	16	31	37	28	13

3. Item Bank and Test Design

Content specialists and psychometricians review all items in the Independent College and Career Readiness (ICCR) item banks with respect to the psychometric properties of the items, content bias, and sensitivity for the state of North Dakota. After these reviews, the selected items were used for the North Dakota operational item pool. In this section, we describe the characteristics of the spring 2021 operational item pool for the ELA and mathematics computer-adaptive tests (CATs). These characteristics include both content (e.g., item types) and statistical summaries. Test design and methodology of field-testing new items in spring 2021 are also discussed.

3.1 Overview of Item Development

All operational items used on the North Dakota State Assessments (NDSA) test forms are drawn from the ICCR item bank. Volume 2 is a separate, stand-alone report containing complete details on the ICCR item bank. The ICCR is a pre-equated item bank with item parameters estimated under the multiple group item response theory (IRT) framework described in a later section of this volume.

The operational item pool includes an array of item types. Each item type for English language arts (ELA) and mathematics is described in Table 12 and Table 13, respectively. Table 14 and Table 15 show the number of items by item type that were available in the item pool. Examples of item types are available in Volume 2 Appendix E.

Table 12: ELA Item Types and Descriptions

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 13: Mathematics Item Types and Descriptions

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.
Multiple-Choice (MC)	Student selects one correct answer from four options.

Response Type	Description
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 14: ELA Operational Items by Item Type and Grade

Item Type	Grade						
	3	4	5	6	7	8	10
MC	258	277	239	321	295	300	173
MS	22	35	33	47	69	42	17
MI	13	9	16	10	4	7	–
GI	–	–	1	–	–	–	–
ETC	49	58	52	46	49	42	46
Two-part ETC	–	–	1	–	1	–	–
HT	37	39	47	39	41	38	32
Two-part HT	3	5	4	8	1	4	2
EBSR	35	36	29	65	66	53	26
TE	2	4	3	3	5	2	3

Table 15: Mathematics Operational Items by Item Type and Grade

Item Type	Grade						
	3	4	5	6	7	8	10
MC	124	101	103	178	102	171	266
MS	48	94	44	52	17	45	37
GI	83	56	28	40	36	52	39
ETC	1	5	5	1	–	3	4
Two-part ETC	–	2	–	–	–	–	1
HT	–	–	–	–	–	–	18
TI	15	15	11	30	3	8	5
MI	11	31	12	13	9	9	6
EQ	327	343	321	337	283	230	295
Two-part EQ	–	2	–	–	1	2	4

3.2 Field Test

The spring 2020–2021 CAT ELA and mathematics assessments contained new field-test items (stand-alone items) and item clusters in the non-scored embedded field-test (EFT) slots. Clusters consist of several item parts that require students to interact with the item in various ways. To obtain high-quality responses to the EFT items, students were unaware of which items were operational and which were EFT. For reading, 6–9 EFT items or 1 EFT item cluster per test were administered; for mathematics, 8 stand-alone items or 1 cluster per test were administered, except in the segmented grade 6 test. For the segmented grade 6 test, 1 stand-alone and 1 cluster or 3 stand-alone and 1 cluster were administered. For more details regarding EFT items or item clusters administered by grade refer to Volume 2.

A total of 344 cluster items and 371 stand-alone items in ELA grades 3–8 and 10 was field-tested during spring 2021 in North Dakota. For mathematics, a total of 119 clusters and 187 stand-alone items were field-tested. For grades 3–8, both stand-alone items and item clusters were administered, and for grade 10, only stand-alone items were administered.

The spring 2021 field-test items were put onto the North Dakota reporting scale by using a fixed-anchor item calibration method. All operational item parameters were fixed to their item bank values and the item parameters of stand-alone items and clusters were freely estimated with a concurrent calibration method.

The spring 2021 ELA and mathematics EFT items were put onto the North Dakota reporting scale by using a fixed anchor item calibration method. The field-test items were administered in multiple states, such as Arizona, New Hampshire, West Virginia, and Wyoming. All of the operational (treated as fixed anchor) and field-test items were merged into a single incomplete data matrix for a multiple group IRT calibration. Operational item parameters were fixed to their item bank values, while field-test item parameters were estimated in a single run. If a calibration run did not converge, the reason was investigated. One or two items with negative item-total correlations were usually the cause. Those items were removed from the calibration and sent to the Cambium Assessment, Inc. (CAI) content team for further action, such as a revision or rejection. The state group means, provided in Appendix J, were obtained during free estimations.

3.3 Operational Test Design

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 NDSA test specifications. In addition, each student's unique test assembled by the algorithm contains the items that best match the student's achievement level, as defined by the blueprint. The details of the adaptive item selection algorithm are presented in Volume 2.

3.4 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 in order to maximize test score precision while meeting blueprint specifications based on the available item pool. The blueprint match was monitored for both the simulation and operational test administrations. The summary of the simulation versus operational blueprint match for spring 2021 is provided in Appendix B. The summary shows that across all grades and subjects the vast majority of tests met the blueprint specifications with a 100% match at the reporting category level in both the simulation and operational test administrations.

The IRT statistical properties of the operational item pool used for the spring 2021 NDSA are summarized in Tables 16–21 for ELA and mathematics. The acronyms 3PL and 2PL refer to the three-parameter logistic model and the two-parameter logistic model, respectively, while GPCM refers to 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 2PL items. For GPCM, step parameters (*b1*, *b2, and b3*) are summarized.

Table 16: 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*	124	0.42	0.89	1.20	1.51	1.88	2.47	3.3
	b	124	−4.61	−3.73	−2.73	−2.34	−1.88	−1.39	−0.75
	c	124	0.01	0.06	0.13	0.19	0.25	0.36	0.59
4	*a*	101	0.24	0.61	0.94	1.21	1.53	1.84	2.97
	b	101	−3.87	−3.10	−2.30	−1.71	−1.18	−0.40	0.62
	c	101	0.05	0.09	0.13	0.18	0.26	0.40	0.6
5	*a*	102	0.22	0.48	0.77	0.95	1.34	1.91	2.33
	b	102	−5.70	−2.46	−1.61	−1.05	−0.45	0.07	1.15
	c	102	0.04	0.07	0.15	0.18	0.24	0.33	0.56
6	*a*	178	0.11	0.44	0.72	0.95	1.15	1.56	4.79
	b	178	−3.16	−2.36	−1.29	−0.27	0.34	1.43	4.73
	c	178	0.01	0.06	0.12	0.18	0.23	0.35	0.4
7	*a*	102	0.10	0.38	0.58	0.80	0.97	1.24	7.62
	b	102	−4.09	−1.72	−0.41	0.65	1.52	2.21	2.91
	c	102	0.02	0.06	0.10	0.18	0.25	0.35	0.48
8	a	171	0.08	0.36	0.52	0.74	0.93	1.26	2.76
	b	171	−2.15	−1.41	−0.12	1.06	2.05	3.13	5.9
	c	171	0.02	0.05	0.12	0.19	0.26	0.38	0.51
10	*a*	266	0.07	0.24	0.52	0.72	0.99	1.44	2.98
	b	266	−0.41	0.82	2.42	3.25	4.20	5.84	9.9
	c	266	0.01	0.05	0.13	0.20	0.26	0.35	0.49

Table 17: 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*	473	0.27	0.76	1.22	1.52	1.77	2.13	2.6
	b	473	−5.61	−3.26	−2.71	−2.30	−1.86	−1.22	1.25
4	*a*	517	0.35	0.68	0.98	1.22	1.48	1.78	2.29
	b	517	−3.42	−2.77	−2.07	−1.54	−1.04	−0.32	0.84
5	*a*	411	0.20	0.56	0.83	1.04	1.25	1.55	2.06
	b	411	−4.11	−2.44	−1.47	−0.94	−0.42	0.44	2.72
6	*a*	450	0.10	0.52	0.75	0.95	1.13	1.44	1.92
	b	450	−4.04	−2.12	−0.86	−0.09	0.58	1.53	6.97
7	*a*	331	0.16	0.43	0.65	0.89	1.11	1.43	2.47
	b	331	−1.75	−0.93	−0.02	0.75	1.60	2.60	3.85
8	*a*	329	0.10	0.38	0.58	0.75	0.89	1.16	1.72
	b	329	−5.51	−0.18	1.14	1.95	2.58	3.89	6.69
10	*a*	389	0.15	0.34	0.55	0.75	0.93	1.31	3.09
	b	389	−2.52	1.18	2.72	3.87	5.03	6.75	9.54

Table 18: 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*	12	0.80	0.85	0.98	1.16	1.44	1.64	1.66
	b1	12	−3.41	−3.07	−2.46	−1.99	−1.69	−1.02	−0.68
	b2	12	−2.85	−2.79	−2.68	−2.01	−1.31	−0.46	−0.19
4	*a*	24	0.46	0.53	0.67	0.92	1.05	1.20	1.34
	b1	24	−4.02	−3.12	−2.19	−1.87	−1.55	0.01	0.40
	b2	24	−3.18	−2.90	−2.02	−1.64	−1.01	−0.32	0.54
	b3	24	−1.77	−1.73	−1.56	−1.35	−1.14	−0.98	−0.94
5	*a*	18	0.49	0.51	0.63	0.77	0.79	1.14	1.19
	b1	18	−2.10	−2.09	−1.80	−1.19	−0.43	0.24	0.47
	b2	18	−2.69	−2.31	−1.24	−0.32	0.02	0.45	0.85
6	*a*	23	0.49	0.51	0.72	0.79	0.84	0.92	1.06
	b1	23	−2.08	−1.81	−0.99	−0.54	0.25	2.09	2.38
	b2	23	−2.01	−0.79	−0.39	0.09	0.66	2.23	3.38
7	*a*	18	0.50	0.51	0.55	0.65	0.69	1.04	1.22
	b1	18	−1.17	−0.78	0.22	0.62	1.27	1.98	3.67
	b2	18	−0.33	−0.18	0.43	0.99	1.52	2.54	2.86
8	*a*	20	0.22	0.29	0.39	0.57	0.67	0.78	0.79
	b1	20	−1.50	−1.39	−0.64	0.15	2.00	3.01	4.75
	b2	20	−3.15	−0.89	1.31	2.22	2.76	4.30	6.94
	b3	20	−0.18	−0.18	−0.18	−0.18	−0.18	−0.18	−0.18
10	*a*	20	0.25	0.29	0.42	0.47	0.59	0.90	0.92
	b1	20	−1.89	0.19	1.23	1.44	2.84	5.31	7.21
	b2	20	−0.28	0.53	1.76	2.73	4.27	5.44	5.49

Table 19: 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*	258	0.30	0.58	0.90	1.19	1.52	2.13	2.81
	b	258	−2.36	−1.86	−1.26	−0.85	−0.40	0.30	1.89
	c	258	0.03	0.07	0.14	0.19	0.24	0.32	0.59
4	*a*	277	0.19	0.40	0.72	0.99	1.31	1.78	2.44
	b	277	−2.84	−1.84	−1.28	−0.76	−0.11	0.74	1.98
	c	277	0.01	0.04	0.10	0.16	0.22	0.29	0.37
5	*a*	239	0.23	0.43	0.73	0.99	1.30	1.76	2.49
	b	239	−2.03	−1.44	−0.74	−0.26	0.23	1.06	2.54
	c	239	0.03	0.06	0.13	0.18	0.23	0.31	0.42
6	*a*	321	0.18	0.39	0.70	0.96	1.26	1.65	3.73
	b	321	−2.33	−1.04	−0.39	0.10	0.66	1.52	5.67
	c	321	0.01	0.06	0.12	0.18	0.24	0.32	0.42
7	*a*	295	0.11	0.43	0.67	0.88	1.14	1.55	2.76
	b	295	−1.98	−1.04	−0.21	0.37	0.85	1.86	7.40
	c	295	0.01	0.03	0.10	0.17	0.24	0.33	0.40
8	a	300	0.05	0.39	0.69	0.90	1.14	1.45	2.04
	b	300	−1.39	−0.87	−0.08	0.42	1.20	2.30	3.85
	c	300	0.00	0.04	0.11	0.18	0.25	0.32	0.43
10	*a*	173	0.13	0.24	0.43	0.67	0.94	1.44	4.04
	b	173	−1.36	−0.47	0.24	0.92	1.66	3.10	8.07
	c	173	0.00	0.03	0.11	0.18	0.24	0.35	0.41

Table 20: 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*	138	0.03	0.41	0.68	0.87	1.06	1.35	1.92
	b	138	−4.99	−2.55	−1.36	−0.59	−0.15	1.24	2.40
4	*a*	164	0.04	0.34	0.48	0.68	0.89	1.26	1.59
	b	164	−2.96	−2.02	−1.17	−0.53	0.28	1.99	5.31
5	*a*	156	0.19	0.36	0.56	0.74	0.95	1.24	1.47
	b	156	−2.15	−1.58	−0.82	−0.11	0.86	1.98	4.98
6	*a*	192	0.12	0.29	0.52	0.71	0.88	1.13	2.19
	b	192	−2.13	−1.58	−0.19	0.43	1.21	3.34	6.75
7	*a*	211	0.19	0.29	0.48	0.68	0.86	1.25	1.43
	b	211	−2.31	−1.16	−0.16	0.57	1.29	2.67	4.91
8	*a*	160	0.06	0.29	0.47	0.63	0.81	1.04	1.22
	b	160	−4.60	−1.30	−0.01	0.78	1.48	3.15	5.82
10	*a*	110	0.12	0.22	0.42	0.57	0.74	0.97	1.30
	b	110	−1.38	−0.93	0.21	0.98	2.08	3.83	8.16

Table 21: 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*	27	0.25	0.29	0.45	0.79	1.13	1.56	1.58
	b1	27	−3.70	−3.26	−2.50	−2.19	−1.95	0.29	1.08
	b2	27	−4.31	−1.86	−1.47	−0.73	−0.25	1.58	1.93
	b3	27	−2.44	−2.06	−1.06	0.73	0.78	0.81	0.81
4	*a*	26	0.17	0.32	0.39	0.59	0.86	1.47	1.49
	b1	26	−3.45	−3.29	−2.44	−2.25	−1.77	−0.67	−0.37

Grade	Parameter	*N* Item	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b2	26	−1.85	−1.50	−0.97	−0.27	0.84	2.18	3.95
	b3	26	−1.64	−0.99	1.58	1.68	2.05	2.21	2.25
5	*a*	34	0.24	0.27	0.43	0.49	0.70	1.41	1.50
	b1	34	−3.20	−2.79	−2.22	−1.86	−1.38	−0.09	1.02
	b2	34	−1.49	−1.09	−0.50	−0.11	0.64	1.48	4.30
	b3	34	−1.95	−1.80	−1.12	−0.67	1.62	2.39	2.74
6	*a*	30	0.28	0.29	0.39	0.50	0.62	1.58	1.70
	b1	30	−4.71	−3.28	−2.08	−1.68	−0.39	2.24	3.86
	b2	30	−2.75	−2.00	−0.42	0.15	1.14	1.89	3.20
	b3	30	−0.76	−0.62	0.38	2.20	2.42	2.68	2.75
7	*a*	29	0.21	0.27	0.35	0.49	0.80	1.64	1.74
	b1	29	−3.14	−2.19	−1.72	−1.07	0.07	2.84	4.20
	b2	29	−1.25	−0.90	0.05	0.48	1.66	2.88	3.37
	b3	29	−0.34	0.10	1.85	1.90	2.83	3.01	3.05
8	*a*	32	0.24	0.30	0.37	0.55	0.73	1.31	1.35
	b1	32	−3.07	−2.64	−1.70	−1.21	0.07	2.02	2.93
	b2	32	−2.22	−1.55	−0.45	−0.08	0.87	2.08	2.65
	b3	32	−0.94	−0.76	0.40	2.23	2.58	2.75	2.76
10	*a*	20	0.26	0.27	0.36	0.45	1.07	1.26	1.27
	b1	20	−2.28	−2.13	−1.33	−1.08	−0.74	−0.02	0.62
	b2	20	−0.94	−0.84	0.77	1.26	1.69	3.80	5.05
	b3	20	2.64	2.64	2.65	2.72	2.89	3.10	3.16

4. FIELD-TEST CLASSICAL ANALYSES OVERVIEW

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) stand-alone field-test items, and reponse time analysis is performed for item clusters. Any items flagged for out-of-range statistics are reviewed by the Cambium Assessment, Inc. (CAI) content and psychometric staff, and poorly performing items are removed from the item bank. The flagging rules are defined by the item statistics computed from the Independent College and Career Readiness (ICCR) states' data where the field-test items are administered. Furthermore, for the computation of DIF statistics, the data from all states' operational and field-test items are combined in order to obtain a sufficient number of students for each demographic group.

For item clusters, the classical item statistics are computed for individual assertions, and the business rules for flagging are defined at the item level. The criteria for flagging and reviewing items are provided in Table 22. A description of the classical statistics is provided in the following subsections.

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. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for MC items was calculated as the biserial correlation between the item score and the ability estimate for students.

4.2 DISTRACTOR ANALYSIS

Distractor analysis for MC items was 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. For MC items, the correct response should have been the most frequently selected option by high-scoring students. The discrimination value of the correct response should have been substantial and positive, and the discrimination values for distractors should have been lower and, generally, negative.

4.3 ITEM DIFFICULTY

Items that were either extremely difficult or extremely easy were flagged for review but were not necessarily removed if they were grade-level appropriate and aligned with the test specifications. For MC items, the proportion of students in the sample selecting the correct answer (the p-value) was computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty was 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). 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 22.

Table 22: Thresholds for Flagging in Classical Item Analysis

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.
DIF	Item DIF categorization of "C" in either direction.

4.4 Differential Item Functioning Analysis

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) document provides a guideline for when sample sizes permitting subgroup differences in performance should be examined and when appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors. To identify such potential problems, all field-tested ICCR summative items are evaluated in terms of DIF statistics prior to becoming operational in the bank. DIF statistics for items field-tested in spring 2021 are available in Appendix H.

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. DIF-flagged items are further examined by content experts who are asked to reexamine each flagged item to determine whether the item should be excluded from the 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.

At CAI, DIF analysis is conducted to detect potential item bias across major gender, ethnic, and special population groups. A minimum sample of 200 responses (Zwick, 2012) per item in each subgroup is required for DIF analyses, and the following groups were investigated for inclusion in the study:

- White/African American
- White/Hispanic
- White/Asian or Pacific Islander
- White/American Indian or Alaskan Native
- White/Multi-racial
- Special Education (SPED)/Non-SPED

Table 23 and Table 24 illustrate the minimum to maximum number of field test item responses for each group.

Table 23: Range of ELA Field Test Item Responses by DIF Group

Grade	Group	Non-SPED/SPED	Male/Female	White/Asian	White/American Indian	White/African American	White/Hispanic	White/Multiracial
3	Reference	430-980	502-1362	716-2269	716-2269	716-2269	716-2269	716-2269
	Focal	63-177	440-1344	11-29	31-78	25-106	53-129	49-135
4	Reference	738-1028	560-1530	871-2514	871-2514	871-2514	871-2514	871-2514
	Focal	124-169	573-1449	14-39	58-79	37-118	81-141	52-126
5	Reference	381-992	300-2808	471-4752	471-4752	471-4752	471-4752	471-4752
	Focal	51-176	272-2721	4-43	38-81	25-209	29-178	31-232
6	Reference	986-1236	687-1956	1102-3299	1102-3299	1102-3299	1102-3299	1102-3299
	Focal	156-199	655-1919	13-33	68-91	32-148	110-195	56-146
7	Reference	594-1145	588-2086	901-3395	901-3395	901-3395	901-3395	901-3395
	Focal	87-156	578-1880	11-37	54-92	32-133	98-189	38-151
8	Reference	462-1238	558-2544	879-4350	879-4350	879-4350	879-4350	879-4350
	Focal	59-167	543-2436	9-48	50-79	21-193	65-156	35-190
10	Reference	965-1956	539-1118	837-1706	837-1706	837-1706	837-1706	837-1706
	Focal	112-223	538-1064	9-20	28-211	8-39	148-159	23-50

Table 24: Range of Math Field Test Item Responses by DIF Group

Grade	Group	Non-SPED/SPED	Males/Females	White/Asian	White/American Indian	White/African American	White/Hispanic	White/Multiracial
3	Reference	318-634	1780-2024	3082-3459	3082-3459	3082-3459	3082-3459	3082-3459
	Focal	49-120	1673-1971	19-33	22-56	114-153	78-145	145-195
4	Reference	192-245	2067-2245	3561-3818	3561-3818	3561-3818	3561-3818	3561-3818
	Focal	30-49	1967-2112	22-37	34-48	140-187	96-140	160-202
5	Reference	354-658	1983-2248	3476-3814	3476-3814	3476-3814	3476-3814	3476-3814
	Focal	38-105	1950-2169	22-38	36-62	128-175	95-148	155-193
6	Reference	416-1037	599-2381	858-3996	858-3996	858-3996	858-3996	858-3996
	Focal	59-180	556-2253	12-41	46-94	43-189	72-191	44-175
7	Reference	564-650	2142-2579	3778-4321	3778-4321	3778-4321	3778-4321	3778-4321
	Focal	67-95	2035-2431	23-43	6-65	145-192	68-164	139-202
8	Reference	233-291	1955-2183	3344-3724	3344-3724	3344-3724	3344-3724	3344-3724
	Focal	26-44	1780-2040	21-42	28-51	119-165	69-129	122-161
10	Reference	796-4314	458-2435	681-3854	681-3854	681-3854	681-3854	681-3854
	Focal	87-552	417-2419	2-50	106-148	13-55	35-634	13-133

A generalized Mantel-Haenszel (MH) method is applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's raw 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 $MH\chi^2$ DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

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 $k = \{1, 2, \ldots K\}$ for the strata, $n_{R1k}$is the number of 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 total number of correct responses, $n_{R+k}$is the number of students in the reference group, and n_{++k} is the number of students in stratum k, and the variance is calculated as

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

where 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

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

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

The 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.

Items are classified into three categories (A, B, or C), ranging from "no evidence of DIF" to "severe DIF." DIF classification rules are illustrated in Table 25. Items are also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African American, Hispanic, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., white or male). If DIF statistics fall into category C for any group, the item shows significant DIF and is reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts review all items flagged based on DIF statistics. The experts are encouraged to discuss these items and are asked to decide whether each item should be excluded from the pool of potential operational items. The spring 2021 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 item bank to examine them for the presence of DIF for accommodated versus non-accommodated students. The results of these analyses are presented in Volume 7.

Table 25: DIF Classification Rules

Dichotomous Items	
Category	*Rule*
C	MH_{χ^2} is significant and $\lvert\hat{\Delta}_{MH}\rvert$ ≥1.5
B	MH_{χ^2} is significant and 1 ≤ $\lvert\hat{\Delta}_{MH}\rvert$ <1.5

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

4.5 Classical Analyses Results

This section presents a summary of results from the classical item analysis for the 2021 ICCR field-test items. Table 24 through Table 27 provide the summary of the *p*-values and biserial correlations for the field-tested items in four states, New Hampshire, North Dakota, West Virginia and Wyoming for ELA and mathematics, respectively. The statistics were computed using all four states data.

*Table 26: 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	89	0.12	0.23	0.35	0.46	0.54	0.65	0.72
4	72	0.12	0.22	0.34	0.49	0.61	0.80	0.91
5	80	0.03	0.16	0.31	0.46	0.58	0.80	0.92
6	51	0.09	0.18	0.35	0.48	0.54	0.73	0.91
7	63	0.07	0.14	0.28	0.42	0.62	0.75	0.89
8	84	0.06	0.22	0.35	0.44	0.58	0.72	0.78
10	46	0.19	0.25	0.36	0.47	0.52	0.68	0.80

**Results presented excluded flagged items.*

*Table 27: 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	89	0.01	0.11	0.23	0.37	0.51	0.62	0.68
4	72	0.03	0.12	0.42	0.47	0.55	0.66	0.72
5	80	0.07	0.18	0.38	0.47	0.56	0.67	0.70
6	51	0.07	0.16	0.35	0.41	0.54	0.67	0.71
7	63	0.10	0.14	0.32	0.41	0.50	0.60	0.68
8	84	0.06	0.19	0.35	0.44	0.51	0.61	0.63
10	46	0.08	0.11	0.28	0.44	0.48	0.56	0.64

**Results presented excluded flagged items*

Table 28: 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	41	0.05	0.12	0.24	0.41	0.52	0.78	0.86
4	37	0.02	0.08	0.25	0.40	0.53	0.86	0.92
5	38	0.03	0.08	0.19	0.27	0.44	0.58	0.71
6	44	0.04	0.12	0.22	0.39	0.53	0.74	0.80
7	34	0.06	0.07	0.20	0.28	0.38	0.65	0.71
8	40	0.01	0.03	0.17	0.31	0.46	0.66	0.85
10	36	0.01	0.01	0.08	0.20	0.34	0.56	0.63

Results presented excluded flagged items.

Table 29: 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	41	0.06	0.13	0.45	0.61	0.73	0.78	0.82
4	37	0.32	0.39	0.59	0.65	0.71	0.76	0.77
5	38	0.04	0.19	0.38	0.52	0.68	0.80	0.82
6	44	0.03	0.25	0.42	0.57	0.65	0.75	0.76
7	34	0.11	0.14	0.33	0.44	0.58	0.76	0.81
8	40	0.25	0.29	0.36	0.60	0.67	0.73	0.81
10	36	0.29	0.32	0.48	0.68	0.79	0.84	0.84

Results presented excluded rejected flagged items.

5. 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) items. IRT is a general framework that models test responses resulting from an interaction between students and 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 model (Bock & Zimowski, 1997) is used to calibrate all ICCR items.

5.1 Item Response Theory Methods

All individuals in the calibration sample are considered to have the observed responses z_{ijk_j} corresponding to examinee j, in group k to the ith item. The multiple group IRT 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 (3PL) model (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 examinee j answering item i correctly, 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 $(z_{ijk_j} + 1)$th category to the ith item as

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

where b_{il} is the lth 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

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

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

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

All item parameter estimates were obtained with IRTPRO version 4.1 (Cai, Thissen, & du Toit, 2011). IRTPRO uses marginal maximum likelihood estimation. 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.

5.2 Equating to the Scale

Equating to the established reporting scale is done using the Stocking-Lord procedure (Stocking & Lord, 1983). The methods are implemented by calibrating the item response data using the same multiple group IRT model as described previously 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 multiple group IRT 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 $\mathrm{z}_i=\{1,\ldots,\mathrm{m}_i\}$ 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

$$p\left(z_{i,I}^{*}\middle|\theta\right)=\frac{\exp\left(\sum_{l=1}^{z_i}D\frac{a_{i,J}}{A}\left(\theta-\left(b_{il,J}*A+B\right)\right)\right)}{1+\sum_{h=0}^{m_i}\exp\sum_{l=0}^{h}D\frac{a_{i,J}}{A}\left(\theta-\left(b_{il,J}*A+B\right)\right)}.$$

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 coefficients A and B.

The symmetric approach uses the reverse transformation 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,

$$p\left(z_{i,I}^{*}\middle|\theta\right)=\frac{\exp\left(\sum_{l=0}^{z_i}DAa_i\left(\theta-\frac{\left(b_{il,I}-B\right)}{A}\right)\right)}{1+\sum_{h=1}^{m_i}\exp\sum_{l=0}^{h}DAa_{i,I}\left(\theta-\frac{\left(b_{il,I}-B\right)}{A}\right)}.$$

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

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

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

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

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

5.2.1 Establishing the Initial ICCR Item Bank

Establishing the initial set of item parameters and equating the items over the years they were used is described within this section. Initially, the ICCR item bank spanned three different years (2015–2017) of field testing with multiple states. Each grade and subject were calibrated separately within a given year using multiple group IRT. 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 where a set of common items were administered in the item pools each year. All common items in the item pool were used unless the item's *A* parameter was less than 0.1 or greater than 3 and the absolute *B* parameter was larger than 6, then the item was not included. Table 28 displays year-to-year equating constants.

Table 30: Linking Across Years Results

			2015 to 2016			2016 to 2017	
Subject	**Grade**	**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
	7	163	1.1462	−0.0069	178	1.0259	0.1838
	8	135	0.9785	−0.1097	155	1.0279	−0.1285
	9	149	0.9156	−0.0281	174	0.9401	−0.0262
	10	166	0.9786	−0.0578	173	1.0667	−0.1436
	11	148	0.9749	−0.1287	163	1.0544	−0.2303
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
	9	171	1.1250	−0.0787	257	1.1200	−0.0268
	10	116	1.0697	−0.1756	216	1.1852	−0.1731
	11	164	0.9782	−0.0526	226	1.0043	0.0089

5.2.2 Linking Initial ICCR Item Bank to SAGE Item Bank

The methods described previously were used to calibrate and equate the ICCR item bank. Once that item bank was established, these items were then linked to the Utah Student Assessment of Growth and Excellence (SAGE) item bank, which provides a vertical reporting scale for the North Dakota State Assessment (NDSA). Linking the ICCR and SAGE item banks also used the Stocking-Lord procedure (Stocking & Lord, 1983), using the same common-item non-equivalent groups design. Table 29 shows linking constants for each grade and subject between the initial ICCR item bank and the SAGE item bank. These linking constants were used to put the initial ICCR item bank onto the SAGE item bank on-grade level scale.

Appendix D documents the design and results of the vertical linking study that was implemented to develop the SAGE English language arts (ELA) and mathematics item bank.

Table 31: 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
	9	236	1.0421	0.0642
	10	186	1.0084	0.2021
	11	231	0.9889	0.1200
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
	9	217	0.8693	0.3189
	10	213	1.0592	0.2563
	11	183	0.9826	0.5578

Table 30 and Table 31 display the number of students in each participating state contributing to the ICCR multiple group IRT model.

Table 32: Number of Students Used in ICCR Multiple Group IRT Calibration for ELA

Grade	Year	Utah	Florida	Arizona	Oregon (2015)/Ohio (2016)
3	2015	39279	–	33687	9323
	2016	46901	–	62242	85972
	2017	47317	–	72754	–
4	2015	39753	–	33091	11858
	2016	43190	207867	61065	95211
	2017	45537	206341	73195	–
5	2015	38976	35780	32398	8398
	2016	36196	199326	60210	97451
	2017	43825	209984	72289	–
6	2015	38340	42565	33114	8234
	2016	38106	196409	57635	101799
	2017	39662	200039	69837	–
7	2015	36082	56752	30911	10688
	2016	45469	193186	58050	105249
	2017	45484	197752	69754	–
8	2015	36445	82159	32277	13590
	2016	42530	195125	57349	104360
	2017	42018	197269	69481	–
9	2015	36867	97690	23036	–
	2016	40489	199657	51004	–
	2017	40165	197807	62956	–
10	2015	31619	132712	19635	–
	2016	39407	191764	46817	–
	2017	37477	195673	58182	–
11	2015	32100	–	21510	1674
	2016	35888	–	41487	–
	2017	9716	–	54018	–

Table 33: Number of Students Used in ICCR Multiple Group IRT Calibration for Mathematics

Grade	Year	Utah	Florida	Arizona	Oregon (2015) / Ohio (2016)
3	2015	48473	–	43543	27642
	2016	49762	–	62586	94869
	2017	49688	185609	72857	–
4	2015	47088	–	43464	27102
	2016	48367	–	61384	95765
	2017	49727	173825	73438	–
5	2015	47098	87436	42419	26957
	2016	46702	201278	60448	97308
	2017	48021	212008	72428	–
6	2015	46160	87831	40512	27550
	2016	46380	193158	57868	101015
	2017	46263	195425	70034	–
7	2015	43517	79949	39887	26753
	2016	43718	170453	57467	102933
	2017	43623	171940	68366	–
8	2015	43745	60958	39997	26969
	2016	43377	125120	49781	78629
	2017	44035	120321	59171	–
Geometry	2015	32430	65306	23911	–
	2016	40058	201299	45782	89001
	2017	37274	175871	56135	–
Algebra 1	2015	39923	85572	31251	–
	2016	43942	218008	53721	105895
	2017	42838	201800	66695	–
Algebra 2	2015	27288	71562	21125	–
	2016	28212	137337	39249	–
	2017	9763	120631	50063	–

5.2.3 Linking 2021 ICCR Field-Test Items

The spring 2021 ELA and mathematics embedded field-test items were put onto the North Dakota reporting scale by using a fixed anchor item calibration method. The field-test items were administered in multiple ICCR states, such as Arizona, New Hampshire, West Virginia, and Wyoming. All of the operational (treated as fixed anchor) and field-test items were put into a single incomplete data matrix for a multiple group IRT 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, then the reason was investigated. One or two items with negative item-total correlations were usually the cause. Those items were removed from the calibration and sent to the Cambium Assessment, Inc. (CAI) content team for further action, such as revision or rejection. The state group means, provided in Appendix B, were free estimations.

6. Scoring and Reporting

6.1 Maximum Likelihood Estimation

Ability estimates were generated using pattern scoring, a method that scores students depending on how they answer individual items. Scoring details are provided in the following sections.

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* indexes item, *h* indexes 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 $\arg\max_{\theta} log\big(L(\theta)\big)$ 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 θ_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 (3PL) model or a two-parameter logistic (2PL) model.

6.1.3 Standard Errors of Estimate

When the MLE is available, the standard error 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

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{h=1}^{m_i} hexp\left(\sum_{l=1}^{j} Da_i\left(\hat{\theta} - b_{il}\right)\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i\left(\hat{\theta} - b_{il}\right)\right)} \right)^2 - \frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} Da_i\left(\hat{\theta} - b_{il}\right)\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i\left(\hat{\theta} - b_{il}\right)\right)} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(p_i - c_i) q_i}{(1 - c_i)^2} \left(1 - \frac{z_i c_i}{p_i^2}\right),$$

where N_{CR} is the number of items that are scored using the GPCM, 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. In addition, when a student's raw score is lower than the expected raw score due to guessing the likelihood is not identified. For the North Dakota State Assessments (NDSA) scoring, the extreme cases were handled as follows:

i. Assign the lowest observable theta (LOT) score value of –4 to a raw score of 0.
ii. Assign the highest observable theta (HOT) score 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 the NDSA uses a vertical score for scoring, the truncated LOT and HOT are converted to the vertical scale before being applied. These truncated LOT and HOT scores in the vertical scale and the associated scale scores for each grade and subject are provided in Table 32 and Table 33.

Table 34: Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates—ELA

Grade	LOT	HOT	Lowest Observable Scale Score (LOSS)	Highest Observable 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
10	−4.04	5.00	480	900

Table 35: Theta and Corresponding Scaled Score Limits for Extreme Ability Estimates—Mathematics

Grade	LOT	HOT	LOSS	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
10	−4.76	8.20	350	960

6.1.5 Standard Error of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the standard error (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 student *s*. The NDSA included items that were scored using the 3PL model, 2PL model, and GPCM from IRT. The 2PL can be visualized as either a 3PL item with no pseudo-guessing parameter or a dichotomously scored GPCM item. The test information was calculated as

$$I(\theta_s) = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} D a_i(\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} D a_i(\theta_s - b_{il})\right)} - \left(\frac{\sum_{h=1}^{m_i} hexp\left(\sum_{l=1}^{h} D a_i(\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} D a_i(\theta_s - b_{il})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{q_i}{p_i} \left[\frac{p_i - c_i}{1 - c_i} \right]^2 \right),$$

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

For standard error 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. The truncated standard error of measurement (SEM) on a vertical scale is provided in Table 34.

Table 36: SEM Truncation Values for Each Grade and Subject

Subject	Grades	SEM Truncation Values on Theta Metric	SEM Truncation Values on Vertical Scale
ELA	3	1.5	1.25
ELA	4	1.5	1.34
ELA	5	1.5	1.34
ELA	6	1.5	1.35
ELA	7	1.5	1.41
ELA	8	1.5	1.52
ELA	10	1.5	1.70
Mathematics	3	1.5	0.90
Mathematics	4	1.5	1.11
Mathematics	5	1.5	1.28
Mathematics	6	1.5	1.49
Mathematics	7	1.5	1.52
Mathematics	8	1.5	1.86
Mathematics	10	1.5	2.43

6.2 Transforming Vertical Scores to Reporting Scale Scores

For spring 2021, the NDSA scale scores were reported for each student who took the English language arts (ELA) and mathematics assessments. The scale scores were based on the operational items presented to the student and did not include any field-test or linking items. Independent College and Career Readiness (ICCR) item parameters were converted to a vertical scale in the item bank, and a single scale across all grades was created 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 at 50 and 550, respectively. 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 the NDSA scale scores for each test is provided in Appendix E, and the summary of scale scores for each reporting category is provided in Appendix F.

6.3 Overall Performance Classification

Each student was assigned an overall performance category in accordance with his or her overall scale score. Table 35 and Table 36 provide the scale score range for performance standards for ELA and mathematics, respectively. The lower bound of Level 3, Proficient, marks the minimum cut score for proficiency.

Table 37: Proficiency Levels for ELA by Grade

Grade	Level 1 Novice	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
3	420–559	560–584	585–620	621–750
4	430–571	572–599	600–638	639–790
5	450–594	595–621	622–660	661–810
6	460–609	610–637	638–670	671–830
7	470–610	611–640	641–679	680–850
8	480–615	616–649	650–701	702–870
10	480–626	627–666	667–712	713–900

Table 38: Proficiency Levels for Mathematics by Grade

Grade	Level 1 Novice	Level 2 Partially Proficient	Level 3 Proficient	Level 4 Advanced
3	300–409	410–427	428–462	463–550
4	310–436	437–464	465–500	501–610
5	320–445	446–483	484–522	523–660
6	330–469	470–512	513–557	558–720
7	340–502	503–549	550–597	598–750
8	350–518	519–579	580–639	640–830
10	350–593	594–649	650–692	693–960

6.4 Reporting Category Performance Classification

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

- If $(\theta_{tt} < \theta_{Proficient} - 1.5 \times SE_{RC})$, then performance was classified as Below Standard;
- If $(\theta_{Proficient} - 1.5 \times SE_{RC} \leq \theta_{tt} < \theta_{Proficient} + 1.5 \times SE_{RC})$, then performance was classified as At or Near Standard; and
- If $(\theta_{tt} \geq \theta_{Proficient} + 1.5 * SE_{RC})$, then performance was 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 SE of the given subscale. Zero and perfect scores were assigned Below Standard and Above Standard, respectively.

6.5 Strengths and Weaknesses Scores

For an individual student, strengths and weaknesses with reporting categories were 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 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 target scores of individual students with different abilities who received different items that measure the same reporting category at different levels of difficulty,

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

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 from 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 $\bar{\delta}_{Tg}$) in teaching a given target.

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

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

6.6 LEXILE AND QUANTILE SCORES

The NDSA reports Lexile and Quantile measures with summative ELA and mathematics test scores. MetaMetrics provides conversion tables between ELA summative test scores and Lexile measures and between mathematics summative test scores and Quantile measures for each grade and subject. A linking study for mathematics and ELA took place at the end of June 2018 to determine final conversions. The study report can be found in Volume 7. Lexile and Quantile measures (not a range) are reported for all summative assessments in all modes (e.g., online, braille).

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.

7.1 Quality Assurance Reports

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 marginal maximum likelihood estimations using the procedures described within this report.

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

Table 39: 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 rate issues
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

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 is provided in Table 38, and a description of the statistics is provided later in this section.

Table 40: 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.

Item Discrimination

As described in Section 4.1, 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 3.5% 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.

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 aligned with the test specifications. For further detail, refer back to Section 4.2. Most of the operational items had *p*-values within the expected range, but 24 items, across all test grades and subjects, were flagged for a *p*-value of zero. CAI content experts and psychometricians verified that this item behaved as expected and was scored correctly.

Distractor Analysis

As discussed in Section 4.3, distractor analysis for multiple-choice items was 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 issues.

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) for ELA and mathematics. 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 test administrations in which the blueprint requirements were not met; and, for test 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 (ELA and mathematics) 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 shows the detailed comparison for simulation and operational blueprint match for ELA and mathematics. This summary shows that, across all grades and subjects, the vast majority of tests met the blueprint specifications with a 100% match at the reporting category level in both simulation 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. Item overexposure often 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, most test items were administered to 20% or fewer test takers across all grades and subjects. Appendix G shows the item exposure rates for the operational test administrations for ELA and mathematics.

7.1.4 Cheating Detection Analysis

The CAI QA system can also provide a forensics report to identify possible irregularities in the 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, including changes in test scores across test administrations, item response times, and 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 maximum likelihood estimates for ELA and mathematics using the procedures described within this report.

8. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

Bock R. D., & Zimowski M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433–448). Springer, New York, NY.

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

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.

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

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.

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.

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.